ACCURATE LINEAR PARAMETER ESTIMATION IN COLORED NOISE

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

Computational NeuroEngineering Lab, University of Florida, Gainesville, FL 32611
{yadu,deniz,principe}@cnel.ufl.edu

ABSTRACT

Estimation of the parameters of an unknown system is an important problem in signal processing. The classical Mean Squared Error (MSE) criterion and its variants have been widely used to solve this problem. However, it is well known that MSE criterion produces biased parameter estimates when the signals of interest (especially the input) are corrupted with additive noise having arbitrary or no coloring (white). Alternative approaches require additional system constraints and explicit estimation of the noise covariances. Recently, we proposed a new criterion called the Error Whitening Criterion (EWC) along with associated algorithms that solved the problem when the additive disturbances are white. However, the performance of EWC is not satisfactory when the disturbances are correlated (colored). In this paper, we propose a method based on the principles of EWC that can consistently estimate the parameters of an unknown arbitrary linear system in colored input noise *without* estimating the noise covariances. We then present a novel stochastic gradient algorithm that estimates the optimal parameters in an on-line fashion. We will briefly discuss the convergence of this algorithm and present extensive simulation results to show the superiority of this criterion over MSE.

1. INTRODUCTION

Parameter estimation or system identification is a very important problem in signal processing and control. The framework for the conventional approaches that solve this problem is typically built around the popular Mean Squared Error (MSE) criterion [1]. This criterion offers cost-efficient stochastic (LMS) and fast converging recursive algorithms (RLS) that iteratively estimate the unknown system parameters. However, MSE has a genuine limitation that can seriously limit its applicability. The estimates obtained with MSE are biased when the signals of interest (input and output) are corrupted with additive noise with arbitrary coloring. The recently proposed *Error Whitening Criterion* (EWC) extends the MSE cost function and has been shown to produce unbiased parameter estimates when the additive noise is *white* [2], [3]. If the whiteness assumption is relaxed, EWC fails to give an improvement over MSE. There are other methods that attempt to solve this problem. Regalia gave a conceptual treatment for the IIR filter estimation based on equation-error techniques with the monic constraint replaced by a unit-norm constraint [4]. Douglas *et al*. extended the work to colored noise case in [5]. However, these methods require estimation of the noise covariances from the data, which is not desirable. The Instrumental Variable (IV) technique is traditionally limited to white noise, and the generalizations to the colored noise require additional prewhitening filters [6]. In this paper, we propose a method to estimate the unknown system parameters without computing the input noise covariance matrices under the assumption that the noise on the desired signal is white. Firstly, we will present the cost function and then derive the analytical solution that provides an unbiased estimate of the underlying system parameters. We restrict ourselves to the case of unknown linear FIR systems in this paper. Generalizations to the IIR filter estimation and the associated stability issues will be dealt in a later paper.

2. CRITERION

A traditional setting of the system identification problem is shown in fig 1. Suppose noisy training data pair $(\hat{\mathbf{x}}_k, \hat{d}_k)$ is provided, where $\hat{\mathbf{x}}_k \in \Re^N = \mathbf{x}_k + \mathbf{v}_k$ and $\hat{d}_k \in \Re^1 = d_k + u_k$ with $\mathbf{x}_k$ as the noise-free input vector at discrete time index *k*, $\mathbf{v}_k$, the additive noise vector with arbitrary covariance $\mathbf{V} = E[\mathbf{v}_k \mathbf{v}_k^T]$ 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 independent from the data pair and also independent from 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*. We will assume that the length of $\mathbf{w}$, the estimated weight vector is *N* (sufficient order case). Then, the error sample $\hat{e}_k$ is simply given by $\hat{e}_k = \hat{d}_k - \mathbf{w}^T \hat{\mathbf{x}}_k$. Consider the cost function in (1).

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

Consider a single term in 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,

$$\begin{aligned} E[\hat{e}_k \hat{d}_{k-\Delta}] &= \mathbf{w}_T^T E[\mathbf{x}_k \mathbf{x}_{k-\Delta}^T] \mathbf{w}_T - \mathbf{w}_T^T E[\mathbf{x}_k \mathbf{x}_{k-\Delta}^T] \mathbf{w} + E[u_k u_{k-\Delta}] \\ E[\hat{e}_{k-\Delta} \hat{d}_k] &= \mathbf{w}_T^T E[\mathbf{x}_{k-\Delta} \mathbf{x}_k^T] \mathbf{w}_T - \mathbf{w}_T^T E[\mathbf{x}_{k-\Delta} \mathbf{x}_k^T] \mathbf{w} + E[u_{k-\Delta} u_k] \end{aligned} \tag{2}$$

If we assume that the noise u_k is white, then $E[u_k u_{k-\Delta}] = 0$, and (2) reduces to functions of only the clean input and the weights. The input noise never multiplies itself; hence it gets eliminated. Further, the cost function in (1) simplifies to

$$J(\mathbf{w}) = \sum_{\Delta=1}^{N} \left| \mathbf{w}_T^T \mathbf{R}_\Delta \mathbf{w}_T - \mathbf{w}_T^T \mathbf{R}_\Delta \mathbf{w} \right| \tag{3}$$

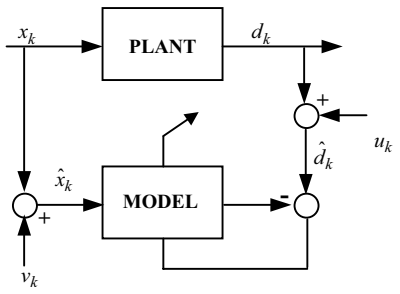

Figure 1. System Identification block diagram

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

$$\mathbf{R}_\Delta = E[\mathbf{x}_k \mathbf{x}_{k-\Delta}^T + \mathbf{x}_{k-\Delta}\mathbf{x}_k^T] \quad (4)$$

The matrix $\mathbf{R}_\Delta$ is symmetric, but indefinite and hence can have mixed eigenvalues. Also, observe that the cost function in (3) is *linear* in the weights **w**. If, for instance, we had 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 single unique solution $\mathbf{w}_*$ for the equation $J(\mathbf{w}_*) = 0$ and $\mathbf{w}_* = \mathbf{w}_T$.

Proof. For $J(\mathbf{w}) = 0$, $\mathbf{w}_T^T \mathbf{R}_\Delta \mathbf{w}_T - \mathbf{w}_T^T \mathbf{R}_\Delta \mathbf{w}$ must be zero for all selected 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

$$\begin{bmatrix} \mathbf{w}_T^T \mathbf{R}_1 \\ \mathbf{w}_T^T \mathbf{R}_2 \\ .... \\ \mathbf{w}_T^T \mathbf{R}_N \end{bmatrix} \mathbf{w} = 2 \begin{bmatrix} E[d_k d_{k-1}] \\ E[d_k d_{k-2}] \\ .... \\ E[d_k d_{k-N}] \end{bmatrix} \quad (5)$$

If the rows of the composite matrix on the left of **w** in (5) 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 true 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 $\boldsymbol{\varepsilon} = \mathbf{0}$ and this completes the proof. □

Note that each term inside the summation of equation (1) can be perceived as a constraint on the cross correlation between the desired response and the error signal. 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,

$$\mathbf{w}_* = 2 \begin{bmatrix} E[\hat{d}_k \hat{\mathbf{x}}_{k-1}^T + \hat{d}_{k-1}\hat{\mathbf{x}}_k^T] \\ E[\hat{d}_k \hat{\mathbf{x}}_{k-2}^T + \hat{d}_{k-2}\hat{\mathbf{x}}_k^T] \\ .... \\ E[\hat{d}_k \hat{\mathbf{x}}_{k-N}^T + \hat{d}_{k-N}\hat{\mathbf{x}}_k^T] \end{bmatrix}^{-1} \begin{bmatrix} E[\hat{d}_k \hat{d}_{k-1}] \\ E[\hat{d}_k \hat{d}_{k-2}] \\ .... \\ E[\hat{d}_k \hat{d}_{k-N}] \end{bmatrix} \quad (6)$$

Each row of the composite matrix can be estimated using simple correlators having linear complexity. 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 convenient matrix inversion lemma efficiently [7] to reduce the complexity of matrix inversion. This motivates the development of a low cost stochastic algorithm to compute and track the optimal solution given by (6).

3. STOCHASTIC ALGORITHM

Taking the expectation operator out of the cost function in (1), we obtain an instantaneous cost given by,

$$J(\mathbf{w}_k) = \sum_{\Delta=1}^{N} \left| \hat{e}_k \hat{d}_{k-\Delta} + \hat{e}_{k-\Delta} \hat{d}_k \right| \quad (7)$$

The direction of the stochastic gradient of (7) will then depend on the instantaneous cost and the resulting weight update equation is given by,

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \eta \sum_{\Delta=1}^{N} sign(\hat{e}_k \hat{d}_{k-\Delta} + \hat{e}_{k-\Delta}\hat{d}_k)(\hat{\mathbf{x}}_k \hat{d}_{k-\Delta} + \hat{\mathbf{x}}_{k-\Delta}\hat{d}_k) \quad (8)$$

where, $\eta > 0$ is a small 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 regular LMS type stochastic updates. We will now briefly discuss the convergence of this algorithm to the optimal solution both in the noisy as well as noise-free scenarios.

Lemma 2**.** In the noise-free case, (8) converges to the stationary point $\mathbf{w}_* = \mathbf{w}_T$ provided that the step size satisfies the following inequality at every update.

$$0 < \eta < \frac{2J(\mathbf{w}_k)}{\left\| \nabla J(\mathbf{w}_k) \right\|^2} \quad (9)$$

Proof. It is obvious from the previous discussions that the cost function in (8) 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 (8), we get,

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

Taking the norm of this error vector and allowing the error vector norm to decay asymptotically by forcing $\left\| \boldsymbol{\varepsilon}_{k+1} \right\|^2 < \left\| \boldsymbol{\varepsilon}_k \right\|^2$, we obtain the bound in (9). The error vector will eventually converge to zero by design, and since the gradient becomes null at the true solution: $\lim_{k\to\infty} \left\| \boldsymbol{\varepsilon}_k \right\|^2 \to 0$, thus $\lim_{k\to\infty} \mathbf{w}_k \to \mathbf{w}_* = \mathbf{w}_T$. □

Lemma 3. In the noisy data case, the stochastic algorithm in (8) converges to the stationary point $\mathbf{w}_* = \mathbf{w}_T$ in the mean provided that the step size is bound by the inequality

$$\eta < \frac{2 \sum_{\Delta=1}^{N} \left| E[\hat{e}_k \hat{d}_{k-\Delta} + \hat{e}_{k-\Delta} \hat{d}_k] \right|}{E\left\| \nabla J(\mathbf{w}_k) \right\|^2} \quad (10)$$

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 lemma. 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,

$$\|\boldsymbol{\varepsilon}_{k+1}\|^2 = \|\boldsymbol{\varepsilon}_k\|^2 - 2\eta \sum_{\Delta=1}^{N} sign(\hat{z}_{k,\Delta})\boldsymbol{\varepsilon}_k^T(\hat{\mathbf{x}}_k \hat{d}_{k-\Delta} + \hat{\mathbf{x}}_{k-\Delta}\hat{d}_k) + \eta^2 \|\nabla J(\mathbf{w}_k)\|^2 \quad (11)$$

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 (11) and letting $E\|\boldsymbol{\varepsilon}_{k+1}\|^2 < E\|\boldsymbol{\varepsilon}_k\|^2$ as in the previous case results in the following inequality.

$$\eta E\|\nabla J(\mathbf{w}_k)\|^2 < 2E\sum_{\Delta=1}^{N} \boldsymbol{\varepsilon}_k^T(\hat{\mathbf{x}}_k \hat{d}_{k-\Delta} + \hat{\mathbf{x}}_{k-\Delta}\hat{d}_k) sign(\hat{z}_{k,\Delta}) \quad (12)$$

Simplifying further, we get,

$$\eta E\|\nabla J(\mathbf{w}_k)\|^2 < 2E\sum_{\Delta=1}^{N} \left|\hat{e}_k \hat{d}_{k-\Delta} + \hat{e}_{k-\Delta}\hat{d}_k\right| \quad (13)$$

Using Jensen's inequality, (13) can be reduced further to result in a loose upper bound on the step-size.

$$\eta E\|\nabla J(\mathbf{w}_k)\|^2 < 2\sum_{\Delta=1}^{N} \left|E[\hat{e}_k \hat{d}_{k-\Delta} + \hat{e}_{k-\Delta}\hat{d}_k]\right| \quad (14)$$

Notice that the RHS of (14) now resembles the cost function in (1). Rearranging the terms, we get the upper bound in (10). □ The important point is that the bound is practical as it can be numerically computed without any knowledge of the actual filter or the noise statistics.

4. SIMULATIONS

System Identification: We will show the results obtained using the new criterion in the problem of 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. During each trial, a different random coloring filter as well as input/desired data was generated. We computed the Wiener solution for MSE as well as the optimal solution given by (6). The performance measure for the comparison was chosen as the error vector norm given by,

$$error\ \ norm = 20\log 10\left[\|\mathbf{w}_T - \mathbf{w}_*\|\right] \quad (15)$$

where, $\mathbf{w}_*$ is the optimal solution estimated using samples and $\mathbf{w}_T$ is the true weight vector. Fig. 2 shows the histograms of the error vector norms for the proposed method as well as MSE. The inset plots in fig. 2 show the summary of the histograms for each method. Clearly, the performance of the new criterion is superior 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.

Stochastic Algorithm: We will now analyze the performance of the stochastic gradient algorithm given by (8) in the same framework of system identification. 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 1e-5 and 8e-4 respectively. 100 Monte Carlo runs were performed and the averaged weight tracks over iterations are plotted for both algorithms in fig 3. 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 done 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.

In order to verify the local stability of the stochastic algorithm, we performed another experiment. This time, the four taps of the true system were [0.5, -0.5, 1, -1]. The initial weights for both LMS and the gradient algorithm in (8) were set to the true parameters. Both input and output SNR levels were kept at 10dB and the step-sizes were the same as in the previous experiment. Figure 4 shows the weight tracks for LMS and the proposed gradient algorithm. Notice that LMS diverges from this point immediately and converges to a biased solution. In comparison, the proposed algorithm shows very little displacement from the optimal solution (stable stationary point).

In the above experiments with system identification, we assumed that the filter order is at least equal to the true system. However, in many cases, this a priori knowledge is unavailable. In such cases, the problem becomes even harder with the presence of noise. In order to understand the behavior of the proposed method in the under-modeling case, we performed a simple experiment. We chose a 4-tap FIR system and tried to model it with a 2-tap adaptive filter. Figure 5 shows the weight tracks for both LMS and the stochastic algorithm. Surprisingly, the gradient algorithm converged to a solution that matched closely with the first two coefficients of the actual system. This encourages us to state (speculatively) that the criterion will try to find a solution that matches the actual system in some sense. However, there is still not enough evidence to claim that the proposed method can provide exact "coefficient matching." To the best of our knowledge, none of the techniques have the exact matching property given noisy data.

5. CONCLUSIONS

In this paper, we proposed a new criterion to solve the problem of system identification in the presence of colored input noise. Existing techniques either result in a biased solution or require explicit estimation of the noise covariance matrices to obtain an unbiased estimate of the unknown system. The new criterion exploits the correlations between the error and the desired signals at different lags and does not require the estimation of

the noise covariances. We further proved that the optimal solution with this cost function is always unique and approaches the underlying system under the sufficient order assumption. We then derived a simple stochastic gradient algorithm to estimate the optimal solution in an online manner. Brief discussions on the convergence were presented. Simulation studies showed the effectiveness of this criterion as well as the stochastic gradient algorithm. In this paper, we limited our focus to the sufficient order scenario only. In cases, when the model order is unknown the problem becomes more difficult and has been seldom addressed in literature. Currently, we are working on the theoretical aspects pertaining to the under-modeling case and the conditions under which the estimates obtained by the proposed method match with the actual system. Future work will also be focused around extending this method for handling colored noise in the desired signal.

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

6. REFERENCES

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

[2] J. C. Principe, Y.N. Rao, D. Erdogmus, "Error Whitening Wiener Filters: Theory and Algorithms," (Chapter 10) in *Least-Mean-Square Adaptive Filters*, S. Haykin, B. Widrow (eds), Wiley, Sep 2003.

[3] 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, Jun 2003.

[4] P. Regalia, "An Unbiased Equation Error Identifier and Reduced-Order Approximations," IEEE Trans. Signal Proc., vol. 42, no. 6, June 1994.

[5] S.C. Douglas, M. Rupp, "On bias removal and unit-norm constraints in equation-error adaptive filters," 30th Ann. Asilomar Conf. Sig., Syst., Comput., Pacific Grove, CA, Nov 1996.

[6] T. Söderström, P. Stoica. *System Identification*, Prentice-Hall, London, United Kingdom, 1989.

[7] G.H. Golub, C.F. van Loan, *Matrix Computations*, Baltimore, MD, Johns Hopkins Univ. Press, 1989.

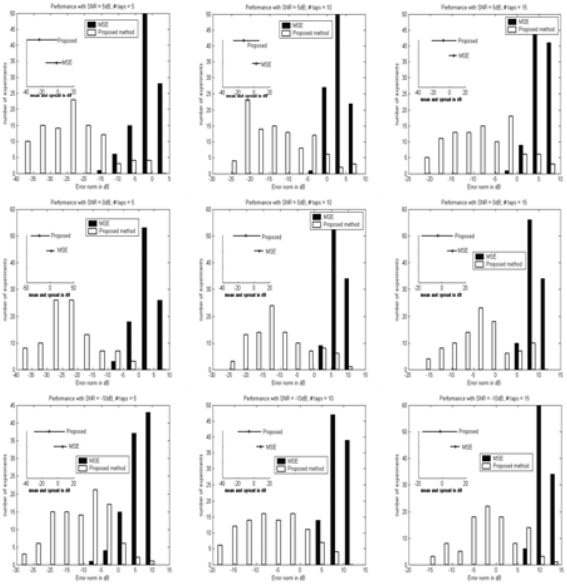

Figure 2- Histogram plots showing the error vector norm in dB for the proposed and MSE criteria.

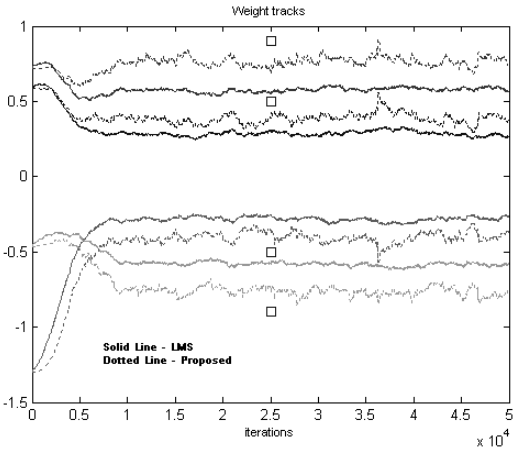

Figure 3- Weight tracks for LMS and the stochastic gradient algorithm in the system identification example.

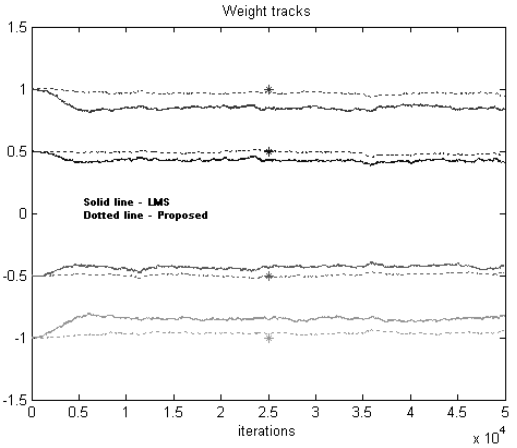

Figure 4- Weight tracks for LMS and the stochastic gradient algorithm showing stability around the optimal solution.

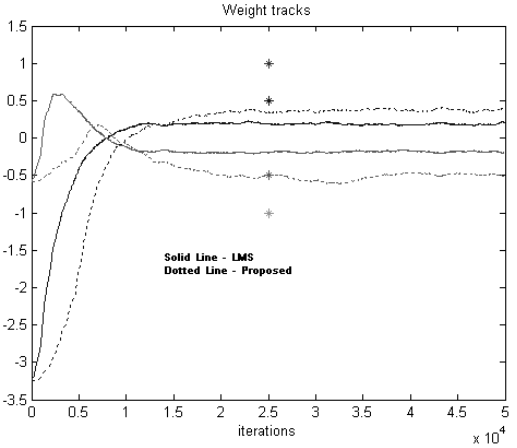

Figure 5- Weight tracks for LMS and the stochastic gradient algorithm in the case of undermodeling.