Error Whitening Criterion For Linear Filter Estimation

Yadunandana N. Rao, Deniz Erdogmus, Jose C. Principe
Computational NeuroEngineering Laboratory
Department of Electrical and Computer Engineering
University of Florida, Gainesville, FL-32611

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 Error Whitening Criterion (EWC) to tackle the problem of linear system identification in the presence of additive white disturbances. We will motivate the theory behind the new criterion and derive an online stochastic gradient algorithm based on EWC. Convergence proof of the stochastic gradient algorithm is derived making mild assumptions. Simulation results show the effectiveness of this criterion. We will compare its performance with MSE as well as the powerful Total-Least Squares method.**

I. 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 [1]. 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 [1]. 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 [1]. 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 [2]. 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) [3] 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 [2], [4]. In this paper, we will present a completely different approach that would partially whiten the error sequence at the output of an adaptive filter even in the presence of noisy inputs. A new criterion is formulated that enforces zero autocorrelation of the error signal beyond a certain lag; hence the name *Error Whitening Criterion* (EWC). In the next section, we will motivate the theory of EWC and state some of its interesting properties

II. ERROR WHITENING CRITERION THEORY

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 autocorrelation is always biased by the noise power. Instead, we propose to analyze the error autocorrelation at a non-zero lag. 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 on the input, d_k being the noise-free desired signal and u_k being the additive white noise added to the desired signal. We further assume that the noises $\mathbf{v}_k$ 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 *N*. Since $d_k = \mathbf{x}_k^T \mathbf{w}_T$, the error is $\hat{e}_k = \mathbf{x}_k^T(\mathbf{w}_T - \mathbf{w}) + u_k - \mathbf{v}_k{}^T\mathbf{w}$. Error autocorrelation at some arbitrary lag L is given by,

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

From (1), it is obvious that if $L \geq N$, where N is the length of the true filter $\mathbf{w}_T$, $E[\mathbf{v}_k \mathbf{v}_{k-L}^T] = \mathbf{0}$. 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 $L \geq N$ 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 $L \ge N$. 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 [5], we can write the product $\hat{e}_k \hat{e}_{k-L}$ as,

$$\hat{e}_k \hat{e}_{k-L} = \frac{1}{2}[\hat{e}_k^2 + \hat{e}_{k-L}^2 - (\hat{e}_k - \hat{e}_{k-L})^2] \quad (2)$$

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

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

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

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

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

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

The form in (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 (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. The stochastic instantaneous gradient of the EWC cost function in (5) is,

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

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

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

where, $\eta > 0$ is a small step-size parameter. Note that when $\beta = 0$, (7) reduces to the renowned LMS algorithm [6]. We are specifically interested in using (7) with $\beta = -0.5$. In the next section, we will present the convergence analysis of (7) and derive some useful results.

III. Convergence Analysis

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

$$0 < \eta < \frac{2\left| e_k^2 + \beta \dot{e}_k^2 \right|}{\left\| e_k \mathbf{x}_k + \beta \dot{e}_k \dot{\mathbf{x}}_k \right\|^2} \forall k \quad (8)$$

***Proof*.** From the arguments presented in the previous section and owing to the quadratic nature of the EWC performance surface, it is clear that the EWC-LMS algorithm in (7) has a single stationary point (global minimum) $\mathbf{w}_* = \mathbf{w}_T$. 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 (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, $\|\boldsymbol{\varepsilon}_{k+1}\|^2 = \|\boldsymbol{\varepsilon}_k\|^2 - 2\eta sign(e_k^2 + \beta \dot{e}_k^2)\boldsymbol{\varepsilon}_k^T(e_k \mathbf{x}_k + \beta \dot{e}_k \dot{\mathbf{x}}_k) + \eta^2 \|e_k \mathbf{x}_k + \beta \dot{e}_k \dot{\mathbf{x}}_k\|^2$. In case of noise-free data, $\boldsymbol{\varepsilon}_k^T \mathbf{x}_k = e_k$ and $\boldsymbol{\varepsilon}_k^T \dot{\mathbf{x}}_k = \dot{e}_k$. Using these two equations we get,

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

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 (8). The error vector will eventually converge to zero, i.e., $\lim_{k\to\infty} \|\boldsymbol{\varepsilon}_k\|^2 \to 0$, which implies that $\lim_{k\to\infty} \mathbf{w}_k \to \mathbf{w}_* = \mathbf{w}_T$. □

Observe that when $\beta = 0$, the upper bound on the step-size in (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 II: In the noisy data case, EWC-LMS given in (7) with $\beta = -0.5$ converges to the stationary point $\mathbf{w}_* = \mathbf{w}_T$ in the mean provided that the step size is bound by the inequality

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

Proof. Again, it is clear that the only stationary point of (7) with $\beta = -0.5$ is $\mathbf{w}_* = \mathbf{w}_T$ even in the presence of noise where $\mathbf{w}_T$ 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,

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

Apply the expectation operator on both sides of (11) and let $E\|\boldsymbol{\varepsilon}_{k+1}\|^2 < E\|\boldsymbol{\varepsilon}_k\|^2$ as in the previous case. With further simplifications and assuming that $\eta > 0$, we get

$$\frac{\eta}{2}E\left\|\hat{e}_k\hat{\mathbf{x}}_k + \beta\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k\right\|^2 < E\left|\boldsymbol{\varepsilon}_k^T(\hat{e}_k\hat{\mathbf{x}}_k + \beta\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k)\right| \tag{12}$$

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

$$0 < \eta < 2\frac{\left|E[\boldsymbol{\varepsilon}_k^T(\hat{e}_k\hat{\mathbf{x}}_k + \beta\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k)]\right|}{E\left\|\hat{e}_k\hat{\mathbf{x}}_k + \beta\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k\right\|^2} \tag{13}$$

Since we switched the order of expectation operator and absolute value function, we can further simplify the bound in (13). The evaluation of the terms $E(\boldsymbol{\varepsilon}_k^T\hat{e}_k\hat{\mathbf{x}}_k)$ and $E(\boldsymbol{\varepsilon}_k^T\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k)$ are tedious and is omitted here. It can be shown that,

$$\begin{aligned}E(\boldsymbol{\varepsilon}_k^T\hat{e}_k\hat{\mathbf{x}}_k) &= \boldsymbol{\varepsilon}_k^T\mathbf{R}\boldsymbol{\varepsilon}_k - \boldsymbol{\varepsilon}_k^T\mathbf{V}\mathbf{w}_k \\ E(\boldsymbol{\varepsilon}_k^T\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k) &= \boldsymbol{\varepsilon}_k^T(2\mathbf{R} - \mathbf{R}_L)\boldsymbol{\varepsilon}_k - 2\boldsymbol{\varepsilon}_k^T\mathbf{V}\mathbf{w}_k\end{aligned} \tag{14}$$

where, $\mathbf{R} = E[\mathbf{x}_k\mathbf{x}_k^T], \mathbf{R}_L = E[\mathbf{x}_k\mathbf{x}_{k-L}^T + \mathbf{x}_{k-L}\mathbf{x}_k^T], \mathbf{V} = E[\mathbf{v}_k\mathbf{v}_k^T]$. 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 $\beta = -0.5$,

$$E(\boldsymbol{\varepsilon}_k^T\hat{e}_k\hat{\mathbf{x}}_k) - 0.5E(\boldsymbol{\varepsilon}_k^T\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k) = \frac{1}{2}\boldsymbol{\varepsilon}_k^T\mathbf{R}_L\boldsymbol{\varepsilon}_k \tag{15}$$

Using $\boldsymbol{\varepsilon}_k = \mathbf{w}_* - \mathbf{w}_k = \mathbf{w}_T - \mathbf{w}_k$ and $d_k = \mathbf{x}_k^T\mathbf{w}_T$, (15) can be further reduced to

$$E(\boldsymbol{\varepsilon}_k^T\hat{e}_k\hat{\mathbf{x}}_k) - 0.5E(\boldsymbol{\varepsilon}_k^T\hat{\hat{e}}_k\hat{\hat{\mathbf{x}}}_k) = E(e_k e_{k-L}) = E(\hat{e}_k^2 - 0.5\hat{\hat{e}}_k^2) \tag{16}$$

Substituting the numerator of (13) with the above result, we immediately get the upper bound in (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. Therefore, $\lim_{k\to\infty} E\|\boldsymbol{\varepsilon}_k\|^2 \to 0$ and $\lim_{k\to\infty} E(\mathbf{w}_k) \to \mathbf{w}_* = \mathbf{w}_T$. □

We would like to mention that the upper bound on step-size given by (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.

We can deduce a normalized EWC-LMS algorithm from this result in (10) that would give a constant data independent upper bound on the step-size. The normalized EWC-LMS algorithm can be derived using the principle of minimum norm updates [1] and will be provided in a later paper.

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 [1].

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

$$|\rho_{\hat{e}}(L)| \leq \frac{\eta}{2}E(\hat{e}_\infty^2)[Tr(\mathbf{R}+\mathbf{V})] + 2\eta[\sigma_u^2 + \|\mathbf{w}_\infty\|\|\mathbf{w}_*\|Tr(\mathbf{V})] \tag{17}$$

where $\mathbf{R} = E[\mathbf{x}_k\mathbf{x}_k^T]$, and $\mathbf{V} = E[\mathbf{v}_k\mathbf{v}_k^T]$ and $Tr(\bullet)$ 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 (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 (11) will give,

$$E\|\boldsymbol{\varepsilon}_{k+1}\|^2 = E\|\boldsymbol{\varepsilon}_k\|^2 - 2\eta E[sign(\hat{e}_k^2 - 0.5\hat{\dot{e}}_k^2)\boldsymbol{\varepsilon}_k^T(\hat{e}_k\hat{\mathbf{x}}_k - 0.5\hat{\dot{e}}_k\hat{\dot{\mathbf{x}}}_k)]$$
$$+\eta^2 E\left\|\hat{e}_k\hat{\mathbf{x}}_k - 0.5\hat{\dot{e}}_k\hat{\dot{\mathbf{x}}}_k\right\|^2 \quad (18)$$

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

$$E\|\boldsymbol{\varepsilon}_{k+1}\|^2 = E\|\boldsymbol{\varepsilon}_k\|^2 + \eta^2 E\left\|\hat{e}_k\hat{\mathbf{x}}_k - 0.5\hat{\dot{e}}_k\hat{\dot{\mathbf{x}}}_k\right\|^2$$
$$+2\eta E[sign(\hat{e}_k^2 - 0.5\hat{\dot{e}}_k^2)[\mathbf{w}_*^T(\mathbf{v}_k\mathbf{v}_k^T - 0.5\dot{\mathbf{v}}_k\dot{\mathbf{v}}_k^T)\mathbf{w}_k]]$$
$$+2\eta E[sign(\hat{e}_k^2 - 0.5\hat{\dot{e}}_k^2)(u_k^2 - 0.5\dot{u}_k^2)] - 2\eta E\left|\hat{e}_k^2 - 0.5\hat{\dot{e}}_k^2\right| \quad (19)$$

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

$$E\left|(\hat{e}_k^2 - 0.5\hat{\dot{e}}_k^2)\right| = \frac{\eta}{2}E\left\|\hat{e}_k\hat{\mathbf{x}}_k - 0.5\hat{\dot{e}}_k\hat{\dot{\mathbf{x}}}_k\right\|^2 + \eta \times$$
$$E\left[sign(\hat{e}_k^2 - 0.5\hat{\dot{e}}_k^2)[\mathbf{w}_*^T(\mathbf{v}_k\mathbf{v}_k^T - 0.5\dot{\mathbf{v}}_k\dot{\mathbf{v}}_k^T)\mathbf{w}_k + u_k^2 - 0.5\dot{u}_k^2]\right] \quad (20)$$

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

$$\left|\rho_{\hat{e}}(L)\right| \leq \frac{\eta}{2}E\left\|\hat{e}_k\hat{\mathbf{x}}_k - 0.5\hat{\dot{e}}_k\hat{\dot{\mathbf{x}}}_k\right\|^2 + \eta \times$$
$$E\left[sign(\hat{e}_k^2 - 0.5\hat{\dot{e}}_k^2)[\mathbf{w}_*^T(\mathbf{v}_k\mathbf{v}_k^T - 0.5\dot{\mathbf{v}}_k\dot{\mathbf{v}}_k^T)\mathbf{w}_k + u_k^2 - 0.5\dot{u}_k^2]\right] \quad (21)$$

Note that, we used the relation $\rho_{\hat{e}}(L) = \left|E(\hat{e}_k^2 - 0.5\hat{\dot{e}}_k^2)\right|$ in the above equation. The first term on the RHS of (21) can be easily evaluated by invoking the assumption that $\|\hat{\mathbf{x}}_k\|^2$ and $\hat{e}_k^2$ are uncorrelated in steady state as,

$$E\left\|\hat{e}_k\hat{\mathbf{x}}_k - 0.5\hat{\dot{e}}_k\hat{\dot{\mathbf{x}}}_k\right\|^2 = E(\hat{e}_k^2)[Tr(\mathbf{R}) + Tr(\mathbf{V})] \quad (22)$$

The above assumption is commonly used in computing the steady state excess-MSE for stochastic LMS algorithm [7]. Importantly, this assumption is less restrictive and more natural when compared to the independence theory that was frequently used in the past [1]. The second term in RHS of (21) is involved and has no closed form expression even with Gaussianity assumptions that are typically made in the analysis of sign-LMS algorithm [7]. Even the validity of Gaussianity assumption is questionable as discussed by Eweda [8] who proposed additional, reasonable constraints on the noise probability density function to overcome the Gaussianity and independence assumptions [8] 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. With this closing argument, we proceed by rewriting (21) using the identity $E[sign(a)b] \leq E|b|$ as,

$$\left|\rho_{\hat{e}}(L)\right| \leq \frac{\eta}{2}E(\hat{e}_k^2)[Tr(\mathbf{R}) + Tr(\mathbf{V})] + \eta$$
$$E\left|[\mathbf{w}_*^T(\mathbf{v}_k\mathbf{v}_k^T - 0.5\dot{\mathbf{v}}_k\dot{\mathbf{v}}_k^T)\mathbf{w}_k + u_k^2 - 0.5\dot{u}_k^2]\right| \quad (23)$$

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

$$E\left|[u_k^2 - 0.5\dot{u}_k^2]\right| \leq E(u_k^2) + 0.5E(\dot{u}_k^2) = 2\sigma_u^2 \quad (24)$$

Similarly,

$$E\left|\mathbf{w}_*^T(\mathbf{v}_k\mathbf{v}_k^T - 0.5\dot{\mathbf{v}}_k\dot{\mathbf{v}}_k^T)\mathbf{w}_k\right| \leq E\left|\mathbf{w}_*^T\mathbf{v}_k\mathbf{v}_k^T\mathbf{w}_k\right| +$$
$$0.5E\left|\mathbf{w}_*^T\dot{\mathbf{v}}_k\dot{\mathbf{v}}_k^T\mathbf{w}_k\right| \quad (25)$$

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

$$\mathbf{w}_*^T\mathbf{v}_k\mathbf{v}_k^T\mathbf{w}_k \leq \|\mathbf{w}_k\|\|\mathbf{w}_*\|\|\mathbf{v}_k\|^2$$
$$\mathbf{w}_*^T\dot{\mathbf{v}}_k\dot{\mathbf{v}}_k^T\mathbf{w}_k \leq \|\mathbf{w}_k\|\|\mathbf{w}_*\|\|\dot{\mathbf{v}}_k\|^2 \quad (26)$$

We know that $E[\mathbf{v}_k\mathbf{v}_{k-L}^T]$ is $\mathbf{0}$ for $L \geq N$. Therefore,

$$E\left|\mathbf{w}_*^T\mathbf{v}_k\mathbf{v}_k^T\mathbf{w}_k\right| + 0.5E\left|\mathbf{w}_*^T\dot{\mathbf{v}}_k\dot{\mathbf{v}}_k^T\mathbf{w}_k\right| \leq 2\|\mathbf{w}_*\|\|\mathbf{w}_k\|Tr(\mathbf{V}) \quad (27)$$

Using (24) and (27) in (23), and letting $k \to \infty$, we see that,

$$\left|\rho_{\hat{e}}(L)\right| \leq \frac{\eta}{2}E(\hat{e}_\infty^2)[Tr(\mathbf{R}+\mathbf{V})] + 2\eta[\sigma_u^2 + \|\mathbf{w}_\infty\|\|\mathbf{w}_*\|Tr(\mathbf{V})] \quad (28)$$

The term $E(\hat{e}_\infty^2)$ represents the residual MSE and is given by,

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

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

It is important to observe that by reducing the step-size, one can arbitrarily reduce the steady state excess error autocorrelation at lag $L \geq N$. We have confirmed this fact by extensive simulations and the results have been reported in [9].

IV. CASE STUDY

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 50000 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 –10dB, 0dB and 10dB. 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,

$$error\ \ norm = 20\log10\left[\left\|\mathbf{w}_T - \mathbf{w}_*\right\|\right] \tag{29}$$

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 (29). Fig. 1 shows the histograms of the error vector norms for all three methods. The inset plots in Fig. 1 show the summary of the histograms for each method. EWC-LMS performs significantly better than LMS at low SNR values (-10dB and 0dB), while performances are on par for SNR greater than 10dB. 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.

V. CONCLUSIONS

In this paper, we proposed a new criterion called the Error Whitening Criterion (EWC), which includes MSE as a special case. MSE and Total Least Squares (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 $O(N)$ complexity. 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. Extensive Monte-Carlo simulations were carried out to show the superiority of the new criterion in a FIR system identification problem. 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

[1] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, Upper Saddle River, New Jersey, 1996.

[2] S.C. Douglas, "Analysis of an Anti-Hebbian Adaptive FIR Filtering Algorithm". IEEE Trans. Circuits and Systems-II: Analog and Digital Signal Processing, vol. 43, no. 11, 1996.

[3] G.H. Golub, C.F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1989.

[4] Y.N. Rao, J.C. Principe, "Efficient Total Least Squares Method for System Modeling Using Minor Component Analysis," Proceedings of NNSP'02, pp. 259-268, 2002.

[5] F.J. Taylor, J. Mellott, *Hands-On Digital Signal Processing*, McGraw-Hill, 1998.

[6] B. Widrow, S. D. Stearns. *Adaptive Signal Processing*, Prentice-Hall, New Jersey, 1985.

[7] T. Y. Al-Naffouri and A. H. Sayed, "Adaptive Filters with Error Nonlinearities: Mean-Square Analysis and Optimum Design," EURASIP Jour. App. Sig. Proc., no. 4, pp. 192-205, 2001.

[8] E. Eweda, "Convergence Analysis of the Sign Algorithm Without the Independence and Gaussian Assumptions," IEEE Trans. Sig. Proc., vol. 48, no. 9, September 2000.

[9] J.C. Principe, Y.N. Rao, D. Erdogmus, "Error Whitening Wiener Filters: Theory and Algorithms", to appear as chapter 10 in Advances in LMS Filters, ed., S. Haykin, B. Widrow.

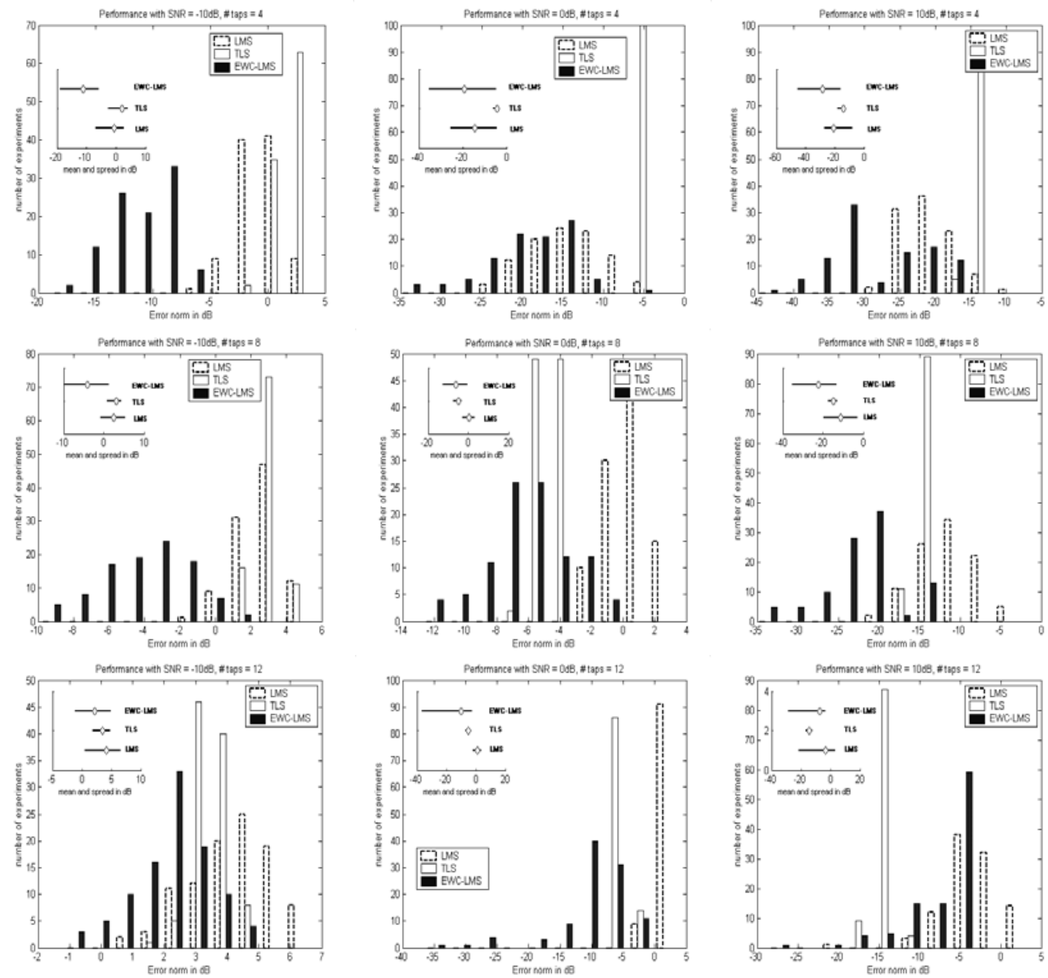

Figure 1- Histogram plots showing the error vector norm for EWC-LMS, LMS algorithms and the numerical TLS solution.