FAST ERROR WHITENING ALGORITHMS FOR SYSTEM IDENTIFICATION AND CONTROL WITH NOISY DATA

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

Computational NeuroEngineering Laboratory

Electrical and Computer Engineering Department

University of Florida, NEB 486, Gainesville, FL 32611-6130

[yadu, deniz, geethak, principe]@cnel.ufl.edu

Keywords: Error whitening, RLS, TLS, GTLS, System Identification, Control

Abstract. Linear system identification with noisy input/output is a critical problem in signal processing and control. Conventional techniques based on the Mean Squared-Error (MSE) criterion can at best provide a biased parameter estimate of the unknown system being modeled. Recently, we proposed a new criterion called the Error Whitening Criterion (EWC) to solve the problem of linear parameter estimation in the presence of additive white noise. Accordingly, the central idea is to partially whiten the error signal beyond a predetermined correlation lag. In the first half of the paper, we will derive a fast Quasi-Newton type recursive algorithm to compute the optimal EWC solution in an online manner. The algorithm has O(N^2) complexity where, N represents the length of the parameter vector to be estimated. One of the primary limitations of EWC is the assumption that the input noise must be white. In the second half of this paper, we will introduce a modified cost function that overcomes this assumption and allows the noise in the input to be colored. The analysis of this modified cost function is then presented

followed by a sample-by-sample stochastic gradient algorithm to optimally compute the analytical solution. Finally, we will show the experimental results with EWC as well as the modified criterion in system identification and controller design problems.

I. Introduction

Mean-squared Error (MSE) criterion has been around for many years and has been widely applied in a variety of signal processing and control problems [1], [2]. Inverse control and system identification are some of the key applications in automatic control where MSE plays a vital role. System identification is the problem of estimating the parameters of an unknown system using the observed input and output (desired) sequences [3]. The objective of inverse control is to design a controller that would work in tandem with the actual system to produce a desired reference output [2]. The existence of cost effective and efficient algorithms like the stochastic Least Mean-Squares (LMS) [4] and the Recursive Least Squares (RLS) [1] has benefited the extensive application of the MSE criterion for system identification and control. However, in the presence of additive disturbances (both correlated and white) on the input and the desired signals of interest, MSE can at best provide a biased solution.[1] Further, the MSE based solutions change with changing noise statistics, which is highly undesirable. Noise-free data are seldom available in many real-world applications and significant amount of research over the past few decades has resulted in many engineering solutions, some of which are outlined in this paper. However, most of these methods have inherent drawbacks as we will see in the next section.

[1] Wiener MSE solution with noise-free data gives *unbiased* parameter estimates. We refer to this mismatch in the parameters obtained with and without noise as the *bias* introduced by noise.

Recently, we proposed a new criterion called the *Error Whitening Criterion* (EWC), which can produce unbiased parameter estimates of a linear system in the presence of additive *white noise* [5,6]. Instead of minimizing the mean-squared error, the EWC formulation enforces zero autocorrelation of the error signal beyond a certain lag, and hence the name *Error Whitening Criterion*. In this paper, we will first derive a fast, recursive algorithm to solve for the optimal EWC solution and show its applicability in a controller design problem. One of the limitations of the EWC is that it cannot handle colored disturbances. In the second half of the paper, we will propose a modified cost function that allows input noise to be colored and derive a computable analytical solution as well as a stochastic gradient algorithm.

This paper is organized as follows. In the next section, we will briefly discuss some of the existing methods in the literature followed by an introduction to the Error Whitening Criterion in section III. In section IV, the Quasi-Newton EWC algorithm is presented. In section V, we will propose a modified cost function to handle colored noise and discuss a stochastic gradient method to estimate the optimal solution. Section VI has the simulation results followed by the discussions and conclusions in section VII.

II. Existing Methods

A powerful class of solutions is based on input preprocessing primarily aimed at signal enhancement. Subspace Wiener filtering [1] is a data conditioning technique based on Principal Subspace Analysis (PSA). The idea is to first project the noisy input onto the signal subspace and then derive optimal Wiener filters in the reduced dimensional space. This method can be cumbersome especially when the data dimensionality is very high and also when the Signal-to-Noise (SNR) ratio is low. The latter will make the signal and noise subspaces indistinguishable.

The well-known Total Least Squares (TLS) method can provide bias-free solutions with noisy data [7,8]. Recall that the TLS can be formulated as the problem of solving an over-determined set of linear equations of the form $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A} \in \Re^{m \times n}$ is the data matrix, $\mathbf{b} \in \Re^{m}$ is the desired data vector, $\mathbf{x} \in \Re^{n}$ is the parameter vector and m denotes the number of different observation vectors each of dimension n [8]. The optimal TLS solution is then obtained by computing the eigenvector corresponding to the smallest eigenvalue of the augmented data matrix $[\mathbf{A}, \mathbf{b}]$. The result is unbiased only when both the noise in the input and the desired data are independent and identically distributed (i.i.d.) with the same variance. Further, when the noise is Gaussian-distributed, the TLS solution is also the maximum likelihood solution. However, the assumption of equal noise variances is very restrictive. The Generalized TLS (GTLS) problem [8] specifically deals with cases where the noise (still assumed to be i.i.d.) variances are different. But the caveat is that the ratio of noise variances is assumed to be known which is rarely valid. In order to overcome the i.i.d. assumption, Mathews and Cichocki have proposed the Extended TLS (ETLS) [9] that allows the noise to be colored. We will briefly describe the approach they adopted. Let the augmented input matrix $[\mathbf{A}, \mathbf{b}]$ be represented as, $\overline{\mathbf{H}} = [\mathbf{A}, \mathbf{b}]$. The matrix $\overline{\mathbf{H}}^T \overline{\mathbf{H}}$ can then be written as a combination of the clean data matrix $\mathbf{H}^T\mathbf{H}$ and the noise covariance matrix $\mathbf{R}_N$.

$$\overline{\mathbf{H}}^T \overline{\mathbf{H}} = \mathbf{H}^T \mathbf{H} + \mathbf{R}_N \tag{1}$$

The above equation is true when the noise is uncorrelated with the clean data. This assumption is reasonable as the noise processes in general are unrelated to (hence independent from) the physical sources that produced the data. Assume that there exists a matrix transformation $\widetilde{\mathbf{H}}$, such that

$$\widetilde{\mathbf{H}} = \overline{\mathbf{H}}\mathbf{R}_N^{-1/2} \tag{2}$$

The transformed data correlation matrix of $\widetilde{\mathbf{H}}$ is simply

$$\widetilde{\mathbf{H}}^T\widetilde{\mathbf{H}} = \mathbf{R}_N^{-1/2}\overline{\mathbf{H}}^T\overline{\mathbf{H}}\mathbf{R}_N^{-1/2} + \mathbf{I} \tag{3}$$

From (3), the overall problem reduces to the regular TLS with transformed data, the solution for which is once again obtained by estimating the minor eigenvector of the matrix $\widetilde{\mathbf{H}}^T\widetilde{\mathbf{H}}$. In other words, the optimal ETLS solution for colored noise is given by the generalized eigenvector corresponding to the smallest generalized eigenvalue of the matrix pencil ($\overline{\mathbf{H}}^T\overline{\mathbf{H}}$, $\mathbf{R}_N$). Thus, in order to compute the ETLS solution, we require the full knowledge of the correlation matrix of the noise $\mathbf{R}_N$, which is unreasonable.

The Instrumental Variables (IV) method proposed as an extension to the Least-Squares (LS) has been previously applied for estimating parameters in white noise [3]. This method requires choosing a set of *instruments* that are uncorrelated with the noise in the input. It can be shown that the IV solution is a special case of the Quasi-Newton EWC algorithm [10] that is detailed in this paper. Although, the IV methods have been successfully applied to many engineering problems, the approach itself has limitations when the noise is colored. The generalizations to the colored noise require additional pre-whitening filters or knowledge of the noise correlation depth, both of which are not practical. More details on IV can be found in [3] and the references therein.

III. Error Whitening Criterion: A Review

Consider the problem of identifying a linear system characterized by the parameter vector $\mathbf{w}_T \in \Re^N$ as shown in Fig 1. Let $(\mathbf{x}_k, d_k)$ denote the actual input and output of the system. Further, we will model the measurement errors and system disturbances by uncorrelated additive white noise sequences u_k and v_k (with unknown variances) that appear at the output and input of the system respectively. The problem of system identification can now be stated as follows: Given the noisy data pair $(\hat{\mathbf{x}}_k, \hat{d}_k)$ where $\hat{\mathbf{x}}_k = \mathbf{x}_k + \mathbf{v}_k \in \Re^N$ and $\hat{d}_k = d_k + u_k \in \Re^1$, determine the parameter vector $\mathbf{w} \in \Re^M$ that best describes the underlying system. Without loss of generality, assume that the length of $\mathbf{w}$ is at least *N*, the number of parameters in the actual system or $M \geq N$. Since $d_k = \mathbf{x}_k^T \mathbf{w}_T$, the error is $\hat{e}_k = \mathbf{x}_k^T(\mathbf{w}_T - \mathbf{w}) + u_k - \mathbf{v}_k{}^T\mathbf{w}$. Defining a vector $\boldsymbol{\varepsilon} = \mathbf{w}_T - \mathbf{w}$, the error autocorrelation at some arbitrary lag *L* is given by

$$\rho_{\hat{e}}(L) = \boldsymbol{\varepsilon}^T E[\mathbf{x}_k \mathbf{x}_{k-L}^T]\boldsymbol{\varepsilon} + \mathbf{w}^T E[\mathbf{v}_k \mathbf{v}_{k-L}^T]\mathbf{w} \tag{4}$$

If the chosen lag $L \geq M$, it is obvious that $E[\mathbf{v}_k \mathbf{v}_{k-L}^T] = \mathbf{0}$. Also, if the matrix $E[\mathbf{x}_k \mathbf{x}_{k-L}^T]$ is full rank, $\rho_{\hat{e}}(L) = 0$ when $\mathbf{w} = \mathbf{w}_T$ [5,6]. Therefore, if we make the error autocorrelation at any lag $L \geq M$ zero, then the estimated weight vector will be exactly equal to the true weight vector. In other words, the criterion tries to *whiten* the error signal for lags greater than or equal to the adaptive filter length, i.e., $\rho_{\hat{e}}(L) = 0$ for $L \geq M$ and hence the name *Error Whitening Criterion*. Defining $\dot{\hat{e}}_k = (\hat{e}_k - \hat{e}_{k-L})$, equation (4) can be rewritten in a convenient form as [5,6]

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

where, β is a constant. It is easy to see that when $\beta = -0.5$, (5) reduces to the error autocorrelation $\rho_{\hat{e}}(L)$. The goal is then to find the weight vector $\mathbf{w}$ that would make

$J(\mathbf{w}) = 0$. Note that when $\beta = 0$, (5) reduces to the MSE cost function. Therefore, MSE becomes a special case of EWC. Further, EWC has very interesting properties which are not discussed here for the sake of clarity. Refer [5,10] for details.

An online, stochastic gradient algorithm to compute the optimal EWC solution has been proposed in [6]. This stochastic gradient algorithm referred to as EWC-LMS resembles the MSE based LMS algorithm both in structure and computational complexity. Being a stochastic gradient method, it also suffers from the well-known limitations of stochastic gradient optimization, viz., step-size dependent convergence and sensitivity to the eigenvalue spread of the Hessian matrix.[2] This motivates the need of fast converging algorithms that can accurately track the optimal EWC solution. In the next section, we will derive a Quasi-Newton type algorithm that shows improved convergence behavior when compared to the stochastic gradient based EWC-LMS. As expected, the improved rate of convergence and robustness is obtained at an increase in computational requirements.

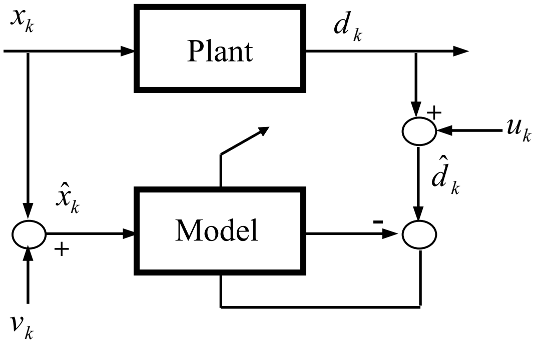


Figure 1. System Identification block diagram.

[2] The Hessian matrix is the second derivative of the cost function with respect to the weight vector **w**. It is well-known that stochastic gradient algorithms have convergence issues when this matrix has very high eigenspread.

IV. Quasi-Newton Type Recursive EWC Algorithm

We will begin this section by defining some matrices that will be used in the rest of the paper. Define input correlation matrices as $\mathbf{R} = E[\mathbf{x}_k \mathbf{x}_k^T]$, $\hat{\mathbf{R}} = E[\hat{\mathbf{x}}_k \hat{\mathbf{x}}_k^T]$, $\mathbf{R}_L = E[\mathbf{x}_{k-L}\mathbf{x}_k^T + \mathbf{x}_k \mathbf{x}_{k-L}^T]$, 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 (denoted by capped variables). Further, the input noise vector autocorrelation matrices are $\mathbf{V} = E[\mathbf{v}_k \mathbf{v}_k^T]$ and $\mathbf{V}_L = E[\mathbf{v}_{k-L}\mathbf{v}_k^T + \mathbf{v}_k \mathbf{v}_{k-L}^T]$. Additionally, we will define the matrices $\mathbf{S} = E[\dot{\mathbf{x}}_k \dot{\mathbf{x}}_k^T]$ and $\hat{\mathbf{S}} = E[\hat{\dot{\mathbf{x}}}_k \hat{\dot{\mathbf{x}}}_k^T]$. The dot is used to symbolize difference between the current and L^{th} previous sample vector/scalar, for example, $\dot{\mathbf{x}}_k = \mathbf{x}_k - \mathbf{x}_{k-L}$. Further, define cross-correlation vectors between the input vector and the desired signal as $\mathbf{P} = E[\mathbf{x}_k d_k]$, $\hat{\mathbf{P}} = E[\hat{\mathbf{x}}_k \hat{d}_k]$, $\mathbf{P}_L = E[\mathbf{x}_{k-L} d_k + \mathbf{x}_k d_{k-L}]$, and $\hat{\mathbf{P}}_L = E[\hat{\mathbf{x}}_{k-L}\hat{d}_k + \hat{\mathbf{x}}_k \hat{d}_{k-L}]$ for both noise-free and noisy data. Also, we will define vectors $\mathbf{Q} = E[\dot{\mathbf{x}}_k \dot{d}_k]$ and $\hat{\mathbf{Q}} = E[\hat{\dot{\mathbf{x}}}_k \hat{\dot{d}}_k]$. Using the above definitions, we can rewrite $J(\mathbf{w})$ in (5) as

$$J(\mathbf{w}) = E[d_k^2 + \beta \dot{d}_k^2] + \mathbf{w}^T(\mathbf{R} + \beta\mathbf{S})\mathbf{w} - 2(\mathbf{P} + \beta\mathbf{Q})^T\mathbf{w} \tag{6}$$

The above equation can be easily derived by substituting $e_k = \hat{d}_k - \hat{\mathbf{x}}_k^T\mathbf{w}$ and $\hat{\dot{e}}_k = \hat{\dot{d}}_k - \hat{\dot{\mathbf{x}}}_k^T\mathbf{w}$ in (5). Taking the gradient with respect to $\mathbf{w}$ and equating to zero, we get

$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = 2(\hat{\mathbf{R}} + \beta\hat{\mathbf{S}})\mathbf{w} - 2(\hat{\mathbf{P}} + \beta\hat{\mathbf{Q}}) = \mathbf{0} \tag{7}$$

Then the optimal weight vector $\mathbf{w}_*$ is given by

$$\mathbf{w}_* = (\hat{\mathbf{R}} + \beta\hat{\mathbf{S}})^{-1}(\hat{\mathbf{P}} + \beta\hat{\mathbf{Q}}) \tag{8}$$

When $\beta = 0$, (8) reduces to the RLS algorithm for MSE. Simple calculations show that $\hat{\mathbf{S}} = 2\hat{\mathbf{R}} - \hat{\mathbf{R}}_L$ and $\hat{\mathbf{Q}} = 2\hat{\mathbf{P}} - \hat{\mathbf{P}}_L$. Also, the noisy correlation matrices are related to the noise-free signal and noise correlation matrices through the following set of equations.

$$\begin{aligned}
\hat{\mathbf{R}} &= E[\hat{\mathbf{x}}_k\hat{\mathbf{x}}_k^T] = \mathbf{R}+\mathbf{V} \\
\hat{\mathbf{S}} &= E[(\hat{\mathbf{x}}_k-\hat{\mathbf{x}}_{k-L})(\hat{\mathbf{x}}_k-\hat{\mathbf{x}}_{k-L})^T] = 2(\mathbf{R}+\mathbf{V})-\mathbf{R}_L-\mathbf{V}_L \\
\hat{\mathbf{P}} &= E[\hat{\mathbf{x}}_k\hat{d}_k] = \mathbf{P} \\
\hat{\mathbf{Q}} &= E[(\hat{\mathbf{x}}_k-\hat{\mathbf{x}}_{k-L})(\hat{d}_k-\hat{d}_{k-L})] = 2\mathbf{P}-\mathbf{P}_L
\end{aligned} \tag{9}$$

Using (9), the expression for the optimal weight vector $\mathbf{w}_*$ can be further simplified as

$$\mathbf{w}_* = \left[(1+2\beta)(\mathbf{R}+\mathbf{V})-\beta(\mathbf{R}_L+\mathbf{V}_L)\right]^{-1}\left[(1+2\beta)\mathbf{P}-\beta\mathbf{P}_L\right] \tag{10}$$

When $L \geq M$ and $\beta = -0.5$, it is obvious that all the noise matrices in the above equation cancel out and the optimal solution reduces to

$$\mathbf{w}_* = \mathbf{R}_L^{-1}\mathbf{P}_L \tag{11}$$

which is nothing but the true weight vector $\mathbf{w}_T$ provided $M \geq N$. We will now derive a Quasi-Newton type algorithm called the Recursive Error Whitening (REW) algorithm to adaptively estimate the optimal solution in (11). For the sake of notational simplicity, we will consider the noise-free case to derive the Recursive Error Whitening (REW) algorithm. With $\mathbf{Z}_k = \mathbf{R}_k + \beta\mathbf{S}_k$ and $\boldsymbol{\theta}_k = \mathbf{P}_k + \beta\mathbf{Q}_k$, a recursive relation for $\mathbf{Z}_k$ can be easily derived as

$$\mathbf{Z}_k = \mathbf{Z}_{k-1} + (2\beta\mathbf{x}_k - \beta\mathbf{x}_{k-L})\mathbf{x}_k^T + \mathbf{x}_k(\mathbf{x}_k - \beta\mathbf{x}_{k-L})^T \tag{12}$$

Recall the Sherman-Morrison-Woodbury identity, also known as the matrix inversion lemma [11].

$$(\mathbf{A}+\mathbf{BCD}^T)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1}+\mathbf{D}^T\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{D}^T\mathbf{A}^{-1} \tag{13}$$

Define $\mathbf{A} = \mathbf{Z}_k$, $\mathbf{B} = [2\beta\mathbf{x}_k - \beta\mathbf{x}_{k-L} \quad \mathbf{x}_k]$, $\mathbf{C} = \mathbf{I}_{2x2}$, a 2x2 identity matrix, and $\mathbf{D} = [\mathbf{x}_k \quad (\mathbf{x}_k - \beta\mathbf{x}_{k-L})]$. Then (12) reduces to

$$\mathbf{Z}_k^{-1} = \mathbf{Z}_{k-1}^{-1} - \mathbf{Z}_{k-1}^{-1}\mathbf{B}(\mathbf{I}_{2x2} + \mathbf{D}^T\mathbf{Z}_{k-1}^{-1}\mathbf{B})^{-1}\mathbf{D}^T\mathbf{Z}_{k-1}^{-1} \tag{14}$$

Notice that this recursion for the inverse of $\mathbf{Z}_k$ is different than the conventional RLS algorithm. It requires the inversion of a 2x2 matrix $(\mathbf{I}_{2x2} + \mathbf{D}^T\mathbf{Z}_{k-1}^{-1}\mathbf{B})^{-1}$, which is still trivial. With this, we are able to reduce the complexity of inverting a sum of two matrices from O(N^3) to O(N^2). The recursive estimator for $\boldsymbol{\theta}_k$ is much simpler and can be expressed as

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} + [(1+2\beta)d_k\mathbf{x}_k - \beta d_k\mathbf{x}_{k-L} - \beta d_{k-L}\mathbf{x}_k] \tag{15}$$

From (14) and (15), the optimal solution $\mathbf{w}_*$ is given by

$$\mathbf{w}_k = \mathbf{Z}_k^{-1}\boldsymbol{\theta}_k \tag{16}$$

To convert equation (16) into a recursive form, define a gain matrix (analogous to the Kalman gain in the RLS algorithm) as

$$\boldsymbol{\kappa}_k = \mathbf{Z}_{k-1}^{-1}\mathbf{B}\left(\mathbf{I}_{2x2} + \mathbf{D}^T\mathbf{Z}_{k-1}^{-1}\mathbf{B}\right)^{-1} \tag{17}$$

Using (17) in (14), we get

$$\mathbf{Z}_k^{-1} = \mathbf{Z}_{k-1}^{-1} - \boldsymbol{\kappa}_k\mathbf{D}^T\mathbf{Z}_{k-1}^{-1} \tag{18}$$

Multiplying (17) from the right by $\left(\mathbf{I}_{2x2} + \mathbf{D}^T\mathbf{Z}_{k-1}^{-1}\mathbf{B}\right)$, and using (18), we obtain

$$\boldsymbol{\kappa}_k\left(\mathbf{I}_{2x2} + \mathbf{D}^T\mathbf{Z}_{k-1}^{-1}\mathbf{B}\right) = \mathbf{Z}_{k-1}^{-1}\mathbf{B} \Rightarrow \boldsymbol{\kappa}_k = \mathbf{Z}_k^{-1}\mathbf{B} \tag{19}$$

Substituting (15) in (16),

$$\mathbf{w}_k = \mathbf{Z}_k^{-1}\boldsymbol{\theta}_{k-1} + \mathbf{Z}_k^{-1}[(1+2\beta)d_k\mathbf{x}_k - \beta d_k\mathbf{x}_{k-L} - \beta d_{k-L}\mathbf{x}_k] \tag{20}$$

which can be further simplified as

$$\mathbf{w}_k = \mathbf{w}_{k-1} - \boldsymbol{\kappa}_k\mathbf{D}^T\mathbf{w}_{k-1} + \mathbf{Z}_k^{-1}[(1+2\beta)d_k\mathbf{x}_k - \beta d_k\mathbf{x}_{k-L} - \beta d_{k-L}\mathbf{x}_k] \tag{21}$$

From the definition of $\mathbf{B}$, $(1+2\beta)d_k\mathbf{x}_k - \beta d_k\mathbf{x}_{k-L} - \beta d_{k-L}\mathbf{x}_k = \mathbf{B}[d_k; d_k - \beta d_{k-L}]$, where

$[d_k; d_k - \beta d_{k-L}]$ is a column vector with elements d_k and $d_k - \beta d_{k-L}$. Therefore, the update equation can then be written as

$$\mathbf{w}_k = \mathbf{w}_{k-1} - \boldsymbol{\kappa}_k \mathbf{D}^T \mathbf{w}_{k-1} + \boldsymbol{\kappa}_k [d_k; d_k - \beta d_{k-L}] \tag{22}$$

Note that the product $\mathbf{D}^T \mathbf{w}_{k-1} = [y_k \quad y_k - \beta y_{k-L}]^T$, where $y_k = \mathbf{x}_k^T \mathbf{w}_{k-1}$, and $y_{k-L} = \mathbf{x}_{k-L}^T \mathbf{w}_{k-1}$ represent the outputs with the weights of the previous iteration. Defining an *apriori error vector* $\mathbf{e}_k$ as

$$\mathbf{e}_k = [d_k - y_k; d_k - y_k - \beta(d_{k-L} - y_{k-L})] = [e_k; e_k - \beta e_{k-L}] \tag{23}$$

we can simplify (22) to give us the REW update equation

$$\mathbf{w}_k = \mathbf{w}_{k-1} + \boldsymbol{\kappa}_k \mathbf{e}_k \tag{24}$$

A summary of the REW algorithm is shown in Table 1.

Table 1. Summary of the REW Algorithm

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

Note that although the REW algorithm achieves partial error whitening for β = -0.5, the algorithm outlined in Table 1 can be used for any value of β. Further, the update equation in (24) implies that the REW algorithm tracks the optimal EWC solution at every iteration. Another noteworthy feature is the fact the REW algorithm converges in a finite number of steps unlike the stochastic gradient methods that converge only in the limiting sense. This is in agreement with other Newton type gradient optimization methods.

Theorem 1. (No input noise case): The REW algorithm with β = -0.5 converges in the mean to the optimal solution in (8) in a finite number of steps.

Proof. Recall that the REW algorithm with β = -0.5 is simply

$$\mathbf{w}(n) = \mathbf{R}_L^{-1}(n)\hat{\mathbf{P}}_L(n) \tag{25}$$

where, the sample estimates of the matrix $\mathbf{R}_L(n)$ and vector $\hat{\mathbf{P}}_L(n)$ are given by

$$\begin{aligned}\mathbf{R}_L(n) &= \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_{i-L}^T + \mathbf{x}_{i-L}\mathbf{x}_i^T \\ \hat{\mathbf{P}}_L(n) &= \sum_{i=1}^{n} \mathbf{x}_i \hat{d}_{i-L} + \mathbf{x}_{i-L}\hat{d}_i\end{aligned} \tag{26}$$

The above expressions are true for all $n > n_{\min}$, where, $n_{\min}$ is the smallest positive number for which the matrix $\mathbf{R}_L(n)$ is full-rank. Let the optimal EWC solution (which is also the true weights) be $\mathbf{w}_*$. Therefore, the noisy desired signal can be expressed as $\hat{d}_i = \mathbf{x}_i^T \mathbf{w}_* + u_i$. With this definition, the vector $\hat{\mathbf{P}}_L(n)$ can be rewritten as

$$\begin{aligned}\hat{\mathbf{P}}_L(n) &= \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_{i-L}^T + \mathbf{x}_{i-L}\mathbf{x}_i^T \right)\mathbf{w}_* + \sum_{i=1}^{n} u_i \mathbf{x}_{i-L} + u_{i-L}\mathbf{x}_i \\ &= \mathbf{R}_L(n)\mathbf{w}_* + \sum_{i=1}^{n} u_i \mathbf{x}_{i-L} + u_{i-L}\mathbf{x}_i\end{aligned} \tag{27}$$

Multiplying the above equation from the left by $\mathbf{R}_L^{-1}(n)$ we get,

$$\mathbf{w}(n) = \mathbf{w}_* + \mathbf{R}_L^{-1}(n)\sum_{i=1}^{n} u_i \mathbf{x}_{i-L} + u_{i-L}\mathbf{x}_i \tag{28}$$

Taking the expected value on both sides of (28) and realizing that $E(x) = E(E(x \mid y))$, we get

$$E[\mathbf{w}(n)] = \mathbf{w}_* + E\left[E\left\{ \mathbf{R}_L^{-1}(n) \sum_{i=1}^{n} u_i \mathbf{x}_{i-L} + u_{i-L}\mathbf{x}_i \middle| \mathbf{x}_i ,\ i = 1 - L,...,n \right\} \right] \tag{29}$$

It is easy to see that $\mathbf{R}_L(n)$ is constant with respect to the inner conditional expectation as it is completely determined by the sample vectors $\mathbf{x}_i$, $i = 1 - L,...,n$. Since the noise u_i is zero mean and uncorrelated with the input, (29) reduces to

$$E[\mathbf{w}(n)] = \mathbf{w}_* \tag{30}$$

It is important to note that (30) is true for all $n > n_{\min}$ and hence the REW algorithm converges to the optimal EWC solution in the mean within a finite number of iterations.■

Corollary 1. Even in the presence of input noise, the REW algorithm with β = -0.5 converges in the mean to the optimal solution in (8).

Proof. With noisy data, the REW algorithm becomes

$$\mathbf{w}(n) = \hat{\mathbf{R}}_L^{-1}(n)\hat{\mathbf{P}}_L(n) \tag{31}$$

However, the noisy matrix $\hat{\mathbf{R}}_L(n)$ is exactly the same as its noiseless equivalent $\mathbf{R}_L(n)$ owing to the white noise assumption. Thus, the previous proof can be immediately applied following this observation and hence $E[\mathbf{w}(n)] = \mathbf{w}_*$. ■

Convergence of the REW algorithm in the mean does not fully address its overall behavior. In our experiments, we have observed that the transient response of the REW algorithm is dependent on the input data correlation and hence the eigenvalues of the data correlation matrix $\hat{\mathbf{R}}_L(n)$. This is in concurrence with the fact that the MSE based RLS algorithm is susceptible to the smallest eigenvalue of the input covariance matrix during the initial stages of adaptation.

Theorem 2. Trace of REW error covariance matrix $E[\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n^T]$, where $\boldsymbol{\varepsilon}_n = \mathbf{w}(n) - \mathbf{w}_*$ is always bound from above during the initial stages of adaptation.

Proof. For simplicity, we will assume noise-free input. By the arguments listed in corollary 1, this proof can be extended to the noisy data case. Recall that the REW estimate at time index (iteration) *n* is given by,

$$\mathbf{w}(n) = \mathbf{w}_* + \mathbf{R}_L^{-1}(n)\sum_{i=1}^{n} u_i\mathbf{x}_{i-L} + u_{i-L}\mathbf{x}_i \tag{32}$$

The error vector is then given by,

$$\boldsymbol{\varepsilon}_n = \mathbf{w}(n) - \mathbf{w}_* = \mathbf{R}_L^{-1}(n)\sum_{i=1}^{n} u_i\mathbf{x}_{i-L} + u_{i-L}\mathbf{x}_i \tag{33}$$

From (33), we can compute the error covariance as

$$E[\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n^T] = E\left[\mathbf{R}_L^{-1}(n)\sum_{j=1}^{n}\sum_{i=1}^{n}(u_i\mathbf{x}_{i-L} + u_{i-L}\mathbf{x}_i)(u_j\mathbf{x}_{j-L}^T + u_{j-L}\mathbf{x}_j^T)\mathbf{R}_L^{-1}(n)\right] \tag{34}$$

Expanding the terms inside the summations, we get

$$E[\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n^T] = E\left[\mathbf{R}_L^{-1}(n)\sum_{j=1}^{n}\sum_{i=1}^{n}(u_{i-L}u_{j-L}\mathbf{x}_i\mathbf{x}_j^T + u_iu_{j-L}\mathbf{x}_{i-L}\mathbf{x}_j^T + u_{i-L}u_j\mathbf{x}_i\mathbf{x}_{j-L}^T + u_iu_j\mathbf{x}_{i-L}^T\mathbf{x}_{j-L}^T)\mathbf{R}_L^{-1}(n)\right] \tag{35}$$

Again, by utilizing the relationship $E(x) = E(E(x \mid y))$ and the fact that the white noise u_i is zero mean and uncorrelated with the input, we can further simplify the above equation.

$$E[\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n^T] = \sigma_u^2 E\left[\mathbf{R}_L^{-1}(n)\sum_{i=1}^{n}(\mathbf{x}_i\mathbf{x}_i^T + \mathbf{x}_{i-L}\mathbf{x}_{i+L}^T + \mathbf{x}_{i+L}\mathbf{x}_{i-L}^T + \mathbf{x}_{i-L}\mathbf{x}_{i-L}^T)\mathbf{R}_L^{-1}(n)\right] \tag{36}$$

Define the matrices $\mathbf{R}_0(n)$ and $\mathbf{R}_{2L}(n)$ as in (37). Note that both the matrices are symmetric and further, $\mathbf{R}_0(n)$ is positive definite and is similar to the covariance matrix.

$$\mathbf{R}_0(n) = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T + \mathbf{x}_{i-L}\mathbf{x}_{i-L}^T$$
$$\mathbf{R}_{2L}(n) = \sum_{i=1}^{n} \mathbf{x}_{i-L} \mathbf{x}_{i+L}^T + \mathbf{x}_{i+L}\mathbf{x}_{i-L}^T \tag{37}$$

With the above definitions the expression for the error covariance becomes

$$E[\boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^T] = \sigma_u^2 E[\mathbf{R}_L^{-1}(n)\{\mathbf{R}_0(n) + \mathbf{R}_{2L}(n)\}\mathbf{R}_L^{-1}(n)] \tag{38}$$

In general, there is no tractable closed form solution for the above expression.[3] Instead, we will attempt to derive an approximate upper bound on the trace of the error covariance matrix. In order to do so, assume that the matrices defined in (37) and the matrix $\mathbf{R}_L(n)$ are constant, i.e., $\mathbf{R}_0(n) = \mathbf{R}_0$, $\mathbf{R}_{2L}(n) = \mathbf{R}_{2L}$ and $\mathbf{R}_L(n) = \mathbf{R}_L$. This basically implies that the matrices are the same irrespective of the *n*-length block of data which is a valid assumption for a reasonably large *n*.

$$Tr\left(E[\boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^T]\right) = E[\boldsymbol{\varepsilon}_n^T \boldsymbol{\varepsilon}_n] = \sigma_u^2 Tr[\mathbf{R}_L^{-1}(\mathbf{R}_0 + \mathbf{R}_{2L})\mathbf{R}_L^{-1}] \tag{39}$$

Using the property of the trace invariance under cyclic permutations,[4] (39) reduces to

$$E[\boldsymbol{\varepsilon}_n^T \boldsymbol{\varepsilon}_n] = \sigma_u^2 Tr[\mathbf{R}_L^{-2}(\mathbf{R}_0 + \mathbf{R}_{2L})] = \sigma_u^2 [Tr(\mathbf{R}_L^{-2}\mathbf{R}_0) + Tr(\mathbf{R}_L^{-2}\mathbf{R}_{2L})] \tag{40}$$

By using the singular values instead of the eigenvalues, equation (40) is bound as shown.

$$E[\boldsymbol{\varepsilon}_n^T \boldsymbol{\varepsilon}_n] \le \sigma_u^2 \sum_{i=1}^{M} s_i(\mathbf{R}_L^{-2}\mathbf{R}_0) + s_i(\mathbf{R}_L^{-2}\mathbf{R}_{2L}) \tag{41}$$

where, s_i denotes the i^{th} singular value. In order to simplify further, we exploit the arithmetic-geometric mean inequality for singular values by Bhatia and Kittaneh [12]. Accordingly, if **A** and **B** are two Hermitian matrices, and $s_i(\mathbf{Z})$ denotes the i^{th} singular value of a matrix **Z**, then, arithmetic-geometric mean inequality states that,

[3] For the RLS algorithm, the error covariance is estimated by invoking the Gaussianity assumptions which allows the use of Wishart distribution and the results therein [1].

[4] $Tr[\mathbf{ABC}] = Tr[\mathbf{CAB}] = Tr[\mathbf{BCA}]$.

$$2s_i(\mathbf{A}^H\mathbf{B}) \le s_i(\mathbf{A}^H\mathbf{A} + \mathbf{B}^H\mathbf{B}) \tag{42}$$

From our matrix definitions, $\mathbf{R}_0$, $\mathbf{R}_{2L}$ and $\mathbf{R}_L$ are all symmetric matrices with real entries. Therefore, applying the above inequality to these matrices in equation (41) we get

$$E[\boldsymbol{\varepsilon}_n^T\boldsymbol{\varepsilon}_n] \le \frac{\sigma_u^2}{2}\sum_{i=1}^{M} s_i(\mathbf{R}_L^{-4} + \mathbf{R}_0^2) + s_i(\mathbf{R}_L^{-4} + \mathbf{R}_{2L}^2) \tag{43}$$

Realizing that the singular values coincide with the eigenvalues for the above symmetric matrices in (43), we can rewrite (43) as

$$\begin{aligned} E[\boldsymbol{\varepsilon}_n^T\boldsymbol{\varepsilon}_n] &\le \frac{\sigma_u^2}{2}\sum_{i=1}^{M} \lambda_i(\mathbf{R}_L^{-4} + \mathbf{R}_0^2) + \lambda_i(\mathbf{R}_L^{-4} + \mathbf{R}_{2L}^2) \\ &\le \frac{\sigma_u^2}{2}\left[Tr\left(\mathbf{R}_L^{-4} + \mathbf{R}_0^2\right) + Tr\left(\mathbf{R}_L^{-4} + \mathbf{R}_{2L}^2\right)\right] \\ &\le \sigma_u^2\left[Tr\left(\mathbf{R}_L^{-4}\right) + \frac{1}{2}Tr\left(\mathbf{R}_0^2 + \mathbf{R}_{2L}^2\right)\right] \end{aligned} \tag{44}$$

From the above bound, it is clear that the quantity $E[\boldsymbol{\varepsilon}_n^T\boldsymbol{\varepsilon}_n]$ is mainly affected by the smallest eigenvalue of the matrix $\mathbf{R}_L$ estimated using *n* data samples.■

Although this result is intuitively satisfying, as a consequence, the REW algorithm can produce noisier transient response when compared to the RLS algorithm. It further emphasizes the fact the clean input data must have sufficient correlation depth to result in a well conditioned data matrix $\mathbf{R}_L$. However, as in the case of RLS, extensive experiments have revealed that the sensitivity of the REW algorithm fades with increasing number of iterations as the matrix $\mathbf{R}_L$ becomes well conditioned.

An alternative recursive algorithm that truly tracks the EWC solution can be derived by using minor component analysis. By reformulating the EWC as a problem of solving an over-determined set of linear equations, we can effectively apply the computational principles of TLS. It can be shown that, the optimal EWC solution is obtained by estimating the minor eigenvector corresponding to the zero eigenvalue of the augmented data matrix **G** which is given by

$$\mathbf{G} = \begin{bmatrix} \mathbf{R}_L & \mathbf{P}_L \\ \mathbf{P}_L^T & 2\rho_d(L) \end{bmatrix} \tag{45}$$

where, the term $\rho_d(L)$ denotes the autocorrelation of the desired signal at lag L. Although the eigenvectors of **G** are real, the matrix itself is indefinite and can have mixed eigenvalues. Most of the existing methods for computing the minor eigenvector assume that the matrix is positive-definite and hence cannot be used in our case. However, the classical inverse iteration method [11] can be utilized to solve the problem. It is beyond the scope of this paper to outline the details of the method. For a detailed derivation and description of the algorithm, see [10,13].

V. Parameter Estimation in Colored Input Noise

In the theory of error whitening criterion, we made a crucial assumption that the input noise is uncorrelated with itself or is *white*. Although, in many problems, the white noise model holds, this assumption can be certainly restrictive in other applications. From the discussions in the previous sections, it is clear that EWC fails to remove the bias in the parameter estimates when the input noise is correlated or *colored*. Our goal in this section is to derive a method to accurately estimate the parameters of a linear system in the presence of colored input noise by exploiting the signal correlations at different lags similar to the EWC. In this paper, we will only consider the case wherein the input noise can be correlated whereas the desired signal is either noise-free or assumed to be corrupted with white noise. The general case of having colored noise in both the input and desired data will be dealt in a later paper.

Consider the system identification framework shown in Fig 1. The additive input noise v_k can now have an arbitrary covariance matrix $\mathbf{V} = E[\mathbf{v}_k \mathbf{v}_k^T]$, whereas, the noise in the desired signal u_k is assumed to be white. Also, the noises v_k and u_k are independent from the data pair and independent of each other. Further, we will assume sufficient order for the model i.e., $\mathbf{w} \in \Re^N$. All other quantities that appear in this section have the same definitions as before. Consider the cost function in equation (46).

$$J(\mathbf{w}) = \sum_{\Delta=1}^{N} \left| E[\hat{e}_k \hat{d}_{k-\Delta} + \hat{e}_{k-\Delta} \hat{d}_k] \right| \tag{46}$$

where, Δ denotes a lag. Consider a single term in the summation of the above equation. It is easy to see that the cross products $E[\hat{e}_k \hat{d}_{k-\Delta}]$ and $E[\hat{e}_{k-\Delta} \hat{d}_k]$ are given by

$$\begin{aligned} E[\hat{e}_k \hat{d}_{k-\Delta}] &= \mathbf{w}_T^T E[\mathbf{x}_k \mathbf{x}_{k-\Delta}^T] \mathbf{w}_T - \mathbf{w}_T^T E[\mathbf{x}_k \mathbf{x}_{k-\Delta}^T] \mathbf{w} + E[u_k u_{k-\Delta}] \\ E[\hat{e}_{k-\Delta} \hat{d}_k] &= \mathbf{w}_T^T E[\mathbf{x}_{k-\Delta} \mathbf{x}_k^T] \mathbf{w}_T - \mathbf{w}_T^T E[\mathbf{x}_{k-\Delta} \mathbf{x}_k^T] \mathbf{w} + E[u_{k-\Delta} u_k] \end{aligned} \tag{47}$$

Since the noise u_k is assumed to be white, $E[u_k u_{k-\Delta}] = 0$, and (47) reduces to a function of only the clean data (input and desired) and the weights. The input noise never multiplies itself; hence it gets eliminated. Further, the cost function in (46) simplifies to

$$J(\mathbf{w}) = \sum_{\Delta=1}^{N} \left| \mathbf{w}_T^T \mathbf{R}_\Delta \mathbf{w}_T - \mathbf{w}_T^T \mathbf{R}_\Delta \mathbf{w} \right| \tag{48}$$

where, the matrix $\mathbf{R}_\Delta$ is given by

$$\mathbf{R}_\Delta = E[\mathbf{x}_k \mathbf{x}_{k-\Delta}^T + \mathbf{x}_{k-\Delta} \mathbf{x}_k^T] \tag{49}$$

The matrix $\mathbf{R}_\Delta$ is symmetric, but indefinite and hence can have mixed eigenvalues. Also, observe that the cost function in (49) is *linear* in the weights **w**. If for instance, there was a single term in the summation, and we force $J(\mathbf{w}) = 0$, then it is easy to see that one of the solutions for **w** will be the true parameter vector $\mathbf{w}_T$. However, when the number of terms in the summation becomes equal to the length of our estimated filter, there is always a unique solution for **w**, which will be the true vector $\mathbf{w}_T$.

Lemma 1. For suitable choices of lags, there is a unique solution $\mathbf{w}_*$ for the equation $J(\mathbf{w}_*) = 0$ and $\mathbf{w}_* = \mathbf{w}_T$.

Proof. If $J(\mathbf{w}) = 0$, $\mathbf{w}_T^T \mathbf{R}_\Delta \mathbf{w}_T - \mathbf{w}_T^T \mathbf{R}_\Delta \mathbf{w}$ must be zero for all selected lags Δ. For simplicity assume $\Delta = 1,..., N$. Therefore, we have N linear equations in **w** given by, $[\mathbf{w}_T^T \mathbf{R}_\Delta]\mathbf{w} = \mathbf{w}_T^T \mathbf{R}_\Delta \mathbf{w}_T$. This system of equations can be compactly written as

$$\begin{bmatrix} \mathbf{w}_T^T \mathbf{R}_1 \\ \mathbf{w}_T^T \mathbf{R}_2 \\ .... \\ \mathbf{w}_T^T \mathbf{R}_N \end{bmatrix} \mathbf{w} = 2 \begin{bmatrix} E[d_k d_{k-1}] \\ E[d_k d_{k-2}] \\ .... \\ E[d_k d_{k-N}] \end{bmatrix} \tag{50}$$

If the rows of the composite matrix on the left of **w** in (50) are linearly independent (full-rank matrix), then there is a unique inverse and hence $J(\mathbf{w}) = 0$ has a unique solution. We will prove that this unique solution has to be $\mathbf{w}_T$ by contradiction. Let the optimal solution be $\mathbf{w}_* = \mathbf{w}_T + \boldsymbol{\varepsilon}$. Then, $J(\mathbf{w}_*) = 0$ implies $\mathbf{w}_T^T \mathbf{R}_\Delta \boldsymbol{\varepsilon} = 0$ for all Δ which is possible only when $\boldsymbol{\varepsilon} = \mathbf{0}$ (composite matrix if full rank) and this completes the proof. ■

Note that each term inside the summation of equation (46) can be perceived as a constraint on the cross correlation between the desired data and the error. By forcing these sums of cross correlations at *N* different lags to simultaneously approach zero, we can obtain an unbiased estimate of the true filter.

The optimal solution for the proposed criterion in terms of the noisy input and the desired responses is given in (51). Each row of the composite matrix can be estimated using simple correlators having linear complexity.

$$\mathbf{w}_* = 2\begin{bmatrix} E[\hat{d}_k\hat{\mathbf{x}}_{k-1}^T + \hat{d}_{k-1}\hat{\mathbf{x}}_k^T] \\ E[\hat{d}_k\hat{\mathbf{x}}_{k-2}^T + \hat{d}_{k-2}\hat{\mathbf{x}}_k^T] \\ .... \\ E[\hat{d}_k\hat{\mathbf{x}}_{k-N}^T + \hat{d}_{k-N}\hat{\mathbf{x}}_k^T] \end{bmatrix}^{-1} \begin{bmatrix} E[\hat{d}_k\hat{d}_{k-1}] \\ E[\hat{d}_k\hat{d}_{k-2}] \\ .... \\ E[\hat{d}_k\hat{d}_{k-N}] \end{bmatrix} \tag{51}$$

Also, a recursive relationship for the evolution of this matrix over iterations can be easily derived. However, this recursion does not involve simple reduced rank updates and hence it is not possible to use the matrix inversion lemma efficiently to reduce the complexity of matrix inversion. The overall complexity of the recursive solution in equation (51) is $O(N^3)$. This necessitates the development of a low cost stochastic algorithm to compute and track the optimal solution given by equation (51). The derivation of the stochastic gradient algorithm is similar to that of the EWC-LMS algorithm in [6].

Consider the cost function in (52). It is easy to see that (52) corresponds to the stochastic version of the cost $J(\mathbf{w}) = \sum_{\Delta=1}^{N} E[|\, \hat{e}_k\hat{d}_{k-\Delta} + \hat{e}_{k-\Delta}\hat{d}_k \,|]$ which is an upper bound on the actual objective function in (46).

$$J(\mathbf{w}_k) = \sum_{\Delta=1}^{N}\left|\hat{e}_k\hat{d}_{k-\Delta} + \hat{e}_{k-\Delta}\hat{d}_k\right| \tag{52}$$

The goal is to find the minimum of the above function. Notice that the gradient direction depends on the instantaneous cost and therefore, the weight update is given by

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \eta\sum_{\Delta=1}^{N} sign(\hat{e}_k\hat{d}_{k-\Delta} + \hat{e}_{k-\Delta}\hat{d}_k)(\hat{\mathbf{x}}_k\hat{d}_{k-\Delta} + \hat{\mathbf{x}}_{k-\Delta}\hat{d}_k) \tag{53}$$

where, $\eta > 0$ is a small step-size. The step-size has been chosen to be a constant in the above update equation, but it is possible to have a time-varying step-size. Owing to the presence of multiple terms (constraints) in the gradient, the complexity of the update is

$O(N^2)$ which is higher than that of the regular LMS type stochastic updates. However, we are still at a gain because the complexity is significantly lower when compared to the recursive solution in (51). We will now discuss the convergence of this stochastic gradient algorithm to the optimal solution in the noisy as well as noise-free scenarios.

Theorem 3. In the noise-free case, (53) converges to the stationary point $\mathbf{w}_* = \mathbf{w}_T$ provided that the step size satisfies the following inequality at every iteration.

$$0 < \eta < \frac{2J(\mathbf{w}_k)}{\|\nabla J(\mathbf{w}_k)\|^2} \tag{54}$$

Proof. It is obvious from the previous discussions that the cost function in (52) has a single stationary point $\mathbf{w}_* = \mathbf{w}_T$. The weight update becomes zero only when the cost goes to zero thereby zeroing the gradient. Consider the weight error vector defined as $\boldsymbol{\varepsilon}_k = \mathbf{w}_* - \mathbf{w}_k$. From (53), we get

$$\boldsymbol{\varepsilon}_{k+1} = \boldsymbol{\varepsilon}_k - \eta \sum_{\Delta=1}^{N} sign(e_k d_{k-\Delta} + e_{k-\Delta} d_k)(\mathbf{x}_k d_{k-\Delta} + \mathbf{x}_{k-\Delta} d_k) \tag{55}$$

Taking the norm of this error vector on both sides gives,

$$\|\boldsymbol{\varepsilon}_{k+1}\|^2 = \|\boldsymbol{\varepsilon}_k\|^2 - 2\eta \sum_{\Delta=1}^{N} sign(e_k d_{k-\Delta} + e_{k-\Delta} d_k)\boldsymbol{\varepsilon}_k^T(\mathbf{x}_k d_{k-\Delta} + \mathbf{x}_{k-\Delta} d_k) + \eta^2 \|\nabla J(\mathbf{w}_k)\|^2 \tag{56}$$

Observe that in the noiseless case, $\boldsymbol{\varepsilon}_k^T \mathbf{x}_k = e_k$ and $\boldsymbol{\varepsilon}_k^T \mathbf{x}_{k-\Delta} = e_{k-\Delta}$. Hence (56) can be simplified further to

$$\|\boldsymbol{\varepsilon}_{k+1}\|^2 = \|\boldsymbol{\varepsilon}_k\|^2 - 2\eta \sum_{\Delta=1}^{N} |(e_k d_{k-\Delta} + e_{k-\Delta} d_k)| + \eta^2 \|\nabla J(\mathbf{w}_k)\|^2 \tag{57}$$

By allowing the error vector norm to decay asymptotically i.e., $\|\boldsymbol{\varepsilon}_{k+1}\|^2 < \|\boldsymbol{\varepsilon}_k\|^2$, we obtain the bound in (54). The error vector will eventually converge to zero by design, and since the gradient becomes null at the true solution, $\lim_{k\to\infty}\|\boldsymbol{\varepsilon}_k\|^2 \to 0$, and hence $\lim_{k\to\infty}\mathbf{w}_k \to \mathbf{w}_* = \mathbf{w}_T$. ■

Theorem 4. In the noisy data case, the stochastic algorithm in (53) converges to the stationary point $\mathbf{w}_* = \mathbf{w}_T$ in the mean provided that the step size is bound as below.

$$\eta < \frac{2\sum_{\Delta=1}^{N}\left|E[\hat{e}_k\hat{d}_{k-\Delta} + \hat{e}_{k-\Delta}\hat{d}_k]\right|}{E\|\nabla J(\mathbf{w}_k)\|^2} \tag{58}$$

Proof. Again, the facts about the uniqueness of the stationary point and it being equal to the true filter hold even for the noisy data case. The convergence to this stationary point in a stable manner will be proved in this theorem. Following the same steps as in the proof of the previous lemma, the dynamics of the error vector norm can be determined by the difference equation

$$\|\boldsymbol{\varepsilon}_{k+1}\|^2 = \|\boldsymbol{\varepsilon}_k\|^2 - 2\eta\sum_{\Delta=1}^{N} sign(\hat{z}_{k,\Delta})\boldsymbol{\varepsilon}_k^T(\hat{\mathbf{x}}_k\hat{d}_{k-\Delta} + \hat{\mathbf{x}}_{k-\Delta}\hat{d}_k) + \eta^2\|\nabla J(\mathbf{w}_k)\|^2 \tag{59}$$

where, $\hat{z}_{k,\Delta} = \hat{e}_k\hat{d}_{k-\Delta} + \hat{e}_{k-\Delta}\hat{d}_k$. Applying the expectation operator on both sides of (59) and letting $E\|\boldsymbol{\varepsilon}_{k+1}\|^2 < E\|\boldsymbol{\varepsilon}_k\|^2$ as in the previous case results in the following inequality.

$$\eta E\|\nabla J(\mathbf{w}_k)\|^2 < 2E\sum_{\Delta=1}^{N}\boldsymbol{\varepsilon}_k^T(\hat{\mathbf{x}}_k\hat{d}_{k-\Delta} + \hat{\mathbf{x}}_{k-\Delta}\hat{d}_k)sign(\hat{z}_{k,\Delta}) \tag{60}$$

Simplifying further, we get,

$$\eta E\|\nabla J(\mathbf{w}_k)\|^2 < 2E\sum_{\Delta=1}^{N}\left|\hat{e}_k\hat{d}_{k-\Delta} + \hat{e}_{k-\Delta}\hat{d}_k\right| \tag{61}$$

Using Jensen's inequality, (61) can be reduced further to result in a loose upper bound on the step-size.

$$\eta E\|\nabla J(\mathbf{w}_k)\|^2 < 2\sum_{\Delta=1}^{N}\left|E[\hat{e}_k\hat{d}_{k-\Delta}+\hat{e}_{k-\Delta}\hat{d}_k]\right| \tag{62}$$

Notice that the RHS of (62) now resembles the cost function in (46). Rearranging the terms, we get the upper bound in (58). ■

The important point is that this upper bound is practical as it can be numerically computed without any knowledge of the actual filter or the noise statistics. Further, the upper bound itself can be included in the update equation to result in a normalized stochastic gradient algorithm with improved speed of convergence.

VI. Simulation Results

Until this point, we presented new criteria and their associated algorithms to accurately estimate the parameters of a linear system in the presence of input noise. In this section, we will present some simulation results that validate the claims made earlier.

A. Estimation of System Parameters in White Noise using REW

The REW algorithm can be used effectively to solve the system identification problem in noisy environments. As we have seen before, by setting the value of $\beta = -0.5$, noise immunity can be gained for parameter estimation. We performed several experiments with different length filters and input SNRs. The results have been summarized in [10,13]. Accordingly, the REW algorithm significantly outperforms the RLS and the analytical TLS methods.

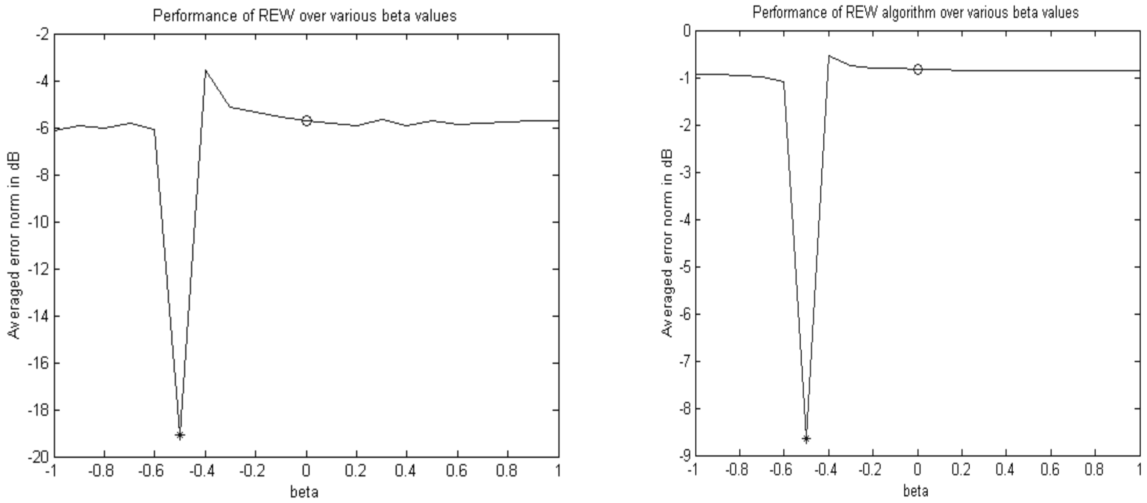
B. Effect of β and Weight Tracks of REW Algorithm

Recall that the REW algorithm can be used for any value of β. Interestingly, when β is positive, the EWC cost function is always positive (also convex) and the $E(\dot{e}^2)$ is nothing but the sample derivative of the error. Therefore, minimizing EWC with β>0 is equivalent to the joint minimization of MSE along with a smoothness constraint on the error. The benefits of such an augmented MSE minimization can perhaps be significant when modeling physical plants with constrained error dynamics. Yet another interesting aspect is the relationship of the EWC cost to the sample error entropy. It can be shown that the error derivative term $E(\dot{e}^2)$ is proportional to a sample estimate of the error entropy [5]. Thus minimization of EWC cost with β>0 implies a simultaneous minimization of MSE and error entropy. It is well-known that the error entropy optimization produces superior model estimates of nonlinear systems when compared to MSE [14].

However, when there is noise in the data and the objective is to obtain an unbiased parameter estimate of the underlying model, we claimed that $\beta = -0.5$ gives the best (unbiased) solution. We will now show the effect of β on the EWC parameter estimates. The SNR of the input signal was fixed at values 0dB and –10dB, the number of filter taps was set to 4 and the desired signal was noise free. We performed 100 Monte Carlo experiments and analyzed the average error vector norm defined in (63) values for $-1 \le \beta \le 1$. The results of the experiment are shown in Fig 2.

$$error = 20\log 10\left[\frac{\|\mathbf{w}_T - \mathbf{w}_*\|}{\|\mathbf{w}_T\|}\right] \tag{63}$$

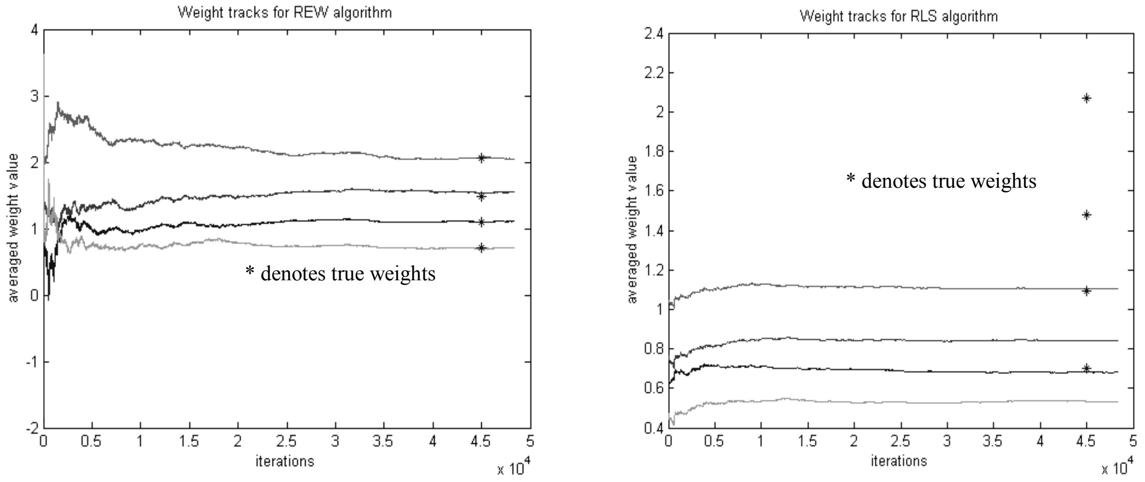
Notice that there is a dip at $\beta = -0.5$ (indicated by a "*" in the figure) and this clearly gives us the minimum parameter bias. This corresponds to the EWC solution. For $\beta = 0$, (indicated by a "o" in the figure) the REW algorithm reduces to the regular RLS giving a fairly significant bias in the parameter values.



(a) (b)

Figure 2. Performance of REW with different values of β when input SNR is (a) 0dB (b) -10dB.

Next the parameter β is set to –0.5 and SNR to 0dB, and the weight tracks are estimated for the REW and the RLS algorithms. Fig 3 shows the averaged weight tracks for both REW and RLS algorithms averaged over 50 Monte Carlo trials. Asterisks on the plots indicate the true parameters. The tracks for the RLS algorithm are smoother, but they converge to wrong values, which we have observed quite consistently. The weight tracks for the REW algorithm are noisier compared to those of the RLS, but they eventually converge to values very close to the true weights. Also, note that the REW weight tracks are noisier only during the initial stages of adaptation. This is in agreement with the theoretical arguments presented in section IV according to which the REW algorithm is sensitive to the smallest eigenvalue of the matrix $\mathbf{R}_L(n)$. Typically, the eigenvalues of $\mathbf{R}_L(n)$ can be smaller than those of $\mathbf{R}(n)$ because the latter is a diagonally dominant matrix unlike $\mathbf{R}_L(n)$. This is true because the diagonal of $\mathbf{R}_L(n)$ has $2\rho_x(L)$ whereas the regular covariance $\mathbf{R}(n)$ has $\rho_x(0) > |\rho_x(L)|$ on the diagonal.



(a) (b)

Figure 3. Weight tracks of the (a) REW algorithm and (b) RLS algorithm.

C. System Identification with Colored Input Noise

The experimental setup is similar to the block diagram shown in Fig 1. We generated 50000 samples of correlated clean input signal and passed it through an unknown random FIR filter to create a clean desired signal. Gaussian random noise was passed through a random coloring filter (FIR filter with 400 taps) and then added to the clean input signal. Three different input SNR values of 5, 0 and -10dB and three different true filter lengths of 5, 10 and 15 taps were used in the experiment. For each combination of SNR value and number of taps, 100 Monte Carlo runs were performed. For each trial, a different random coloring filter as well as input/desired data was generated. For the purpose of comparison, we computed the Wiener solution for MSE as well as the optimal solution given by equation (51). As before, we will utilize the error vector norm (dB) to quantify the performance of the methods. Fig 4 shows the histograms of the error vector norms for the proposed method as well as MSE. The inset plots in the figure show the summary of the histograms for each method. Clearly, the performance of the new criterion is superior

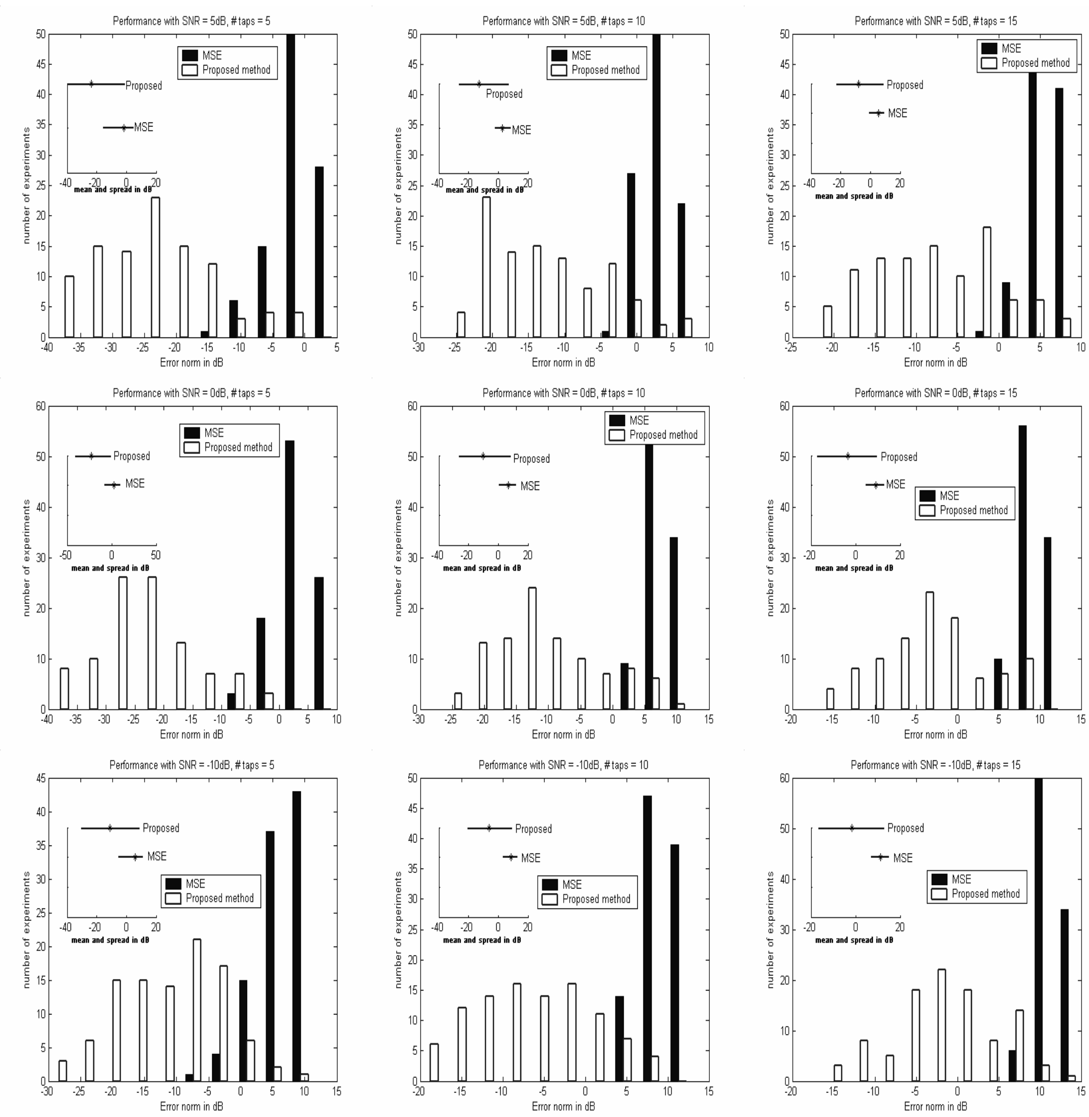


Figure 4. Histograms of the error vector norm obtained using MSE and the proposed criterion.

in every experiment given the fact that the criterion neither requires any knowledge of the noise statistics nor does it try to estimate the same from data.

D. System Identification with Stochastic Gradient Algorithm

We will use the stochastic gradient algorithm given by equation (53) to identify the parameters of a FIR filter in the presence of correlated input noise. A random four tap FIR filter was chosen as the true system. The input SNR (colored noise) was fixed at 5dB and the output SNR (white noise) was chosen to be 10dB. The step-sizes for the proposed method and the classical LMS algorithm were fixed at 1*e*-5 and 8*e*-4 respectively. One hundred Monte Carlo runs were performed and the averaged weight tracks over iterations are plotted for both algorithms in Fig 5. Note that our method gives a better estimate of the true parameters (shown by the square markers) than the LMS algorithm. The weight tracks of the proposed gradient method are noisier compared to those of LMS. One of the difficulties with the stochastic gradient method is the right selection of step-size. We have observed that in cases when the noise levels are very high, we require a very small step-size and hence the convergence time can be high. Additional gradient normalizations can be included to speed up the convergence. Also, the shape of the performance surface is dependent on the correlations of the input and the desired signals at different lags. If the performance surface is relatively flat around the optimal solution, we have observed that including a trivial momentum term in the update equation increases the speed of convergence. Additional experiments on the local stability of the method and the effects of undermodeling can be found in [15].

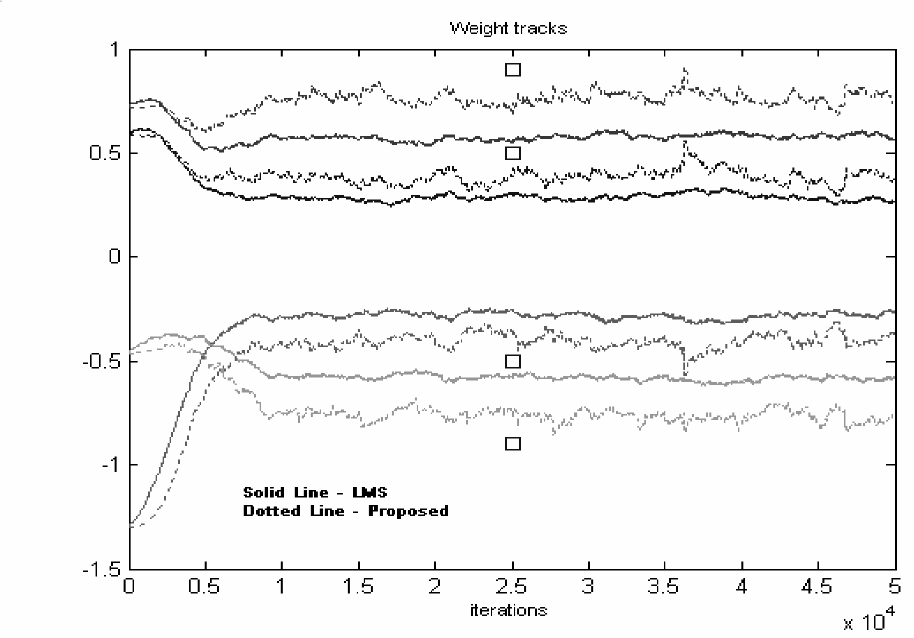


Figure 5. Weight tracks of the LMS and the proposed method.

E. Inverse Modeling and Control Using REW Algorithm

We will show the application of EWC for designing a model reference inverse controller. Fig 6 shows a block diagram of model reference inverse control [2]. Clearly, we require the plant parameters (which are typically unknown) to devise the controller. Once we have a model for the plant, the controller can be easily designed using conventional MSE minimization techniques. In this example, we will assume that the plant (AR system) transfer function is $P(z) = 1/(1 + 0.8z^{-1} - 0.5z^{-2} - 0.3z^{-3})$. The reference model is chosen to be an FIR filter with 5 taps. The block diagram for the plant identification is shown in Fig 7. Notice that the output of the plant is noise corrupted with white noise due to measurement errors. The SNR at the plant output was set to 0dB. We then ran the REW and RLS algorithms to estimate the model parameters given the noisy input and desired signals. The model parameters thus obtained are used to derive the controller (see Fig 6) using standard backpropagation of error. We then tested the adaptive controller-plant pair for trajectory tracking by feeding a random time series and observing the responses. Ideally, the controller-plant pair must follow the trajectory generated by the reference model. Fig 8 shows a histogram of the tracking errors. Notice

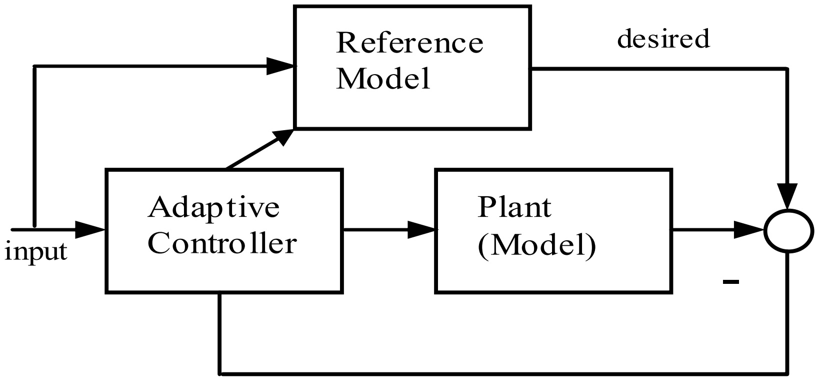


Figure 6. Block diagram for model reference inverse control.

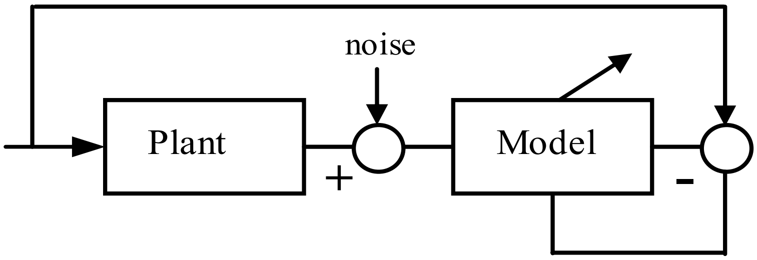


Figure 7. Block diagram for inverse modeling.

that the errors with REW controller are all concentrated around zero, giving an almost perfect controller for the plant. In contrast, the errors produced by the MSE based controller are high and could become worse if the SNR levels drop further.

VII. Discussion and Conclusions

Accurate parameter estimation with noisy data is a difficult problem that unfortunately becomes critical in many practical applications. Conventional MSE based methods have been shown to give biased parameters with noisy data. As a matter of fact, for the linear parameter estimation problem, the optimal Wiener solution for MSE

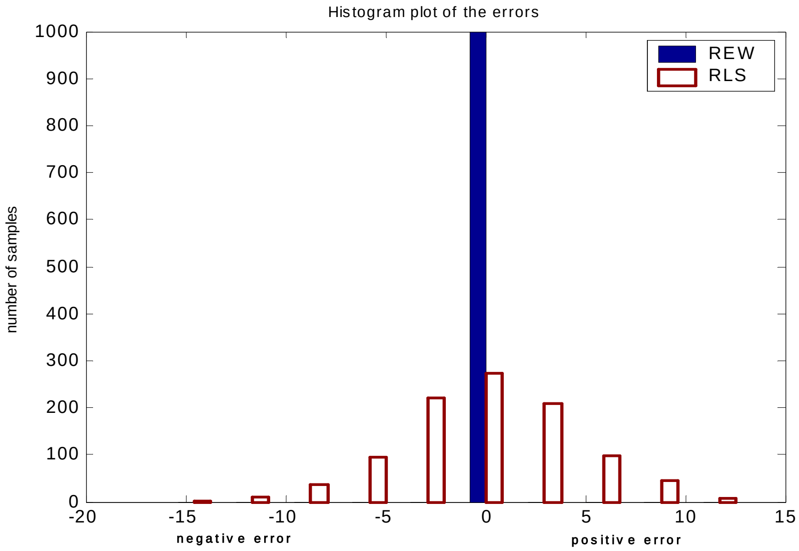


Figure 8. Histogram of tracking errors.

changes with noise statistics which is highly unacceptable. Other methods like TLS, and its extended versions as well as the Instrumental Variables (IV) have been widely used to solve this problem; but the underlying assumptions restrict their applicability. We recently proposed a new criterion called the *Error Whitening Criterion* (EWC) that can produce unbiased parameter estimates for linear systems even the data is corrupted by additive white noise. The criterion works with the error correlation instead of the error energy and achieves partial whiteness of the error which in turn results in an unbiased parameter estimate even with noisy data. In this paper, we first presented a Quasi-Newton type algorithm to solve for the optimal EWC solution. This algorithm is a truly fixed-point type method with O(N^2) complexity similar to the RLS algorithm. A detailed analysis of the algorithm was also presented.

EWC gives unbiased parameters only when the input noise is white. In the later half of this paper, we proposed another criterion which again exploits data correlations to accurately estimate the parameters when the input noise is colored. A stochastic gradient algorithm was developed to estimate the optimal solution. Brief convergence analysis of this gradient algorithm was presented.

Lastly, we showed the advantages of the proposed algorithms in the problem of system identification with noisy data. The algorithms can be used in many applications that require accurate parameter estimation which is exemplified in the design of a model based inverse controller described in this paper.

The *Error Whitening Criterion* and the modified criterion for colored noise coupled with the fast algorithms presented in this paper form a powerful tool that can be used in several engineering applications requiring accurate parameter estimation. The extensions of these methods to estimation problems involving colored input and colored output noise scenarios are currently being studied and preliminary successes are reported in [16].

Acknowledgements: This work was partially supported by the National Science Foundation NSF grant ECS-0300340.

References

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

2. B. Widrow, E. Walach. “**Adaptive Inverse Control**.” Prentice Hall, New Jersey, 1995.

3. T. Söderström, P. Stoica. “**System Identification**.” Prentice-Hall, London, United Kingdom, 1989.

4. B. Widrow, M.E. Hoff Jr. “Adaptive switching circuits.” **IRE Western Electric Show and Convention Record**, Part 4, pp. 96-104, August 23, 1960.

5. J.C. Principe, Y.N. Rao, D. Erdogmus. “Error Whitening Wiener Filters: Theory and Algorithms.” Chapter-10, **Least-Mean-Square Adaptive Filters**, S. Haykin, B. Widrow, (eds.), John Wiley, New York, 2003.

6. Y.N. Rao, D. Erdogmus, G.Y. Rao, J.C. Principe. “Stochastic Error Whitening Algorithm for Linear Filter Estimation with Noisy Data.” **Neural Networks**, vol. 16, no. 5-6, pp. 873-880, June 2003.

7. Y.N. Rao, J.C. Principe. “Efficient Total Least Squares Method for System Modeling using Minor Component Analysis.” Proceedings of the **IEEE Workshop on Neural Networks for Signal Processing XII**, pp. 259-268, September 2002.

8. S. van Huffel, J. Vanderwalle. “**The Total Least Squares Problem: Computational Aspects and Analysis**.” SIAM, Philadelphia, 1991.

9. J. Mathews, A. Cichocki. “**Total Least Squares Estimation**.” Technical Report, University of Utah, USA and Brain Science Institute Riken, Japan, 2000.

10. Y.N. Rao, D. Erdogmus, J.C. Principe. “Error Whitening Criterion for Adaptive Filtering: Theory and Algorithms.” To appear in **IEEE Transactions on Signal Processing**.

11. G.H. Golub, C.F. van Loan. “**Matrix Computations**.” The John Hopkins University Press, London, UK, 1996.

12. R. Bhatia. “**Matrix Analysis**.” Springer Verlag, New York, 1997.

13. Y.N. Rao, D. Erdogmus, G.Y. Rao, J.C. Principe. “Fast Error Whitening Algorithms for System Identification and Control.” **IEEE Workshop on Neural Networks for Signal Processing XIII**, pp. 309-318, September 2003.

14. D. Erdogmus, J.C. Principe. “An Error-Entropy Minimization Algorithm for Supervised Training of Nonlinear Adaptive Systems.” **IEEE Transactions on Signal Processing**, vol. 50, no. 7, pp. 1780-1786, July 2002.

15. Y.N. Rao, D. Erdogmus, J.C. Principe. “Accurate Linear Parameter Estimation in Colored Noise.” Accepted for publication in **International Conference on Acoustics, Speech and Signal Processing**, Canada, May 2004.

16. Y.N. Rao. “Augmented Error Criterion: Theory, Algorithms and Applications.” Ph.D. Dissertation, **University of Florida**, Gainesville, FL, May 2004.