AN IMPROVED MINIMUM ERROR ENTROPY CRITERION WITH SELF ADJUSTING STEP-SIZE

Seungju Han[*], Sudhir Rao[*], Deniz Erdogmus[†], Jose Principe[*]

[*]CNEL, ECE Department, University of Florida, Gainesville, Florida, USA
[†] CSEE Department, OGI, Oregon Health and Science University, Portland, Oregon, USA

ABSTRACT

In this paper, we propose Minimum Error Entropy with self adjusting step-size (MEE-SAS) as an alternative to the Minimum Error Entropy (MEE) algorithm for training adaptive systems. MEE-SAS has faster speed of convergence as compared to MEE technique for the same misadjustment. We attribute this characteristic to automatic learning rate inherent in MEE-SAS where the changing step size helps the algorithm to take large "jumps" when far away from the optimal solution and small "jumps" when near the solution. We test the performance of both the algorithms for two classic problems of system identification and prediction. However, we show that MEE performs better than MEE-SAS in situations where tracking ability of the optimal solution is required like in the case of non-stationary signals.

1. INTRODUCTION

For many years, adaptive signal processing community had been using mean squared error (MSE) as the cost function to train the adaptive filters [1][3]. Apart from having favorable properties like elegant and tractable mathematical solutions, it is the simplicity of the Least Mean Square (LMS) algorithm which had made this cost function a workhouse and a benchmark standard in adaptive signal processing. Although this criterion had been successfully applied in many real and practical situations, it is clear that MSE only takes into account the second order statistics and is optimal in the case of Gaussian signals with linear filters [3].

In an effort to take into account higher order statistics, Least Mean Fourth (LMF) and its family of cost functions had been devised [5]. The corresponding family of filters had spurred a fresh interest in the field of adaptive filters with applications ranging from echo cancellation [8] to adaptive channel equalization in communications [9]. In his classic paper [5], Widrow shows that there exists a set of rules based on the increasing or decreasing nature of the moments of noise signal, which will help select the optimal filter among this family of adaptive filters.

It has been observed that LMF and its higher order counterparts are stable only in a very narrow range and a proper selection of learning rate μ is crucial. To overcome this difficulty, a linear combination of the cost functions of LMS and LMF filters using a single parameter $0 \leq \lambda \geq 1$ has been proposed [6][7]. Many variations of these filters have already been developed by adaptively estimating the optimal parameter λ or by recursive estimating the cost function [6][10].

Although the mixed norm family of filters helps extend the theory of adaptive filters using one or two higher order statistics, it is evident that constraining all the moments of the error signal at the same time will be an ideal solution. It is a well known fact that knowledge of all moments of a signal completely characterizes its pdf [15]. Thus a cost function based on better descriptors of error pdf is a productive research direction. Entropy, first introduced by Shannon [4], quantifies the average information contained in a pdf. Minimization of the cost function based on entropy constrains all the moments of the error pdf and is clearly an elegant way of extending MSE.

Information Theoretic Learning has become quite popular in recent years with the introduction of smooth sample estimators of entropy that do not require an explicit estimation of the pdf as proposed by Principe and collaborators [2][14]. The goal of entropy supervised learning follows the MSE framework. Given a set of input-desired signal pairs, the entropy of the output error over the training dataset is minimized. The procedure can be shown equivalent to minimizing the Csiszar distance between the probability distributions of the desired and system outputs [13].

Extensive comparisons have already been done between MEE and MSE techniques by Erdogmus et al. [11][14] and this is not the goal of this paper. Instead we extend the MEE by proposing a new technique called minimum error entropy with self adjusting step-size (MEE-SAS) which provides a faster speed of convergence for the same misadjustment by automatically selecting the step-size during learning. We provide a thorough comparison of these two search techniques in the case of system identification and prediction problems.

The paper is organized as follows. Section 2 introduces both MEE and MEE-SAS Information Theoretic Criteria. The analysis of the relation between MEE and MEE-SAS is discussed in section 3. Section 4 deals with simulation results and finally we conclude in section 5.

2. INFORMATION THEORETIC CRITERIA

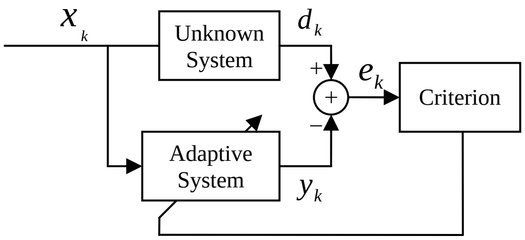

Fig.1. Adaptive System training using information theoretic criterion

Consider the supervised training scheme as shown in Fig.1. Since in practice, we are not given the error entropy, one needs to estimate this quantity nonparametrically from the training data samples. Renyi's quadratic entropy for a random variable e is given in terms of its pdf as

$$H_2(e) = -\log \int f_{e,\mathbf{w}}{}^2(e)\, de \qquad (1)$$

The pdf of a random variable e is estimated using Parzen window estimation with kernel function $\kappa_\sigma(\cdot)$ as shown in eq.2.The information potential is defined as the argument of the log. The minimum value of the entropy (maximum information potential) will be achieved for a δ -distributed random variable ($e_1 = e_2 = ... = e_N = 0$). Hence,

$$\begin{aligned} H_2(e) &= -\log\left[\frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N}\kappa_{\sigma\sqrt{2}}(e_j - e_i)\right] \geq -\log V(0) \\ V(e) &= \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N}\kappa_{\sigma\sqrt{2}}(e_j - e_i) \leq V(0) \end{aligned} \qquad (2)$$

Minimizing the entropy is equivalent to maximizing the information potential since the log is a monotonous function. Thus, the cost function $J(e)$ for Minimum Error Entropy criterion [14] is given as shown in eq.4

$$\text{MEE}: \ J(e) = \min_{\mathbf{w}} V(e) \qquad (3)$$

Since the information potential is smooth and differentiable, a simple search technique to find the maximum is to move in the direction of its gradient. This well known technique called steepest ascent has the form shown below

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu_1 \nabla V(e) \qquad (4)$$

where assuming a Gaussian kernel the gradient is

$$\nabla V(e) = \frac{1}{2N^2\sigma^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(e_i - e_j)G_{\sigma\sqrt{2}}(e_i - e_j)\left(\frac{\partial y_i}{\partial \mathbf{w}} - \frac{\partial y_j}{\partial \mathbf{w}}\right) \qquad (5)$$

As shown in eq.2, $V(e) < V(0)$ always; hence $V(0)$ provides the upper bound on the achievable $V(e)$. Seen from a different perspective, $V(0)$ is the "target" value to be reached in the information potential curve. Thus $(V(0) - V(e))$ is always a positive scalar quantity which does not change the direction of the weight vector but can be used to accelerate the search technique given in eq.4. This modified search technique is named MEE-SAS.

$$\begin{aligned} \mathbf{w}(n+1) &= \mathbf{w}(n) + \mu_2 (V(0) - V(e))\nabla V(e) \\ &= \mathbf{w}(n) + \eta^* \nabla V(e) \end{aligned} \qquad (6)$$

We can further note that there exists a cost function which gives rise to this gradient descent technique given by

$$\text{MEE-SAS}: \ J(e) = \min_{\mathbf{w}} \left[V(0) - V(e)\right]^2 \qquad (7)$$

Maximizing the information potential is equivalent to minimizing the proposed cost function. One intuitive way to understand the MEE-SAS algorithm is to consider it as a variant of the MEE with a variable step size $\eta^* = (V(0) - V(e))\mu_2$. The term $((V(0) - V(e))$ regulates the step size by giving acceleration when far away from the optimal solution and reducing the step size as the solution is approached.

For online training methods, the information potential can be estimated using stochastic version (SIG) [13] as shown in eq.8, where the sum is over the most recent L samples at time k . Thus for a filter order of length M, the complexity of MEE-SAS is similar to MEE and is equal to O(ML) per weight update

$$V(e) \approx \frac{1}{L}\sum_{i=k-L}^{k-1}\kappa_{\sigma\sqrt{2}}(e_k - e_i) \qquad (8)$$

The selection of kernel size σ is an important step in estimating the information potential and is critical to the success of these information theoretic criteria. In practice we anneal the kernel from a large value of 10^{-1} to a small value of 10^{-3} to avoid the problem of local minima.

3. ANALYSIS OF MEE-SAS AROUND THE OPTIMAL SOLUTION FOR LINEAR FILTERS

Suppose that the adaptive system is an FIR structure with a weight vector $\mathbf{w}$. The error samples are $e_k = d_k - \mathbf{w}^T\mathbf{x}_k$, where $\mathbf{x}_k$ is the input vector, formed by feeding the signal to a tapped delay line for the special case of an FIR filter. In order to minimize the cost function, we update the weights along the gradient direction with a certain step size μ.

$$\begin{aligned}\mathbf{w}(n+1) &= \mathbf{w}(n) + \mu\nabla J(e)\\ J(e) &= \left(V(0)-V(e)\right)^2\end{aligned} \tag{9}$$

To continue with our analysis, we consider the Taylor series expansion of the cost function truncated to the linear term of the gradient around the optimal weight vector $\mathbf{w}_o$.

$$\nabla J(e) = \nabla J_{\mathbf{w}_o}(e) + \frac{\partial \nabla J_{\mathbf{w}_o}(e)}{\partial \mathbf{w}}(\mathbf{w}-\mathbf{w}_o) \tag{10}$$

Notice that truncating the gradient at the linear term corresponds to approximating the cost function around the optimal point by a quadratic surface. The Hessian matrix of this cost function is $\mathbf{R}$, where $\mathbf{R}$ is given as

$$\begin{aligned}\mathbf{R} &= \frac{\partial \nabla J_{\mathbf{w}_o}(e)}{\partial \mathbf{w}}\\ &= 2\nabla V_{\mathbf{w}_o}(e)\nabla V_{\mathbf{w}_o}^T(e) - 2\left[V(0)-V_{\mathbf{w}_o}(e)\right]\frac{\partial \nabla V_{\mathbf{w}_o}(e)}{\partial \mathbf{w}}\\ \frac{\partial \nabla V_{\mathbf{w}_o}(e)}{\partial \mathbf{w}} &= \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N}\kappa''_{\sigma\sqrt{2}}(\Delta e_{\mathbf{w}}^{ji})(\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^T\end{aligned} \tag{11}$$

where $\Delta e_{\mathbf{w}}^{ji} = e_j - e_i = (d_j - d_i) - \mathbf{w}^T(\mathbf{x}_j - \mathbf{x}_i)$. Note that since $\mathbf{R}$ is the Hessian of the performance surface evaluated at a local minimum, its eigenvalues are positive, which results in the well known bound for the step size for convergence.

$$0 < \mu < \frac{2}{\max \lambda_k} \tag{12}$$

The expression for time constant is given by

$$\tau_k = \frac{1}{\ln(1+\mu\lambda_k)} \approx \frac{1}{\mu\lambda_k} \tag{13}$$

For large kernel size σ the surface can be approximated as a quadratic surface and the second derivative of the cost function at an arbitrary point is given by

$$\begin{aligned}\frac{\partial \nabla J(e)}{\partial w} &= -2\left(V(0)-V_{\mathbf{w}}(e)\right)\frac{\partial \nabla V_w(e)}{\partial \mathbf{w}} + 2\nabla V_{\mathbf{w}}(e)\nabla V_{\mathbf{w}}^T(e)\\ \mathbf{R} &= -2\left(V(0)-V_{\mathbf{w}}(e)\right)\tilde{\mathbf{R}} + \mathbf{D}\end{aligned} \tag{14}$$

where $\tilde{\mathbf{R}}$ is the Hessian for MEE. Further, large σ helps approximate the kernel as

$$\begin{aligned}\kappa_{\sigma\sqrt{2}}(\Delta e_{\mathbf{w}}^{ji}) &\approx c\left(1-\frac{(\Delta e_{\mathbf{w}}^{ji})^2}{4\sigma^2}\right)\\ \Rightarrow \kappa''_{\sigma\sqrt{2}}(\Delta e_{\mathbf{w}}^{ji}) &= \kappa_{\sigma\sqrt{2}}(\Delta e_{\mathbf{w}}^{ji})\left(\frac{1}{2\sigma^2}+\frac{(\Delta e_{\mathbf{w}}^{ji})^2}{4\sigma^4}\right)\\ &\approx c\left(1-\frac{(\Delta e_{\mathbf{w}}^{ji})^2}{4\sigma^2}\right)\left(\frac{1}{2\sigma^2}+\frac{(\Delta e_{\mathbf{w}}^{ji})^2}{4\sigma^4}\right)\\ &= l_{ji}\end{aligned} \tag{15}$$

Therefore the expression for $\tilde{\mathbf{R}}$ simplifies to a weighted Scatter Matrix given by

$$\tilde{\mathbf{R}} = \frac{\partial \nabla V_{\mathbf{w}}(e)}{\partial \mathbf{w}} \approx \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} l_{ji}(\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^T \tag{16}$$

Using eigendecomposition to simplify eq.14, we see that the eigenvalues of MEE and MEE-SAS are related as

$$\lambda_k \approx -2\left(V(0)-V_{\mathbf{w}}(e)\right)\tilde{\lambda}_k + \beta_k \tag{17}$$

where β_k is the square of the derivative of V with respect to kth weight vector. It is clear that MEE-SAS adaptively and automatically scales the eigenvalues of MEE to accelerate the search for the optimal solution.

4. SIMULATIONS

4.1. System Identification using FIR

The purpose of this simulation is to show the faster convergence of MEE-SAS compared to MEE for the same misadjustment. We consider a simple plant identification model with transfer function given by (order=9) [5]

$$\begin{aligned}H(z) = 0.1 + 0.2z^{-1} + 0.3z^{-2} + 0.4z^{-3} + 0.5z^{-4}\\ + 0.4z^{-5} + 0.3z^{-6} + 0.2z^{-7} + 0.1z^{-8}\end{aligned} \tag{18}$$

The input to both the plant and the adaptive filter is white Gaussian noise with unit power. We analyze this problem for both critical and under fitted models using the stochastic gradient (LMS type adaptation). A standard

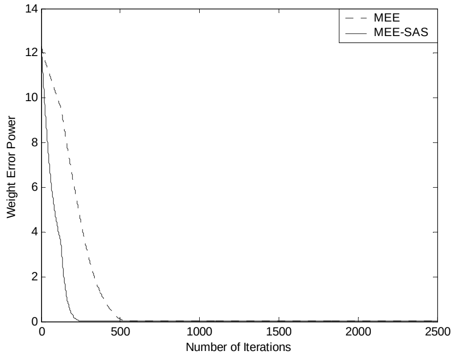

Fig.2. Average weight error power for critically fitted model

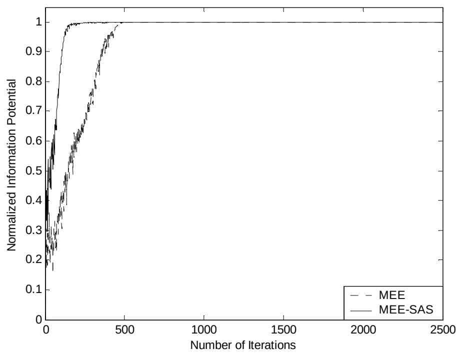

Fig.3. Normalized Information Potential

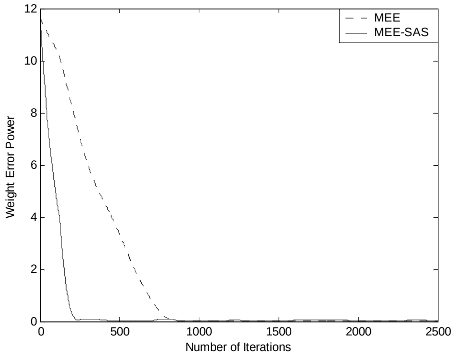

Fig.4. Average weight error power for under fitted model

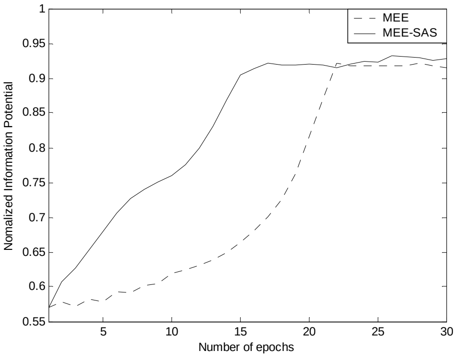

Fig.5. Normalized information potential for sin(x^2) prediction using TDNN

method of comparing the performance in system identification problems is plotting the weight error norm since this is directly related to misadjustment [5]. In each case the power of the weight noise (averaged over 125 samples) was plotted versus the number of iterations performed. The adaptive weights were initialized randomly at each instance. Further, in order to make the result independent of the input and weight initializations, we performed Monte-Carlo simulations with 100 different inputs and 100 different weight initializations for each input. To compare among different experiments we also plot the Normalized Information Potential (NIP) defined as

$$V_N(e) = \frac{V(e)}{V(0)} \tag{19}$$

which has a maximum value of 1.

4.1.1. Critically Fitted Model

Consider the case where the model of the adaptive filter is equal to that of the plant (Model Order = 9). In this case, ideally we can exactly track the output of the plant. Thus the maximum value of the NIP can be achieved (implies error is zero for all inputs). In Fig.2 the weight misadjustment values for the last 100 samples of error are 3.4×10^{-4} and 1.44×10^{-4} for MEE and MEE-SAS respectively (For practical purposes, we consider misadjustment values as zero for values less than 10^{-3}). Thus with the same misadjustment values, it can be observed that MEE-SAS converges in 250 iterations whereas MEE takes 500 iterations to converge. In Fig.3 we show the normalized information potential curve.

4.1.2. Under fitted Model

We analyze here the case where our filter order (N=7) is less than the true filter order. In this case we can expect a non-zero final error and hence the maximum achievable NIP will be less than unity. Fig.4 shows the averaged weight error power. MEE-SAS just takes 250 iterations to converge with a misadjustment of 0.0143 as compared to MEE which takes nearly 600 iterations with misadjustment of 0.0120. These results for linear systems are encouraging.

4.2. Non linear prediction using TDNN

As a second case study, we selected a non linear prediction problem using a Time Delay Neural Network

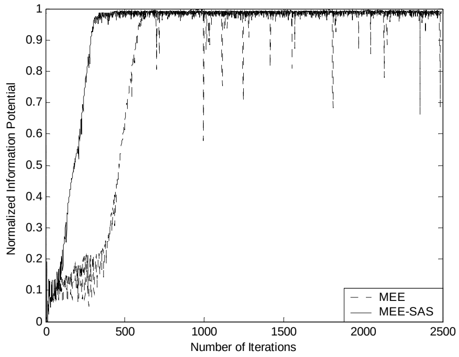

Fig.6. Information potential for online prediction of Mackey Glass time series

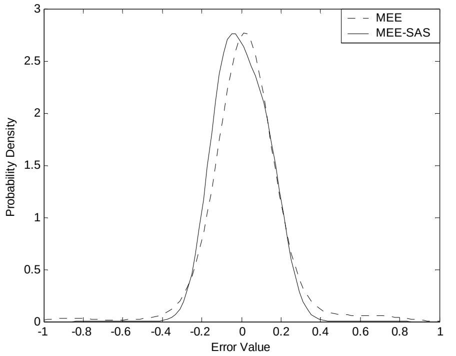

Fig.7. Probability density of error for last 200 samples

(TDNN) trained with the backpropagation algorithm [16]. Two hundred samples of $\sin(x^2)$ were selected for this purpose. Architecture of 20-10-1 TDNN with tanh non-linearity and one linear output PE was chosen. The training was carried out in batch mode for 30 epochs. Once again Monte-Carlo simulations were performed using 20 different weight initializations and the average performance was selected for comparison.

As seen from Fig.5, MEE-SAS has a fast transition towards the solution. It was observed in general that for a single simulation MEE quite often showed a piecewise transitional behavior. From adaptation point of view this has a very interesting interpretation. Recall that the inherent property of MEE-SAS is that it has a large effective step size when the present solution is far from the optimal one leading to large "jumps" in the bowl of the cost function. Since, in the initial phase of adaptation, large kernel size ensures a smoother learning curve surface, thus large transitions in these surfaces helps MEE-SAS to avoid most local solutions and reach directly in the vicinity of the global solution. Hence, although both algorithms reported a similar mean square error of 10^{-4}, there is much faster learning in the case of MEE-SAS than in the case of MEE.

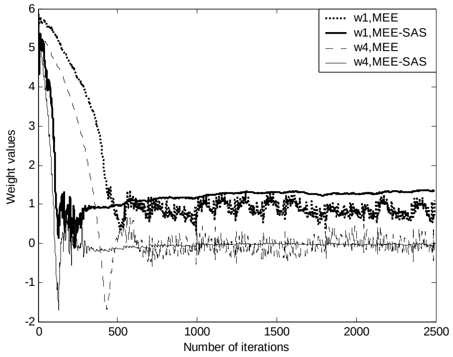

Fig.8. Two of the six weight tracks for MEE and MEE-SAS

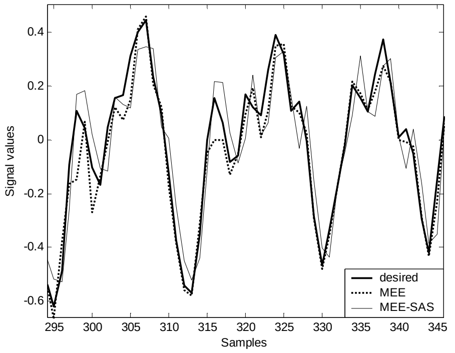

Fig.9. Prediction performance of the Mackey Glass time series

4.3. Chaotic Time Series prediction using a FIR filter

Finally, we consider a FIR filter for single-step prediction of the Mackey-Glass (MG) time series using the SIG estimation of information potential. The MG time series is generated by an MG system with delay parameter $\tau = 30$. The input vector consists of 6 (tap) consecutive samples of the MG time series.

We used the non stationary MG time series to compare the weight tracking ability of MEE and MEE-SAS algorithms. Due to online mode of simulation, SIG results in some misadjustment and variation about the optimal solution. In order to compare two algorithms, we find the step size for each algorithm to be such that it produces similar probability densities of error (e_k) for both cases within a window length of $L = 200$ as shown in Fig.7.

In Fig.6, MEE-SAS converges in about 400 iterations whereas MEE need 700 iterations to achieve the same level of performance. Note the large fluctuations in the information potential curve of MEE as compared to MEE-SAS. To investigate the effect of these large fluctuations, we plot two weight tracks and the predicted outputs of both the algorithms in Fig.8 and 9. The fluctuations in the MEE information potential curve translate into an ability to track the changes in the FIR optimal solution. This is confirmed in Fig.9. MEE performs better especially near high peaks and variations in MG signal.

The loss of “sensitivity” of MEE-SAS can be attributed to the extremely small value of $((V(0)-V(e))$ near the optimal solution which suppresses the transfer of information from the information potential gradient to the weight vectors. In non-stationary signals tracking these small changes in the location of the weight vector is crucial for good prediction. Therefore, MEE-SAS still suffers from a tradeoff between speed of convergence and tracking of the optimal solution. A compromise is to use MEE-SAS for faster convergence and then switch to MEE technique when the information potential is close to unity (which is achieved near the optimal solution). In this way we can double the speed of convergence as well as retain the ability to track the changes in weight vector.

5. CONCLUSIONS

In this paper, an information-theoretic supervised learning criterion for adaptive systems, namely, minimum error entropy with squared error (MEE-SAS) has been proposed. We demonstrated that MEE-SAS extends MEE by using an automatic adaptive step size to accelerate the search for the optimal solution.

Three different case studies were presented. The first one was a linear system identification problem using FIR. The second one extended this investigation to non-linear prediction using TDNN. In both these case studies it is clear that MEE-SAS converges much faster and avoids most of the local solutions compared to MEE.

Finally we tested the performance of these two algorithms on the adaptation of FIR filter for the short-term prediction of MG chaotic time series where tracking of optimal solution is crucial. We conclude that although MEE-SAS converges much faster than MEE, the lack of sensitivity near the optimal solution hinders tracking ability. Future direction of research includes overcoming this drawback by combining MEE and MEE-SAS so as to converge fast using MEE-SAS and then retain the tracking ability by switching to MEE. A detailed theoretical analysis of these two techniques needs further investigation.

Acknowledgments: This work was partially supported by NSF grant ECS-0300340. We thank the anonymous reviewers for their constructive suggestions which will be useful in further investigating this novel algorithm.

6. REFERENCES

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

[2] J.C. Principe, D. Xu and J. Fisher, “Information Theoretic Learning” in S. Haykin, *Unsupervised Adaptive Filtering*, Wiley, Newyork, vol I, pp 265-319, 2000.

[3] Simon Haykin, *Adaptive Filter Theory*, Prentice Hall, Upper Saddle River, 4th edition, 2001.

[4] C.E. Shannon, “A mathematical theory of communications,” Bell Syst. Tech. Journal, vol.27, pp. 379-423, 1948.

[5] E. Walach and B. Widrow, “The Least Mean Fourth(LMF) Adaptive Algorithm and its Family,” *IEEE trans. of Inf. Theory,* vol. IT 30, No.2, pp. 275-283, March 1984.

[6] O. Tanrikulu and J.A. Chambers, “ Convergence and steady-state properties of the least-mean mixed norm (LMMN) adaptive algorithm,” *IEEE Proc. of Vision, Image, Signal Processing,* vol.143, pp. 137-142, June, 1996.

[7] D.I. Pazaitis and A.G. Constantinides, “LMS+F algorithm,” *Electronic Letters*, 31, (17), pp. 1423-1424, 1995.

[8] A. Zerguine, C.F.N. Cowan and M. Bettayeb, “Adaptive Echo Cancellation using Least Mean Mixed-Norm Algorithm,” *IEEE trans. of Signal Processing,* vol.45, No.5, May 1997.

[9] C.F.N. Cowan, “Channel Equalization” in N. Kalouptsidis and S. Theodorids, *Adaptive System Identification and Signal Processing Algorithms*, Prentice-Hall, pp.388-406, 1993.

[10] C.F.N. Cowan and C. Rusu, “ Adaptive echo cancellation using cost function adaptation,” Conference Digest of Fourth IMA International Conference on Mathematics in Signal Processing, Warwick, UK, Dec.1996.

[11] D. Erdogmus and J.C. Principe, “Generalized Information Potential Criterion for Adaptive System Training,” *IEEE Trans. of Neural Networks*, vol.13, no.5, pp. 1035-1044, Sep. 2002.

[12] D. Erdogmus, J.C. Principe, K.E.Hild II, “Online entropy manipulation: Stochastic Information Gradient,” *IEEE Signal Processing Letters*, vol. 10, no. 8, pp. 242-245, Aug. 2003.

[13] D. Erdogmus, J.C. Principe, “An Entropy Minimization algorithm for Supervised Training of Nonlinear Systems,” *IEEE trans. of Signal Processing*, vol.50, No. 7, pp. 1780-1786, July 2002.

[14] D. Erdogmus, “Information Theoretic Learning: Renyi’s Entropy and its Applications to Adaptive System Training,” Ph.D Dissertation, University of Florida, Gainesville, FL, 2002.

[15] A. Papoulis and S.U. Pillai, *Probability, Random Variables and Stochastic Processes.* McGraw-Hill, New York, 2002.

[16] Simon Haykin, *Neural Networks-A Comprehensive Foundation*, Prentice Hall, Upper Saddle River, New Jersey, 1999.