Error Whitening Criterion for Adaptive Filtering: Theory and Algorithms

Yadunandana N. Rao, Deniz Erdogmus, *Member, IEEE*, and Jose C. Principe, *Fellow, IEEE*

Abstract—Mean squared error (MSE) has been the dominant criterion in adaptive filter theory. A major drawback of the MSE criterion in linear filter adaptation is the parameter bias in the Wiener solution when the input data are contaminated with noise. In this paper, we propose and analyze a new augmented MSE criterion called the Error Whitening Criterion (EWC). EWC is able to eliminate this bias when the noise is white. We will determine the analytical solution of the EWC, discuss some interesting properties, and develop stochastic gradient and other fast algorithms to calculate the EWC solution in an online fashion. The stochastic algorithms are locally computable and have structures and complexities similar to their MSE-based counterparts (LMS and NLMS). Convergence of the stochastic gradient algorithm is established with mild assumptions, and upper bounds on the step sizes are deduced for guaranteed convergence. We will briefly discuss an RLS-like Recursive Error Whitening (REW) algorithm and a minor components analysis (MCA) based EWC-total least squares (TLS) algorithm and further draw parallels between the REW algorithm and the Instrumental Variables (IV) method for system identification. Finally, we will demonstrate the noise-rejection capability of the EWC by comparing the performance with MSE criterion and TLS.

Index Terms—Error whitening, LMS, MSE, noisy system identification, RLS.

I. INTRODUCTION

THE MEAN squared error (MSE) criterion has been the workhorse of adaptive filter and the Wiener filter theory due to the simple and analytically tractable structure of the linear least squares solution [1], [2]. Although the Wiener solution is optimal in the least squares sense, the input covariance matrix in the presence of additive white input noise yields a bias in the optimal parameters when compared with the estimation obtained with noise-free data [1], [2]. This is a major drawback, since noise is omnipresent in practical scenarios. Significant research efforts to improve MSE-based solutions in noisy cases yielded many modifications [3]–[7]. Total-least-squares (TLS) is one such method, which is quite powerful in eliminating the bias due to noise [8]–[10]. Unfortunately, it involves singular value decomposition (SVD) calculations that limit its applicability. Even though fast TLS algorithms have been developed [11]–[13], some inherent drawbacks of TLS include equal input

and output noise power constraint and exact model order knowledge, which prevent its application in many problems [10], [14]. The instrumental variables (IV) method proposed as an extension to the least-squares (LS) has been previously applied for parameter estimation in white noise [15]. This method requires choosing a set of *instruments* that are uncorrelated with the noise in the input [15]. Yet another classical approach is the subspace Wiener filtering [1], [2]. This approach tries to suppress the bias by performing a PCA subspace projection and then training a filter in the reduced input space. Subspace Wiener filtering improves the SNR when the signal power is already greater than the noise power [16]. Otherwise, it fails since noise and signal subspaces cannot be distinguished.

In this paper, we will propose a new criterion called the error whitening criterion (EWC) to solve the problem of parameter estimation in white noise. We introduce a correction term to the MSE, based on the time differences of the error signal that leads to an augmented cost function with interesting properties. Instead of minimizing the mean squared error, the EWC formulation enforces zero autocorrelation of the error signal beyond a certain lag; hence, the name *Error Whitening Criterion*. We will then derive adaptive algorithms that estimate and track the optimal EWC solution. The rest of the paper is organized as follows.

In the next section, we will propose the error whitening criterion and discuss its properties in Section III. In Section IV, we will derive a stochastic gradient algorithm and the conditions for guaranteed convergence. In Section V, the Recursive Error Whitening (REW) algorithm will be presented, and we will contrast this with the IV method and briefly discuss a minor components based algorithm for EWC. Section VI contains simulation results that evaluate the performance of EWC and other methods in parameter-estimation problems. Discussion on the effects of under modeling will be presented followed by conclusions.

II. ERROR WHITENING CRITERION

Suppose that noise-free training data of the form $(\mathbf{x}_k, d_k)$, generated by a linear system with weight vector $\mathbf{w}_T$ (the true weight vector) through $d_k = \mathbf{x}_k^T \mathbf{w}_T$, is provided. Assume that the input vector $\mathbf{x}_k \in \Re^m$, the weight vector $\mathbf{w}_T \in \Re^m$, and the desired output $d_k \in \Re$. Suppose that the weight vector of the adaptive linear system is sufficiently long $(n \geq m)$, which allows us to assume without loss of generality that the length of $\mathbf{w}_T$ is n as well since we can extend $\mathbf{w}_T$ by padding zeros. The quadratic form in (1) gives the cost function of EWC, and its unique stationary point (i.e., the weight vector that makes the gradient zero) gives the sought optimal solution. In (1), $\beta \in \Re$ and the following definitions hold: $e_k = d_k - \mathbf{x}_k^T \mathbf{w}, \dot{e}_k = \dot{d}_k -$

The authors are with the Computational NeuroEngineering Laboratory, Electrical and Computer Engineering Department, University of Florida, Gainesville, FL 32611 USA (e-mail: yadu@cnel.ufl.edu; deniz@cnel.ufl.edu; principe@cnel.ufl.edu).

$\dot{\mathbf{x}}_k^T\mathbf{w}$, where $\dot{d}_k = d_k - d_{k-L}, \dot{\mathbf{x}}_k = \mathbf{x}_k - \mathbf{x}_{k-L}$, and $L \geq 0$ is an integer.

$$J(\mathbf{w}) = E\left[e_k^2\right] + \beta E\left[\dot{e}_k^2\right]. \tag{1}$$

Throughout the rest of the paper, we will use the following notations: The noisy input vector $\hat{\mathbf{x}}_k$, the noise-free input vector $\mathbf{x}_k$, and the input noise vector $\mathbf{v}_k$ obey $\hat{\mathbf{x}}_k = \mathbf{x}_k + \mathbf{v}_k$; the noisy desired signal $\hat{d}_k$, the noise-free desired signal d_k, and noise u_k are related by $\hat{d}_k = d_k + u_k$. In the case of an adaptive linear neuron (ADALINE), $\mathbf{v}_k$ is temporaly white, although its covariance matrix is arbitrary. In the case of an adaptive FIR filter, $\mathbf{v}_k$ is composed of delayed versions of a temporally white noise signal v_k. Similarly, u_k is assumed to be white. We define the following symmetric matrices: The input autocorrelation matrices $\mathbf{R} = E[\mathbf{x}_k\mathbf{x}_k^T], \hat{\mathbf{R}} = E[\hat{\mathbf{x}}_k\hat{\mathbf{x}}_k^T], \mathbf{R}_L = E[\mathbf{x}_{k-L}\mathbf{x}_k^T + \mathbf{x}_k\mathbf{x}_{k-L}^T]$, and $\hat{\mathbf{R}}_L = E[\hat{\mathbf{x}}_{k-L}\hat{\mathbf{x}}_k^T + \hat{\mathbf{x}}_k\hat{\mathbf{x}}_{k-L}^T]$ for noise-free and noisy signals; the input noise autocorrelation matrices $\mathbf{V} = E[\mathbf{v}_k\mathbf{v}_k^T]$, and $\mathbf{V}_L = E[\mathbf{v}_{k-L}\mathbf{v}_k^T + \mathbf{v}_k\mathbf{v}_{k-L}^T]$; the input derivative autocorrelation matrices $\mathbf{S} = E[\dot{\mathbf{x}}_k\dot{\mathbf{x}}_k^T]$ and $\hat{\mathbf{S}} = E[\dot{\hat{\mathbf{x}}}_k\dot{\hat{\mathbf{x}}}_k^T]$ for noise-free and noisy signals. We also define crosscorrelation matrices following the same structure: the correlation between the input vector and the desired signal $\mathbf{P} = E[\mathbf{x}_k d_k], \hat{\mathbf{P}} = E[\hat{\mathbf{x}}_k\hat{d}_k], \mathbf{P}_L = E[\mathbf{x}_{k-L}d_k + \mathbf{x}_k d_{k-L}]$, and $\hat{\mathbf{P}}_L = E[\hat{\mathbf{x}}_{k-L}\hat{d}_k + \hat{\mathbf{x}}_k\hat{d}_{k-L}]$; the correlation between the input vector and desired signal derivatives $\mathbf{Q} = E[\dot{\mathbf{x}}_k\dot{d}_k]$, and $\hat{\mathbf{Q}} = E[\dot{\hat{\mathbf{x}}}_k\dot{\hat{d}}_k]$. The following theorems and lemmas address the analytical solution of EWC for linear adaptive systems and the structure of the performance surface as a function of the parameter β.

Lemma 1: In the noise-free training data case, i.e., with $(\mathbf{x}_k, d_k)$, the true weight vector satisfies the equation $\mathbf{R}_L\mathbf{w}_T = \mathbf{P}_L$.

Proof: This result follows immediately from the substitution of $d_k = \mathbf{x}_k^T\mathbf{w}_T$ and $d_{k-L} = \mathbf{x}_{k-L}^T\mathbf{w}_T$ in the definitions of $\mathbf{R}_L$ and $\mathbf{P}_L$. Notice that the Wiener–Hopf equations correspond to the special case when $L = 0$. □

Theorem 1 (Analytical EWC Solution): Suppose the training data are noisy, i.e., we have $(\hat{\mathbf{x}}_k, \hat{d}_k)$. The following are equivalent expressions for the stationary point of the EWC performance surface given in (1).

$$\hat{\mathbf{w}}_* = (\hat{\mathbf{R}} + \beta\hat{\mathbf{S}})^{-1}(\hat{\mathbf{P}} + \beta\hat{\mathbf{Q}}) \tag{2}$$

$$\hat{\mathbf{w}}_* = [(1 + 2\beta)\hat{\mathbf{R}} - \beta\hat{\mathbf{R}}_L]^{-1}[(1 + 2\beta)\hat{\mathbf{P}} - \beta\hat{\mathbf{P}}_L]. \tag{3}$$

Proof: Substituting $e_k = \hat{d}_k - \hat{\mathbf{x}}_k^T\mathbf{w}$ and $\hat{e}_k = \dot{\hat{d}}_k - \dot{\hat{\mathbf{x}}}_k^T\mathbf{w}$ yields the following explicit expression for $J(\mathbf{w})$:

$$J(\mathbf{w}) = E\left[\hat{d}_k^2\right] + \beta E\left[\dot{\hat{d}}_k^2\right]$$
$$+ \mathbf{w}^T(\hat{\mathbf{R}} + \beta\hat{\mathbf{S}})\mathbf{w} - 2(\hat{\mathbf{P}} + \beta\hat{\mathbf{Q}})^T\mathbf{w}. \tag{4}$$

Taking the gradient with respect to $\mathbf{w}$ and equating to zero

$$\frac{\partial J(\mathbf{w})}{\partial\mathbf{w}} = 2(\hat{\mathbf{R}} + \beta\hat{\mathbf{S}})\mathbf{w} - 2(\hat{\mathbf{P}} + \beta\hat{\mathbf{Q}}) = \mathbf{0}$$
$$\Rightarrow \hat{\mathbf{w}}_* = (\hat{\mathbf{R}} + \beta\hat{\mathbf{S}})^{-1}(\hat{\mathbf{P}} + \beta\hat{\mathbf{Q}}). \tag{5}$$

In order to show the equivalence of (3) with (2), we notice the following identities: $\hat{\mathbf{S}} = 2\hat{\mathbf{R}} + \hat{\mathbf{R}}_L$ and $\hat{\mathbf{Q}} = 2\hat{\mathbf{P}} + \hat{\mathbf{P}}_L$. Substituting these in (2) yields (3). Notice that MSE corresponds to the special case $\beta = 0$. □

Lemma 2: In the noise-free case, regardless of the specific value of β, the optimal solution of EWC is equal to the true weight vector, i.e., $\mathbf{w}_* = \mathbf{w}_T$.

Proof: In the noise-free case, the optimal solution of EWC is given by $\mathbf{w}_* = [(1 + 2\beta)\mathbf{R} - \beta\mathbf{R}_L]^{-1}[(1 + 2\beta)\mathbf{P} - \beta\mathbf{P}_L]$. Once again, directly substituting the definitions $d_k = \mathbf{x}_k^T\mathbf{w}_T$ and $d_{k-L} = \mathbf{x}_{k-L}^T\mathbf{w}_T$ yields $\mathbf{w}_* = \mathbf{w}_T$. □

It is easy to show that the noisy correlation matrices are related to the noise-free signal and noise correlation matrices through the following set of equations:

$$\hat{\mathbf{R}} = E\left[\hat{\mathbf{x}}_k\hat{\mathbf{x}}_k^T\right] = \mathbf{R} + \mathbf{V}$$
$$\hat{\mathbf{S}} = E[(\hat{\mathbf{x}}_k - \hat{\mathbf{x}}_{k-L})(\hat{\mathbf{x}}_k - \hat{\mathbf{x}}_{k-L})^T]$$
$$= 2(\mathbf{R} + \mathbf{V}) - \mathbf{R}_L - \mathbf{V}_L$$
$$\hat{\mathbf{P}} = E[\hat{\mathbf{x}}_k\hat{d}_k] = \mathbf{P}$$
$$\hat{\mathbf{Q}} = E[(\hat{\mathbf{x}}_k - \hat{\mathbf{x}}_{k-L})(\hat{d}_k - \hat{d}_{k-L})^T] = 2\mathbf{P} - \mathbf{P}_L$$
$$\hat{\mathbf{R}}_L = \mathbf{R}_L + \mathbf{V}_L$$
$$\hat{\mathbf{P}}_L = \mathbf{P}_L. \tag{6}$$

Theorem 2 (Noise Rejection With EWC): If $\beta = -1/2$, $\mathbf{V}_L = \mathbf{0}$, and $\mathbf{R}_L$ is invertible, then the optimal solution of EWC obtained using noisy data is equal to the true weight vector of the reference model that generated the data.

Proof: Clearly, when the specified conditions in the theorem are met, the given analytical solution $\hat{\mathbf{w}}_* = [(1 + 2\beta)\hat{\mathbf{R}} - \beta\hat{\mathbf{R}}_L]^{-1}[(1 + 2\beta)\hat{\mathbf{P}} - \beta\hat{\mathbf{P}}_L]$ reduces to $\mathbf{w}_* = \mathbf{R}_L^{-1}\mathbf{P}_L$, which is identical to the true weight vector due to Lemma 1. □

In the FIR filter case, since the input noise vector consists of delayed versions of the additive white contaminating noise signal, selecting $L \geq n$, where n is the length of the adaptive filter, will guarantee that $\mathbf{V}_L = \mathbf{0}$. In the ADALINE case, $L \geq 1$ is sufficient since the input noise vector is assumed to be temporally white.

III. SOME PROPERTIES OF EWC ADAPTATION

1) Performance Surface: The performance surface of the MSE criterion for linear adaptive filtering is a smooth convex function with a single global minimum. However, the shape of the performance surface for EWC cost function changes with β and is also dependent on the eigenvalues of the matrices $\hat{\mathbf{R}}$ and $\hat{\mathbf{R}}_L$. Note that if $\beta \geq 0$, the performance surface will remain convex because $\hat{\mathbf{R}} + \beta\hat{\mathbf{S}}$ is a symmetric positive definite matrix. When $\beta < 0$, the stationary point can be a global maximum, global minimum, or a saddle point as shown in the contour plots of Fig. 1.

2) Orthogonality of Error to Input: In the case of MSE with noise-free input and sufficiently long filter (the effects of undermodeling will be dealt later), the optimal solution results in an error signal that is orthogonal to the input vector, i.e., $E[e_k\mathbf{x}_k] = \mathbf{0}$ [1], [2]. Similarly, at the optimal solution of EWC, the error vector and the input vector satisfy the relationship $\beta E[e_k\mathbf{x}_{k-L} + e_{k-L}\mathbf{x}_k] = (1 + 2\beta)E[e_k\mathbf{x}_k]$ for all $L \geq n$,
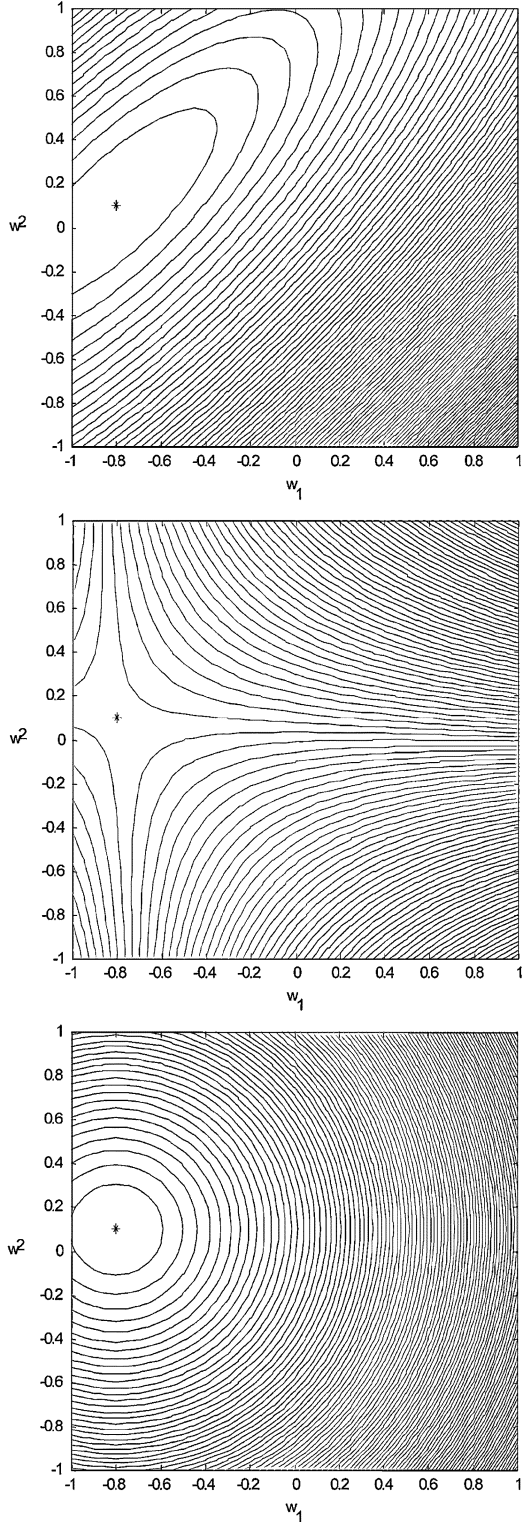
Fig. 1. Contour plots of the EWC performance surface in two dimensions (2-D) with $\beta = -1/2$ showing stationary point as (left) global minimum, (center) saddle point, (right) and global maximum.

as long as the adaptive filter is longer than the true filter. For $\beta = -1/2$, we obtain $E[e_k \mathbf{x}_{k-L} + e_{k-L} \mathbf{x}_k] = \mathbf{0}$. This result reveals an interesting insight if interpreted in terms of Newtonian physics, which we state here: The optimal solution of EWC tries to decorrelate the residual error from the estimated future value of the input vector (see Appendix A for details).

3) Why This Criterion Is Called EWC: Consider linear filter adaptation with noisy data. The autocorrelation of the error signal for lag $L \geq n$ is given by (7). Clearly, if the input autocorrelation matrix at lag L is nonsingular, then the error autocorrelation at lag L becomes zero if the weight vector is identically equal to the true weight vector. In other words, the criterion tries to partially *whiten* the error signal for lags greater than or equal to the length of our adaptive filter. However, the fact that $\rho_e(L) = 0$ for all $L \geq n$ is true only for the case when the adaptive filter length is longer than the true filter length (otherwise, $\rho_e(L) = 0$ only for the specified lag). We will revisit this property again in Section VI in our discussions on undermodeling effects.

$$
\begin{aligned}
\rho_e(L) &= E[e_k e_{k-L}] \\
&= E\left[\left(\mathbf{x}_k^T(\mathbf{w}_T - \mathbf{w}) + u_k - \mathbf{v}_k^T \mathbf{w} \right) \right. \\
&\quad \times \left. \left(\mathbf{x}_{k-L}^T(\mathbf{w}_T - \mathbf{w}) + u_{k-L} - \mathbf{v}_{k-L}^T \mathbf{w} \right) \right] \\
&= E\left[(\mathbf{w}_T - \mathbf{w})^T \mathbf{x}_k \mathbf{x}_{k-L}^T (\mathbf{w}_T - \mathbf{w}) \right. \\
&\quad \left. + \left(u_k - \mathbf{v}_k^T \mathbf{w} \right) \left(u_{k-L} - \mathbf{v}_{k-L}^T \mathbf{w} \right) \right] \\
&= (\mathbf{w}_T - \mathbf{w})^T E\left[\mathbf{x}_k \mathbf{x}_{k-L}^T \right] (\mathbf{w}_T - \mathbf{w}).
\end{aligned} \tag{7}
$$

In the next section, we will propose an online stochastic gradient-based method to compute the optimal EWC solution. We will derive the conditions for guaranteed convergence and show the expressions for excess error autocorrelations at lags greater than the length of the filter.

IV. STOCHASTIC ERROR WHITENING: EWC-LMS ALGORITHM

Consider the EWC cost function and its gradient given in (4) and (5). This exact gradient can be approximated using the stochastic instantaneous gradient by removing the expectation operators, which yields

$$
\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} \approx -2(\hat{e}_k \hat{\mathbf{x}}_k + \beta \hat{\hat{e}}_k \hat{\mathbf{x}}_k). \tag{8}
$$

Traditionally, optimization problems involve minimization or maximization of an objective function. In the case of EWC, the stationary point might sometimes be a saddle point, and hence, the traditional fixed-sign gradient approach (for maximization or minimization) will fail to converge in a stable manner. On the other hand, if we utilize additional information regarding the local curvature of the performance surface, we can modify the gradient accordingly to converge to the desired saddle point. For offline training, the convexity of a local parabola passing through the current weight vector and extending along the gradient direction could be used to determine the sign. If this parabola is convex (i.e., if it has a positive coefficient for the quadratic term), then negative gradient must be employed. Otherwise, the update must proceed along the positive gradient direction. This evaluation requires keeping track of the input covariance matrices $\mathbf{R}$ and $\mathbf{S}$, which is not attractive. Alternatively, we can determine the sign at every update by a stochastic estimate of the EWC cost computed using

the most recent weights. This yields the EWC-LMS algorithm given below.

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \eta_k \text{sign}\left(\hat{e}_k^2 + \beta \hat{\dot{e}}_k^2\right)(\hat{e}_k \hat{\mathbf{x}}_k + \beta \hat{\dot{e}}_k \hat{\dot{\mathbf{x}}}_k). \quad (9)$$

Theorem 3: In the noise-free case, EWC-LMS given in (9) converges to the stationary point in (5), provided that the step size satisfies the following inequality at every update.

$$|\eta_k| < \frac{2\left|e_k^2 + \beta \dot{e}_k^2\right|}{\|e_k \mathbf{x}_k + \beta \dot{e}_k \dot{\mathbf{x}}_k\|^2}. \quad (10)$$

Proof: First, notice that in the noise-free case, the stationary point $\mathbf{w}_*$ given by (5) is the same as the true weight vector $\mathbf{w}_T$. This is because at the true weight vector, the error and its time derivative become zero making the update zero. In order to prove convergence to this solution, consider the weight error vector defined as $\varepsilon_k = \mathbf{w}_* - \mathbf{w}_k$. Subtracting both sides of (9) from $\mathbf{w}_*$, we get the weight error dynamics as $\varepsilon_{k+1} = \varepsilon_k - \eta_k \text{sign}(e_k^2 + \beta \dot{e}_k^2)(e_k \mathbf{x}_k + \beta \dot{e}_k \dot{\mathbf{x}}_k)$. Taking the norm of both sides, we get $\|\varepsilon_{k+1}\|^2 = \|\varepsilon_k\|^2 - 2\eta_k \text{sign}(e_k^2 + \beta \dot{e}_k^2)\varepsilon_k^T(e_k \mathbf{x}_k + \beta \dot{e}_k \dot{\mathbf{x}}_k) + \eta_k^2 \|e_k \mathbf{x}_k + \beta \dot{e}_k \dot{\mathbf{x}}_k\|^2$. If we allow the error vector norm to decay asymptotically by making $\|\varepsilon_{k+1}\|^2 < \|\varepsilon_k\|^2$, then the upper bound on the magnitude of the step-size is given by

$$|\eta_k| < \frac{2\left|\varepsilon_k^T(e_k \mathbf{x}_k + \beta \dot{e}_k \dot{\mathbf{x}}_k)\right|}{\|e_k \mathbf{x}_k + \beta \dot{e}_k \dot{\mathbf{x}}_k\|^2}. \quad (11)$$

Since we assumed noise-free data, we have $\varepsilon_k^T \mathbf{x}_k = e_k$ and $\varepsilon_k^T \dot{\mathbf{x}}_k = \dot{e}_k$. Therefore, the upper bound reduces to (10). If the step size satisfies this instantaneously computable bound, the error vector norm is a monotonically decreasing sequence, which eventually converges to the stationary point, which is a zero vector, i.e., $\lim_{k\to\infty}\|\varepsilon_k\|^2 \to 0$, and this implies $\lim_{k\to\infty}\mathbf{w}_k \to \mathbf{w}_* = \mathbf{w}_T$ □. See Appendix B for details.

As an observation, see that if $\beta = 0$, the bound reduces to $0 < \eta_k < 2/\|\mathbf{x}_k\|^2$. If this is included in the update (9), we obtain normalized LMS (NLMS) algorithm [1]. We will now derive the necessary conditions on the step size to ensure convergence in the noisy case when $\beta = -1/2$.

Theorem 4: In the noisy case, the EWC-LMS rule with $\beta = -1/2$, which is given in (9), converges to the solution given by $\hat{\mathbf{w}}_* = (\hat{\mathbf{R}} - \hat{\mathbf{S}}/2)^{-1}(\hat{\mathbf{P}} - \hat{\mathbf{Q}}/2)$ in the mean if the step size satisfies the following inequality:

$$0 < \eta_k < \frac{2\left|E\left(\hat{e}_k^2 - 0.5\hat{\dot{e}}_k^2\right)\right|}{E\|\hat{e}_k \hat{\mathbf{x}}_k - 0.5\hat{\dot{e}}_k \hat{\dot{\mathbf{x}}}_k\|^2}. \quad (12)$$

Proof: As before, let the error vector be given by $\hat{\varepsilon}_k = \hat{\mathbf{w}}_* - \mathbf{w}_k$. Error vector dynamics are then given by the difference equation $\|\hat{\varepsilon}_{k+1}\|^2 = \|\hat{\varepsilon}_k\|^2 - 2\eta_k \text{sign}(\hat{e}_k^2 + \beta \hat{\dot{e}}_k^2)\hat{\varepsilon}_k^T(\hat{e}_k \hat{\mathbf{x}}_k + \beta \hat{\dot{e}}_k \hat{\dot{\mathbf{x}}}_k) + \eta_k^2 \|\hat{e}_k \hat{\mathbf{x}}_k + \beta \hat{\dot{e}}_k \hat{\dot{\mathbf{x}}}_k\|^2$. Taking the expectations on both sides and letting $E|\hat{\varepsilon}_{k+1}|^2 < E|\hat{\varepsilon}_k|^2$, we immediately see that the upper bound on the positive step size $\eta_k > 0$ is

$$\eta_k < 2\frac{E\left|\hat{\varepsilon}_k^T\left(\hat{e}_k \hat{\mathbf{x}}_k + \beta \hat{\dot{e}}_k \hat{\dot{\mathbf{x}}}_k\right)\right|}{E\|\hat{e}_k \hat{\mathbf{x}}_k + \beta \hat{\dot{e}}_k \hat{\dot{\mathbf{x}}}_k\|^2}. \quad (13)$$

Using Jensen's inequality for convex functions $E|X| \geq |E(X)|$, we can deduce a loose upper bound for the step size as

$$\eta_k < 2\frac{\left|E\left[\hat{\varepsilon}_k^T(\hat{e}_k \hat{\mathbf{x}}_k + \beta \hat{\dot{e}}_k \hat{\dot{\mathbf{x}}}_k)\right]\right|}{E\|\hat{e}_k \hat{\mathbf{x}}_k + \beta \hat{\dot{e}}_k \hat{\dot{\mathbf{x}}}_k\|^2}. \quad (14)$$

Since the lag L is chosen to be longer than the filter length, i.e., $L \geq n$, and owing to the fact that the input noise vector $\mathbf{v}_k$ is white, we have $E[\mathbf{v}_k \mathbf{v}_{k-L}^T] = E[\mathbf{v}_{k-L}\mathbf{v}_k^T] = 0$ and $E[\mathbf{v}_k \mathbf{v}_k^T] = \mathbf{V}$, which is a nonzero matrix. We can then further simplify $E[\hat{\varepsilon}_k^T \hat{e}_k \hat{\mathbf{x}}_k]$ and $E[\hat{\varepsilon}_k^T \hat{\dot{e}}_k \hat{\dot{\mathbf{x}}}_k]$ to yield

$$E\left[\hat{\varepsilon}_k^T \hat{e}_k \hat{\mathbf{x}}_k\right] = E\left(e_k^2\right) - \mathbf{w}_*^T \mathbf{V}\mathbf{w}_k + \mathbf{w}_k^T \mathbf{V}\mathbf{w}_k \quad (15)$$

$$E\left[\hat{\varepsilon}_k^T \hat{\dot{e}}_k \hat{\dot{\mathbf{x}}}_k\right] = E\left(\dot{e}_k^2\right) - 2\mathbf{w}_*^T \mathbf{V}\mathbf{w}_k + 2\mathbf{w}_k^T \mathbf{V}\mathbf{w}_k. \quad (16)$$

Substituting (15) and (16) in the numerator of (14) and using the fact that $E(e_k^2 - 0.5\dot{e}_k^2) = E(e_k e_{k-L}) = E(\hat{e}_k^2 - 0.5\hat{\dot{e}}_k^2)$, all the terms with the error vector $\hat{e}_k$ vanish, and we obtain the bound in (12). Observe that this upper bound can be easily computed without any knowledge of the optimal weight vector $\mathbf{w}_*$. This was possible only because we chose $\beta = -1/2$, which helped remove the quadratic products $\mathbf{w}_*^T \mathbf{V}\mathbf{w}_k, \mathbf{w}_k^T \mathbf{V}\mathbf{w}_k$. For any other nonzero value of β, it is impossible to estimate an upper bound on the step size that is computable without the knowledge of the noise variances or $\mathbf{w}_*$. This again validates our claim that $\beta = -1/2$ removes the bias in the weights. Thus, if the step size satisfies the upper bound, the mean error vector norm decays asymptotically to zero (stable stationary point), and hence, the EWC-LMS rule with $\beta = -1/2$ converges to its stationary point, i.e., $\lim_{k\to\infty}\hat{\mathbf{w}}_k \to \mathbf{w}_* = \mathbf{w}_T$. □

In the next theorem, we will show that the asymptotic excess error correlation at lags $L \geq n$ is always bounded from above and can be arbitrarily reduced by controlling the step size.

Theorem 5: With $\beta = -1/2$, the steady-state excess error autocorrelation at lag $L \geq n$, i.e., $|\rho_{\hat{e}}(L)|$ is always bound by

$$|\rho_{\hat{e}}(L)| \leq \frac{\eta}{2}\left[\text{Tr}(\mathbf{R}) + \sigma_v^2 \mathbf{I}\right]\left[E\left(e_a^2(k)\right) + \sigma_u^2 + \|\mathbf{w}\|_\infty^2 \sigma_v^2\right] \quad (17)$$

where $\mathbf{R} = E[\mathbf{x}_k \mathbf{x}_k^T]$, and $\text{Tr}(\cdot)$ denotes the matrix trace. The term $E(e_a^2(k))$ denotes the excess MSE, which is $(\mathbf{w}_\infty - \mathbf{w}_T)^T \mathbf{R}(\mathbf{w}_\infty - \mathbf{w}_T)$. The noise variances in the input and desired signals are represented by σ_v^2 and σ_u^2, respectively. Note that the term $\|\mathbf{w}\|_\infty^2$ is always bound because of the step-size bound.

Proof: We will start from the equation describing the dynamics of the error vector norm given below.

$$\|\hat{\varepsilon}_{k+1}\|^2 = \|\hat{\varepsilon}_k\|^2 - 2\eta\text{sign}\left(\hat{e}_k^2 + \beta \hat{\dot{e}}_k^2\right)\hat{\varepsilon}_k^T(\hat{e}_k \hat{\mathbf{x}}_k + \beta \hat{\dot{e}}_k \hat{\dot{\mathbf{x}}}_k) + \eta^2 \|\hat{e}_k \hat{\mathbf{x}}_k + \beta \hat{\dot{e}}_k \hat{\dot{\mathbf{x}}}_k\|^2 \quad (18)$$

where we have assumed a constant step-size that satisfies the upper bound in (12). Letting $E\|\hat{\varepsilon}_{k+1}\|^2 = E\|\hat{\varepsilon}_k\|^2$ as $k \to \infty$, we see that

$$\frac{\eta}{2}E\|\hat{e}_k \hat{\mathbf{x}}_k + \beta \hat{\dot{e}}_k \hat{\dot{\mathbf{x}}}_k\|^2 = E\left|\hat{\varepsilon}_k^T(\hat{e}_k \hat{\mathbf{x}}_k + \beta \hat{\dot{e}}_k \hat{\dot{\mathbf{x}}}_k)\right|. \quad (19)$$

Once again, we use Jensen's inequality to reduce (19), further yielding

$$\frac{\eta}{2}E\|\hat{e}_k\hat{\mathbf{x}}_k + \beta\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k\|^2 \geq \left|E\left[\hat{\boldsymbol{\varepsilon}}_k^T(\hat{e}_k\hat{\mathbf{x}}_k + \beta\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k)\right]\right|. \quad (20)$$

The noisy error term is given by $\hat{e}_k = e_a(k) + u_k - \mathbf{w}^T\mathbf{v}_k$, where the excess error $e_a(k) = \boldsymbol{\varepsilon}^T\mathbf{x}_k$. Using the expressions in (15), (16) and $\beta = -0.5$, we can immediately recognize that the RHS of (20) is simply the steady-state excess error autocorrelation at lag $L \geq n$, i.e., $|\rho_{\hat{e}}(L)|$. In order to evaluate the LHS of (20), we will assume that the terms that $\|\hat{\mathbf{x}}_k\|^2$ and $\eta\hat{e}_k^2$ are uncorrelated in the steady state. This assumption (more relaxed than the independence assumptions) has been used in the computation of the steady-state excess MSE for the stochastic LMS algorithm [17] and is more realistic for long filters. Using this, we get

$$\frac{\eta}{2}E\|\hat{e}_k\hat{\mathbf{x}}_k - 0.5\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k\|^2 = \frac{\eta}{2}E\left(\hat{e}_k^2\right)\left[\text{Tr}(\mathbf{R}) + \sigma_v^2\mathbf{I}\right] \quad (21)$$

where $E(\hat{e}_k^2) = E(e_a^2(k)) + \sigma_u^2 + \|\mathbf{w}\|_\infty^2\sigma_v^2$. Using (21) in (20), we get the inequality in (17). □

As a special case, consider $L = 0$ and the noise-free input. Then, (17) is true with the equality sign, and $|\rho_{\hat{e}}(L)|$ will be the same as $E(e_a^2(k))$, which is nothing but the excess MSE (as $k \to \infty$) of the LMS algorithm. In other words, (17) reduces to

$$E\left(e_a^2(k)\right) = \frac{\eta}{2}\text{Tr}(\mathbf{R})\left[E\left(e_a^2(k)\right) + \sigma_u^2\right] \quad (22)$$

from which the excess MSE for the LMS algorithm [2] can be deduced as

$$E\left(e_a^2(k)\right) = \frac{\eta\sigma_u^2\text{Tr}(\mathbf{R})}{2 - \eta\text{Tr}(\mathbf{R})} \quad (23)$$

which will become $\eta\sigma_u^2\text{Tr}(\mathbf{R})/2$ for very small step sizes. If the adaptive filter is long enough, the excess error $e_a(k)$ will be Gaussian, and we can easily show that the excess MSE is bounded by $\text{Tr}(\mathbf{R})E[\|\boldsymbol{\varepsilon}_0\|^2]/4$, where $\boldsymbol{\varepsilon}_0$ denotes the error due to the initial condition [17].

Other Versions of Stochastic EWC-LMS: It is easy to see that for mean convergence, the condition is $|I - \eta\lambda_k(\mathbf{R} + \beta\mathbf{S})| < 1$ for all k, where $\lambda_k(\mathbf{R} + \beta\mathbf{S})$ denotes the kth eigenvalue of the matrix $(\mathbf{R} + \beta\mathbf{S})$. This gives an upper bound on the step size as $|\eta| < 2/|\lambda_{\max}(\mathbf{R} + \beta\mathbf{S})|$. From the triangle inequality [11], $\|\mathbf{R}\|_2 + |\beta|\|\mathbf{S}\|_2 \leq \sqrt{\lambda_{\max}(\mathbf{R})} + |\beta|\sqrt{\lambda_{\max}(\mathbf{S})}$, where $\|\cdot\|_2$ denotes the matrix norm. Since both $\mathbf{R}$ and $\mathbf{S}$ are positive-definite matrices, we can write

$$\|\mathbf{R}\|_2 + |\beta|\|\mathbf{S}\|_2 \leq \sqrt{\text{Tr}(\mathbf{R})} + |\beta|\sqrt{\text{Tr}(\mathbf{S})}$$
$$\leq \sqrt{E\|\mathbf{x}_k\|^2} + |\beta|\sqrt{E\|\dot{\mathbf{x}}_k\|^2}. \quad (24)$$

In a stochastic framework, we can include this in the update equation in (9) to give us a step-size normalized EWC-LMS update rule given by

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \frac{\eta\text{sign}\left(\hat{e}_k^2 + \beta\hat{\hat{e}}_k^2\right)(\hat{e}_k\hat{\mathbf{x}}_k + \beta\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k)}{(\|\hat{\mathbf{x}}_k\| + |\beta|\|\hat{\hat{\mathbf{x}}}_k\|)^2}. \quad (25)$$

Note that when $\beta = 0$, (25) reduces to the well-known NLMS algorithm [2]. Alternatively, we can normalize by the norm squared of the gradient, and this gives the following modified update rule:

$$\mathbf{w}_{k+1} = \mathbf{w}_k + 2\frac{\left(\hat{e}_k^2 + \beta\hat{\hat{e}}_k^2\right)(\hat{e}_k\hat{\mathbf{x}}_k + \beta\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k)}{(\|\hat{e}_k\hat{\mathbf{x}}_k + \beta\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k\|^2 + \delta)}. \quad (26)$$

The term δ, which is 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 noise-less case, even with finite number of samples. Once again, we would like to state that with $\beta = 0$, (26) defaults to the NLMS algorithm. However, the caveat is that both (25) and (26) do not satisfy the principle of minimum disturbance, unlike the NLMS [2]. Nevertheless, the algorithms in (25) and (26) can be used to provide faster convergence at the expense of increased misadjustment (in the error correlation sense) in the final solution. A formal proof of this assertion is omitted here for the sake of space and clarity.

The stochastic EWC algorithms have linear complexity, but like any gradient method, these algorithms suffer from misadjustment when the step size does not asymptotically decay to zero. In addition, the step size plays an important role affecting both the speed of convergence and the accuracy of the asymptotic solution. The algorithm presented in the next section attempts to overcome some limitations of the gradient methods.

V. RECURSIVE ERROR WHITENING (REW) ALGORITHM

In this section, we will derive a quasi-Newton type algorithm to estimate the optimal EWC solution. This is truly a fixed-point algorithm that tracks the asymptotic solution in (5) and exhibits faster convergence to the same when compared with the stochastic gradient versions discussed before. The price paid for fast convergence is increased complexity. The complexity of REW is $O(n^2)$ (which is the same as RLS) compared with $O(n)$ of EWC-LMS.

For the sake of simplicity, assume noise-free data, and consider the optimal solution in (5). Let $\mathbf{Z}_k = \mathbf{R}_k + \beta\mathbf{S}_k$ and $\boldsymbol{\theta}_k = \mathbf{P}_k + \beta\mathbf{Q}_k$. The following recursion can be derived using the sample estimates for the correlation matrices:

$$\mathbf{Z}_k = \mathbf{Z}_{k-1} + (2\beta\mathbf{x}_k - \beta\mathbf{x}_{k-L})\mathbf{x}_k^T + \mathbf{x}_k(\mathbf{x}_k - \beta\mathbf{x}_{k-L})^T. \quad (27)$$

Recall the Sherman–Morrison–Woodbury identity, which is also known as the matrix inversion lemma [10].

$$(\mathbf{A} + \mathbf{B}\mathbf{C}\mathbf{D}^T)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{D}^T\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{D}^T\mathbf{A}^{-1}. \quad (28)$$

In (28), we define $\mathbf{A} = \mathbf{Z}_k, \mathbf{B} = [2\beta\mathbf{x}_k - \beta\mathbf{x}_{k-L} \quad \mathbf{x}_k], \mathbf{C} = \mathbf{I}_{2\times2}$, a 2×2 identity matrix, and $\mathbf{D} = [\mathbf{x}_k \quad (\mathbf{x}_k - \beta\mathbf{x}_{k-L})]$. With these definitions, we can obtain the inverse of $\mathbf{Z}_k$ as

$$\mathbf{Z}_k^{-1} = \mathbf{Z}_{k-1}^{-1} - \mathbf{Z}_{k-1}^{-1}\mathbf{B}\left(\mathbf{I}_{2\times2} + \mathbf{D}^T\mathbf{Z}_{k-1}^{-1}\mathbf{B}\right)^{-1}\mathbf{D}^T\mathbf{Z}_{k-1}^{-1}. \quad (29)$$

TABLE I
SUMMARY OF THE REW ALGORITHM

Initialize $\mathbf{Z}^{-1}(0) = c\mathbf{I}$, c is a large positive constant
$\quad \mathbf{w}(0) = \mathbf{0}$
At every iteration, compute
$\quad \mathbf{B} = [(2\beta\mathbf{x}(n) - \beta\mathbf{x}(n-L)) \quad \mathbf{x}(n)]$ and $\mathbf{D} = [\mathbf{x}(n) \quad (\mathbf{x}(n) - \beta\mathbf{x}(n-L))]$
$\quad \boldsymbol{\kappa}(n) = \mathbf{Z}^{-1}(n-1)\mathbf{B}\left(\mathbf{I}_{2x2} + \mathbf{D}^T\mathbf{Z}^{-1}(n-1)\mathbf{B}\right)^{-1}$
$\quad y(n) = \mathbf{x}^T(n)\mathbf{w}(n-1)$ and $y(n-L) = \mathbf{x}^T(n-L)\mathbf{w}(n-1)$
$\quad \mathbf{e}(n) = \begin{bmatrix} d(n) - y(n) \\ d(n) - y(n) - \beta(d(n-L) - y(n-L)) \end{bmatrix} = \begin{bmatrix} e(n) \\ e(n) - \beta e(n-L) \end{bmatrix}$
$\quad \mathbf{w}(n) = \mathbf{w}(n-1) + \boldsymbol{\kappa}(n)\mathbf{e}(n)$
$\quad \mathbf{Z}^{-1}(n) = \mathbf{Z}^{-1}(n-1) - \boldsymbol{\kappa}(n)\mathbf{D}^T\mathbf{Z}^{-1}(n-1)$

Notice that this recursion for the inverse of $\mathbf{Z}_k$ is different from the conventional RLS algorithm. It requires the inversion of a 2×2 matrix (owing to the rank-2 update) $(\mathbf{I}_{2\times 2} + \mathbf{D}^T\mathbf{Z}_{k-1}^{-1}\mathbf{B})^{-1}$, which is still trivial. The recursive estimator for $\boldsymbol{\theta}_k$ is much simpler.

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} + [(1+2\beta)d_k\mathbf{x}_k - \beta d_k\mathbf{x}_{k-L} - \beta d_{k-L}\mathbf{x}_k]. \quad (30)$$

Using (29) and (30), the optimal weight vector can be estimated as

$$\mathbf{w}_k = \mathbf{Z}_k^{-1}\boldsymbol{\theta}_k. \quad (31)$$

The above expression can be further simplified as outlined in [24]. A summary of REW is given in Table I. The derivation above assumes stationary signals. For nonstationary signals, a forgetting factor is required for tracking [1]. Inclusion of this factor in the derivation is trivial and is left out in this paper.

1) Relationship With IV Method: The REW algorithm is similar to the IV method, which was proposed as an extension to Least Squares regression [15] and can be used to estimate the parameters in white noise once the model order is known. The fundamental principle is to choose delayed regression vectors known as *instruments* that are uncorrelated with the additive white noise. Mathematically speaking, the IV method computes the solution $\mathbf{w} = E[\hat{\mathbf{x}}_k\hat{\mathbf{x}}_{k-\Delta}^T]^{-1}E[\hat{\mathbf{x}}_{k-\Delta}\hat{d}_k]$. On the other hand, the REW solution is given by $\mathbf{w}_* = \mathbf{R}_L^{-1}\mathbf{P}_L$. Notice that in REW solution, the matrix $\mathbf{R}_L$ is symmetric and Toeplitz, which is very desirable, and we exploit this fact to derive an elegant minor components based algorithm in the next subsection. Thus, in effect, the IV method can be considered to be a special case of the REW algorithm by removing the symmetric terms in $\mathbf{R}_L$ and $\mathbf{P}_L$. We will compare the performances of REW and IV methods later in this paper.

2) REW With TLS Framework: In the beginning of this paper, we mentioned that the TLS method can be used for parameter estimation when the observations are corrupted with additive white noise. However, TLS poses an inherent restriction in that the variances of the disturbances must be equal. Some extensions of TLS allow the variances to be different, but they require the knowledge of the ratio of the variances [22]. In this subsection, we will demonstrate how EWC can be reformulated using the TLS principles. Basically, TLS solves an overdetermined set of linear equations of the form $\mathbf{A}\mathbf{x} = \mathbf{b}$, where

$\mathbf{A} \in \Re^{m \times n}$ is the data matrix, $\mathbf{b} \in \Re^m$ is the desired vector, $\mathbf{x} \in \Re^n$ is the parameter vector, and m denotes the number of different observation vectors each of dimension n [18]. Alternatively, the linear equations can be written in the form $[\mathbf{A};\mathbf{b}][\mathbf{x}^T; -1] = \mathbf{0}$, where $[\mathbf{A};\mathbf{b}]$ denotes an augmented data matrix. Let $\mathbf{S}$ be the SVD of the augmented data matrix $[\mathbf{A};\mathbf{b}]$ such that $\mathbf{S} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, where $\mathbf{U}^T\mathbf{U} = \mathbf{I}_m, \mathbf{V}^T\mathbf{V} = \mathbf{I}_{n+1}$, and $\boldsymbol{\Sigma} = [\text{diag}(\sigma_1, \sigma_2, \sigma_3, \sigma_4, \ldots, \sigma_{n+1}); \mathbf{0}_{(m-n-1 \times n+1)}]$ with all singular values $\sigma_k > 0$. If $[\mathbf{A};\mathbf{b}][\mathbf{x}^T; -1] = \mathbf{0}$, the smallest singular value must be zero. This is possible only if $[\mathbf{x}^T - 1]$ is a singular vector of $[\mathbf{A};\mathbf{b}]$ (corresponding to the zero singular value) normalized such that its $(n+1)$th element value is -1. Instead of computing the smallest singular value of $[\mathbf{A};\mathbf{b}]$, we can compute the minor eigenvector of the equivalent square matrix that reduces the analytical TLS solution to

$$[\mathbf{x}; -1] = -\mathbf{v}_{n+1}/\mathbf{v}_{n+1,n+1} \quad (32)$$

where $\mathbf{v}_{n+1,n+1}$ is the last element of the minor eigenvector $\mathbf{v}_{n+1}$. The TLS technique can be easily applied to estimate the optimal MSE solution using fast minor components estimation algorithms [11], [19]. In the case of EWC, it is easy to show that the equivalent square matrix of $[\mathbf{A};\mathbf{b}]$ is

$$\mathbf{G} = \begin{bmatrix} \mathbf{R}_L & \mathbf{P}_L \\ \mathbf{P}_L^T & 2\rho_d(L) \end{bmatrix}. \quad (33)$$

The term $\rho_d(L)$ in (33) denotes the autocorrelation of the desired signal at lag L. It is important to note that the matrix (33) is square symmetric due to the symmetry of $\mathbf{R}_L$. Hence, the eigenvectors of $\mathbf{G}$ are all real which is highly desirable. In addition, it is important to note the fact that (33) still holds even with noisy data as the entries of $\mathbf{G}$ are unaffected by the noise terms. In the infinite-sample case, the matrix $\mathbf{G}$ is not full rank, and we can immediately see that one of the eigenvalues of (33) is zero. In the finite-sample case, the goal would be to find the eigenvector corresponding to the minimum absolute eigenvalue (finite samples also imply that $\mathbf{G}^{-1}$ exists). Since the eigenvalues of $\mathbf{G}$ can be both positive and negative, a typical iterative gradient or even some fixed-point type algorithms tend to become unstable [11]. A workaround would be to use the matrix $\mathbf{G}^2$ instead of $\mathbf{G}$. This will obviate the problem of having mixed eigenvalues while still preserving the eigenvectors. The squaring operation is good if the eigenvalues of $\mathbf{G}$ are well separated. Otherwise, the smaller eigenvalues blend together, making the estimation of the minor component of $\mathbf{G}^2$ more difficult. In addition, the squaring operation creates additional overhead, thereby negating any computational benefits offered by the fixed point type solutions as in [11]. Therefore, we propose to use the inverse iteration method to estimate the minor eigenvector of $\mathbf{G}$ [10]. If $\mathbf{w}_k \in \Re^{n+1}$ denotes the estimate of the minor eigenvector corresponding to the smallest absolute eigenvalue at time instant k, then the estimate at the $(k+1)$th instant is given by

$$\mathbf{w}_{k+1} = \mathbf{G}_{k+1}^{-1}\mathbf{w}_k$$
$$\mathbf{w}_{k+1} = \frac{\mathbf{w}_{k+1}}{\|\mathbf{w}_{k+1}\|}. \quad (34)$$

The term $\mathbf{G}_{k+1}$ denotes the estimate of the augmented data matrix $\mathbf{G}$ [in (33)] at the $(k+1)$th instant. It is easy to see that
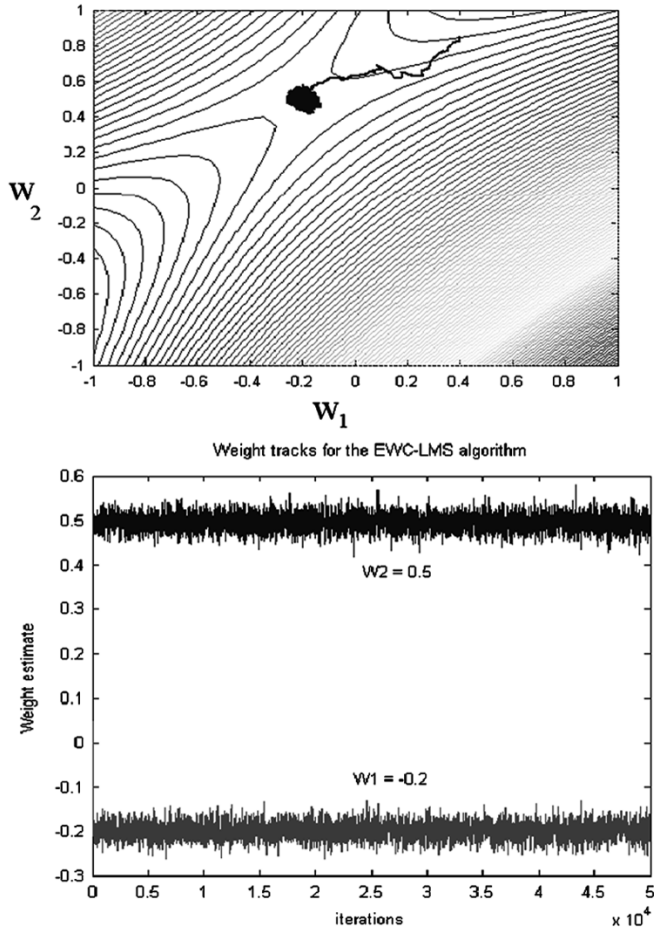
Fig. 2. Contour plot of the EWC performance surface (2-D) (top) with weight tracks and (bottom) steady-state weight tracks for the EWC-LMS algorithm.

Fig. 3. (Top) EWC performance surface (in 2-D). (Bottom) Weight tracks overlaid on the contours of EWC performance surface.

$\mathbf{G}_k$ can be recursively estimated as $\mathbf{G}_k = \mathbf{G}_{k-1} + \psi_k \psi_{k-L}^T + \psi_{k-L} \psi_k^T$, where $\psi_k = [\mathbf{x}_k; d_k]$ is the concatenated vector of the input and desired response. Now, we can invoke the inversion lemma as before and obtain a recursive $\mathrm{O}(n^2)$ estimate for matrix inversion in (34). The details of this derivation are trivial and are omitted here. Once the minor component estimate converges, i.e., $\mathbf{w}_k \to \mathbf{v}$, the EWC-TLS solution is simply given by (32). Thus, the overall complexity of the EWC-TLS algorithm is still $\mathrm{O}(n^2)$, which is the same as the REW algorithm. However, we have observed through simulations that the EWC-TLS method converges faster than the EWC-REW while preserving the accuracy of the parameter estimates.

VI. SIMULATION RESULTS AND DISCUSSION

1) Weight Tracks and Convergence of EWC-LMS: First, we will experimentally verify the convergence of EWC-LMS. For this experiment, we consider a two-tap filter with data corrupted by additive white noise. The SNR of the input signal is 10 dB, and the desired output is noise-free (since input noise is what causes bias in MSE). The top of Fig. 2 shows the contour plot of the EWC surface. Obviously, the stationary point is a saddle; in addition, the optimal solution is $\mathbf{w}_T = [-0.2 \, 0.5]^T$. A constant step size of 0.001 is used for EWC-LMS with $\beta = -1/2$. The weight tracks on the contour plot clearly show convergence to
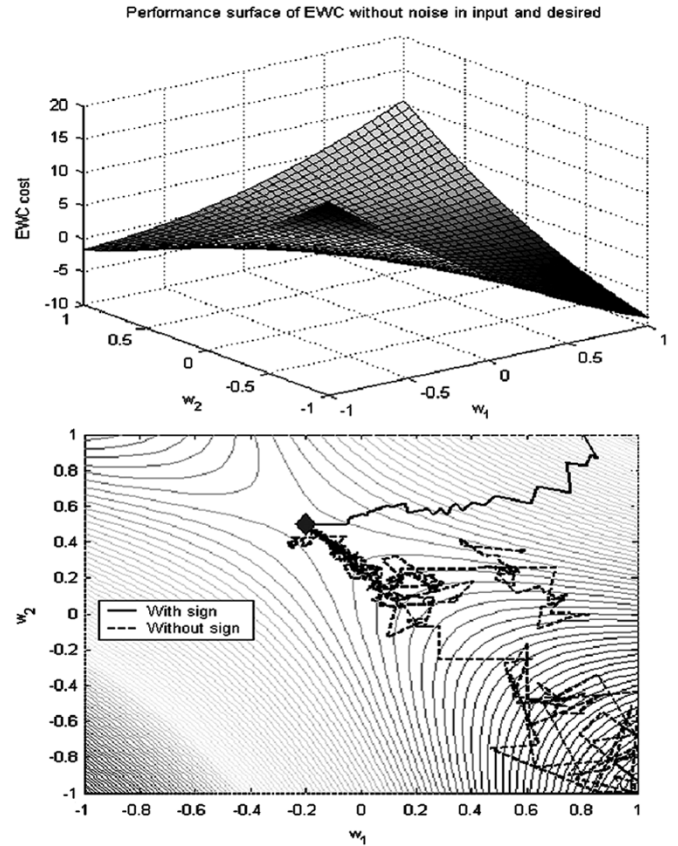
the saddle point, and the same behavior can be observed for different initial conditions. The final misadjustment can be easily controlled by the step-size similar to the LMS algorithm. The bottom of Fig. 2 shows the weights versus iterations. In order to stress the importance of the sign information for convergence to the saddle point, we performed another experiment with the same setup, this time using noise-free input and desired signals and removing the sign term from EWC-LMS. The top of Fig. 3 shows the noise-free EWC surface, and the bottom of Fig. 3 shows the weight tracks on the contours. Clearly, the weights do not converge to the desired saddle point, even in the absence of noise. On the other hand, using the sign information leads the weights to the saddle point in a stable manner. In the noise-free case, the misadjustment becomes zero.

2) Estimation of System Parameters in White Noise Using EWC-LMS: A noise-free colored input signal of 50 000 samples is passed through an *unknown* system to form the noise-free desired signal. An uncorrelated purely white noise of length 50 000 samples is added to the input signal for training. The desired training output is left noise-free since this is not critical to the performance when it is uncorrelated with the other signals. We varied the input SNR in the range [−10 dB, 10 dB]. The number of unknown and adaptive filter coefficients was varied from 4 to 12, and we performed 100 Monte Carlo runs calculating the error vector norm for each case using

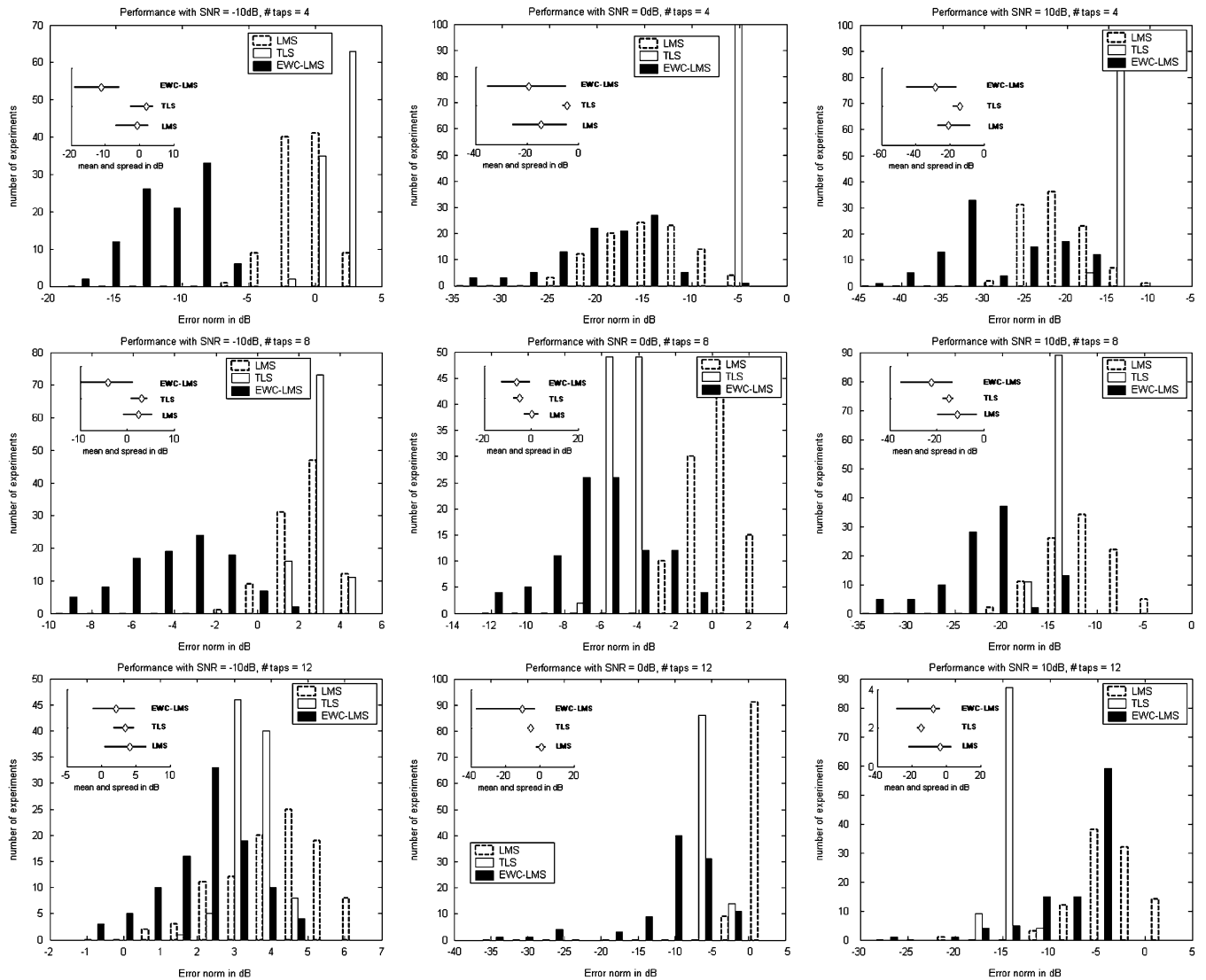$$\text{error norm} = 20 \log 10[\|\mathbf{w}_T - \mathbf{w}_*\|] \qquad (35)$$

Fig. 4.   Histogram plots showing the error vector norm for EWC-LMS, LMS algorithms and the numerical TLS solution.

where $\mathbf{w}_*$ is the solution given by EWC-LMS after one complete presentation of the training data, and $\mathbf{w}_T$ is the true weight vector. It is possible to use other statistical measures instead of the error norm, but the performance measure in (35) is sufficient to demonstrate the bias removal ability of EWC-LMS. For comparison purposes, we computed the solutions with LMS as well as the numerical TLS (regular TLS) methods. A time varying step size was chosen for EWC-LMS satisfying the bounds derived in the previous sections. For LMS, the step-size that gave the least error norm in each trial was used. The histograms of the error norms are shown in Fig. 4. The inset plots in Fig. 4 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 performing equally well for 10-dB SNR. The input noise variances for -10, 0, and 10 dB SNR values are 10, 1, and 0.1, respectively. Thus, we expect (and observe) TLS results to be worst for -10 dB and best for 10 dB. As per theory, we observe that TLS performance drops when the noise variances are not the same in the input and desired signals.

3) Estimation of System Parameters in White Noise Using REW: In order to show the performance of the EWC-REW

algorithm, we will use the same experimental setup as in the EWC-LMS case. The input noise SNR was again varied between -10, 0, and 10 dB, and we estimate the performance with different lengths of the unknown filter viz., 4, 8, and 12 taps. For every combination of SNR and filter length, 100 Monte Carlo runs were performed, and the error vector norms defined in (35) are computed. For comparison purposes, RLS and numerical TLS solutions are also evaluated. The inset plots in each subfigure of Fig. 5 summarize the mean and the spread of the error norms for all the three algorithms, whereas the error norm histograms are also provided. Clearly, REW outperforms RLS for low SNR values, and their performances become similar as SNR increases. In fact, RLS yields a useless $\mathbf{w} = \mathbf{0}$ (zero weight vector) when the SNR is lower than -10 dB. Again, TLS performs well only for similar input and output noise variances.

4) Performance Comparisons Between REW, EWC-TLS, and IV Methods: In this example, we will contrast the performances of the REW, EWC-TLS, and the IV method in a 4-tap system identification problem with noisy data. The input signal is colored and corrupted with white noise (input SNR was set at 5 dB), and the desired signal SNR is 10 dB. For the IV method, we
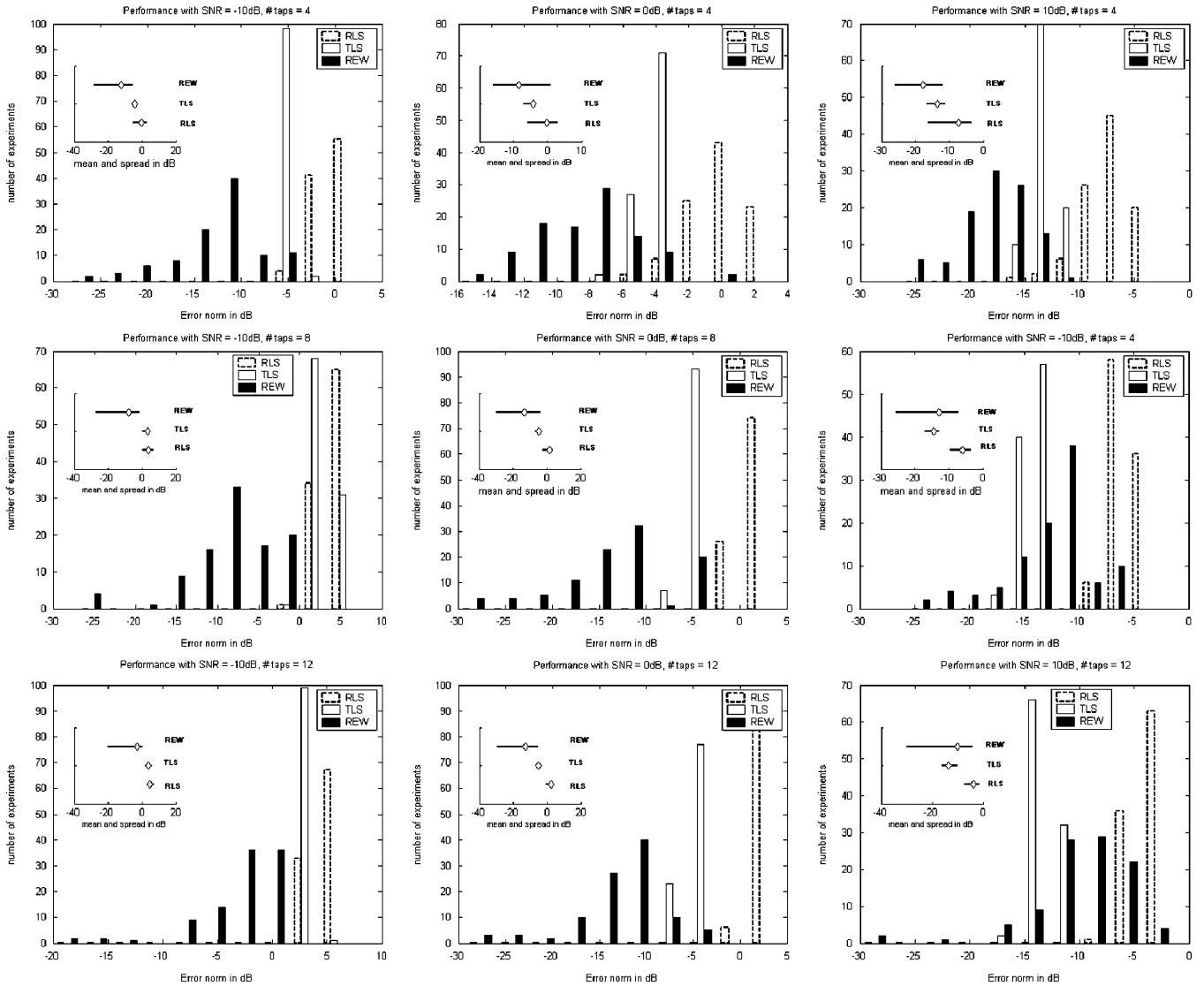
Fig. 5.   Histogram plots showing the error vector norms for REW, RLS, and numerical TLS algorithms.

chose the delayed input vector $\mathbf{x}_{k-\Delta}$ as the instrument, and the lag Δ was chosen to be four, which is the length of the true filter. Once again, the performance metric was chosen as the error vector norm in decibels given by (35). Fig. 6 shows the error histograms for REW, EWC-TLS, IV, and the optimal Wiener solutions. EWC-TLS and REW algorithms outperform the Wiener MSE solution. The IV method also produces better results than the Wiener solution. Amongst the EWC solutions, we obtained better results with the EWC-TLS algorithm [(32) and (34)] than REW. However, both EWC-TLS and REW outperformed IV method. This may be partially attributed to the conditioning of the matrices involved in the estimation of the REW and IV methods. Further theoretical analysis is required to quantify the effects of conditioning and symmetric Toeplitz structure of $\mathbf{R}_L$. In Fig. 7, we show the angle between the estimated minor eigenvector and the true eigenvector of the augmented data matrix $\mathbf{G}$ for a random single trial in scenarios with and without noise. Notice that the rates of convergence are very different. It is well known that the rate of convergence for inverse iteration method is given by the ratio $|\lambda_1/\lambda_2|$, where
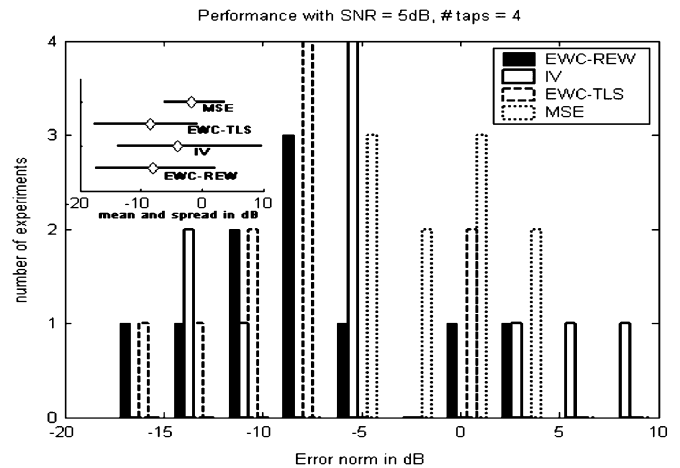

Fig. 6.   Histogram plots showing the error vector norms.

$|\lambda_1|^{-1}$ is the largest eigenvalue of $\mathbf{G}^{-1}$, and $|\lambda_2|^{-1}$ is the second largest eigenvalue of $\mathbf{G}^{-1}$ [10]. Faster convergence can be seen in the noiseless case owing to the huge $|\lambda_1/\lambda_2|$ ratio.
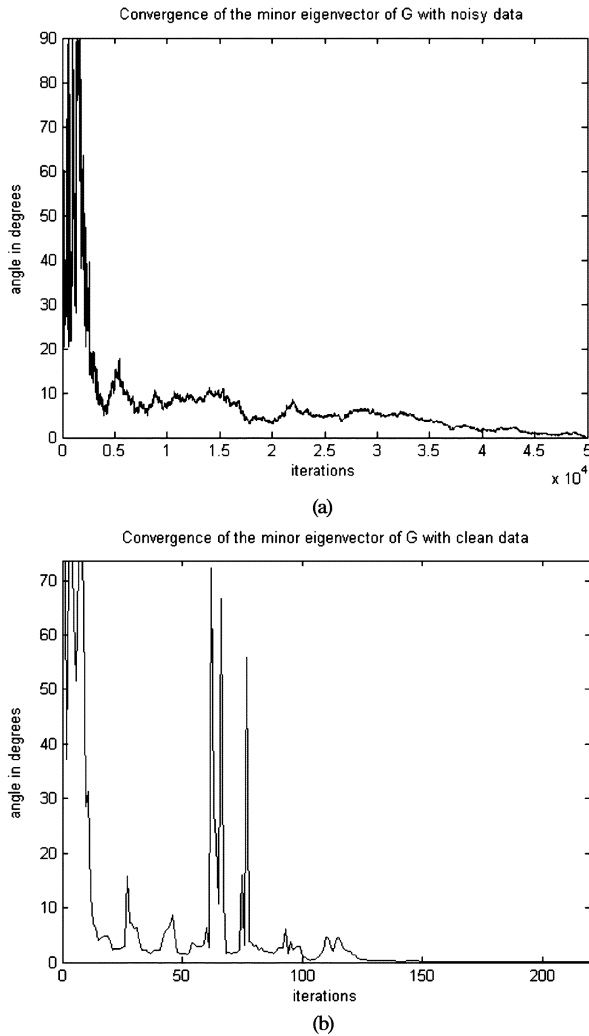
Fig. 7. Convergence of the minor eigenvector of **G** with (a) noisy data and (b) clean data.

5) Effects of Under Modeling: Estimating the model order of an unknown system is a nontrivial problem that has not had a satisfactory solution until now. In the EWC discussions and simulations, we always assumed prior information about the model order. In this section, we will briefly explore the effects of undermodeling and contrast the performance of the EWC with MSE. Theoretically speaking, in order to get an unbiased parameter estimate, we require the estimated filter to be equal to or longer than the unknown system. In the case of undermodeling, it would not be useful to compare the estimated model with the true model. However, depending on the criterion used, the parameters estimated will represent the true system in some fashion. Let us first look at what the MSE criterion would do in the case of undermodeling with and without noise in the input data. Consider noise-free input case. If we train adaptive systems with MSE, the model-mismatch error monotonically reduces to zero as we approach the true model order. At the same time, the output MSE will monotonically reduce as the filter order is increased. However, there is no one-to-one correspondence (weight matching) between the estimated model and the true system. Exact weight matching will only happen when the

input signals are white, which is neither interesting nor practical. Another fact is that the error signal with undermodeling will be uncorrelated with the input only at the lags smaller than the filter length. When the filter length is increased beyond the true system (exact or overestimation), the error becomes uncorrelated with input for all lags. Now, consider the case with noisy inputs. We know that the solution given by MSE is biased. This means that increasing the filter order will not reduce the effect of the bias. A more serious limitation of the MSE is that the solution changes with varying input noise variance. Thus, the monotonic reduction in the model mismatch is no longer true, even though the MSE decreases with increasing filter lengths. The decorrelation property of the MSE will also not hold for lags greater than the adaptive filter length (it is well known that decorrelation property is true for all lags only when the error is white).

Now, consider EWC adaptation in the undermodeling scenario. Again, the noise-free case is treated first. EWC tries to find the solution that gives zero error correlation at the specified lag. As in the case of MSE, there is no exact matching between the actual system and the estimated model. In the noisy case, EWC will still try to decorrelate the error at a specific lag. However, the error correlations at higher lags are nonzero. When the length of the filter is increased, the values of the error correlations at the higher lags decay down to zero, and the model mismatch decreases. This is a very interesting property that can be exploited to determine the correct model order of an unknown system. In the discussion on the properties of EWC, we mentioned the fact that EWC tries to orthogonalize the error with the lagged input, i.e., $E[e_k \mathbf{x}_{k-L} + e_{k-L} \mathbf{x}_k] = \mathbf{0}$. In an undermodeling scenario, this orthogonalization will be true only for the specified lag (similar to MSE). As the filter order is increased, EWC will make the error orthogonal to the input at all lags greater than the chosen lag L. This is again another property of EWC, which is not matched by MSE. Thus, we conclude that some of the nice properties of MSE are carried over by EWC (in a slightly different framework), even in the presence of input noise.

We performed some undermodeling experiments to verify the claims made in this section. Figs. 8(a) and (b) show the averaged weight error norms of MSE and EWC with input SNR values at 5 and 0 dB, respectively. The unknown system has four taps, and we can clearly see that both EWC and MSE solutions are biased when the number of filter taps is less than four. For four or more taps, EWC produces a much better estimate than the MSE (the minor variations in the error norms are because of different input/output pairs used in the Monte Carlo trials), whereas performance of MSE does not improve by increasing the order (unlike the noiseless case). Consider another example of an unknown system with six taps, and the problem is to estimate a 2-tap model for this system. The input SNR is fixed at 0 dB. The top of Fig. 9 shows a plot of the crosscorrelation between input and the error. Note that the crosscorrelation is zero for only two lags in the case of MSE, and with EWC, the error and the input are orthogonal only at the specified lag $L = 5$ (arbitrarily chosen) in this example. The bottom of Fig. 9 shows the same plot in an overestimation scenario. The key observation is that with the MSE criterion, the error is uncorrelated with the
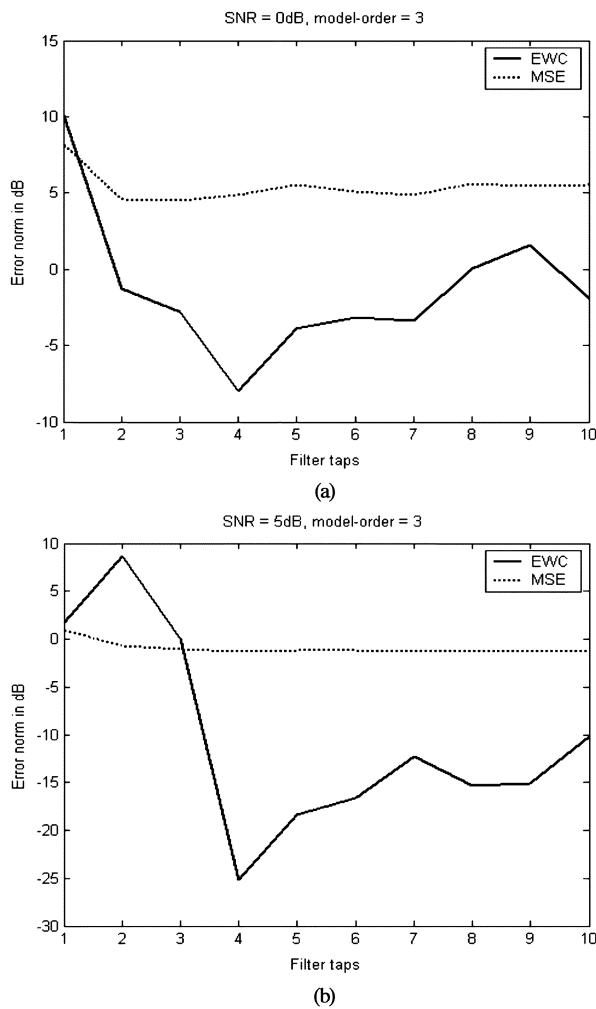
Fig. 8. Undermodeling effects with (a) input SNR $= 0$ dB and (b) input SNR $= 5$ dB.

Fig. 9. Crosscorrelation plots for EWC and MSE for (top) undermodeling and (bottom) overestimation.

input only for a few lags, whereas in the case of EWC, error and the input are uncorrelated for all lags greater than filter length. Fig. 10 shows the normalized error autocorrelation at higher lags in the overestimation case for both EWC and MSE. Notice that the error autocorrelations for EWC are very small for higher lags.

VII. CONCLUSION

MSE has been the popular criterion for adaptation owing to the efficient recursive and stochastic algorithms that work on individual samples. A serious drawback of MSE is its biased parameter estimate when the signals are corrupted with additive white noise (AWN). Several signal processing solutions exist to enhance the solution, including Total Least-Squares (TLS), that nullifies the bias if the input and output signals have equal noise variances or the extended TLS that requires the knowledge of the ratio of the noise variances. Here, we proposed a new criterion called the *Error Whitening Criterion* (EWC) that provides unbiased parameter estimates in the presence of arbitrary power AWN in both input and desired signals of interest. The theoretical foundations of EWC were laid out, and some interesting properties of the EWC cost function were explored.
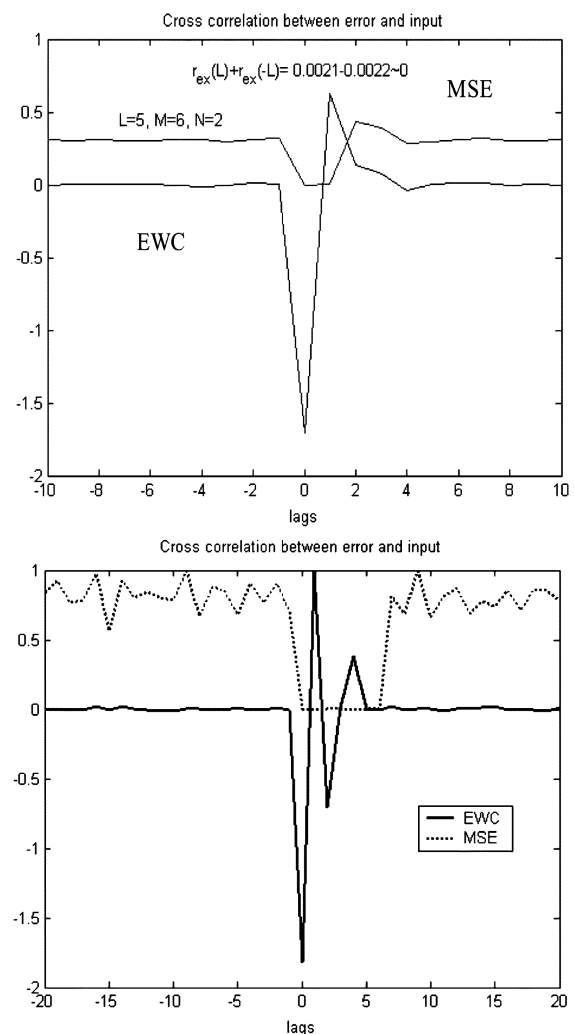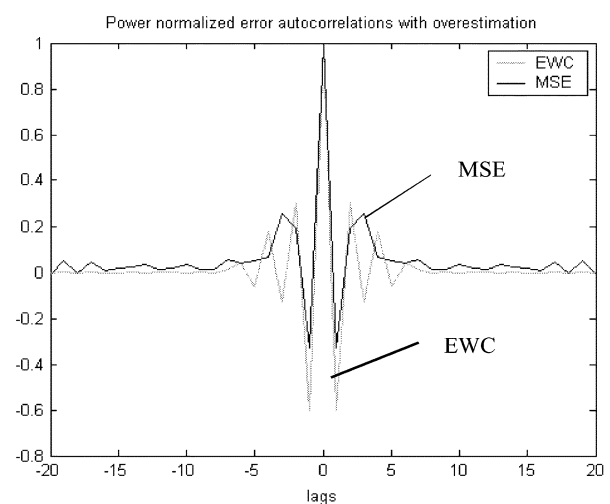

Fig. 10. Power normalized error crosscorrelation for EWC and MSE with overestimation.

We also proposed an online stochastic gradient algorithm called EWC-LMS, which is $O(n)$ in complexity and proved that the algorithm converges to the optimal EWC solution under

some constraints on the adaptation step size. We also showed that the steady-state excess error correlation was always bound from above by a positive quantity that can be made arbitrarily small by controlling the step size. The MSE criterion turns out to be a special case of the EWC, and therefore, the Least Mean Squares (LMS) algorithm is a special case of EWC-LMS. The convergence conditions and upper bounds on the excess MSE for LMS can, hence, be easily deduced from the corresponding EWC equations. We then designed a quasi-Newton type rule Recursive Error Whitening (REW) algorithm that converges to the optimal solution and tracks it at every iteration similar to the Recursive Least Squares (RLS). The complexity of this algorithm is $O(n^2)$, which is similar to that of the RLS updates. We explored the similarities between the REW algorithm and the *Instrumental Variables* (IV) method, which has been used for parameter estimation in white noise. The REW algorithm has better structure than the recursive IV method and allows the derivation of a fast minor components based EWC-TLS algorithm. The EWC-TLS algorithm adopts the inverse iteration method for estimating the minor component.

Extensive simulations showed the superiority of the EWC methods over the MSE equivalents. We have also successfully applied EWC in an inverse modeling problem and controller design [24]. One of the limitations of the EWC is the assumption that the noise must be white. This can be restrictive in certain applications, and we are working on modifications to EWC to handle colored noise. The initial results have so far been very promising [25]. Another aspect is to generalize the application of EWC to the identification of nonlinear systems. This is a very challenging problem that has not received much attention in signal processing research. Future work will also be directed toward designing hypothesis testing methods that monitor the error autocorrelations at higher lags to make decisions about the true model order.

APPENDIX A

Recall that the optimal solution of EWC satisfies

$$E[(1+2\beta)e_k\mathbf{x}_k - \beta(e_k\mathbf{x}_{k-L} + e_k\mathbf{x}_{k+L})] = 0. \quad \text{(A.1)}$$

Rearranging the terms in (A.1), we obtain

$$E[e_k(\mathbf{x}_k - \beta(\mathbf{x}_{k+L} - 2\mathbf{x}_k + \mathbf{x}_{k-L}))] = 0. \quad \text{(A.2)}$$

Notice that $(\mathbf{x}_{k+L} - 2\mathbf{x}_k + \mathbf{x}_{k-L})$ forms an estimate of the *acceleration* of the input vector $\mathbf{x}_k$. Specifically for $\beta = -1/2$, the term that multiplies e_k becomes a single-step prediction for the input vector $\mathbf{x}_k$ (assuming zero velocity and constant acceleration), according to Newtonian mechanics. Thus, the optimal solution of EWC tries to decorrelate the error signal from the predicted next input vector.

APPENDIX B

The dynamics of the error vector norm is given by

$$\|\varepsilon_{k+1}\|^2 = \|\varepsilon_k\|^2 - 2\eta \text{sign}\left(e_k^2 + \beta\dot{e}_k^2\right)\varepsilon_k^T(e_k\mathbf{x}_k + \beta\dot{e}_k\dot{\mathbf{x}}_k) + \eta^2\|e_k\mathbf{x}_k + \beta\dot{e}_k\dot{\mathbf{x}}_k\|^2. \quad \text{(B.1)}$$

Further, since $\varepsilon_k^T\mathbf{x}_k = e_k$ and $\varepsilon_k^T\dot{\mathbf{x}}_k = \dot{e}_k$, we have

$$\|\varepsilon_{k+1}\|^2 = \|\varepsilon_k\|^2 - 2\eta \left|e_k^2 + \beta\dot{e}_k^2\right| + \eta^2\|e_k\mathbf{x}_k + \beta\dot{e}_k\dot{\mathbf{x}}_k\|^2. \quad \text{(B.2)}$$

Let

$$\varphi = 2 - \eta \frac{\|e_k\mathbf{x}_k + \beta\dot{e}_k\dot{\mathbf{x}}_k\|^2}{|e_k^2 + \beta\dot{e}_k^2|}. \quad \text{(B.3)}$$

If the step-size (positive) upper bound in (10) is satisfied, then $\varphi > 0$ for all k. Therefore, (B.3) reduces to the inequality

$$\|\varepsilon_{k+1}\|^2 \le \|\varepsilon_k\|^2 - \eta\varphi \left|e_k^2 + \beta\dot{e}_k^2\right|. \quad \text{(B.4)}$$

Iterating (B.4) from $k = 0$, we get $\|\varepsilon_k\|^2 \le \|\varepsilon_0\|^2 - \eta\varphi\sum_{t=1}^k |e_t^2 + \beta\dot{e}_t^2|$. In the limit $k \to \infty$, it is easy to see that $\|\varepsilon_\infty\|^2 + \varphi\eta\sum_{t=1}^\infty |e_t^2 + \beta\dot{e}_t^2| \le \|\varepsilon_0\|^2$, which implies $\lim_{t\to\infty}|e_t^2 + \beta\dot{e}_t^2| = 0$ as the summation in the error terms must converge to a finite value given by $\|\varepsilon_0\|^2 - \|\varepsilon_\infty\|^2$. The instantaneous cost $|e_t^2 + \beta\dot{e}_t^2|$ becomes zero only when the weights converge to the true weights $\mathbf{w}_T(\therefore \|\varepsilon_\infty\|^2 = 0)$. In addition, note that the gradient becomes zero at this point.

REFERENCES

[1] B. Farhang-Boroujeny, *Adaptive Filters: Theory and Applications*. New York: Wiley, 1998.
[2] S. Haykin, *Adaptive Filter Theory*. Upper Saddle River, NJ: Prentice-Hall, 1996.
[3] J. A. Cadzow, "Total least squares, matrix enhancement, and signal processing," *Digital Signal Process.*, vol. 4, pp. 21–39, 1994.
[4] P. Lemmerling, "Structured Total Least Squares: Analysis, Algorithms, and Applications," Ph.D. Dissertation, Katholeike Univ. Leuven, Leuven, Belgium, 1999.
[5] A. Yeredor, "The extended least squares criterion: Minimization algorithms and applications," *IEEE Trans. Signal Process.*, vol. 49, no. 1, pp. 74–86, Jan. 2000.
[6] H. C. So, "Modified LMS algorithm for unbiased impulse response estimation in nonstationary noise," *Electron. Lett.*, vol. 35, no. 10, pp. 791–792, 1999.
[7] K. Gao, M. O. Ahmad, and M. N. S. Swamy, "A constrained anti-hebbian learning algorithm for total least squares estimation with applications to adaptive FIR and IIR filtering," *IEEE Trans. Circuits Syst. II*, vol. 41, no. 11, pp. 718–729, Nov. 1994.
[8] D. Z. Feng, Z. Bao, and L. C. Jiao, "Total least mean squares algorithm," *IEEE Trans. Signal Process.*, vol. 46, no. 8, pp. 2122–2130, Aug. 1998.
[9] G. H. Golub and C. F. van Loan, "An analysis of the total least squares problem," *SIAM J. Numer. Anal.*, vol. 17, no. 4, pp. 883–893, 1979.
[10] ——, *Matrix Computations*. Baltimore, MD: Johns Hopkins Univ. Press, 1989.
[11] Y. N. Rao and J. C. Principe, "Efficient total least squares method for system modeling using minor component analysis," in *Proc. IEEE Workshop Neural Networks Signal Process. XII*, 2002, pp. 259–268.
[12] ——, "A fast, on-line algorithm for PCA and its convergence characteristics," in *Proc. IEEE Workshop Neural Networks Signal Process. X*, vol. 1, 2000, pp. 299–307.
[13] Y. N. Rao and J. C. Principe, "Time series segmentation using a novel adaptive eigendecomposition algorithm," *J. VLSI Signal Process.*, vol. 32, pp. 7–17, 2002.

[14] B. de Moor, "Total least squares for affinely structured matrices and the noisy realization problem," *IEEE Trans. Signal Process.*, vol. 42, no. 11, pp. 3104–3113, Nov. 1994.

[15] T. Söderström and P. Stoica, *System Identification*. London, U.K.: Prentice-Hall, 1989.

[16] Y. N. Rao, "Algorithms for Eigendecomposition and Time Series Segmentation," M.S. Thesis, Univ. Florida, Gainesville, FL, 2000.

[17] T. Y. Al-Naffouri and A. H. Sayed, "Adaptive filters with error nonlinearities: Mean-square analysis and optimum design," *EURASIP J. Appl. Signal Process.*, no. 4, pp. 192–205, 2001.

[18] F. Deprettere, *SVD and Signal Processing—Algorithms, Applications, and Architectures*. Amsterdam, The Netherlands: North-Holland, 1988.

[19] C. E. Davila, "An efficient recursive total least squares algorithm for FIR adaptive filtering," *IEEE Trans. Signal Process.*, vol. 42, no. 2, pp. 268–280, Feb. 1994.

[20] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*. New York: Wiley, 2002.

[21] M. Mueller, "Least-squares algorithms for adaptive equalizers," *Bell Syst. Tech. J.*, vol. 60, pp. 1905–1925, 1981.

[22] M. Chansarkar and U. B. Desai, "A robust recursive least squares algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 7, pp. 1726–1735, Jul. 1997.

[23] M. Reuter, K. Quirk, J. Zeidler, and L. Milstein, "Nonlinear effects in LMS adaptive filters," in *Proc. AS-SPCC*, Calgary, AB, Canada, 2000.

[24] Y. N. Rao, D. Erdogmus, G. Y. Rao, and J. C. Principe, "Fast error whitening algorithms for system identification and control," in *Proc. NNSP*, Toulouse, France, Sept. 2003, pp. 309–318.

[25] Y. N. Rao, D. Erdogmus, and J. C. Principe, "Accurate parameter estimation in colored noise," in *Proc. ICASSP*, 2004, to be published.

Yadunandana N. Rao was born in Mysore, India. He received the B.E. degree in electronics and communication engineering degree from the University of Mysore in August 1997 and the M.S. degree in electrical and computer engineering from the University of Florida, Gainesville, in May 2000. He is currently pursuing the Ph.D. degree at the University of Florida, with the Computational NeuroEngineering Laboratory under the guidance of Dr. J. C. Principe.

Between August 1997 and July 1998, he was a software engineer in Bangalore, India. From May 2000 to January 2001, he was a Design Engineer at GE Medical Systems, Waukesha, WI. His research interests include adaptive signal processing theory, algorithms, and analysis, neural networks for signal processing, and biomedical applications.

Deniz Erdogmus (M'00) received the B.S. degree in electrical and electronics engineering and mathematics in 1997 and the M.S. degree in electrical and electronics engineering, with emphasis on systems and control, in 1999, all from the Middle East Technical University, Ankara, Turkey. He received the Ph.D. degree in electrical and computer engineering from the University of Florida, Gainesville, in 2002.

His current research interests include applications of information theory to adaptive systems, signal processing, communications, and control.

Dr. Erdogmus is a member of Tau Beta Pi and Eta Kappa Nu.

Jose C. Principe (F'00) is a Distinguished Professor of Electrical and Computer Engineering and Biomedical Engineering at the University of Florida, Gainesville, where he teaches advanced signal processing, machine learning and artificial neural networks (ANNs) modeling. He is BellSouth Professor and the Founder and Director of the University of Florida Computational NeuroEngineering Laboratory (CNEL). His primary area of interest is processing of time-varying signals with adaptive neural models. The CNEL Lab has been studying signal and pattern recognition principles based on information theoretic criteria (entropy and mutual information).

Dr. Principe is a member of the ADCOM of the IEEE Signal Processing Society, Member of the Board of Governors of the International Neural Network Society, and Editor-in-Chief of the IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING. He is a member of the Advisory Board of the University of Florida Brain Institute and has more than 90 publications in refereed journals, ten book chapters, and 200 conference papers. He has directed 35 Ph.D. dissertations and 45 Master theses. He recently wrote an interactive electronic book entitled *Neural and Adaptive Systems: Fundamentals Through Simulation* (New York: Wiley).