Signal Processing - EURASIP (SUBMITTED)

A Minimum Error Entropy criterion with Self Adjusting Step-size (MEE-SAS)

Seungju Han[*], *Sudhir Rao*[*], *Deniz Erdogmus*[†], *Kyu-Hwa Jeong*[*], *Jose Principe*[*]

Corresponding author: Seungju Han (han@cnel.ufl.edu)

[*]CNEL, ECE Department, University of Florida, Gainesville, Florida, USA

[†] CSEE Department, Oregon Health and Science University, Portland, Oregon, USA

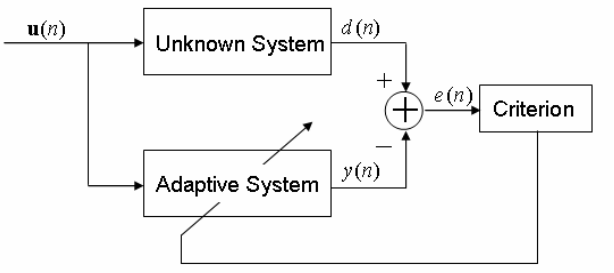
Tel/fax: (352)392-2682/392-0044, Emails: han@cnel.ufl.edu, sudhir@cenl.ufl.edu, derdogmus@ieee.org, khjeong@cnel.ufl.edu, principe@cnel.ufl.edu

Abstract

In this paper, we propose a 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 algorithm for the same misadjustment. We attribute the self adjusting step size property of MEE-SAS to its changing curvature as opposed to MEE which has a constant curvature. Analysis of the curvature shows that MEE-SAS converges faster in noisy scenarios than noise free scenario, thus making it more suitable for practical applications as shown in our simulations. Finally in case of nonstationary environment, MEE-SAS loses its tracking ability due to the "flatness" of the curvature near the optimal solution. We overcome this problem by proposing a switching MEE and MEE-SAS algorithm for non-stationary scenario which effectively combines the speed of MEE-SAS when far from the optimal solution with the tracking ability of MEE when near the solution. We demonstrate the performance of the switching algorithm in system identification in nonstationary environment.

Keywords: Minimum Error Entropy (MEE); MEE-SAS; Renyi's Entropy; Supervised Training.

I. INTRODUCTION



[Figure 1. Adaptive System training using information theoretic criterion]

Many adaptive signal processing problems such as system identification [1], noise canceling [2] and channel equalization [3] are typically solved in the framework of Figure 1, where the aim is to minimize the difference between the desired and the system outputs. For many years, the adaptive signal processing community has been using the mean-square-error (MSE) as the optimality criterion [4,5]. The main reason for the wide use of MSE lies in the various analytical and computational simplicities it brings coupled with the minimization of the error energy, which makes sense in the framework of linear signal processing. However, from a statistical point of view, MSE only takes into account the second order statistics and is therefore only optimal in the case of Gaussian signals and

linear filters.

In an effort to take into account higher order statistics, the mean-fourth-error (MFE) and its family of cost functions had been proposed by Walach and Widrow [6]. MFE and its higher order counterparts have faster adaptation for additive noise having a light-tailed probability distribution function (PDF), but are stable only in a very narrow range and a proper selection of learning rate is very crucial. To overcome this difficulty, a linear combination of the cost functions of LMS and LMF filters using a single parameter $0 \le \lambda \le 1$ has been proposed [7,[8]. Many variations of these filters have already been developed by adaptively estimating the optimal parameter λ or by recursive estimating the cost function [9].

In a statistical learning sense, especially for nonlinear signal processing, a better approach would be to constrain directly the information content of signals rather than simply their energy, if the designer seeks to achieve the best performance in terms of information filtering [10].

Entropy, first defined and proved useful by Shannon [11] and generalized by Alfred Renyi [12], is a scalar quantity that provides a measure for the average information contained in a given PDF. When entropy is minimized, all moments of the error PDF are constrained. The entropy criterion has been utilized as an alternative for MSE in supervised adaptation by Principe, Erdogmus and coworkers. For instance, minimization of error entropy (MEE) had been shown as a more robust criterion for dynamic modeling [13] and an alternative to MSE in other supervised learning applications using nonlinear systems [14].

The MEE cost function can be searched with gradient descent learning [10] or even second order search methods [15]. One of the difficulties with these search algorithms is the computational complexity that arises due to the estimation of entropy. Stochastic gradient algorithms have been derived to alleviate this problem [16]. This paper extends the class of search algorithms for the MEE by taking advantage of the fact that the cost maximizes the argument of the logarithm (a quantity that is called information potential), which is nonlinearly related to the samples. As will be demonstrated in this paper, a self adjusting step size can be defined, which requires only an initial stepsize selection for a more controlled gradient search (apart from the selection of the kernel size for information potential estimation). This new search algorithm will be called minimum error entropy with self adjusting step-size (MEE-SAS).

We can see that μ controls the behavior of the algorithm, and that two important goals are competing: for fast convergence, one would use a large step-size μ, but to achieve low steady-state MSE, a smaller step-size would be better. The ideal step size should decrease or increase as the overall system error decreases or increases. Various schemes for controlling the step-size of LMS have proposed in [17-20]. These schemes provide a “measure of error” to control the step size using the additional parameters. However, MEE-SAS provides a natural “Target” that is available to automatically control the algorithm step size. One intuitive way to understand the MEE-SAS

algorithm is to consider it as a variant to MEE with a variable step-size $\mu(n) = [V(\mathbf{0}) - V(\mathbf{e})]\mu$. When the error is large, adaptation is faster, when the error is small, adaptation is slower, resulting in a fast convergence with small steady-state error. We theoretically study and experimentally demonstrate that it also provides a faster adaptation than MEE for the same misadjustment.

The previous discussion also explains one of the drawbacks of MEE-SAS. When close to the optimal solution the effective step size of MEE-SAS is very small. In nonstationary environment with small perturbations in optimal solution, MEE-SAS looses its tracking ability due to this small effective step size. We overcome this defect, by using a switched MEE/MEE-SAS algorithm for non stationary environment which tracks the changing solutions very effectively.

The paper is organized as follows: First, Section II and III introduce MEE and MEE-SAS Information Theoretic Criteria. The structural analysis of the relation between MEE and MEE-SAS is discussed in section IV. In Section V we introduce the switching MEE and MEE-SAS for non stationary scenario. Section VI deals with simulation results and finally we conclude in section VII.

II. MEE CRITERION AND GRADIENT SEARCH ALGORITHM

Consider the supervised training scheme of Figure 1. For the evaluation of the error entropy, we seek to estimate entropy directly from the error samples. So, we will utilize initially the Parzen estimator of the error probability density function (PDF) $\hat{f}_e(\xi)$ given by

$$\hat{f}_e(\xi) = \frac{1}{N}\sum_{i=1}^{N} G_\sigma(\xi - e(i)) \tag{1}$$

where $G_\sigma(\cdot)$ denotes the Gaussian function with a radially symmetric variance σ^2 for simplicity. This estimator can be substituted in the Renyi's quadratic entropy definition given by

$$\begin{aligned} H(\mathbf{e}) &= -\log \int \left(\hat{f}_e(\xi)\right)^2 d\xi \\ &= -\log \int \left(\frac{1}{N}\sum_{i=1}^{N} G_\sigma(\xi - e(i)) \right)^2 d\xi \\ &= -\log V(\mathbf{e}). \end{aligned} \tag{2}$$

where $\mathbf{e} = [e(1),\ e(2),\ \ldots,\ e(N)]$.

The information potential $V(\mathbf{e})$ is defined as the argument of the log. The maximum value $V(\mathbf{0})$ of the information potential will be achieved for a Dirac $\delta-$distributed random variable $(e(1) = e(2) = \ldots = e(N))$. Using the fact that the integral of the product of two Gaussians is another Gaussian with a variance equal to the sum of the variances this procedure never needs the explicit evaluation of the integral, and yields a simple and effective nonparametric estimator for the information potential. It can be calculated in closed form from the samples using Gaussian kernel as

$$V(\mathbf{e}) = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} G_{\sigma\sqrt{2}}(e(j)-e(i)) \le V(\mathbf{0}). \tag{3}$$

Minimizing the entropy is equivalent to maximizing the information potential since the log is a monotonic function. Therefore, the cost function $J(\mathbf{e})$ for the MEE criterion is given by

$$J_{MEE}(\mathbf{e}) = \max_{\mathbf{w}} V(\mathbf{e}). \tag{4}$$

Since the information potential is smooth and differentiable by the Gaussian kernel properties, we can use its gradient vector to be used in the steepest ascent algorithm shown below

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu \nabla V(\mathbf{e}) \tag{5}$$

where $\nabla V(\mathbf{e})$ denotes the gradient of the information potential and the gradient is

$$\nabla V(\mathbf{e}) = \frac{1}{2N^2\sigma^2}\sum_{j=1}^{N}\sum_{i=1}^{N}[e(j)-e(i)]\cdot G_{\sigma\sqrt{2}}(e(j)-e(i))\cdot\left[\frac{\partial y(j)}{\partial \mathbf{w}} - \frac{\partial y(i)}{\partial \mathbf{w}}\right]. \tag{6}$$

A batch estimation of the gradient over N samples provides a simple estimation of the gradient, but notice that this procedure is $O(N^2)$. For online training methods, the information potential can be estimated using the stochastic information gradient (SIG) as shown in (7). Here the outer summation is dropped to get the stochastic version of the information gradient and the sum is taken over the most recent L samples at time n. Thus for a filter order of length M, the complexity of MEE is equal to $O(ML)$ per weight update.

$$V(\mathbf{e}) \approx \frac{1}{L}\sum_{i=1}^{L} G_{\sigma\sqrt{2}}(e(n)-e(n-i)) \tag{7}$$

The selection of the kernel size σ is an important step in estimating the information potential and is critical to the success of these information theoretic criteria. In particular, increasing the kernel size leads to a stretching effect on the performance surface in the weight space, which results in increased accuracy of the quadratic approximation around the optimal point [21]. So, we use a large enough kernel size during the adaptation process to guarantee that the operating point lies in the convex hull, and anneal it during training [14].

III. MEE-SAS CRITERION AND GRADIENT SEARCH ALGORITHM

As can be easily inferred from (3), $V(\mathbf{e}) \le V(\mathbf{0})$ always; hence $V(\mathbf{0})$ provides an upper bound on the achievable $V(\mathbf{e})$. Seen from a different perspective, $V(\mathbf{0})$ is the ideal "target" value to be reached in the information potential curve. Thus $[V(\mathbf{0}) - V(\mathbf{e})]$ is always a non-negative scalar quantity which does not change the direction of the weight vector but can be used to accelerate the conventional gradient search algorithm given in (5). This modified search algorithm is named MEE-SAS. The weight update in MEE-SAS becomes

$$\begin{aligned}\mathbf{w}(n+1) &= \mathbf{w}(n) + \mu\left[V(\mathbf{0}) - V(\mathbf{e})\right]\nabla V(\mathbf{e}) \\ &= \mathbf{w}(n) + \mu(n)\nabla V(\mathbf{e})\end{aligned} \tag{8}$$

where $\mu(n) = \mu\left[V(\mathbf{0}) - V(\mathbf{e})\right]$.

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

$$J_{MEE-SAS}(\mathbf{e}) = \min_{\mathbf{w}} \left[V(\mathbf{0}) - V(\mathbf{e})\right]^2. \tag{9}$$

Maximizing the information potential is equivalent to minimizing the cost function (9). Taking the gradient of this cost function as shown below gives the gradient descent method of (8).

$$\nabla J_{MEE-SAS}(\mathbf{e}) = -2\left[V(\mathbf{0}) - V(\mathbf{e})\right] \cdot \nabla V(\mathbf{e}) \tag{10}$$

THEOREM 1: **(Preservation of Optimal Solution)**

The stationary points of $f(V(\mathbf{e}))$ and their nature (minima, saddle, maxima) in the $\mathbf{w}$ space are the same as those of $V(\mathbf{e})$ if $f(\cdot)$ is strictly monotonic increasing on the range of $V(\mathbf{e})$.

Proof: $\frac{\partial f(V(\mathbf{e}))}{\partial \mathbf{w}} = \frac{\partial f(V(\mathbf{e}))}{\partial V(\mathbf{e})} \cdot \frac{\partial V(\mathbf{e})}{\partial \mathbf{w}}$. Since $\frac{\partial f(V(\mathbf{e}))}{\partial V(\mathbf{e})} > 0$ for all $V(\mathbf{e})$, $\frac{\partial f(V(\mathbf{e}))}{\partial \mathbf{w}} = \mathbf{0}$ iff $\frac{\partial V(\mathbf{e})}{\partial \mathbf{w}} = \mathbf{0}$. Also, $\frac{\partial^2 f(V(\mathbf{e}))}{\partial \mathbf{w} \partial \mathbf{w}^T} = \left[\frac{\partial^2 f(V(\mathbf{e}))}{\partial V(\mathbf{e})^2} \cdot \frac{\partial V(\mathbf{e})}{\partial \mathbf{w}^T}\right] \cdot \frac{\partial V(\mathbf{e})}{\partial \mathbf{w}} + \frac{\partial f(V(\mathbf{e}))}{\partial V(\mathbf{e})} \cdot \frac{\partial^2 V(\mathbf{e})}{\partial \mathbf{w} \partial \mathbf{w}^T}$. At all stationary points, $\frac{\partial V(\mathbf{e})}{\partial \mathbf{w}} = \mathbf{0}$, so $\frac{\partial^2 f(V(\mathbf{e}))}{\partial \mathbf{w} \partial \mathbf{w}^T} = \frac{\partial f(V(\mathbf{e}))}{\partial V(\mathbf{e})} \cdot \frac{\partial^2 V(\mathbf{e})}{\partial \mathbf{w} \partial \mathbf{w}^T}$. Since $\frac{\partial f(V(\mathbf{e}))}{\partial V(\mathbf{e})} > 0$ for all $V(\mathbf{e})$, the sign of the eigenvalues of the Hessian at the stationary points are unchanged, thus their nature is preserved. **Q.E.D.**

COROLLARY:

In MEE-SAS, $f(V(\mathbf{e})) = \left[V(\mathbf{0}) - V(\mathbf{e})\right]^2$, and the range of $V(\mathbf{e})$ is $[0,\ V(\mathbf{0})]$. In that case, the theorem holds.

Proof: In that case, the theorem holds except $\frac{\partial f(V(\mathbf{e}))}{\partial V(\mathbf{e})} = 0$. If $\frac{\partial f(V(\mathbf{e}))}{\partial V(\mathbf{e})} = 0$, then $V(\mathbf{0}) = V(\mathbf{e})$.

Also, $V(\mathbf{0}) = V(\mathbf{e})$ implies $\frac{\partial V(\mathbf{e})}{\partial \mathbf{w}} = 0$ since $V(\mathbf{e})$ is smooth and differentiable and $V(\mathbf{0}) \geq V(\mathbf{e})$ in (3). **Q.E.D.**

In order to continue with the convergence analysis of MEE-SAS, we consider a quadratic approximation for the information potential $V(\mathbf{e})$ by employing a Taylor series expansion truncated at the linear term for the gradient around the optimal weight vector.

$$V(\mathbf{e}) = V_{\mathbf{w}_*}(\mathbf{e}) + \frac{1}{2}\tilde{\mathbf{w}}(n)^T \mathbf{R}\tilde{\mathbf{w}}(n) \tag{11}$$

where the optimal solution is defined as $\mathbf{w}_* = \arg\max_{\mathbf{w}} V(\mathbf{e})$, and $\tilde{\mathbf{w}}(n) = \mathbf{w}_* - \mathbf{w}(n)$ and $\mathbf{R} := \nabla^2 V(\mathbf{e})$.

In fact, when the kernel size (the width of the window function used in the parzen estimator) tends to infinity, the local minima and maxima of the MEE disappear, leaving a unique, but biased, global minimum. This dilation property of the MEE is shown in [21]. Clearly, any continuous and (twice) differentiable cost function can be represented accurately with a quadratic approximation in some neighborhood of its global optimum. Then, provided that the kernel size is large enough during the adaptation process to guarantee that the operating point lies in the convex hull, one can perform global convergence analyzes of the steepest descent algorithm in the MEE and determine upper bounds on the step size of gradient-based optimization techniques to guarantee stability.

THEOREM 2: **(Step-size for Convergence)**

Assume that $V(e)$ is a quadratic surface with a Taylor series approximation given by $V(\mathbf{e}) = V_{\mathbf{w}_*}(\mathbf{e}) + \frac{1}{2}\tilde{\mathbf{w}}(n)^T \mathbf{R}\tilde{\mathbf{w}}(n)$ where $\tilde{\mathbf{w}}(n) = \mathbf{w}_* - \mathbf{w}(n)$ and $\mathbf{R} := \nabla^2 V(\mathbf{e})$. To ensure convergence of the MEE-SAS algorithm, a necessary condition is

$$0 < \mu(n) < \frac{-2}{\lambda_k}, \tag{12}$$

where $\mu(n) = \mu[V(\mathbf{0}) - V(\mathbf{e})]$ and λ_k is the smallest eigenvalue of the MEE cost function.

Proof: Let $\mathbf{R} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$, where $\mathbf{Q}$ and $\mathbf{\Lambda}$ denote the orthonormal eigenvector and diagonal eigenvalue matrices, respectively. Subtracting both sides of (8) from $\mathbf{w}_*$ and substituting $\nabla V(\mathbf{e}) = -\mathbf{R}\tilde{\mathbf{w}}(n)$ and $\mathbf{R} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$, we get

$$\begin{aligned} \tilde{\mathbf{w}}(n+1) &= \tilde{\mathbf{w}}(n) + \mu(n)\mathbf{R}\tilde{\mathbf{w}}(n) \\ &= \mathbf{Q}\left[\mathbf{I} + \mu(n)\mathbf{\Lambda}\right]\mathbf{Q}^T\tilde{\mathbf{w}}(n). \end{aligned} \tag{13}$$

The weight error along the natural modes ($\mathbf{v}(n) = \mathbf{Q}^T\tilde{\mathbf{w}}(n)$) is thus given by

$$\mathbf{v}(n+1) = \left[\mathbf{I} + \mu(n)\mathbf{\Lambda}\right]\mathbf{v}(n). \tag{14}$$

The expression for the k^{th} mode then becomes,

$$v_k(n+1) = (1+\mu(n)\lambda_k)v_k(n). \tag{15}$$

From (15) for stability, the step size should satisfy the constraint

$$|1+\mu(n)\lambda_k| < 1 \Leftrightarrow 0 < \mu(n) < \frac{-2}{\lambda_k}. \quad \textbf{Q.E.D.} \tag{16}$$

One intuitive way to understand the MEE-SAS algorithm is to consider it as a variant to MEE with a variable step size $\mu(n) = \mu[V(\mathbf{0}) - V(\mathbf{e})]$. The term $[V(\mathbf{0}) - V(\mathbf{e})]$ regulates automatically the step size by giving acceleration when far away from the optimal solution and reducing the step size as the solution is approached. This intuition can be mathematically proved as follows.

IV. STRUCTURAL ANALYSIS OF CONVERGENCE

THEOREM 3: **(Hessian Relation between MEE-SAS and MEE)**

Let $\tilde{\mathbf{R}}$ and $\mathbf{R}$ denote the Hessian of MEE-SAS and MEE respectively. The relation between the Hessian of the MEE and MEE-SAS is the following,

$$\tilde{\mathbf{R}} = -c\mathbf{R} + \left(\tilde{\mathbf{w}}(n)^T \mathbf{R}\tilde{\mathbf{w}}(n)\right)\mathbf{R} + 2\mathbf{R}\,\tilde{\mathbf{w}}(n)\,\tilde{\mathbf{w}}(n)^T\,\mathbf{R}^T. \tag{17}$$

where $c = 2\left[V(\mathbf{0}) - V_{\mathbf{w}_*}(\mathbf{e})\right]$.

Proof: Differentiating (9) twice with respect to the weight vector produces $\tilde{\mathbf{R}} = -2[V(\mathbf{0}) - V(\mathbf{e})]\nabla^2 V(\mathbf{e}) + 2\nabla V(\mathbf{e})\nabla V(\mathbf{e})^T$ and (17) is obtained by substituting (11) and $\nabla V(\mathbf{e}) = -\mathbf{R}\tilde{\mathbf{w}}(n)$. **Q.E.D.**

From the above equation (17), using the eigen-decomposition of MEE-SAS ($\tilde{\mathbf{R}} = \tilde{\mathbf{Q}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{Q}}^T$) and MEE ($\mathbf{R} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$), and transforming the coordinates ($\mathbf{v}(n) = \mathbf{Q}^T\tilde{\mathbf{w}}(n)$) into the natural modes, we obtain,

$$\begin{aligned}
\tilde{\mathbf{Q}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{Q}}^T &= -c\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T + \left(\tilde{\mathbf{w}}(n)^T\,\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T\tilde{\mathbf{w}}(n)\right)\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T + 2\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T\,\tilde{\mathbf{w}}(n)\,\tilde{\mathbf{w}}(n)^T\left(\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T\right)^T \\
&= \mathbf{Q}\left[-c\mathbf{\Lambda} + \left(\tilde{\mathbf{w}}(n)^T\,\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T\tilde{\mathbf{w}}(n)\right)\mathbf{\Lambda} + 2\mathbf{\Lambda}\mathbf{Q}^T\,\tilde{\mathbf{w}}(n)\,\tilde{\mathbf{w}}(n)^T\mathbf{Q}\mathbf{\Lambda}^T\right]\mathbf{Q}^T \\
&= \mathbf{Q}\left[-c\mathbf{\Lambda} + \left(\mathbf{v}(n)^T\,\mathbf{\Lambda}\,\mathbf{v}(n)\right)\mathbf{\Lambda} + 2\mathbf{\Lambda}\,\mathbf{v}(n)\,\mathbf{v}(n)^T\,\mathbf{\Lambda}\right]\mathbf{Q}^T
\end{aligned} \tag{18}$$

where $c = 2[V(\mathbf{0}) - V_{\mathbf{w}_*}(\mathbf{e})]$. If we can determine the eigendecomposition of the matrix $[-c\mathbf{\Lambda} + \left(\mathbf{v}(n)^T\,\mathbf{\Lambda}\,\mathbf{v}(n)\right)\mathbf{\Lambda} + 2\mathbf{\Lambda}\,\mathbf{v}(n)\,\mathbf{v}(n)^T\,\mathbf{\Lambda}]$, which is denoted by $\mathbf{\Sigma}\mathbf{D}\mathbf{\Sigma}^T$, where $\mathbf{\Sigma}$ is orthonormal and $\mathbf{D}$ is diagonal, then (17) becomes

$$\tilde{\mathbf{Q}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{Q}}^T = \mathbf{Q}\mathbf{\Sigma}\,\mathbf{D}\,\mathbf{\Sigma}^T\mathbf{Q}^T. \tag{19}$$

By direct comparison, the eigenvectors and the eigenvalues are determined to be

$$\tilde{\mathbf{Q}} = \mathbf{Q}\boldsymbol{\Sigma}, \quad \tilde{\boldsymbol{\Lambda}} = \mathbf{D}. \tag{20}$$

The entries of $\boldsymbol{\Sigma}\mathbf{D}\boldsymbol{\Sigma}^T$ are found as follows: The i^{th} diagonal entry is $-c\lambda_i + (\sum_{j=1}^{M} \lambda_j v_j^2)\lambda_i + 2\lambda_i^2 v_i^2$ and the $(i, j)^{\text{th}}$ entry is $2\lambda_i \lambda_j v_i v_j$, where λ_i is the i^{th} diagonal entry of $\boldsymbol{\Lambda}$ and v_i the i^{th} entry of $\mathbf{v}(n)$.

However, especially if $\mathbf{v}(n)$ is small, the matrix $[-c\boldsymbol{\Lambda} + \left(\mathbf{v}(n)^T \boldsymbol{\Lambda}\, \mathbf{v}(n)\right)\boldsymbol{\Lambda} + 2\boldsymbol{\Lambda}\, \mathbf{v}(n)\, \mathbf{v}(n)^T \boldsymbol{\Lambda}]$ is diagonally dominant; hence (due to the Gershgorin theorem) its eigenvalues will be close to those of the diagonal portion $-c\boldsymbol{\Lambda}$. In addition, its eigenvectors will also be close to identity (i.e., the eigenvectors of the diagonal portion of the sum).

Consider the special case when we are moving along one of the eigenvectors ($\mathbf{v} = [0, \ldots, v_k, \ldots, 0]^T$). Then the expressions simplify to the following.

$$\tilde{\Lambda} = \begin{bmatrix} -c\lambda_1 + \lambda_k \lambda_1 v_k^2 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & -c\lambda_2 + \lambda_k \lambda_2 v_k^2 & & 0 & & 0 \\ \vdots & & \ddots & & & \vdots \\ 0 & 0 & & -c\lambda_k + 3\lambda_k^2 v_k^2 & & 0 \\ \vdots & & & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & -c\lambda_M + \lambda_k \lambda_M v_k^2 \end{bmatrix} \tag{21}$$

In real scenarios, there exist modes which converge slower than others due to the eigenvalue spread. If we analyze the convergence along the principal axis of $\mathbf{R}$, it is easy to see that we obtain

$$\begin{aligned} \tilde{\lambda}_j &= -2[V(\mathbf{0}) - V_{\mathbf{w}_*}(\mathbf{e})]\lambda_j + \lambda_k \lambda_j v_k^2 \qquad \forall\, j \neq k \\ \tilde{\lambda}_k &= -2[V(\mathbf{0}) - V_{\mathbf{w}_*}(\mathbf{e})]\lambda_k + \lambda_k^2 v_k^2 + 2\lambda_k^2 v_k^2 \\ &= -2[V(\mathbf{0}) - V_{\mathbf{w}_*}(\mathbf{e})]\lambda_k + 3\lambda_k^2 v_k^2 \qquad j = k. \end{aligned} \tag{22}$$

When the weights are close to the optimal solution $v_k^2 \approx 0$, therefore the eigenvalues are proportional to the eigenvalues of the MEE cost which is quadratic. On the other hand, when the weights are far from the solution, v_k^2 is large and thus the second term dominates and the weights are proportional to the square of the original eigenvalues. A consequence of this is that MEE-SAS has the remarkable property of changing curvature. This is attributed to the fact that the eigenvalue of MEE-SAS $\tilde{\lambda}_k$ is quadratically related to the eigenvalues of MEE λ_k when the weights are far from the solution and is linearly related when near the optimal solution. Also note that the convergence along the k^{th} natural mode is faster than other modes due to the extra $2\lambda_k^2 v_k^2$ term when the weights are far from the optimal solution.

For each natural mode v_k in (19), the relationship between the eigenvalue of MEE-SAS ($\tilde{\lambda}_k$)

and that of MEE (λ_k) is

$$\begin{aligned}\tilde{\lambda}_k &= -2\left[V(\mathbf{0}) - V_{\mathbf{w}_*}(\mathbf{e})\right]\lambda_k + 3\lambda_k^2 v_k^2 \\ &= -\lambda_k\left(c - 3\lambda_k v_k^2\right)\end{aligned} \tag{23}$$

where $c = 2[V(\mathbf{0}) - V_{\mathbf{w}_*}(\mathbf{e})]$. Since we *maximize* the cost function $V(\mathbf{e})$ in MEE, the eigenvalues λ_k of its Hessian are negative. Similarly, for MEE-SAS, the *minimization* of its cost function makes $\tilde{\lambda}_k$ positive. The shape of the bowl is quadratic at each natural mode for MEE-SAS and the turning point of curvature occurs when

$$c - 3\lambda_k v_k^2 = 1 \text{ so that } \tilde{\lambda}_k = -\lambda_k. \tag{24}$$

From (21), we analyze specifically,

$$v_k = \pm\sqrt{\frac{c-1}{3\lambda_k}}. \tag{25}$$

Using the non-negative property of *c* and the form of (22), we get

$$0 \le c \le 1. \tag{26}$$

In (26), $c = 0$ implies $V(\mathbf{0}) = V_{\mathbf{w}_*}(\mathbf{e})$, whereas $c = 1$ implies $v_k = 0$ (i.e., $\mathbf{w} = \mathbf{w}_*$).

It is interesting to note that the location of the turning point of curvature depends on c as seen in (25), which means that it depends on the achievable final error. The larger the final error, the faster is the convergence.

THEOREM 4: **(Turning Point of Curvature)**

The point at which the curvature changes from higher than second order to second order is closer to the optimal solution when $V(\mathbf{0}) \neq V_{\mathbf{w}_*}(\mathbf{e})$ than in the case when $V(\mathbf{0}) = V_{\mathbf{w}_*}(\mathbf{e})$.

Proof: When $V(\mathbf{0}) = V_{\mathbf{w}_*}(\mathbf{e})$, the turning point of curvature is $v_k = \pm\sqrt{\frac{-1}{3\lambda_k}}$, while it is $v_k = \pm\sqrt{\frac{c-1}{3\lambda_k}},\ 0 < c \le 1$ when $V(\mathbf{0}) \neq V_{\mathbf{w}_*}(\mathbf{e})$. So, we get the following result, $\sqrt{\frac{-1}{3\lambda_k}} > \sqrt{\frac{c-1}{3\lambda_k}}\ (\because c > 0)$.

Q.E.D.

$c = 0$	$0 < c < 1$	$c = 1$
$v_k = \pm\sqrt{\frac{-1}{3\lambda_k}}$	$v_k = \pm\sqrt{\frac{c-1}{3\lambda_k}}$	$v_k = 0$

[Table 1.Location of the turning point of curvature]

Thus, the turning point v_k of curvature is farther from the optimal solution for the zero error adaptation case than for the non-zero error case. Since this point marks the change of curvature from 4th order to 2nd order, this implies that for practical scenarios (i.e. $V(\mathbf{0}) \neq V_{\mathbf{w}_*}(\mathbf{e})$), the curvature is going to be 4th order, leading to much faster convergence than MEE for the same initial step size.

V. SWITCHING SCHEME BETWEEN ADAPTIVE ALGORITHMS

One disadvantage of MEE-SAS is the insensitivity of the algorithm due to the "flatness" of the surface near the optimal solution. In the case where we need to track small changes in weight vector, this property would hinder the tracking ability of MEE-SAS. This was exactly observed in prediction of non-stationary Mackey-Glass (MG) time series where there is small perturbation of the optimal weight [24]. The loss of "sensitivity" of MEE-SAS can be attributed to the extremely small value of $[V(\mathbf{0}) - V(\mathbf{e})]$ near the optimal solution which suppresses the transfer of information from the information potential gradient to the weight vectors.

We are trying to apply MEE and MEE-SAS combined algorithm for nonstationary signals where tracking is very important. In order to decide the switching time to maximize convergence speed, an analytical criterion needs to be developed.

The dynamics of adaptation can be understood in terms of energy minimization in the context of Lyapunov stability theory [25]. Lyapunov energy function is a method for analyzing the convergence characteristics of dynamic systems. In our case, we are using it to analyze the speed of convergence. Simply, the faster the Lyapunov energy decreases, the faster we are getting towards the optimal solution, especially since our energy function is based on the criterion that needs to be optimized.

Specifically, consider the MEE-SAS criterion as a Lyapunov energy function. For simplicity, suppose that adaptation is being performed in continuous-time (which could be easily approximated by the typical discrete-time update rules used in practice). We have the following energy function and the continuous-time learning rule:

$$J_{MEE-SAS} = [V(\mathbf{0}) - V(\mathbf{e})]^2 \tag{27}$$

$$\dot{\mathbf{w}} = \frac{\partial \mathbf{w}}{\partial t} = -\mu \frac{\partial J_{MEE-SAS}}{\partial \mathbf{w}}^T \tag{28}$$

From this, we obtain the following temporal dynamics for the Lyapunov energy that describes the learning rule:

$$\begin{aligned} \dot{J}_{MEE-SAS} &= -2[V(\mathbf{0}) - V(\mathbf{e})] \frac{\partial V(\mathbf{e})}{\partial \mathbf{w}} \dot{\mathbf{w}} \\ &= -4\mu [V(\mathbf{0}) - V(\mathbf{e})]^2 \left\| \frac{\partial V(\mathbf{e})}{\partial \mathbf{w}} \right\|^2 \end{aligned} \tag{29}$$

On the contrary, the regular MEE rule would have the following energy function and update rule:

$$J_{MEE} = [V(\mathbf{0}) - V(\mathbf{e})] \tag{30}$$

$$\dot{\mathbf{w}} = -\mu \frac{\partial J_{MEE}(\mathbf{e})}{\partial \mathbf{w}}^T \tag{31}$$

This corresponds to the following temporal dynamics for the minimization of energy:

$$\dot{J}_{MEE} = -\mu \frac{\partial V(\mathbf{e})}{\partial \mathbf{w}} \dot{\mathbf{w}} = -\mu \left\| \frac{\partial V(\mathbf{e})}{\partial \mathbf{w}} \right\|^2 \tag{32}$$

From (29) and (32), the general switching time is determined as

$$\left| \dot{J}_{MEE-SAS} \right| = \left| \dot{J}_{MEE} \right| . \tag{33}$$

Therefore, in the region satisfying the condition $\left| \dot{J}_{MEE-SAS} \right| > \left| \dot{J}_{MEE} \right|$, MEE-SAS should be used since MEE-SAS converges faster than MEE, otherwise MEE is used. However, the application of the switching decision expression (33) to the stochastic gradient search, high computational complexity (the computational complexity of both MEE and MEE-SAS) is required due to the parallel computation of both algorithms. Instead, we can modify simply (33) to read

$$\begin{aligned} \left| \dot{J}_{MEE-SAS} \right| > \left| \dot{J}_{MEE} \right| &\Leftrightarrow 4\mu_{MEE-SAS} [V(\mathbf{0}) - V(\mathbf{e})]^2 \left\| \frac{\partial V(\mathbf{e})}{\partial \mathbf{w}} \right\|^2 > \mu_{MEE} \left\| \frac{\partial V(\mathbf{e})}{\partial \mathbf{w}} \right\|^2 \\ &\Leftrightarrow V(\mathbf{e}) < V(\mathbf{0}) - \frac{1}{2} \sqrt{\frac{\mu_{MEE}}{\mu_{MEE-SAS}}} . \end{aligned} \tag{34}$$

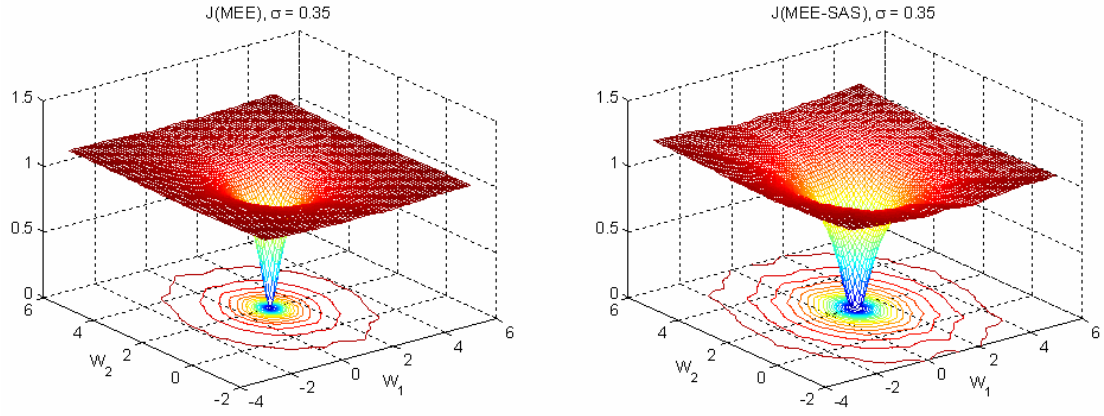
In (34), we need to check just the information potential at each iteration and compare it with a constant, which is evaluated with the learning rates of MEE and MEE-SAS.

VI. SIMULATION RESULTS AND DISCUSSION

In this section, we will present three numerical examples to demonstrate the theoretical conclusions drawn in the preceding section. These include the effect of kernel size and noise on the volume of the region of valid quadratic approximation. In addition, the results of a series of Monte Carlo simulations that illustrate the comparative performance of the proposed MEE-SAS criterion versus MEE in supervised training of moving average models. In our second simulation we present the performance of MEE-SAS and MEE for nonlinear system identification for completeness. Our last simulation shows the tracking ability of the switching algorithm in non stationary environment. To quantify the performance of different algorithms we create this environment using two linear systems by switching suddenly between the two systems. To make the problem more difficult we constantly change the gain of the linear systems in between the switching.

A. First Study: Curvature Analysis of MEE and MEE-SAS

In this simulation, for visualization purposes, we used a two-tap FIR filter for which the training data is also generated by a two-tap FIR filter with weight vector $\mathbf{w}_* = [1, \quad 2]^T$. Thus, in the system identification scheme, both the unknown system and the adaptive system have the same structure. The input to both the unknown system and the adaptive system is white Gaussian noise with unit power.



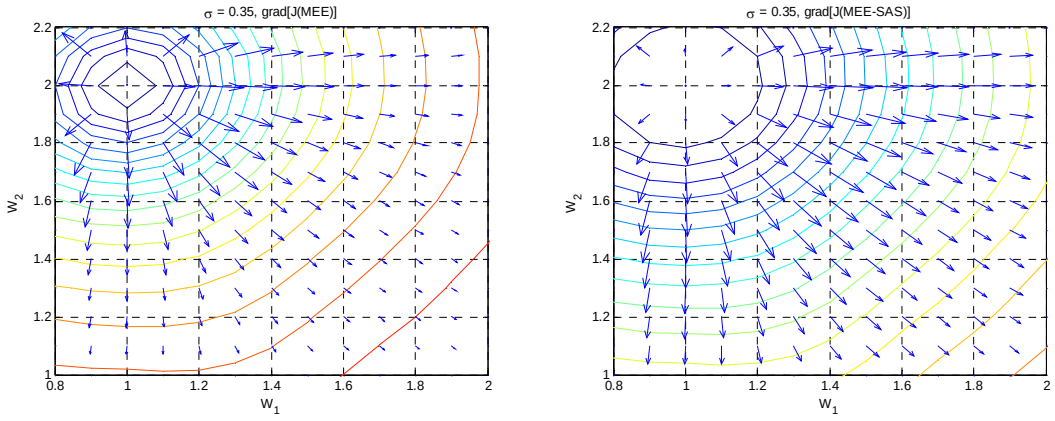
(a) MEE: $[V(\mathbf{0}) - V(\mathbf{e})]$ (b) MEE-SAS: $[V(\mathbf{0}) - V(\mathbf{e})]^2$

[Figure 2. Performance Surface of MEE and MEE-SAS]

1). Effect of Kernel Size on the Performance Surface in zero final error case ($V(\mathbf{0}) = V_{\mathbf{w}_*}(\mathbf{e})$).

This case study aims to illustrate how the performance surface (here represented by its contour and gradient vector plots) of MEE ($V(\mathbf{0}) - V_{\mathbf{w}_*}(\mathbf{e})$) and MEE-SAS ($[V(\mathbf{0}) - V_{\mathbf{w}_*}(\mathbf{e})]^2$) are altered as a

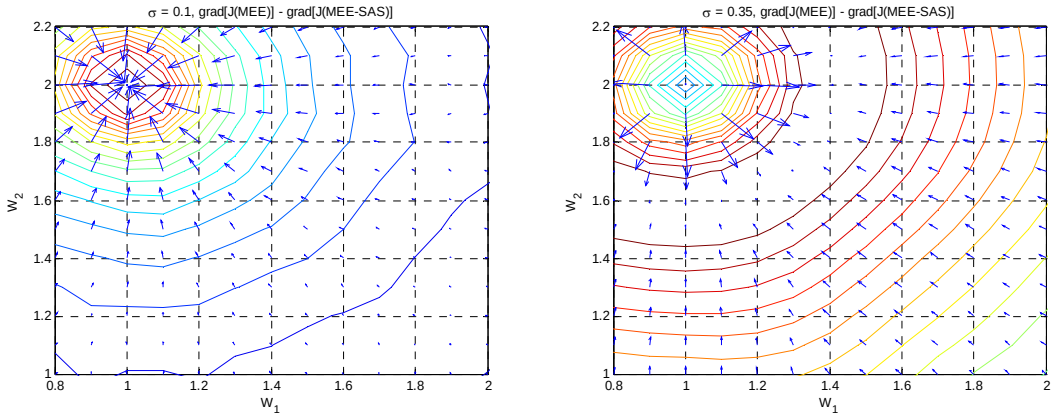
consequence of changing the kernel size in the estimator. In order to avoid excessive computation time requirements, we have utilized 100 noiseless training samples to obtain the contour and gradient vector plots. A kernel size is set to $\sigma = 0.1,\ 0.35,\ 0.6$.



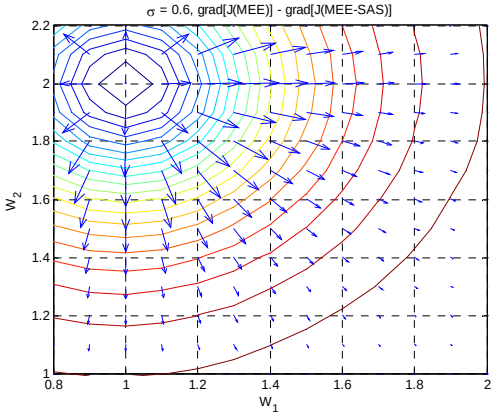
(a) MEE: $\left[V(\mathbf{0})-V(\mathbf{e})\right]$ (b) MEE-SAS: $\left[V(\mathbf{0})-V(\mathbf{e})\right]^2$

[Figure 3. Contour and gradient plot of error information potential in supervised ADALINE training for kernel size=0.35]

In Figure 3, we show that when the current weight is close to optimal solution, the magnitude of gradient vector increases quadratically in a radial direction. Note that the gradient vector decreases when far from the solution, since the performance has an upper bound ($V(\mathbf{0})-V(\mathbf{e})\le V(\mathbf{0})$) unlike MSE (see Figure 2). In order to distinguish the gradient relation between MEE and MEE-SAS, we plot the gradient difference between them.



(a) Kernel Size: 0.1 (b) Kernel Size: 0.35



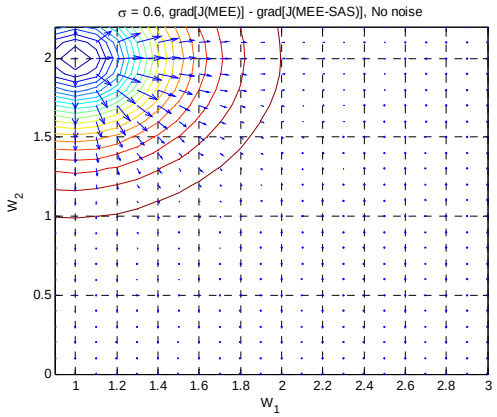
(c) Kernel Size: 0.6

[Figure 4. Contour and gradient difference between MEE and MEE-SAS on error information potential for various choices of kernel size]

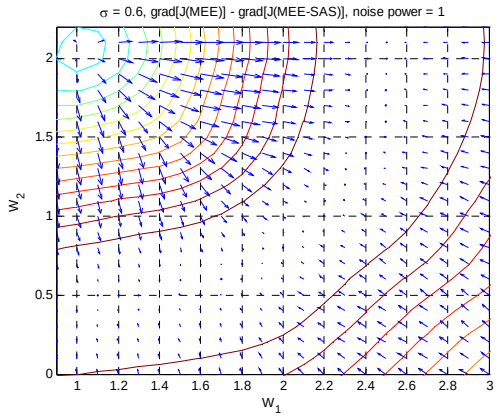
In Figure 4, when using small kernel size ($\sigma = 0.1$), MEE-SAS is superior to MEE with respect to the magnitude of gradient; while for large kernel size ($\sigma = 0.6$), MEE is superior to MEE-SAS. We show that the smaller the kernel we use, the larger is the region over which MEE-SAS is superior to MEE.

2). Effect of non-zero final error on the Performance Surface in ($V(\mathbf{0}) \neq V_{\mathbf{w}_*}(\mathbf{e})$).

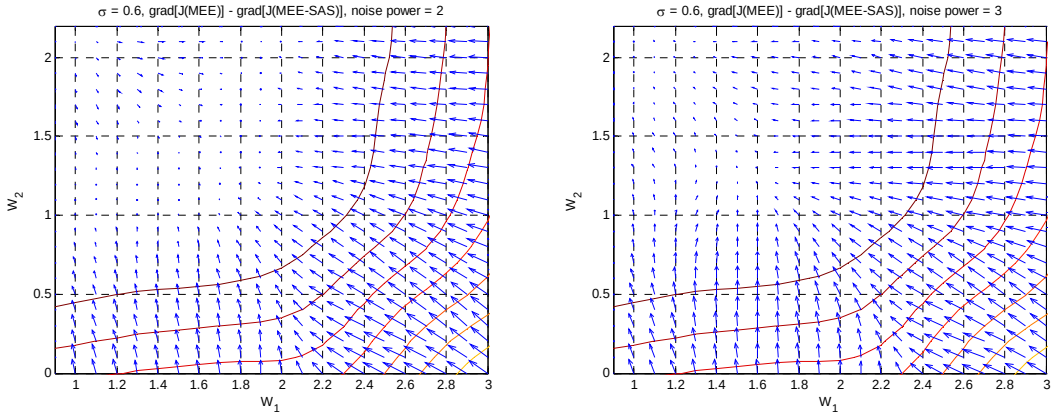
The case of $V(\mathbf{0}) \neq V_{\mathbf{w}_*}(\mathbf{e})$ includes two cases: Measurement Noise case and Error Modeling case. The simulation result of measurement noise case is similar to that of error modeling case, so, we just show the simulation result for the measurement noise case. We add the uniform distributed noise with three different powers (P = 1, 2, and 3) in the above example.



(a) No Noise



(b) Noise Power: 1



(c) Noise Power: 2 (d) Noise Power: 3

[Figure 5. Contour and gradient difference between MEE and MEE-SAS on error information potential for three different measurement noises]

As seen in Figure 5, the higher the noise power, the larger is the region over which MEE-SAS is superior to MEE in terms of gradient magnitude. This means that the point at which the curvature changes from higher than second order to second order is closer to the optimal solution when $V(\mathbf{0}) \neq V_{\mathbf{w}_*}(\mathbf{e})$ than in the case of $V(\mathbf{0}) = V_{\mathbf{w}_*}(\mathbf{e})$ as elucidated by theorem 4. This also means that the larger the final error, the faster is the convergence.

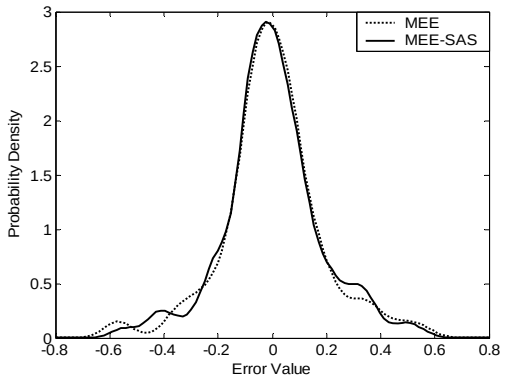
B. Second Study: Nonlinear System Identification

We investigate the performance of the MEE and MEE-SAS criterion in identification of a nonlinear system, whose dynamic equations are given as [23]

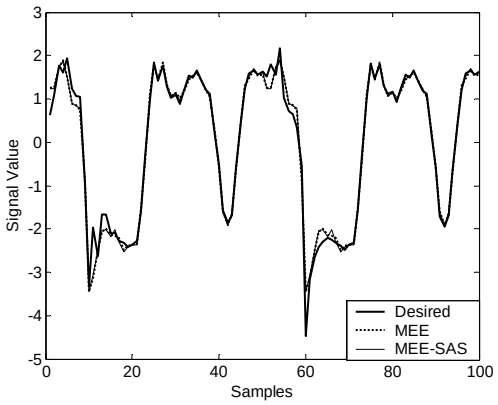
$$\begin{aligned} u(n) &= \sin\left(\frac{2\pi n}{10}\right) + \sin\left(\frac{2\pi n}{25}\right) \\ x_1(n+1) &= \left(\frac{x_1(n)}{1+x_1^2(n)} + 1\right) \cdot \sin\left(x_2(n)\right) \\ x_2(n+1) &= x_2(n) \cdot \cos\left(x_2(n)\right) + x_1(n) \cdot \exp\left(-\frac{x_1^2(n) + x_2^2(n)}{8}\right) + \frac{u^3(n)}{1+u^2(n)+0.5 \cdot \cos\left(x_1(n)+x_2(n)\right)} \\ y(n) &= \frac{x_1(n)}{1+0.5 \cdot \sin\left(x_2(n)\right)} + \frac{x_2(n)}{1+0.5 \cdot \sin\left(x_1(n)\right)}. \end{aligned} \tag{35}$$

A Time Delay Neural Network (TDNN) trained with the backpropagation algorithm will be used. The only difference in the application of the MEE or MEE-SAS to backpropagation is the injected error. We select the TDNN architecture of 10-15-1 with tanh non-linearity processing elements (PE) in the hidden layer and a linear output PE. The training is carried out in batch mode for 2000 epochs. The data set consists of N=100 input-output pairs. The Kernel size is experimentally set at $\sigma = 0.707$.

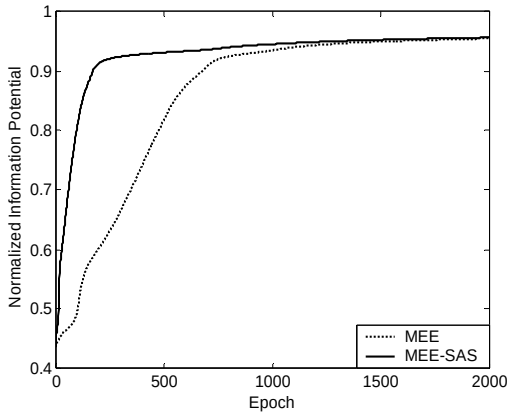
Once again Monte-Carlo simulations are performed using 20 different weight initializations and the average performance is selected for comparison. In order to compare two algorithms, we find the step size for each algorithm to be such that it produces similar probability densities of error within the last epoch as shown in Figure 6. Figure 7 show the identified outputs of both the algorithms.



[Figure 6. Probability density of error for last epoch]



[Figure 7. Identification Performance]



[Figure 8. Normalized Information Potential]

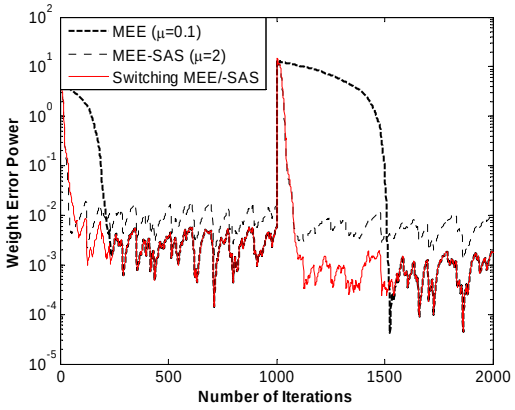
As seen from Figure 8, MEE-SAS converges in about 200 epochs whereas MEE need 750 epochs to achieve the same level of performance. Recall that the inherent property of MEE-SAS is that it has large effective step size when the present solution is far from the optimal leading to large "jumps" in the bowl 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.

C. Third Study: Nonstationary System Identification

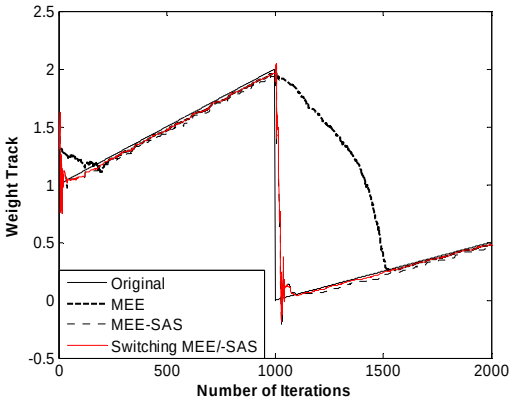
The nonstationary unknown plant transfer functions is given as

$$H(z)=\begin{cases}\left(2+\dfrac{2n}{1000}\right)\cdot\left[\begin{array}{l}0.1+0.2z^{-1}+0.3z^{-2}+0.4z^{-3}+0.5z^{-4}\\+0.4z^{-5}+0.3z^{-6}+0.2z^{-7}+0.1z^{-8}\end{array}\right], & 1\le n\le 1000\\ \left(\dfrac{n}{1000}-1\right)\cdot\left[\begin{array}{l}0.1+0.2z^{-1}+0.3z^{-2}+0.4z^{-3}+0.5z^{-4}\\+0.4z^{-5}+0.3z^{-6}+0.2z^{-7}+0.1z^{-8}\end{array}\right], & 1001\le n\le 2000.\end{cases} \tag{36}$$

The FIR adaptive filter is selected with equal order. The input to both the plant and the adaptive filter is white Gaussian noise with unit variance. We choose a proper kernel size ($\sigma = 0.707$) based on Silverman's rule and set the window length to $L = 50$. The System mismatch (weight error norm) is selected as a performance measure.



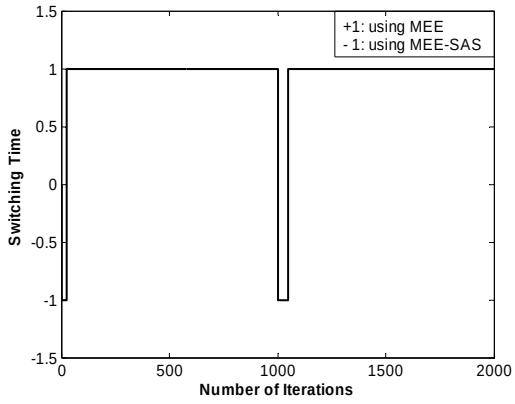
[Figure 9. Weight error power]



[Figure 10. One of the nine weight tracks (W_5)]

Figure 9 depicts a nonstationary system identification problem to show the performance of the switching MEE and MEE-SAS compared to both MEE and MEE-SAS. The convergence performance

of the switching one is the similar to that of MEE-SAS, while better than that of MEE. Also, the tracking performance of the switching one is better than that of both MEE and MEE-SAS in Figure 10. As seen from Figure 11, in abrupt changing part (at initial and at iteration 1000), the switching MEE and MEE-SAS used the MEE-SAS.



[Figure 11. MEE or MEE-SAS used time on the switching MEE and MEE-SAS]

VII. CONCLUSIONS

In this paper, an information-theoretic supervised learning criterion for adaptive systems, namely, minimum error entropy with self adjusting step-size (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.

In structural analysis part, we analytically found the turning point of curvature for MEE-SAS. It was observed that this contour of points in the cost function curvature depends on the SNR of the signal (provided that we have critical model order). Further, MEE-SAS is expected to perform well than MEE in the case where there is non-zero error (due to modeling error or measurement noise) since the turning point of curvature is going to be close to the optimal leading to faster convergence. For the case where zero error is achievable, though MEE-SAS retains its ability to converge faster than MEE, this property is lost as soon as MEE-SAS cross the contour of the turning point of curvature which in this case is farther away from the solution.

The loss of tracking ability of MEE-SAS beyond the turning point of the curvature due to small effective step size hinders the performance of MEE-SAS in nonstationary environment. We solved this problem using a novel switching scheme between MEE and MEE-SAS. Starting with MEE-SAS for faster convergence when far from the solution, the algorithm switches to MEE near the optimal

solution for improved tracking ability. Simulation results in non stationary scenario shows that the proposed switching algorithms outperforms both MEE and MEE-SAS algorithms when used independently and quickly adapts to changing environment.

REFERENCES

[1] N. Kalouptsidis and S. Theodoridis, Adaptive System Identification and Signal Processing Algorithms, Prentice-Hall, 1993.

[2] A. Zerguine, C.F.N. Cowan and M. Bettayeb, Adaptive Echo Cancellation using Least Mean Mixed-Norm Algorithm, IEEE trans. Signal Processing, 45 (5) (May 1997) 1340-1343.

[3] C.F.N. Cowan, Channel Equalization, in: N. Kalouptsidis and S. Theodorids (Ed.), Adaptive System Identification and Signal Processing Algorithms, Prentice-Hall, 1993, pp.388-406.

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

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

[6] E. Walach and B. Widrow, The Least Mean Fourth(LMF) Adaptive Algorithm and its Family, IEEE trans. Information Theory, IT 30 (2) (March 1984) 275-283.

[7] J.A. Chambers, O. Tanrikulu and A.G. Constantinides, Least Mean Mixed-Norm Adaptive Filtering, IEE Electronics Letters, 30 (19) (September 1994) 1574-1575.

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

[9] 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, December 1996.

[10] J.C. Principe, D. Xu and J. Fisher, Information Theoretic Learning, in: S. Haykin (Ed.), Unsupervised Adaptive Filtering, Wiley, Newyork, vol I, 2000, pp 265-319.

[11] C.E. Shannon, A mathematical theory of communications, Bell Syst. Tech. Journal, 27 (1948) 379-423.

[12] A. Renyi, Some Fundamental Questions of Information Theory, Selected Papers of Alfred Renyi, 2, Akademia Kiado, Budapest, 1976, pp. 526-552.

[13] D. Erdogmus and J.C. Principe, Generalized Information Potential Criterion for Adaptive System Training, IEEE Trans. Neural Networks, 13 (5) (September 2002) 1035-1044.

[14] D. Erdogmus, J.C. Principe, An Entropy Minimization algorithm for Supervised Training of Nonlinear Systems, IEEE trans. Signal Processing, 50, (7) (July 2002) 1780-1786.

[15] R.A. Morejon, J.C. Principe, Advanced search algorithms for information-theoretic learning with kernel-based estimators , IEEE trans. Neural Networks, 15 (4) (July 2004) 874-884.

[16] D. Erdogmus, J.C. Principe, K.E.Hild II, Online entropy manipulation: Stochastic Information Gradient, IEEE Signal Processing Letters, 10 (8) (August 2003) 242-245.

[17] R.H. Kwong and E.W. Johnston, A variable step size LMS algorithm, IEEE Trans. Signal Processing, 40 (7) (1992) 1633-1642.

[18] T. Aboulnasr and K. Mayyas, A robust variable step-size LMS-type algorithm: Analysis and simulations, IEEE Trans. Signal Processing, 45 (3) (1997) 631-639.

[19] D.I. Pazaitis and A.G. Constantinides, A novel kurtosis driven variable step-size adaptive algorithm, IEEE Trans. Signal Processing, 47 (3) (1997) 864-872.

[20] H-C. Shin, A.H. Sayed, and W-J. Song, Variable step-size NLMS and affine projection algorithms, IEEE Signal Processing Letters, 2004.

[21] D. Erdogmus, J.C. Principe, Convergence Properties and Data Efficiency of the Minimum Error Entropy Criterion in Adaline Training, IEEE trans. Signal Processing, 51 (7) (July 2003) 1966-1978.

[22] B.W.Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall, London, 1986.

[23] J.C. Principe, N. Euliano, and C. Lefebvre, Neural and Adaptive Systems: Fundamentals Through Simulations. New York: Wiley, 1999.

[24] S. Han, S. Rao, D. Erdogmus, J.C. Principe, An Improved Minimum Error Entropy Criterion with Self-Adjusting Step-Size, IEEE International Workshop on Machine Learning for Signal Processing(MLSP'05), Mystic, Connecticut, Sep 2005, pp. 317-322.

[25] Hassan Khalil, Nonlinear Systems, Macmillan, New York, 1992.