Convergence Properties and Data Efficiency of the Minimum Error Entropy Criterion in Adaline Training

Deniz Erdogmus, *Member, IEEE*, and Jose C. Principe, *Fellow, IEEE*

***Abstract*—Recently, we have proposed the minimum error entropy (MEE) criterion as an information theoretic alternative to the widely used mean square error criterion in supervised adaptive system training. For this purpose, we have formulated a nonparametric estimator for Renyi's entropy that employs Parzen windowing. Mathematical investigation of the proposed entropy estimator revealed interesting insights about the process of information theoretical learning. This new estimator and the associated criteria have been applied to the supervised and unsupervised training of adaptive systems in a wide range of problems successfully. In this paper, we analyze the structure of the MEE performance surface around the optimal solution, and we derive the upper bound for the step size in adaptive linear neuron (ADALINE) training with the steepest descent algorithm using MEE. In addition, the effects of the entropy order and the kernel size in Parzen windowing on the *shape* of the performance surface and the eigenvalues of the Hessian at and around the optimal solution are investigated. Conclusions from the theoretical analyses are illustrated through numerical examples.**

***Index Terms*—Adaptive systems, convergence of numerical methods, linear systems, minimum entropy methods.**

I. Introduction

THE MEAN square error (MSE) has been the workhorse of function approximation due to our knowledge of the mathematical properties of the least square method. Starting with the pioneering work of Wiener [1], MSE has become the fundamental performance criterion in adaptive filtering theory. With the basic adaptive FIR filter structure, MSE yields a simple optimization problem, whose analytical solution is provided by the Wiener–Hopf equation [2]. Following this, algorithms for iteratively approximating the optimal solution including the steepest descent approach and the second-order optimization techniques have been proposed and analyzed [2], [3]. The least square algorithm (LMS) and the recursive least squares (RLS) algorithms are the most widely recognized variants of these algorithms [2]–[4]. Issues of stability and convergence speed have been the main thrusts in these analyses. Due to the quadratic form of the cost function in terms of the weight vector, the convergence analysis of the algorithms could be pursued [2], although some issues still remain unsolved [2], [5].

Recently, we proposed the quadratic Renyi's entropy of the error signal as an alternative criterion for supervised adaptive system training [6] and used a nonparametric estimator based on Parzen windowing with Gaussian kernels to estimate entropy directly from the data samples. The motivation for pursuing the application of Renyi's entropy was the existence of an analytically and computationally simple estimator for Renyi's quadratic entropy, as well as the fact that the commonly used Shannon's entropy is a special case of Renyi's definition. We have proved that when utilizing Renyi's entropy, a system trained with the minimum error entropy (MEE) criterion minimizes the Renyi's distance between the conditional probability density functions (pdf) of the desired and the actual outputs given the input signal (Kullback–Leibler divergence in the case of Shannon's entropy). This theoretical result was also demonstrated in simulations for chaotic time series prediction and nonlinear system identification using feedforward neural networks [7]. Moreover, we have also extended the nonparametric entropy estimator to any entropy order and kernel function [8]. Other successful applications of the proposed nonparametric entropy estimator and MEE include maximally informative subspace projections, blind source separation [9]–[11], and blind deconvolution [12]. These applications showed that the proposed entropy estimator is data efficient, therefore achieving better performance in these problems using a smaller training set compared with alternative algorithms [10].

MEE does not yield to a linear optimization problem even when the finite impulse response (FIR) structure is utilized, which complicates the theoretical understanding of the method in two important ways: First, the nonconvex nature of the performance surface makes the search for the global optimum nontrivial and, in general, disqualifies local search methods like the steepest descent to find consistently the optimal solution. Second, the learning rate, or step size, in steepest descent cannot be set without knowing the eigenstructure of the MEE in the neighborhood of the optimal solution, and the fundamental trade-off between speed of convergence and misadjustment cannot be found.

In spite of these difficulties, we observed and proved [8] a very important dilation property of the MEE estimated with Parzen windows that motivated the convergence analysis of the steepest descent algorithm to be presented in this paper. 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,

Manuscript received December 21, 2001; revised December 31, 2002. This work was supported by the National Science Foundaiton under Grant ECS-9900394. This work is an extended version of the results previously presented at NNSP 2001 [18] and ICA 2001 [19]. The associate editor coordinating the review of this paper and approving it for publication was Prof. Derong Liu.

The authors are with the Computational NeuroEngineering Laboratory, Electrical and Computer Engineering Department NEB 451, University of Florida, Gainesville, FL 32611 USA (e-mail: deniz@cnel.ufl.edu; principe@cnel.ufl.edu).

Digital Object Identifier 10.1109/TSP.2003.812843

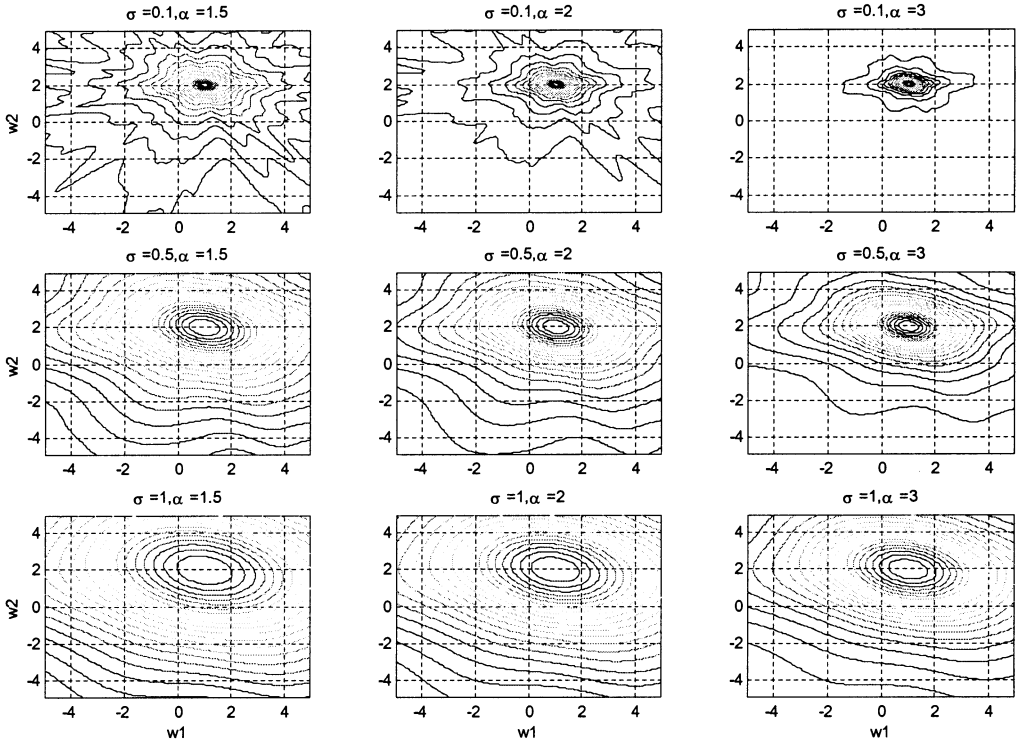


Fig. 1. Contours of error information potential in supervised ADALINE training for various choices of kernel size σ and entropy order α. Kernel size increases from top to bottom, and entropy order increases from left to right.

global minimum.[1] An illustration of this dilation is shown in Fig. 1. Applied to the linear adaptive structures, namely the adaptive linear neuron (ADALINE), the dilation property of the MEE criterion brings about a very useful quality to the performance surface of this cost function. In weight space, the volume of the region where the equilevel contours of the MEE criterion are perfect ellipsoids centered at the global optimum increases when the kernel size is increased. 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. Note that the annealing of the kernel sizes during adaptation reinforces the need for the knowledge of the eigenstructure near the optimum solution and its dependence (if any) on the kernel size.

Hence, practically, convergence to the global solution can be achieved by starting the algorithm with a large kernel size and decreasing this parameter slowly[2] during the course of adaptation, just like in convolution smoothing and stochastic annealing [13]. The step size can also be kept at small values during adaptation to guarantee convergence, but this sacrifices convergence speed and does not provide a principled way to know the tradeoff with misadjustment. Hence, we seek here a more analytical approach to study the convergence of the steepest descent algorithm in MEE.

First, we need to make the entropy cost function continuous and differentiable, but guaranteeing this is a trivial task. One simply selects a kernel function that is continuous and differentiable (up to the order necessary) [8]. Furthermore, if the kernel function of choice has its maximum at the origin, the entropy estimator can be shown to achieve its minimum value when all the samples are equal (say zero) [8]. This is important for MEE because it guarantees that if it is possible to achieve zero training error, the entropy estimator will have its global minimum for the weights that make all the error samples zero over the training set.

Second, we have to guarantee that the quadratic approximation of the performance surface in the neighborhood of the global optimum remains valid, even if the current weight vector is far from the solution. As was mentioned above, this can be satisfied by starting initially with a large kernel size and decrease it slowly enough. The question that will be left for further research relates to how slow the kernel size should be annealed to guarantee this requirement.

In the literature, it is possible to find other examples of information theoretic adaptation rules utilizing similar or different approaches. Entropic learning rules have found natural application in independent components analysis and blind source separation [14], [15], as well as blind deconvolution [16]. Other fields where similar approaches are widely embraced include the broad areas of information theoretic pattern analysis [17]–[19], information content-based signal processing [20], [21], and model complexity analysis [22], [23]. The utility of entropy and other information theoretic measures in these

[1]The reason for the existence of this bias could be understood in light of the convolution smoothing theory [13]. Unfortunately, it cannot be practically estimated.

[2]A general-purpose annealing schedule for the kernel size is not available at this time. However, in conjunction with other stochastic global optimization approaches, an exponential decrease in time could be utilized.

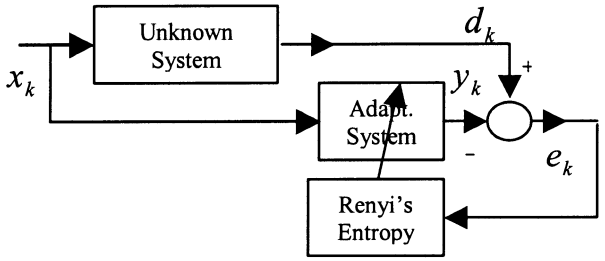


Fig. 2. Illustration of supervised adaptive system training using MEE criterion.

broadly defined areas are endless; however, our main concern is using these measures and their nonparametric estimators in filter and system adaptation.

In the following sections, we will build arguments about the convergence properties of training linear filters with the MEE criterion based on these two key points. Recall that once a continuous and differentiable cost function is approximated by a quadratic form around the global optimum, the results from the theory of convergence for MSE criterion applies immediately in the region of approximation, and therefore, we will simply borrow from these important results to complete the arguments.

II. Minimum Error Entropy Criterion

Consider the supervised training scheme depicted in Fig. 2. Since, in practice, the analytical expression for the error entropy is not available in general, one needs to estimate it nonparametrically from the samples. Renyi's entropy for a random variable e is given in terms of its pdf as [24]

$$H_\alpha(e) = \frac{1}{1-\alpha} \log \int f_e^\alpha(e)\, de \tag{1}$$

where α is referred to as the entropy order. Shannon's entropy is a special case of Renyi's entropy corresponding to order 1. Notice, however, that the parametric entropy family in (1) could be extended to include Shannon's entropy for $\alpha = 1$ since in the limit, it is equal to Shannon's entropy. This can easily be shown using L'Hopital's rule

$$\begin{aligned}\lim_{\alpha\to 1} H_\alpha(e) &= \lim_{\alpha\to 1} \frac{1}{1-\alpha} \log \int f_e^\alpha(e)\, de \\ &= \frac{\lim\limits_{\alpha\to 1} \int \log f_e(e) \cdot f_e^\alpha(e)\, de \Big/ \int f_e^\alpha(e)\, de}{\lim\limits_{\alpha\to 1} -1} \\ &= -\int f_e(e) \cdot \log f_e(e)\, de = H_S(e). \end{aligned} \tag{2}$$

It was shown that minimizing error entropy (MEE) is equivalent to minimizing Renyi's divergence between f_{xd} and f_{xy}, where x, y, and d are the input, output, and desired output signals, respectively; therefore, it determines an optimal model of the unknown system in terms of statistical learning [8]. This result is also related to the information geometry of statistical models and Amari's α-divergence [25].[3]

It is possible to alternatively express Renyi's entropy given in (1) using an expectation operator

$$H_\alpha(e) = \frac{1}{1-\alpha} \log E\left[f_e^{\alpha-1}(e)\right] \tag{3}$$

where $V_\alpha(e) = E[f_e^{\alpha-1}(e)]$ is defined to be the order-α information potential [8], [9] and could replace entropy in the cost functions since "log" is a monotonic function. Approximating the expectation operator with the sample mean (as in [26]) and estimating the pdf using Parzen windowing with kernel $\kappa_\sigma(x)$ [27], we obtain the nonparametric estimator for Renyi's entropy

$$\hat{H}_\alpha(e) = \frac{1}{1-\alpha} \log \frac{1}{N^\alpha} \sum_{j=1}^{N} \left(\sum_{i=1}^{N} \kappa_\sigma(e_j - e_i) \right)^{\alpha-1} \tag{4}$$

where $\{e_1, \ldots, e_N\}$ are the samples and $\sigma > 0$ is the kernel size, which is required to satisfy the scaling property $\kappa_\sigma(x) = \kappa_1(x/\sigma)/\sigma$ for single dimensional random variables. This requirement is a consequence of the requirement that the kernel function in Parzen windowing must be a valid pdf [27]. It must be noted that unlike the MSE criterion, the error entropy cost function has a nonquadratic, nonconvex performance surface. However, near the global minimum the cost function can be accurately approximated by a quadratic Taylor series expansion in the weights, assuming that the selected kernel function satisfies the differentiability conditions we have mentioned above. Furthermore, as we will demonstrate, by controlling the two design parameters, namely, the kernel size and the entropy order of the estimator in (4), it is possible to manipulate the volume and the eigenstructure of this region in the weight space where the quadratic approximation is valid. 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. The structure of the performance surface far away from the global minimum is still under investigation. Some important properties of the estimator in (4) can be listed as follows.

Fact 1: The nonparametric estimator in (4) is mean invariant as well as the actual entropy in (1).

Proof: Let us define a new random variable $\overline{e} = e + m$, where m is a real, deterministic constant. By a change of variables in the integral of (1), the entropy of this new random variable is found as

$$\begin{aligned} H_\alpha(\overline{e}) &= \frac{1}{1-\alpha} \log \int f_{\overline{e}}^\alpha(\overline{e})\, d\overline{e} \\ &= \frac{1}{1-\alpha} \log \int f_e^\alpha(\overline{e} - m)\, d\overline{e} \\ &= \frac{1}{1-\alpha} \log \int f_e^\alpha(e)\, de = H_\alpha(e). \end{aligned} \tag{5}$$

[3]It has been pointed out by an anonymous reviewer that this connection is obvious since Amari's α-divergence is an alternative parameterization of Renyi's divergence. In fact, Amari provides a detailed account on the relationship between his α-divergence and Renyi's mutual information [15].

On the other hand, the samples of $\overline{e}$ are related to those of e by $\overline{e}_k = e_k + m$. Therefore, the entropy estimate of $\overline{e}$ is

$$\begin{aligned}\hat{H}_\alpha(\overline{e}) &= \frac{1}{1-\alpha}\log\frac{1}{N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\kappa_\sigma(\overline{e}_j-\overline{e}_i)\right)^{\alpha-1}\\ &= \frac{1}{1-\alpha}\log\frac{1}{N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\kappa_\sigma(e_j+m-e_i-m)\right)^{\alpha-1}\\ &= \hat{H}_\alpha(e). \end{aligned} \tag{6}$$

Fact 2*:* The identity $H_\alpha(ae) = H_\alpha(e) + \log|a|$ holds for a real scaling factor a in the actual entropy. If we apply the same scaling to the kernel size in the estimator in (4), the same identity also holds for the entropy estimates of the two random variables ae and e.

Proof: Let $\overline{e} = ae$; thus, the samples are $\overline{e}_k = ae_k$. The entropy estimate of $\overline{e}$ using a kernel size of $\overline{\sigma} = a\sigma$ can be written explicitly as

$$\begin{aligned}&\hat{H}_{\alpha,\overline{\sigma}}(\overline{e})\\ &= \frac{1}{1-\alpha}\log\frac{1}{N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\frac{1}{\overline{\sigma}}\kappa\left(\frac{\overline{e}_j-\overline{e}_i}{\overline{\sigma}}\right)\right)^{\alpha-1}\\ &= \frac{1}{1-\alpha}\log\frac{1}{a^{\alpha-1}N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\frac{1}{\sigma}\kappa\left(\frac{ae_j-ae_i}{a\sigma}\right)\right)^{\alpha-1}\\ &= \frac{1}{1-\alpha}\log\frac{1}{N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\frac{1}{\sigma}\kappa\left(\frac{e_j-e_i}{\sigma}\right)\right)^{\alpha-1}\\ &\quad+\log|a| = \hat{H}_{\alpha,\sigma}(e)+\log|a|. \end{aligned} \tag{7}$$

Fact 3*:*

a) $\lim_{N\to\infty}\hat{H}_\alpha(e) = H_\alpha(\hat{e}) \geq H_\alpha(e)$, where $\hat{e}$ is a random variable with the pdf $f_e(.) * \kappa_\sigma(.)$. The equality (in the inequality) occurs if and only if the kernel size is zero. This result is also valid in the mean for the finite-sample case.

b) For independent random variables X and Y, the inequality $H_\alpha(X+Y) \geq \{H_\alpha(X), H_\alpha(Y)\}$ holds. The equality occurs if and only if one of the variables is δ-distributed.

Proof of a): It is well known that the Parzen window estimate of the pdf of e converges consistently to $f_e(.) * \kappa_\sigma(.)$ under some assumptions to guarantee the existence of the integrals. Therefore, the entropy estimator in (4) converges to the actual entropy of this pdf. To prove the inequality, consider

$$\begin{aligned}e^{(1-\alpha)H_\alpha(\hat{e})} &= \int_{-\infty}^{\infty} p_{\hat{e}}^{\alpha}(y)\,dy\\ &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\kappa_\omega(\tau)f_e(y-\tau)\,d\tau\right]^{\alpha} dy. \end{aligned} \tag{8}$$

Using Jensen's inequality for convex and concave cases, we get

$$\begin{aligned}e^{(1-\alpha)H_\alpha(\hat{e})} &\underset{\alpha<1}{\overset{\alpha>1}{\lessgtr}} \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\kappa_\sigma(\tau)\left[f_e(y-\tau)\right]^{\alpha}d\tau\right]dy\\ &= \int_{-\infty}^{\infty}\kappa_\sigma(\tau)\left[\int_{-\infty}^{\infty}\left[f_e(y-\tau)\right]^{\alpha}dy\right]d\tau\\ &= \int_{-\infty}^{\infty}\kappa_\sigma(\tau)V_\alpha(e)\,d\tau\\ &= V_\alpha(e)\cdot\int_{-\infty}^{\infty}\kappa_\sigma(\tau)\,d\tau = V_\alpha(e). \end{aligned} \tag{9}$$

Reorganizing the terms in the last inequality and using the relationship between entropy and information potential, regardless of the value of α and the direction of the inequality, we arrive at the conclusion $H_\alpha(\hat{e}) \geq H_\alpha(e)$. The fact that these results are also valid on the average for the finite-sample case comes from the property $E[\hat{f}_e(.)] = f_e(.) * \kappa_\sigma(.)$ of Parzen windowing.

*Proof of b)*Since for independent variables the pdf of the sum is the convolution of individual pdfs, the preceding derivation yields the result immediately.

Fact 4*:*

a) If the kernel function is symmetric, unimodal, continuous and differentiable, the δ-distribution is a smooth global minimum of (4) in the error space. By a smooth minimum, we mean its gradient is zero, and its Hessian has finite non-negative eigenvalues (there is a zero eigenvalue corresponding to the eigenvector along which only the mean of the error changes). Notice that the differential entropy in (1) approaches $-\infty$ as the pdf approaches a δ-distribution.

b) Consider an ADALINE (without a bias weight), and assume that the kernel function satisfies the conditions in part a). The error signal at the optimal solution (which is denoted by v) is independent of the input, desired output, and the weights, and the inputs are independent from each other. In addition, the inputs x_i and $i \in C$ are Gaussian distributed (C is the set of indices of those inputs that are Gaussian distributed). Then, the Hessian of the MEE evaluated at the global minimum of error entropy has strictly positive eigenvalues in all directions, except for the one that maintains a constant $\sum_{i\in C} w_i^2\sigma_i^2$. Note that in the presence of a number of Gaussian inputs, these solutions form a hypersphere in the weight space on which the error entropy remains constant. In terms of cost, these are all global minima and are equally good, although they differ from a system identification standpoint.

Proof of a): See in [8, Lemma 1, Th. 1].

*Proof of b)*For the ADALINE, any perturbation from the optimal weights will result in the error $\overline{e}$ at the new weights being equal to $\overline{e} = e + v$. Due to the independence assumption, e and v will be independent, which allows Fact 3b to be immediately applicable. The equality $H_\alpha(\overline{e}) = H_\alpha(v)$ is possible if and only if e is δ-distributed, which allows only the mean of the error to change as a result of the perturbation. For only the mean of a linear combination of independent random variables as described above to change, it is necessary that weights are

perturbed only for those inputs with Gaussian densities. Furthermore, for the output pdf to remain constant apart from its mean, the variance of the linear combination of these Gaussian inputs must have constant variance, which gives us the direction defined by the hyper sphere $\sum_{i\in C} w_i^2\sigma_i^2 = \sum_{i\in C} w_{*i}^2\sigma_i^2$, where w_* is the optimal weight vector. Perturbations along any other direction in the weight space will result in the output pdf to change its moments other than its mean; therefore, the entropy will increase. Clearly, if less than two of the inputs are Gaussian, the Hessian matrix is strictly positive definite.

Fact 5*:* In the limit, as the kernel size tends to infinity, the entropy estimator in (4) approaches a nonlinearly scaled version of the sample variance. In particular

$$H_\alpha(e) \approx \frac{1}{1-\alpha}\log\left(\kappa_\sigma(0) + \kappa''_\sigma(0)\cdot(\overline{e^2} - \overline{e}^2)\right) \quad (10)$$

where the over bar indicates the sample mean.

Proof: When the kernel size is very large, the kernel evaluations in (4) can be carried out using the following Taylor series expansion:

$$\kappa_\sigma(\xi) \approx \kappa_\sigma(0) + \kappa'_\sigma(0)\xi + \kappa''_\sigma(0)\xi^2/2 = \kappa_\sigma(0) + \kappa''_\sigma(0)\xi^2/2. \quad (11)$$

Substituting this in (4) immediately yields the desired result, assuming symmetric kernels.

Fact 6*:* Error entropy criterion is robust to additive zero-mean noise, regardless of its pdf.

Proof: Consider the learning process depicted in Fig. 2. Suppose that the desired signal consists of the superposition of a deterministic part and a zero-mean random part, such that $d = g(x)+v$, where $g(.)$ is the unknown function that the adaptive system is trying to identify, and v is the zero-mean noise with pdf $f_v(.)$ independent from x, d, and y. Suppose that the learning system is a parametric family of functions of the form $h(x;\ \mathbf{w})$, where $\mathbf{w}$ is the vector of parameters, called the weight vector. Let $\mathbf{w}_*$ be (one of possibly many) optimal weight vectors that minimize the error entropy, where the error signal is defined as $e = d - y$. Let $\overline{w}_*$ be the optimal weight vector that minimizes the entropy of the *clean* error signal that is defined as $\overline{e} = g(x) - h(x,\ \mathbf{w})$ (if there was no additive noise in the desired signal). Notice that we have the identity $e = \overline{e} + v$. There are two possible situations: First, the unknown function $g(.)$ might lie in the span of the family of functions $h(.,\ \mathbf{w})$, such that there exits a $\overline{w}_*$ that makes $\overline{e}$ have a δ-distribution (possibly centered at zero). Second, $g(.)$ does not lie in this span, and therefore, the optimal solution yields a non-$\delta\overline{e}$ distribution.

In the first case, we notice that the minimum value that the entropy of the noisy error signal can achieve is given by $H_\alpha(v)$ since due to Fact 3b and due to the independence of v, we have $H_\alpha(e) = H_\alpha(\overline{e} + v) \geq H_\alpha(v)$. Recall that the equality is attained if and only if $\overline{e}$ is δ-distributed (since v is not δ-distributed). Consequently, the minimization of the noisy error entropy will result in $\mathbf{w}_* = \overline{\mathbf{w}}_*$.

In the second case, we cannot prove that the situation will be exactly as in the first case. However, since $H_\alpha(e) = H_\alpha(\overline{e} + v) \geq H_\alpha(\overline{e})$ due to Fact 3b, minimization of the noisy error entropy is equivalent to the minimization of an upper bound for the entropy of the *clean* error signal, which is the ultimate objective function that needs to be minimized. Under these circumstances

$$\begin{aligned}\mathbf{w}_* &= \arg\min_{\mathbf{w}} H_\alpha(e(\mathbf{w})) = \arg\min_{\mathbf{w}} H_\alpha\left(\overline{e}(\mathbf{w}) + v\right) \\ &\approx \arg\min_{\mathbf{w}} H_\alpha\left(\overline{e}(\mathbf{w})\right) = \overline{\mathbf{w}}_*. \end{aligned} \quad (12)$$

An alternative proof, which provides better insights about the application of this fact to ADALINE, is provided in the Appendix.

Notice that the MSE criterion satisfies the noise rejection property described in Fact 6 as well. Therefore, the question becomes, which one of these two criteria approaches to this asymptotic total noise rejection property faster when a finite number of samples are used in training for the optimal weights? This question will be addressed in the numerical studies section through a set of Monte Carlo simulations.

The facts stated above all contribute to the practicality of the MEE criterion for information theoretic supervised learning. They also corroborate the use of the proposed nonparametric entropy estimator to this purpose as it possesses many of the properties of the actual entropy and some that are even more advantageous. Specifically, Fact 1 states that the mean of the error must be optimized separately using the output bias term of the adaptive system, if there is one. Fact 4 allows us to utilize numerical nonlinear optimization techniques based on derivatives of the cost function on the estimated entropy as it shares the global optimum of the actual entropy function that we seek. Fact 2 demonstrates how to modify the parameters in the cost function in the case of a scaling of the training data in order to obtain the same solution that one would get before the scaling. Fact 3 basically justifies that the presented entropy estimator provides an upper bound on the average and asymptotically for a quantity that we wish to minimize and that this upper bound is consistent. Fact 5, from a practical point of view, states that MSE is a special case of the proposed entropy estimator corresponding to a large kernel size, and thus, second-order statistics can also be exploited if desired by simply modifying the parameters of the cost function appropriately. Last but not least, Fact 6 provides an important justification to the use of MEE as the learning criterion in supervised training by demonstrating the immunity of the optimal solution to additive noise in the desired signal, which is a common situation in practice.

III. Steepest Ascent Information Potential Training for Adaline

Suppose the adaptive system under consideration in Fig. 2 is an ADALINE structure with a weight vector 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 input signal to a tapped delay line for the special case of FIR filter. When the entropy order is specified, minimizing the error entropy is equivalent to minimizing or maximizing the information potential for $\alpha < 1$ and $\alpha > 1$, respectively. In the following, we consider the choice $\alpha > 1$. Since information potential is the argument of the logarithm in the entropy definition, its estimator is the argument of the logarithm in the entropy estimator given in (4). Then, the gradient

of the information potential estimator with respect to the weight vector is simply

$$\frac{\partial V_\alpha}{\partial \mathbf{w}} = \frac{(\alpha-1)}{N^\alpha} \sum_j \left(\sum_i \kappa_\sigma(e_j - e_i) \right)^{\alpha-2} \cdot \left(\sum_i \kappa'_\sigma(e_j - e_i)(\mathbf{x}_i - \mathbf{x}_j)^T \right). \quad (13)$$

In order to maximize the information potential, we update the weights along the gradient direction with a certain step size η

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \eta \nabla V_\alpha(\mathbf{w}(n)) \quad (14)$$

where $\nabla V_\alpha(\mathbf{w}(n))$ denotes the gradient of V_α evaluated at $\mathbf{w}(n)$. This is a nonlinear update rule, however, as mentioned in the introduction; the selection of a smooth kernel function with a sufficiently large kernel size to allow a quadratic approximation for the cost function to be valid motivates us to employ a Taylor series expansion truncated at the linear term for the gradient, which is evaluated at the optimal weight vector $\mathbf{w}_*$

$$\nabla V_\alpha(\mathbf{w}) = \nabla V_\alpha(\mathbf{w}_*) + \frac{\partial \nabla V_\alpha(\mathbf{w}_*)}{\partial \mathbf{w}} (\mathbf{w} - \mathbf{w}_*). \quad (15)$$

Writing the derivative of the kernel for a given size in terms of the derivative of the unit-size kernel

$$\kappa'_\sigma(x) = \frac{1}{\sigma^2} \kappa' \left(\frac{x}{\sigma} \right) \quad (16)$$

and defining

$$\Delta e_w^{ji} = \left(\frac{(d_j - d_i) - \mathbf{w}^T(\mathbf{x}_j - \mathbf{x}_i)}{\sigma} \right) \quad (17)$$

the gradient can be expressed as

$$\frac{\partial V_\alpha}{\partial \mathbf{w}} = \frac{(\alpha-1)}{\sigma^\alpha N^\alpha} \sum_j \left(\sum_i \kappa(\Delta e_{\mathbf{w}}^{ji}) \right)^{\alpha-2} \cdot \left(\sum_i \kappa'(\Delta e_{\mathbf{w}}^{ji}) \cdot (\mathbf{x}_i - \mathbf{x}_j)^T \right) \quad (18)$$

which leads to the Hessian matrix of $R/2$ for this quadratic surface, where

$$\begin{aligned} R &= \frac{\partial \nabla V_\alpha(\mathbf{w}_*)}{\partial \mathbf{w}} = \frac{\partial^2 V_\alpha(\mathbf{w}_*)}{\partial \mathbf{w}^2} \\ &= \frac{(\alpha-1)}{\sigma^\alpha N^\alpha} \sum_j \left[\sum_i \kappa(\Delta e_{\mathbf{w}_*}^{ji}) \right]^{\alpha-3} \\ &\quad \cdot \left\{ \begin{array}{l} (\alpha-2) \left[\sum_i \kappa'(\Delta e_{\mathbf{w}_*}^{ji}) \cdot (\mathbf{x}_i - \mathbf{x}_j) \right] \\ \cdot \left[\sum_i \kappa'(\Delta e_{\mathbf{w}_*}^{ji}) \cdot (\mathbf{x}_i - \mathbf{x}_j)^T \right] \\ + \left[\sum_i \kappa(\Delta e_{\mathbf{w}_*}^{ji}) \right] \\ \cdot \left[\sum_i \kappa''(\Delta e_{\mathbf{w}_*}^{ji}) \cdot (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \right] \end{array} \right\}. \quad (19) \end{aligned}$$

Note that under the conditions provided in Fact 4b, the Hessian of the information potential evaluated at the global maximum is strictly negative definite (if there is at most one Gaussian input) for the ADALINE structure.

Now that we have a valid quadratic approximation for the cost function and a linear approximation for the weight update equations, we can borrow the well-known convergence analysis results from the MSE convergence theory, only we need to replace the eigenvalues of the input covariance matrix (autocorrelation matrix in the FIR filter case) with the eigenvalues of the Hessian matrix for the entropy criterion given above.

This leads to the following upper bound on the step size of the steepest ascent algorithm for stable convergence to the optimum solution

$$0 < \eta < \frac{1}{\max_i |\lambda_i|}. \quad (20)$$

Similarly, the approximate time constants of convergence of each individual mode (along the eigenvectors of the Hessian matrix) are obtained in terms of the step size and the corresponding eigenvalues as

$$\tau_k = \frac{-1}{\ln(1+\eta\lambda_k)} \approx \frac{-1}{\eta\lambda_k} = \frac{1}{\eta|\lambda_k|}. \quad (21)$$

We remark at this point that although the analysis presented in the above depends on the knowledge of the eigenvalues of the Hessian matrix evaluated at the global maximum of the information potential and is valid only in the region where the cost function is accurately approximated by a quadratic form, using the results of the next section, we will be able to determine the upper bound on the step size for stability using the eigenvalues of the Hessian of the cost function at the current values of the weights. Furthermore, the dilation property of the performance surface (which is demonstrated in Fig. 1 and [28]) guarantees that the approximation will hold.

In addition, within the quadratic region, since the Hessian matrix of the cost function evaluated at any point will be approximately equal to that at the optimal solution, the required eigenvalues can be readily replaced by estimates of the eigenvalues at the current position in the weight space. This approach will render the current estimate of the upper bound on the step size to guarantee stability. The Hessian matrix evaluated at the current estimate of the optimal weight vector is easily obtained using (19); only the evaluation needs to be carried out [not at the optimal solution (since that value is not yet available), but at the current weight values].

IV. Effects of Kernel Size and Entropy Order on the Eigenvalues

Understanding the relationship between the eigenvalues and the kernel size and α is crucial to maintain the convergence of the algorithm under changes in these parameters. One practical case where this relationship becomes important is when we adapt the kernel size during the training. Motivated by the link between the entropy estimator in (4) and the convolution smoothing method of global optimization [8], [25], we suggested starting from a large kernel size and decreasing it to

a nominal value during adaptation. It is then possible to use steepest ascent to maximize the information potential but still guarantee convergence to the global maximum by smoothing of the cost function by convolution by a suitable functional. Since, in this approach, the kernel size is decreased, we need to know how to adapt the step size to achieve faster learning in the initial phase of adaptation (by using a larger step size) and stable convergence in the final phase (by using a smaller step size).

For convenience, we repeat here the reasoning that justifies the dilation in weight space. Consider the information potential estimator (4). It is clear that the introduction of a kernel size other then unity causes the error samples to be treated as if they are divided by σ. Thus, in the error space, the location of the global optimum is scaled along a radial direction from the origin. The exception is the case of zero error because then the global optimum is the origin in the error space, and the optimal solution does not change with kernel size. Since the adaptive topologies used in practice (feedforward systems) are mainly contractive or volume-preserving structures, the dilation/stretching effect is directly translated to the weight space. This property will be observed in the behavior of the eigenvalues under changing kernel size.

As an example, consider the case where we evaluate the quadratic information potential using Gaussian kernels. In this case, the Hessian matrix simplifies to

$$\frac{R}{2} = \frac{1}{2\sigma^2 N^2} \sum_j \left[\sum_i \kappa''(\Delta e^{ji}_{\mathbf{w}_*})(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \right]. \tag{22}$$

Observe from (17) that as σ increases, $\Delta e^{ji}_{\mathbf{w}_*} \to 0$, and therefore, $\kappa''(\Delta e^{ji}_{\mathbf{w}_*}) \to 0^-$ with speed $O(\sigma^{-6})$. This is faster than the reduction rate of the denominator, which is $O(\sigma^{-2})$; hence, overall, the eigenvalues of R approach 0^-. This means that the valley near the global maximum gets wider, and one can use a larger step size in steepest ascent while still achieving stable convergence to it. In fact, this result can be generalized to any kernel function and any α. The dilation effect mentioned in [8] is a direct cause of the increase in eigenvalues toward zero. Notice, however, that this qualitative analysis on the effect of kernel size is valid asymptotically. Yet, it is commonplace to assume (perhaps hope) that the asymptotic behaviors of many statistical quantities satisfactorily represent nonasymptotic behavior.

The analysis of the eigenvalues for varying α is more complicated. In fact, a precise analysis cannot be analytically pursued, but we can still try to predict as to how the eigenvalues of the Hessian behave as this parameter is modified. In order to estimate the behavior of the eigenvalues under changing α, we will exploit the following well-known result from linear algebra relating the eigenvalues of a matrix to its trace. For any matrix R whose eigenvalues are given by the set $\{\lambda_i\}$, the following identity holds:

$$\sum_i \lambda_i = \text{trace}(R). \tag{23}$$

Now, consider the general expression of R given in (19). The trace of R is easily computed to be as given below in (24). The eigenvalues of R are negative, and the dominant component, which introduces this negativity, is the term in the last line of (24). The negativity arises naturally since we use a differentiable symmetric kernel, and since at $\mathbf{w}_*$ the entropy is small, the error samples are close to each other and the second derivative evaluates as a negative coefficient. Now, let us focus on the term that involves the $(\alpha - 3)$-power in the first line of (24). Since all other terms vary linearly with α, this term will dominantly affect the behavior of the trace when α is varied. Consider the case where σ is small enough such that the small entropy causes the kernel evaluations in the brackets to be close to their maximum possible values, and the sum, therefore, exceeds one. In that case, the power of the quantity in the brackets will increase exponentially with increasing α (for $\alpha > 3$); thus, regardless of the terms affected linearly by α, the overall trace value will decrease (increase in absolute value). Consequently, a narrower valley toward the maximum will appear, and the upper bound on the step size for stability will be reduced:

$$\begin{aligned}\text{trace}(R) = \frac{(\alpha-1)}{\sigma^\alpha N^\alpha} \sum_j \left[\sum_i \kappa(\Delta e^{ji}_{\mathbf{w}_*}) \right]^{\alpha-3} \\ \cdot \left\{ \begin{array}{l} (\alpha-2) \sum_k \left[\sum_i \kappa'(\Delta e^{ji}_{\mathbf{w}_*}) \cdot (\mathbf{x}_{ik} - \mathbf{x}_{jk}) \right]^2 \\ + \left[\sum_i \kappa(\Delta e^{ji}_{\mathbf{w}_*}) \right] \\ \cdot \left[\sum_i \kappa''(\Delta e^{ji}_{\mathbf{w}_*}) \cdot \left(\sum_k (\mathbf{x}_{ik} - \mathbf{x}_{jk})^2 \right) \right] \end{array} \right\}.\end{aligned} \tag{24}$$

On the other hand, if the kernel size is large so that the sum in the brackets is less than one, then the $(\alpha - 3)$-power of this quantity will decrease, thus resulting in a wider valley toward the maximum in contrast to the previous case (for $\alpha > 3$). However, in practice, we do not want to use a very small or a very large kernel size as this will increase the variance or increase the bias of the Parzen estimation, respectively [27].

In fact, there is another approach that directly demonstrates how the eigenvalues of R will decrease with increasing α, and vice versa. Consider (19) or (24) again. Since, at the operating point, the error entropy is small and the difference between error samples is close to zero, the sums involving the derivative of the kernel function are approximately zero. Under the conditions mentioned in the previous paragraph, all the terms involving α remain as scalar coefficients that multiply a matrix, whose eigenvalues are negative. With the same arguments on how increasing α increases these coefficients, we conclude that the eigenvalues of the matrix R will increase in absolute value for a small kernel size and decrease for a large kernel size.

In this section, we have investigated the effect of the entropy order α and the kernel size σ on the eigenvalues of the Hessian matrix of the information potential criterion around the optimal solution. We have seen that the entropy order can have different effects, depending on the specific value of the kernel size. As for the effect of kernel size, we have observed that as it increases, the quadratic approximation to the cost function has

larger eigenvalues. This points out a wider region of validity for the linear approximation to the gradient in (18). We remark that our conclusions in this section do not only apply to the eigenvalues of R, but they generalize to how these two parameters affect the volume of the region where our quadratic approximation is valid. These results are imperative from a practical point of view because they explain how the structure of the performance surface can be manipulated by adjusting these parameters. Besides, they identify the procedures to adjust the step size for fast and stable convergence.

In order to summarize the findings of this section, we present the following two facts.

Fact 7*:* Regardless of entropy order, increasing the kernel size results in a wider valley around the global maximum by decreasing the absolute values of the (negative) eigenvalues of the Hessian matrix of the information potential criterion.

Proof: The proof is in the preceding text.

Fact 8*:* The effect of entropy order on the eigenvalues of the Hessian depends on the value of the kernel size. If the kernel size is small, then increasing the entropy order increases the absolute values of the (negative) eigenvalues of the Hessian of the information potential function at the global maximum. This results in a narrower valley. If the kernel size is large, the effect is the opposite, i.e., increasing the entropy order decreases the absolute value of the eigenvalues of the Hessian of the information potential, resulting in a wider valley.

Proof: The proof is in the preceding text.

V. Stochastic Information Gradient

In practice, online training methods are far more valuable than batch mode algorithms since they can provide real-time updates for the weight vector on a sample-by-sample basis as new data arrives, without needing a batch of samples to be collected and stored in memory. The strength of LMS, for example, lies in its ability to determine the optimal solution of the MSE criterion with extremely simple updates on the weight vector, which it computes using only the most recently acquired input signal. In this section, we will derive a stochastic gradient for the MEE criterion, whose batch mode convergence properties are discussed in detail in the preceding sections. This derivation will be motivated by Widrow's approach in deriving the stochastic gradient in LMS, which has proved its merits through the years that followed [4].

Recall that we have defined the order-α information potential for the error as $V_\alpha(e) = E[f_e^{\alpha-1}(e)]$, and since the logarithm is a monotonic function, for $\alpha > 1$, minimization of error entropy is equivalent to maximization of this information potential. Suppose that in an online adaptation scenario, we approximate the information potential *stochastically* by the argument of the expectation operation. This way, dropping $E[.]$ and substituting the required pdf by its Parzen estimate over the most recent L samples, at time k, our information potential estimate becomes

$$V_\alpha(e) \approx \left(\frac{1}{L} \sum_{i=k-L}^{k-1} \kappa_\sigma(e_k - e_i) \right)^{\alpha-1}. \tag{25}$$

Noticing the relationship between entropy and information potential, for an ADALINE structure, the stochastic gradient of entropy with respect to the weight vector is easily computed to be

$$\frac{\partial \hat{V}_{\alpha,k}}{\partial \mathbf{w}} = (\alpha-1) \left(\frac{1}{L} \sum_{i=k-L}^{k-1} \kappa_\sigma(e_k - e_i) \right)^{\alpha-2} \cdot \left(\frac{1}{L} \sum_{i=k-L}^{k-1} \kappa'_\sigma(e_k - e_i)(\mathbf{x}_i - \mathbf{x}_k) \right). \tag{26}$$

For optimizing Renyi's entropy of order $\alpha \neq 1$, the information potential, and, therefore its stochastic gradient given in (26), can be utilized. For Shannon's entropy, the following stochastic gradient exists.

Fact 9*:* The stochastic information gradient (SIG) of Shannon's entropy estimated from the samples using Parzen windowing is given by

$$\frac{\partial \hat{H}_{s,k}}{\partial \mathbf{w}} = \frac{\sum_{i=k-L}^{k-1} \kappa'_\sigma(e_k - e_i)(\mathbf{x}_k - \mathbf{x}_i)}{\sum_{i=k-L}^{k-1} \kappa_\sigma(e_k - e_i).} \tag{27}$$

This is an unbiased estimate of the gradient of Shannon's entropy estimated using Parzen windowing.

Proof: Consider Shannon's entropy of the error given by $H_S(e) = -E[\log f_e(e)]$. Suppose we estimate this quantity by substituting the pdf with its Parzen window estimate over the most recent L samples at time k. Then, this estimate of Shannon's entropy at time k is given by

$$\hat{H}_{S,k}(e) \approx -E\left[\log\left(\frac{1}{L} \sum_{i=k-L}^{k-1} \kappa_\sigma(e_k - e_i) \right) \right]. \tag{28}$$

The gradient of this with respect to the weight vector is easily determined to be

$$\frac{\partial \hat{H}_{s,k}}{\partial \mathbf{w}} = E\left[\frac{\sum_{i=k-L}^{k-1} \kappa'_\sigma(e_k - e_i)(\mathbf{x}_k - \mathbf{x}_i)}{\sum_{i=k-L}^{k-1} \kappa_\sigma(e_k - e_i)} \right] \tag{29}$$

which is simply the expected value of the SIG in (27).

There is an interesting special case of SIG that occurs when the window length $L = 1$ and the kernel function is selected to be a Gaussian function. For Gaussian kernels, the derivative of the kernel can be written in terms of the kernel function itself as $G'_\sigma(x) = -xG_\sigma(x)/\sigma^2$. When these are substituted in (27), the SIG simplifies down to

$$\frac{\partial \hat{H}_{S,k}}{\partial \mathbf{w}} = -\frac{1}{\sigma^2}(e_k - e_{k-1})(\mathbf{x}_k - \mathbf{x}_{k-1}). \tag{30}$$

Notice the structural resemblance between (30) and the LMS update, which is given by $-2e_k\mathbf{x}_k$, for ADALINE. The SIG updates are based on the instantaneous increments of the signal values in this special case, whereas the LMS updates are based on the instantaneous signal values. In relation to anti-Hebbian learning, LMS tries to uncorrelate the error and the input vector,

but SIG tries to uncorrelate the instantaneous increments of these signals. As a final remark on SIG, notice that in (30), the kernel size can be incorporated in the step size of the stochastic adaptation algorithm.

In this section, we have presented SIG (the stochastic gradient of entropy), which can be used to train ADALINE structures online for the minimization of error entropy. SIG is a simple algorithm that allows manipulation of entropy on a sample-by-sample basis with relatively low computational requirements, and specifically, a special case of SIG corresponding to a single-sample-window and Gaussian kernels shows great resemblance to the LMS updates, both in terms of structure and computational complexity.

VI. Numerical Case Studies

In this section, we will present a number of numerical examples to demonstrate the theoretical conclusions drawn in the preceding sections. These include the effect of kernel size and entropy order on the volume of the region of valid quadratic approximation and on the eigenvalues of the Hessian of the cost function at the global optimum solution for ADALINE training. In addition, the results of a series of Monte Carlo simulations that illustrate the noise rejection and data efficiency of the proposed entropy criterion in supervised training of ADALINE are presented, in comparison with MSE.

In all of the simulations below, for visualization purposes, we used a two-tap ADALINE for which the training data is also generated by a two-tap ADALINE with weight vector $\mathbf{w}_* = [1, 2]^T$, except for the SIG examples, where the optimal solution is different. Thus, in the supervised system identification scheme depicted in Fig. 2, both the unknown system and the adaptive system have the same ADALINE structure. For information potential and Renyi's entropy, the estimator in (4) is utilized. Evaluation of associated gradients and Hessians are carried out using the formulas presented in the preceding sections. Note that FIR filters are special cases of the ADALINE structure; therefore, all the conclusions drawn apply to the training of adaptive FIR filters as well.

A. Effect of Entropy Order and Kernel Size on the Performance Surface

This case study aims to illustrate how the performance surface (here represented by its contour plots) of the information potential criterion for supervised training of an ADALINE are altered as a consequence of changing entropy order and kernel size in the estimator. In order to avoid excessive computation time requirements, we have utilized 20 noiseless training samples to obtain the contour plots shown in Fig. 1.

Recall that we have concluded that as the kernel size is increased, the valley around the global maximum becomes wider (allowing larger step size for stable convergence), as well as the volume of the region of quadratic approximation. This is clearly observed in the columns of Fig. 1. As we would expect, as a consequence of Fact 5, the coverage area of the quadratic approximation expands as the cost function approaches the MSE when the kernel size is increased. In Fig. 1, each row represents a constant kernel size $\sigma = 0.1, 0.5, 1$, and each column represents a constant entropy order $\alpha = 1.5, 2, 3$, respectively.

B. Eigenvalues of the Hessian as a Function of Entropy Order and Kernel Size

We have concluded through a theoretical analysis of the analytical expression of the Hessian and its trace that at the optimal solution, the eigenvalues of the Hessian depend on the kernel size and the entropy order. Specifically, in Facts 7 and 8, we saw that as the kernel size is increased, the absolute values of the (negative) eigenvalues of the information potential decrease, and as the entropy order is increased, depending on the current value of the kernel size, the eigenvalues either increase or decrease. For a set of 20 noiseless training samples, we have evaluated the eigenvalues of the Hessian of the information potential at the optimal solution for various values of entropy order and kernel size. The results are shown in Fig. 3. In the subplots presented in the first row of Fig. 3, it is clearly seen that the behavior of the absolute values of the eigenvalues of the information potential is exactly as we have predicted according to the theoretical analysis. Note however, that the logarithms of the eigenvalues are pictured to account for a wider range of values, and this is the reason why the values of the two eigenvalues look similar.

In the lower row of Fig. 3, we have also presented the eigenvalues of the actual entropy evaluated at the optimal point for the same values of kernel size and entropy order. Recall that entropy and information potential are related by $H_\alpha(e) = [\log V_\alpha(e)]/(1-\alpha)$; hence, their Hessians are associated with each other by

$$\frac{\partial^2 H_\alpha(e)}{\partial \mathbf{w}^2} = \frac{1}{1-\alpha} \cdot \frac{V_\alpha(e)\cdot(\partial^2 V_\alpha(e)/\partial \mathbf{w}^2) - (\partial V_\alpha(e)/\partial \mathbf{w})\cdot(\partial V_\alpha(e)/\partial \mathbf{w})^T}{V_\alpha^2(e)}. \tag{31}$$

Since error entropy is minimized (as opposed to the maximization of information potential for $\alpha > 1$), its eigenvalues are already positive.

C. Noise Rejection of MEE and Comparison With MSE

In Fact 6, we mentioned that the error entropy criterion is (ideally) robust to additive noise in the desired signal. This means that if we could obtain analytical expressions of the entropy values for each value of the weight vector and minimize these values, then the additive noise present in the desired signal would not be able to deviate the estimated system parameters from their corresponding actual values. In fact, MSE has the same noise rejection property asymptotically (which can be shown easily).

In this section, we aim to compare the finite-sample performances of MEE and MSE criteria in ADALINE training in noisy conditions. For this purpose, an independent zero-mean Gaussian noise (this specific choice of the noise pdf has no significance) is introduced to the desired response at various

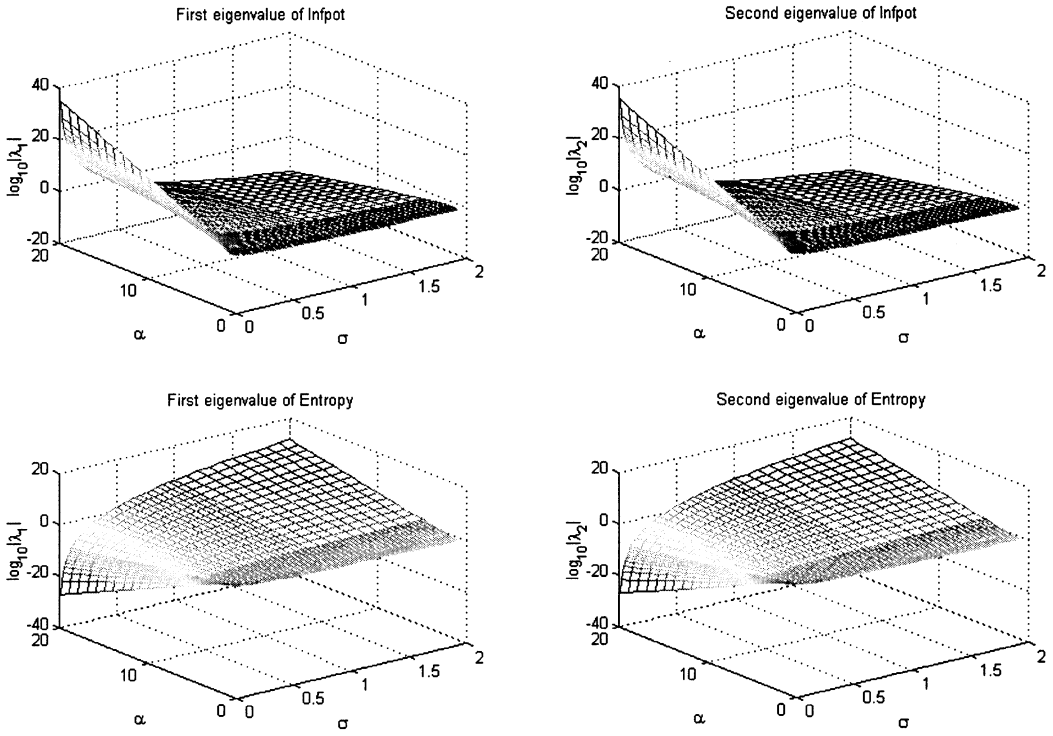


Fig. 3. Eigenvalues of the Hessian of the information potential (upper row) and entropy (lower row) evaluated at the optimal solution, presented as a function of the kernel size (σ) and the entropy order (α). There are two eigenvalues since the ADALINE has two weights.

signal-to-noise-ratio (SNR) levels. Then, for each SNR level (100 Monte Carlo runs using randomly selected training data), the estimated system parameters are obtained through the use of gradient ascent procedure for information potential (with $\alpha = 2$ and $\sigma = 1$ for all simulations) and using the Wiener–Hopf equation for MSE (the covariance of the input and the crosscovariance of the desired signal and the input vector are estimated from the samples). For each run, the distance between the estimated weight vector for the model ADALINE and the actual weight vector of the model is calculated, and then, these are averaged over the 100 runs for each SNR value.

Fig. 4 shows the average deviation of the estimated weight vectors from the actual weight vectors for MEE and MSE as a function of SNR, using training set sizes of $N = 10, 20, 50, 100$. Notice that for small noise power (SNR greater than approximately 5 dB), MEE outperforms MSE in noise rejection consistently. In addition, for high SNR values (greater than 20 dB), MEE is extremely data efficient compared with MSE because it obtains the same level of performance achieved by MSE using fewer samples.

These results indicate that MEE is more robust to noise in the desired signal in a finite-sample case, and furthermore, it extracts the information in the samples efficiently to obtain a better solution with fewer samples. This result is also consistent with those we have obtained in the blind source separation problem using Renyi's entropy [10].

D. Demonstration of SIG in ADALINE Training

We have presented SIG as an online method to manipulate entropy and to minimize error entropy in supervised ADALINE training. Specifically, an oversimplified special case of SIG that is computationally very simple and that resembles LMS in structure is expressed. In the following two sample simulations, we aim to illustrate the performance of SIG in ADALINE training [29]; once again, for visualization purposes, we have chosen two-weight situations, where one is the prediction of a time series from its most recent two samples, where the sequence is generated by $x(t) = \sin 20t + 2\sin 40t + 3\sin 60t$ sampled at 100 Hz. The training set consists of 32 samples, which approximately corresponds to one period of the signal, and are used repeatedly for 150 epochs for both SIG (the oversimplified version) and LMS algorithms. The weight tracks of both algorithms starting from five different initial conditions are shown in Fig. 5(a), along with the contours for the MEE criterion for this training set. In Fig. 5(b), we present the weight tracks of a training scenario for a two-weight frequency-doubler scheme with 20 samples and 1000 epochs. The FIR filter is adapted to approximately generate a sinusoid with double the frequency of the sinusoid at its input. Note that this is just an illustration. In general, an accurate solution of the frequency-doubling problem requires FIR filters with many taps.

VII. Discussions

We have recently introduced minimum error entropy (MEE) as an information theoretic supervised training approach and demonstrated its superiority over the mean square error (MSE) criterion in a variety of applications including channel equalization, chaotic time-series prediction, and nonlinear system identification. In this paper, however, we focused on the investigation of the convergence properties of the steepest descent algorithm for ADALINE training using the MEE criterion. We have shown that adjusting the kernel size and entropy order—the two design parameters in the proposed cost function—can control the

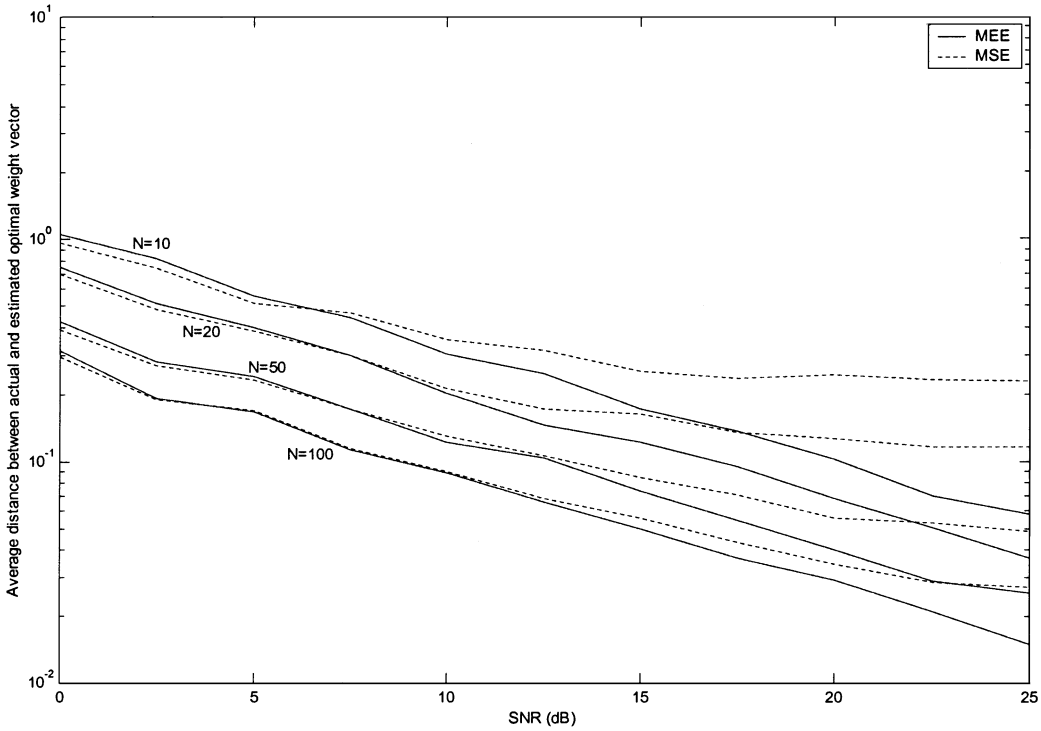
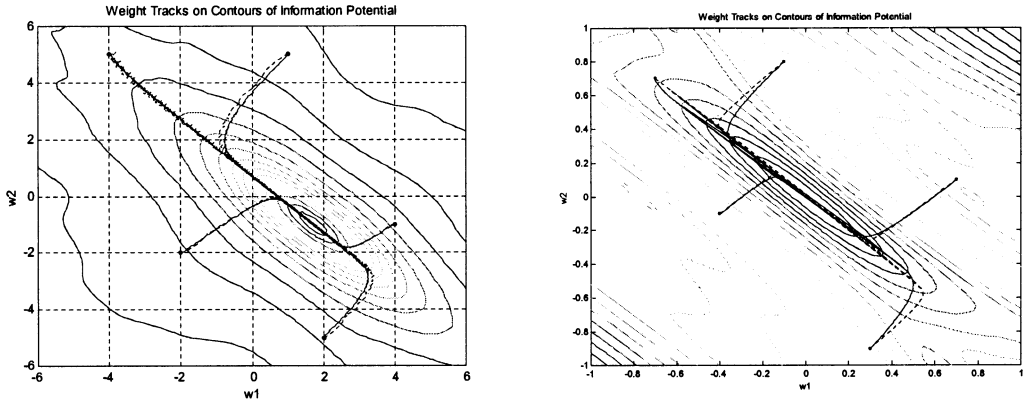


Fig. 4. Average distance between the estimated and actual weight vectors for MEE and MSE as a function of SNR.



(a) (b)

Fig. 5. Weight tracks for SIG (solid) and LMS (dotted) in online ADALINE training for (a) time-series prediction and (b) frequency doubling.

volume of this neighborhood, where the assumed approximations are valid. Borrowing results from the analysis of the MSE criterion in the literature, we have determined the upper bound for the step size for stability and the corresponding time constant that approximately identifies the convergence speed of the algorithm in the vicinity of the solution.

The effects of the two characteristic parameters, namely, the entropy order and the kernel size, on the structure of the performance surface is examined. It was determined theoretically (using analytical techniques) and illustrated through numerical examples that as the kernel size is increased, which is consistent with our conjecture on the equivalence with convolution smoothing, the performance surface smoothens and the eigenvalues of the Hessian of the entropy evaluated at the optimal point gets closer to zero, resulting in a wider valley, thus allowing for larger step size values for stability. Increasing the entropy order, however, depending on the value of the kernel size, may have differing effects on the eigenvalues of the Hessian of entropy. When the kernel size is small, the eigenvalues decrease (toward zero), and when the kernel size is large, the eigenvalues increase with increasing entropy order.

Motivated by the asymptotic total noise rejection properties of MEE and mean square error criteria, we have also compared their robustness in determining the actual parameter values under noise for various sizes of finite-sample training sets. Monte Carlo simulations performed over a range of signal-to-noise ratios revealed that, given a fixed size data set, MEE is more robust to additive noise in the desired signal than mean square error in the range of practically encountered noise levels. In addition, although both criteria are totally robust to additive noise asymptotically as the number of samples goes to infinity, the simulations suggested that the entropy criterion converges faster to this asymptotic result as a function of the number of samples.

Finally, we have introduced the *stochastic information gradient* for minimization of the error entropy online with relatively much less computational requirements at each iteration of the learning algorithm. This stochastic information theoretic learning algorithm was derived from the error entropy, following the stochastic approximation guidelines. We have demonstrated the operation of SIG by portraying its weight tracks imposed on contour plots of the MEE criterion. These simulations, for two different training data sets with multiple initial conditions for each, confirmed that learning with SIG, linear adaptive systems could learn minimum entropy solutions on a sample-by-sample basis.

In conclusion, this paper mainly dealt with the batch convergence properties of the MEE criterion in supervised linear adaptive system training while proposing a simple stochastic learning rule for online information-theoretic adaptation purposes, whose average behavior is also governed by the same dynamics as batch learning. Future research is in order to extend these results to supervised and unsupervised training of nonlinear adaptive systems that use the proposed entropy estimator as an integral part of their learning criteria. In addition, a methodology to determine the annealing schedule for the kernel size for any given problem will benefit the algorithms that utilize MEE.

APPENDIX

Alternative Proof for Fact 6: Assume that a clean desired signal is generated by $\overline{d} = g(x; \mathbf{w}^*)$ and that the noisy desired signal is obtained from this signal with $d = \overline{d} + n$. Let the adaptive system output be $y = g(\mathbf{x}; \mathbf{w})$ for an arbitrary set of weights in $\mathbf{w}$. In addition, let $p_n(\eta)$ be the zero-mean noise pdf for n and $p_{\mathbf{x}}(\xi)$ be the pdf of the input signal $\mathbf{x}$. Suppose that the conditional pdf of $\overline{d}$ given x is $p_{\overline{d}|\mathbf{x}}(\overline{\delta}|\xi; \mathbf{w}^*)$ and that the noise is independent from the input. Then, the conditional pdf of noisy desired given the input is

$$p_{d|\mathbf{x}}(\delta|\xi; \mathbf{w}^*) = p_{\overline{d}|\mathbf{x}}(\delta|\xi; \mathbf{w}^*) * p_n(\delta). \tag{A.1}$$

The error is defined as $e(\mathbf{x}; \mathbf{w}, \mathbf{w}^*) = d - y = [g(\mathbf{x}; \mathbf{w}^*) - g(\mathbf{x}; \mathbf{w})] + n \triangleq m(\mathbf{x}; \mathbf{w}, \mathbf{w}^*) + n$. For an ADALINE structure, the mapping g is given by $g(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T\mathbf{x}$, and therefore, $m(\mathbf{x}; \mathbf{w}, \mathbf{w}^*) = (\mathbf{w}^* - \mathbf{w})^T\mathbf{x}$. The pdf of m becomes $p_{m|\mathbf{x}}(\mu|\xi; \mathbf{w}, \mathbf{w}^*) = p_{\overline{d}|\mathbf{x}}(\mu|\xi; \mathbf{w}^* - \mathbf{w})$. Using this, we write the conditional pdf of the error as

$$p_{e|\mathbf{x}}(\varepsilon|\xi; \mathbf{w}, \mathbf{w}^*) = p_{m|\mathbf{x}}(\varepsilon|\xi; \mathbf{w}, \mathbf{w}^*) * p_n(\varepsilon). \tag{A.2}$$

The probability of error is then written as the product of this conditional pdf and the input pdf

$$\begin{aligned} p_e(\varepsilon) &= \int_{-\infty}^{\infty} p_{e|\mathbf{x}}(\varepsilon|\xi; \mathbf{w}, \mathbf{w}^*) p_{\mathbf{x}}(\xi)\, d\xi \\ &= \int_{-\infty}^{\infty} \left[p_{m|\mathbf{x}}(\varepsilon|\xi; \mathbf{w}, \mathbf{w}^*) * p_n(\varepsilon)\right] p_{\mathbf{x}}(\xi)\, d\xi \\ &= \int_{-\infty}^{\infty} \left[p_{m|\mathbf{x}}(\varepsilon|\xi; \mathbf{w}, \mathbf{w}^*) p_{\mathbf{x}}(\xi)\right] * p_n(\varepsilon)\, d\xi \\ &= \int_{-\infty}^{\infty} \left[p_{\overline{d}|\mathbf{x}}(\varepsilon|\xi; \mathbf{w}^* - \mathbf{w}) p_{\mathbf{x}}(\xi)\right] d\xi * p_n(\varepsilon). \end{aligned} \tag{A.3}$$

Notice that if $\mathbf{w} = \mathbf{w}^*$, the error distribution becomes $p_e(\varepsilon) = \delta(\varepsilon) * p_n(\varepsilon) = p_n(\varepsilon)$, and thus, the error entropy is equal to the noise entropy. Otherwise, the error probability is a convolution of the noise pdf with some other pdf that depends on the current weight vector and the optimal weight vector. In that case, we know by Fact 4 that the entropy of the error will be greater than the entropy of noise. Are there any other weight vectors that may lead to a δ-distributed error? In the ADALINE case, as long as the number of training samples is greater than or equal to the number of weights, the answer is "no" because the weight vector that yields zero error over all samples is determined as the solution to a linear system of equations, and these have unique solutions. This proves that the actual weight vector $\mathbf{w}^*$ is the only global minimum of the MEE criterion, even in the case of noisy desired signal.

REFERENCES

[1] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications*. Cambridge, MA: MIT Press, 1949.

[2] S. Haykin, *Adaptive Filter Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[3] B. Farhang-Boroujeny, *Adaptive Filters: Theory and Applications*. New York: Wiley, 1998.

[4] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1985.

[5] A. H. Sayed, "Advances and challenges in adaptive filtering," in *Proc. Int. Conf. Acoust., Speech, Signal Process., Tutorial Lecture Notes 4*, 2002.

[6] D. Erdogmus and J. C. Principe, "Comparison of entropy and mean square error criteria in adaptive system training using higher order statistics," in *Proc. ICA*, Helsinki, Finland, 2000, pp. 75–80.

[7] ——, "An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems," *IEEE Trans. Signal Processing*, vol. 50, pp. 1780–1786, July 2002.

[8] ——, "Generalized information potential criterion for adaptive system training," *IEEE Trans. Neural Networks*, vol. 13, pp. 1035–1044, Sept. 2002.

[9] J. C. Principe, D. Xu, and J. Fisher, "Information theoretic learning," in *Unsupervised Adaptive Filtering, vol.I: Blind Source Separation*, S. Haykin, Ed. New York: Wiley, 2000, pp. 265–319.

[10] K. E. Hild, II, D. Erdogmus, and J. C. Principe, "Blind source separation using Renyi's mutual information," *IEEE Signal Processing Lett.*, vol. 8, pp. 174–176, June 2001.

[11] D. Erdogmus, K. E. Hild, II, and J. C. Principe, "Blind source separation using Renyi's α-marginal entropy," *Neurocomput.—Special Issue on Blind Source Separation*, vol. 49, no. 1, pp. 25–38, 2002.

[12] D. Erdogmus, J. C. Principe, and L. Vielva, "Blind deconvolution with minimum Renyi's entropy," in *Proc. EUSIPCO*, vol. 2, Toulouse, France, 2002, pp. 71–74.

[13] R. Y. Rubinstein, *Simulation and the Monte Carlo Method*. New York: Wiley, 1981.

[14] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, pp. 1129–1159, 1995.

[15] D. T. Pham, "Mutual information approach to blind separation of stationary sources," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1935–1946, Aug. 2002.

[16] S. Haykin, Ed., *Blind Deconvolution*. Englewood Cliffs, NJ: Prentice-Hall, 1994.

[17] A. O. Hero, B. Ma, O. J. J. Michel, and J. Gorman, "Applications of entropic spanning graphs," *IEEE Signal Processing Mag.*, vol. 19, pp. 85–95, 2002.

[18] K. Torkkola, "Visualizing class structure in data using mutual information," in *Proc. NNSP*, Sydney, Australia, 2000, pp. 376–385.

[19] A. K. C. Wong and P. K. Sahoo, "A gray-level threshold selection method based on maximum entropy principle," *IEEE Trans. Syst., Man, Cybern.*, vol. 19, pp. 866–971, July 1989.

[20] R. G. Baraniuk, P. Flandrin, A. J. E. M. Janssen, and O. J. J. Michel, "Measuring time-frequency information content using the Renyi entropies," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1391–1409, July 2001.

[21] J. W. Fisher, "Nonlinear extensions to the minimum average correlation energy filter," Ph.D. dissertation, Univ. Florida, Gainesville, FL, 1997.
[22] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, 1974.
[23] A. P. Liavas and P. A. Regalia, "On the behavior of information theoretic criteria for model order selection," *IEEE Trans. Signal Processing*, vol. 49, pp. 1689–1695, Aug. 2001.
[24] A. Renyi, *Probability Theory*. New York: Elsevier, 1970.
[25] S. I. Amari, *Differential-Geometrical Methods in Statistics*. Berlin, Germany: Springer-Verlag, 1985.
[26] P. Viola, N. N. Schraudolph, and T. J. Sejnowski, "Empirical entropy manipulation for real-world problems," in *Proc. Advances Neural Inform. Processing Syst. VIII*, 1996, pp. 851–857.
[27] E. Parzen, "On estimation of a probability density function and mode," in *Time Series Analysis Papers*. San Francisco, CA: Holden-Day, 1967.
[28] D. Erdogmus and J. C. Principe, "Convergence analysis of the information potential criterion in adaline training," in *Proc. NNSP*, Falmouth, MA, 2001, pp. 123–132.
[29] D. Erdogmus and J. C. Principe, "An on-line adaptation algorithm for adaptive system training with minimum error entropy: Stochastic information gradient," in *Proc. ICA*, CA, 2001, pp. 7–12.

Deniz Erdogmus (M'02) received the B.S. degree in electrical and electronics engineering and mathematics in 1997 and the M.S. degree in electrical and electronics engineering, with emphasis on systems and control in 1999, both from the Middle East Technical University, Ankara, Turkey. He received the Ph.D. degree in electrical and computer engineering from the University of Florida, Gainesville, in 2002.

He was a Research Engineer with the Defense Industries Research and Development Institute (SAGE), Ankara, from 1997 to 1999. Since 1999, he has been with the Computational NeuroEngineering Laboratory, University of Florida, working under the supervision of Dr. J. C. Principe. His current research interests include information theory and its applications to adaptive systems and adaptive systems for signal processing, communications, and control.

Dr. Erdogmus is a member of Tau Beta Pi and Eta Kappa Nu.

Jose C. Principe (F'00) is a Distinguished Professor of electrical and computer engineering and biomedical engineering with the University of Florida, Gainesville, where he teaches advanced signal processing, machine learning, and artificial neural networks (ANNs) modeling. He is BellSouth Professor and the Founder and Director of the University of Florida Computational NeuroEngineering Laboratory (CNEL). His primary area of interest is processing of time varying signals with adaptive neural models. The CNEL Lab has been studying signal and pattern recognition principles based on information theoretic criteria (entropy and mutual information). He has more than 90 publications in refereed journals, ten book chapters, and 200 conference papers. He has directed 35 Ph.D. dissertations and 45 Master theses. He recently wrote an interactive electronic book entitled *Neural and Adaptive Systems: Fundamentals Through Simulation* (New York: Wiley, 2002)

Dr. Principe is a member of the ADCOM of the IEEE Signal Processing Society, Member of the Board of Governors of the International Neural Network Society, and Editor-in-Chief of the IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING. He is a member of the Advisory Board of the University of Florida Brain Institute.