An Error-Entropy Minimization Algorithm for Supervised Training of Nonlinear Adaptive Systems

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

***Abstract*—This paper investigates error-entropy-minimization in adaptive systems training. We prove the equivalence between minimization of error's Renyi entropy of order α and minimization of a Csiszar distance measure between the densities of desired and system outputs. A nonparametric estimator for Renyi's entropy is presented, and it is shown that the global minimum of this estimator is the same as the actual entropy. The performance of the error-entropy-minimization criterion is compared with mean-square-error-minimization in the short-term prediction of a chaotic time series and in nonlinear system identification.**

***Index Terms*—Minimum error entropy, Renyi's entropy.**

I. Introduction

STARTING with the early work of Wiener [1] on optimal filtering, the mean square error (MSE) has been a popular criterion in the training of all adaptive systems including artificial neural networks [2]. The two main reasons behind this choice are analytical tractability and the assumption that real-life random phenomena may be sufficiently described by second-order statistics. The Gaussian probability density function (pdf) is determined only by its first- and second-order statistics, and the effect of linear systems on low order statistics is well known [3]. Under these linearity and Gaussianity assumptions, further supported by the central limit theorem, MSE, which solely constrains second-order statistics, would be able to extract all possible information from a signal whose statistics are solely defined by its mean and variance.

Although Gaussianity and linear modeling provide successful engineering solutions to most practical problems, it has become evident that when dealing with nonlinear systems, this approach needs to be refined [12]. Therefore, criteria that not only consider the second-order statistics, but that also take into account the higher order statistical behavior of the systems and signals, are much desired. Recent papers have addressed this issue both in the control literature [4] and in the signal processing/machine learning literature [5]–[7].

Entropy, which is introduced by Shannon [8], is a scalar quantity that provides a measure for the average information contained in a given probability distribution function. By definition, information is a function of the pdf; hence, entropy as an optimality criterion extends MSE. When entropy is minimized, all moments of the error pdf (not only the second moments) are constrained. The entropy criterion can generally be utilized as an alternative for MSE in supervised adaptation, but it is particularly appealing in dynamic modeling [9].

Manuscript received September 11, 2000; revised April 2, 2002. This work was supported in part by the National Science Foundation under Grant ECS-9900394 and the Office of Naval Research under Contract N00014-01-1-0405. The associate editor coordinating the review of this paper and approving it for publication was Prof. Colin F. N. Cowan.

The authors are with the Computational NeuroEngineering Laboratory, University of Florida, Gainesville, FL 32611 USA.

Publisher Item Identifier S 1053-587X(02)05659-3.

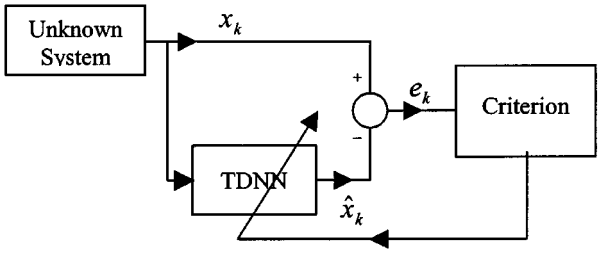

Fig. 1. TDNN prediction scheme; an example of supervised learning.

The goal in dynamic modeling is to identify the nonlinear dynamical system that produced the given input–output mapping. This is traditionally achieved in a predictive framework (see Fig. 1) using a nonlinear adaptive system, whose parameters are adapted with the MSE between the desired output and the system output. Minimization of MSE is, however, simply constraining the square difference between the original trajectory and the trajectory created by the adaptive system, which does not guarantee the capture of all the details of the underlying dynamics. Hence, we propose minimization of error entropy (MEE) as a more robust criterion for dynamic modeling and an alternative to MSE in other supervised learning applications using nonlinear systems such as nonlinear system identification with neural networks.

Application of the entropy criterion to supervised learning is conceptually straightforward. Given samples from an input–output mapping, the entropy of the output error over the training data set must be minimized. In the following, we show that minimizing the error entropy is equivalent to minimizing the distance between the probability distributions of the desired and system outputs. These distance measures, from the information-geometry point of view, are directly related to the divergence of the statistical models in probability spaces [10].

Nonparametric estimation of the probability density function (pdf) of a random variable, which is necessary for the evaluation of its entropy, is required since an analytical expression is not available in most cases. Parzen windowing is an efficient way to approximate the pdf of a given sample distribution, particularly in low-dimensional spaces [11]. In Parzen windowing, the pdf is approximated by a sum of kernels whose centers are translated to the sample points. A suitable and commonly used kernel function is the Gaussian, but others can also be utilized, e.g., Laplacian. The Gaussian function is a preferred choice for adaptation purposes because it is continuously differentiable, i.e., the estimated pdf is continuously differentiable on the space of real vec-

tors. The Gaussian kernel, in addition to these nice features, provides a computational simplification in the algorithm design [12].

The organization of the paper is as follows. First, the equivalence of entropy minimization and pdf matching is established. Second, an analytical proof shows that the global minimum of the entropy is still a minimum of the Parzen window estimated entropy when Gaussian kernels are employed. Then, the backpropagation algorithm for both Shannon's and Renyi's entropy of order 2 are given for the one-dimensional (1-D) case. Finally, two case studies where the entropy criterion is applied to the short-term prediction of a chaotic time series and to the identification of a nonlinear system are presented. The performances of MSE-trained and entropy-trained time delay neural networks (TDNN) built from multiplayer perceptrons (MLPs) are compared in terms of their accuracy in approximating the pdf of the desired output.

II. Error Entropy Minimization and Probability Density Matching

Consider the error between the desired and the actual outputs of the adaptive system (Fig. 1) $e = d - y$. From this, we can deduce the pdf of the error as

$$f_{e,w}(e) = f_{y|x,w}(d - e|x) \tag{1}$$

where the subscript w expresses dependence on the weights of the adaptive system. Minimizing Renyi's order-α error entropy [13] thus becomes

$$\begin{aligned}\min_w \frac{1}{1-\alpha}\log\int f_{e,w}^{\alpha}(e)de &= \frac{1}{1-\alpha}\log\int f_{y|x,w}^{\alpha}(d-e|x)de \\ &= \frac{1}{1-\alpha}\log\int -f_{y|x,w}^{\alpha}(y|x)dy\end{aligned} \tag{2}$$

after the variable change of $y = d - e$. Since we will be concerned with Renyi's quadratic entropy in this paper ($\alpha = 2$), consider the case where entropy order-α is greater than one. Since multiplying the cost function with a factor independent of the weights of the adaptive system will not affect the solution of the problem, we introduce the integral of the power-α of the pdf of the input signal in (2) to obtain the equivalent minimization problem in (3).

$$\begin{aligned}&\equiv \min_w \int f_{y|x,w}^{\alpha}(y|x)dy \cdot \int f_x^{\alpha}(x)dx \\ &= \iint f_{xy,w}^{\alpha}(x,y)dxdy \\ &\equiv \iint f_{xy,w}^{\alpha}(x,y)dxdy \cdot \iint f_{xd}^{1-\alpha}(x,y)dxdy \\ &= \min_w \iint f_{xy,w}(x,y)\left(\frac{f_{xd}(x,y)}{f_{xy,w}(x,y)}\right)^{1-\alpha}dxdy.\end{aligned} \tag{3}$$

We recognize this last expression in (3) as the Csiszar distance [14] with the convex function chosen to be $(.)^{1-\alpha}$. In general, the Csiszar distance between two densities $p(x)$ *and* $q(x)$ is given by

$$D_C(p;q) = \int q(x) f\left(\frac{p(x)}{q(x)}\right)dx \tag{4}$$

where $f(.)$ is convex [14]. Kullback–Leibler divergence [15] is a special case of this divergence corresponding to the choice $-\log(.)$. Consequently, we infer that minimizing Renyi's error entropy results in the minimization of the divergence between the joint pdfs of input-desired and input–output signal pairs. This readily guarantees the matching of the marginal pdfs of the desired and the output signals.

It is interesting to note that for Shannon's entropy, the distance measure in (3) also reduces to the Kullback-Leibler divergence. To see this, we start by modifying the minimization problem by taking the log and dividing by $\alpha - 1$

$$\min_w \frac{1}{\alpha-1}\log\iint f_{xy,w}(x,y)\left(\frac{f_{xd}(x,y)}{f_{xy,w}(x,y)}\right)^{1-\alpha}dxdy. \tag{5}$$

Now, taking the limit of this expression as $\alpha \to 1$ using L'Hopital's rule, we obtain the Kullback–Leibler divergence

$$\begin{aligned}\lim_{\alpha\to 1}\frac{1}{\alpha-1}\log\iint f_{xy,w}(x,y)\left(\frac{f_{xd}(x,y)}{f_{xy,w}(x,y)}\right)^{1-\alpha}dxdy \\ = \iint f_{xy,w}(x,y)\log\left(\frac{f_{xy,w}(x,y)}{f_{xd}(x,y)}\right)dxdy.\end{aligned} \tag{6}$$

Since Shannon's entropy is the limiting case of Renyi's entropy when $\alpha \to 1$ (this fact can also be observed using L'Hopital's rule as Renyi's entropy has a singularity at this value of α), we conclude that specifically, minimizing Shannon's error entropy minimizes the Kullback–Leibler divergence between the joint densities of the input-desired and input–output pairs.

III. Nonparametric Entropy Estimator Preserves the Global Minimum of Actual Entropy

Now, we proceed with proving that the global minimum of the entropy is still a minimum of the nonparametrically estimated entropy for both Shannon's and Renyi's definitions when Parzen windowing with Gaussian kernels is utilized. In practical applications, the pdf of the random process is often unknown *a priori*. Hence, we will utilize the Parzen window method to estimate the pdf directly from the samples. The Parzen estimator of the error pdf $f_e(\xi)$ is given by

$$\hat{f}_e(\xi) = \frac{1}{N}\sum_{i=1}^{N}\kappa(\xi - e_i, \sigma^2) \tag{7}$$

where κ denotes the multidimensional Gaussian function with a radially symmetric variance σ^2 for simplicity. This estimator can then be substituted in the Renyi's entropy definition given in the first line of (2) or Shannon's entropy given in (8).

Shannon's Entropy: We can estimate Shannon's error entropy [8] by substituting the Parzen pdf estimate in place of the actual error pdf, yielding

$$H_S(e) = -\int_{-\infty}^{\infty}\hat{f}_e(\xi)\log\hat{f}_e(\xi)d\xi. \tag{8}$$

Clearly, the global minimum of Shannon's entropy is achieved when the pdf of error is a Dirac-δ function. Since entropy is independent of the mean of the random variable, without loss of generality, we can concentrate on the case where the mean

of e is zero. The gradient of entropy estimated for the Gaussian kernel is given in

$$\frac{\partial H_S}{\partial e_j} = -\frac{\partial}{\partial e_j}\int_{-\infty}^{\infty}\hat{f}_e(\xi)\log\hat{f}_e(\xi)d\xi = -\int_{-\infty}^{\infty}\frac{1}{N\sigma^2}(\xi-e_j)\,\kappa(\xi-e_j)\log\hat{f}_e d\xi. \quad (9)$$

Evaluating this gradient at zero error over the complete set of data $e = [e_1 \quad \cdots \quad e_N] = 0$, we get the integral of an odd function

$$\left.\frac{\partial H_S}{\partial e_j}\right|_{e=0} = -\int_{-\infty}^{\infty}\frac{1}{N\sigma^2}\xi\kappa(\xi)\log\kappa(\xi)d\xi = 0. \quad (10)$$

Hence, $e=0$ is a stationary point of $H_S(e)$. Computation of the Hessian is necessary to see if it is, in fact, a minimum. Using the same approach as above, the diagonal and off-diagonal entries of the Hessian are found to be

$$\left.\frac{\partial^2 H_S}{\partial e_j^2}\right|_{e=0} = \frac{\partial}{\partial e_j}\left(\frac{\partial H_S}{\partial e_j}\right)\Bigg|_{e=0} = \frac{N-1}{N^2\sigma^2}$$
$$\left.\frac{\partial^2 H_S}{\partial e_k \partial e_j}\right|_{e=0} = \frac{\partial}{\partial e_k}\left(\frac{\partial H_S}{\partial e_j}\right)\Bigg|_{e=0} = \frac{-1}{N^2\sigma^2}. \quad (11)$$

The eigenvalues of the Hessian matrix can then be computed as $\lambda_0 = 0$ with multiplicity 1, with a corresponding eigenvector $\vec{e}_0 = [1 \quad 1 \quad \cdots \quad 1]^T$, and $\lambda_i = 1/(N\sigma^2)$, with multiplicity $(N-1)$; hence, the Hessian is positive semi-definite. The eigenvector corresponding to the zero eigenvalue lies along the direction on which the mean remains constant, that is, the value of the entropy is constant. This is expected since the entropy is independent of the mean. This can be easily shown by a simple change of variables in the entropy definition. Therefore, we conclude that Shannon's entropy estimated by Parzen windowing with Gaussian kernels has minima in the directions where all the error samples are identical over the whole data set.

Renyi's Entropy: Renyi's entropy is defined by (2) and is known to approach Shannon's entropy as α approaches 1 [16]. Like Shannon's entropy, it is also independent of the mean of e. In practical situations, we will have to work with an estimator. Here, we will still be using the Parzen estimator with a Gaussian kernel in (7). The gradient of Renyi's entropy in the case of Gaussian kernels, evaluated at $e=0$, is

$$\frac{\partial H_{R\alpha}}{\partial e_j} = \frac{\alpha}{N\sigma^2(1-\alpha)\int_{-\infty}^{\infty}\kappa^{\alpha}d\xi}\int_{-\infty}^{\infty}\xi\kappa^{\alpha}d\xi = 0. \quad (12)$$

Hence, this is a stationary point. Following steps similar to those in Shannon's entropy case, the diagonal and off-diagonal elements of the Hessian matrix evaluated at $e=0$ are found to be

$$\left.\frac{\partial^2 H_{R\alpha}}{\partial e_j^2}\right|_{e=0} = \frac{N-1}{N^2\sigma^2}$$
$$\left.\frac{\partial^2 H_{R\alpha}}{\partial e_k \partial e_j}\right|_{e=0} = \frac{-1}{N^2\sigma^2}. \quad (13)$$

Note that the Hessian matrix for Renyi's entropy computed at the optimal solution is independent of α, and its second-order partial derivatives are identical to those of Shannon's entropy. Hence, it has the same eigenvalues as the Hessian matrix for Shannon's entropy. Similarly, the eigenvector corresponding to the zero eigenvalue is equal as well, and therefore, all the related arguments are valid for Renyi's entropy. Thus, we conclude that Renyi's entropy approximated by Parzen windowing with Gaussian kernels has minima along the line where the error is completely constant over the whole data set.

This analysis, however, only shows that $e=0$ is a local minimum of Renyi's entropy estimator. In order to prove globalness, we need some further analysis. Consider specifically the nonparametric estimator for Renyi's quadratic entropy, which is much simpler to estimate using Gaussian kernels, compared with Shannon's entropy and other orders of Renyi's entropy [12], [17]. When we substitute the Parzen estimator in (2) with the quadratic entropy expression ($\alpha = 2$), we obtain

$$H_2 = -\log\int\left(\frac{1}{N}\sum_{i=1}^{N}\kappa(\xi-e_i,\sigma^2)\right)^2 d\xi = -\log V(e). \quad (14)$$

The argument of the log, $V(e)$ is called the *information potential* [12]. It can be calculated in closed form from the samples using Gaussian kernels as

$$V(e) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\kappa(e_i-e_j,2\sigma^2). \quad (15)$$

This expression is obtained from (14) by interchanging the order of summations and the integral. Then, we notice that the integral of a product of Gaussian kernels is another Gaussian function with twice the variance. Using this expression, one can estimate the value of entropy for $e=0$. It is simply

$$H_2(e=0) = -\log\kappa(0;2\sigma^2). \quad (16)$$

To complete the proof of globalness, we need to show that any other combination of error sample values results in a larger value of entropy, i.e.,

$$-\log\frac{1}{N^2}\sum_j\sum_i\kappa(e_j-e_i;2\sigma^2) \geq -\log\kappa(0;2\sigma^2) \quad (17)$$

or equivalently

$$\sum_j\sum_i\kappa(e_j-e_i;2\sigma^2) \leq N^2\kappa(0;2\sigma^2). \quad (18)$$

This inequality is readily satisfied since, for the Gaussian kernels (with zero mean), the maximum value is achieved at zero. This shows that our nonparametric entropy estimator preserves the global minimum of the actual entropy.

IV. Backpropagation for TDNN Using Entropy Criterion

A typical prediction scheme with a TDNN built from a delay line and an MLP [18] is shown in Fig. 1. The training criterion characterizes the learning process and determines the overall prediction performance. The purpose of this scheme is to find the TDNN weights that optimize the criterion of interest. Although TDNNs are specifically mentioned in this section, it should be noted that the gradient search presented here and MEE criterion applies to the supervised training of any adaptive system with a smooth input–output map.

If the adaptation criterion is chosen to be the minimization of the MSE and the optimization procedure is the steepest descent, then the training algorithm is the well-known backpropagation algorithm [19]. However, if the adaptation criterion is picked to be the minimization of Shannon's entropy of the error due to the reasons stated before, with steepest descent approach, the training algorithm becomes [2]

$$w(n+1) = w(n) - \eta \frac{\partial \hat{H}_S(e)}{\partial w}. \tag{19}$$

Here, the pdf estimator for the error (7) is employed. The gradient of the entropy with respect to the weights is calculated to be

$$\frac{\partial \hat{H}_S(e)}{\partial w} = \frac{1}{N\sigma^2} \sum_{i=1}^{N} \frac{\partial \hat{x}_i}{\partial w} \cdot \int_{-\infty}^{\infty} (\xi - e_i)\kappa(\xi - e_i) \log \hat{f}_e(\xi) d\xi. \tag{20}$$

The term $\partial \hat{x}_i / \partial w$ can be computed as in the standard backpropagation algorithm [2]. The computational drawback of this algorithm is the requirement of the numerical evaluation of a complicated integral over the real line. Therefore, this algorithm is extremely slow and computationally inefficient. Employing Renyi's entropy with $\alpha = 2$, on the other hand, leads to the closed-form nonparametric estimator in (15), simplifying the computational load significantly [12], [17]. Since Renyi's quadratic entropy is a monotonic function of the information potential, we can equivalently maximize information potential instead of minimizing Renyi's entropy and further simplify the adaptation algorithm. The gradient vector to be used in the steepest ascent algorithm for the maximization of the information potential is

$$\frac{\partial \hat{V}(e)}{\partial w} = \frac{1}{2N^2\sigma^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (e_j - e_i) \cdot \kappa(e_i - e_j, 2\sigma^2) \cdot \left[\frac{\partial \hat{x}_j}{\partial w} - \frac{\partial \hat{x}_i}{\partial w} \right]. \tag{21}$$

One important point to note in training with entropy is that since entropy does not change with the mean of the distribution, the algorithm will converge to a set of optimal weights, which may not yield zero-mean error [6]. However, this can be easily corrected by properly modifying the bias of the output processing element (PE) of the MLP to yield zero mean error over the training data set just after training ends. It must also be noted that the optimization of the TDNN with an entropy cost function may display local minima, as it does for the MSE cost. This could be experimentally verified, and it creates well-known difficulties to gradient-based algorithms [2].

V. Simulation Results

As the first case study, the short-term prediction of the Mackey–Glass chaotic time series [20] with parameter $\tau = 30$ using both MSE-trained and MEE (Renyi's)-trained TDNNs is presented. The TDNN inputs consist of the current value of the sequence and six delayed values and a single linear output PE, whereas the number of PEs in the hidden layer is varied from 3 to 10. The nonlinearity used is the $\tanh$ function. The size of the input delay line is consistent with the embedding dimension suggested by Taken's embedding theorem for the Mackey–Glass series [21]. The sampling period is chosen as 0.1 s.

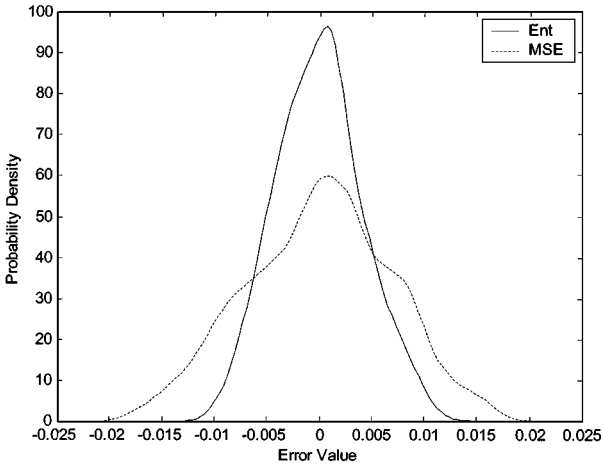

Fig. 2. Probability densities for errors of MSE (dotted) and entropy (solid) trained TDNNs.

All TDNNs are trained with a segment of 200 samples. For each network, 1000 randomly chosen initial weights were tried (Monte Carlo approach) in order to avoid local solutions. The training algorithm utilized backpropagation using a variable step-size gradient algorithm [22] for efficiency. The stopping criteria was experimentally determined and consisted of 100 iterations for MSE-TDNNs and 30 iterations for MEE-TDNNs. At the end of the mentioned Monte Carlo training, the best set of weights (that yield minimum cost function values for entropy and MSE) obtained by each of the criteria are taken and checked for further improvement by employing a very small constant step size to make sure convergence of each criterion to its global minimum is achieved. The kernel size used to estimate the entropy was experimentally set at $\sigma = 0.01$ after a preliminary analysis of the final error dynamic range (however, this value is not critical to the final performance if set properly in a wide range given by $[0.001, 0.1]$ for this example). The general rule of thumb we use is to select the kernel size so that on average, ten samples are covered by each kernel function. Finally, after training, the bias weights of the output PEs are adjusted to yield zero error mean over the training set.

The trained networks are tested on an independently generated test data set of length 10 000 since the goal is to learn the chaotic attractor rather than the specific trajectory. In Fig. 2, the error pdf estimates for the two TDNNs with six hidden PEs (which is the best solution among all MSE-TDNNs) are shown. Clearly, the error distribution of the MEE-TDNN is more concentrated around zero. Fig. 3 depicts the estimated probability densities of the actual Mackey–Glass data set and the predictions by the two TDNNs of interest. It is clear from these plots that the density of the predictions made by the MEE- TDNN is much closer to that of the test data compared with the distribution of the predictions made by the MSE–TDNN. This is expected due to the minimization of Csiszar distance when the

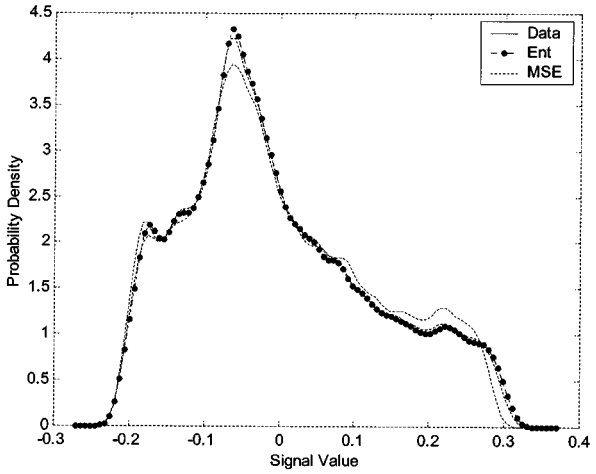

Fig. 3. Probability densities of MG30 test series and its predictions by entropy-trained and MSE-trained TDNNs; desired (solid), entropy-trained (dots), MSE-trained (dotted).

entropy criterion is used in training. We know that variance minimization can produce locally large errors, and this is clearly seen in Fig. 3. Entropy training produces a more uniform match between the two estimated amplitude densities, resembling an L_1-norm fit. Thus, we are lead to believe that the entropy criterion in this case is better than MSE in extracting more information about the pdf of the desired signal distribution. Note, however, that around the signal amplitude of approximately -0.18, both MSE-trained and entropy-trained networks fail to approximate accurately the distribution of the desired signal. Since both criteria fail to model this portion of the distribution accurately, we hypothesize that the two most likely possible causes for this behavior are the insufficiency of the network topology to capture the specific dynamics involved in that region of the attractor and/or the inadequate representation of that dynamic behavior in the trajectories due to the short training sequence used. Finally, in Fig. 4, we present the central moments of the desired and predicted signals for all sizes of TDNNs over the test data. All TDNNs with number of hidden neurons ranging from three to ten are trained, starting from the 1000 initial conditions using both MSE and MEE criteria. Clearly, for all cases, entropy achieves a better fit to the distribution of the desired signal compared with MSE.

As a second case study, we investigate the performance of the MEE criterion in identification of a nonlinear system, whose dynamic equations are given in (23). Once again, a TDNN will be used. The sought-after mapping in this case is from the delayed values of the input and the output of the unknown system to its current output. The training set can be represented as follows:

$$\left\{\left(u_k \quad \cdots \quad u_{k-L} \quad y_{k-1} \quad \cdots \quad y_{k-M}\right)^T, y_k\right\} \qquad k = M, \ldots, M+N-1. \quad (22)$$

Specifically, the number of input samples is chosen to be seven $(L = 6)$, and the number of output samples is chosen to be six $(M = 6)$. A TDNN with seven hidden PEs is assumed, following the suggestion in [23]. The nonlinear system that is utilized has the state dynamics and the output mapping provided in (23).

$$\begin{aligned} x_{1,k+1} &= \left(\frac{x_{1,k}}{1+x_{1,k}^2}+1\right)\cdot \sin x_{2,k} \\ x_{2,k+1} &= x_{2,k}\cdot\cos x_{2,k} + \exp\left(-\frac{x_{1,k}^2+x_{2,k}^2}{8}\right) \\ &\quad + \frac{u_k^3}{1+u_k^2+0.5\cdot\cos\left(x_{1,k}+x_{2,k}\right)} \\ y_k &= \frac{x_{1,k}}{1+0.5\cdot\sin x_{2,k}} + \frac{x_{2,k}}{1+0.5\cdot\sin x_{1,k}}. \end{aligned} \quad (23)$$

The training set consists of $N = 100$ input–output pairs, and the TDNN is trained starting from 50 different initial conditions using both MEE and MSE criteria. The output bias is then set to yield zero error mean over the training set. The performances of the optimal weights obtained from the two criteria are then compared on an independently generated 10 000-sample test set. Fig. 5 shows the error pdfs for the two criteria on this test set. The MSE of training set errors are 0.0676 and 0.0587, and the information potentials for the same samples are 0.996 and 0.989 for MEE and MSE trained weights, respectively. As expected, the training MSE is lower for MSE-trained TDNN, and information potential is higher for entropy-trained TDNN.

This case study demonstrates nicely the basic difference between the entropy and variance minimization. Entropy prefers a larger and more concentrated peak centered at zero error with a number of small peaks at larger error values, whereas the variance (MSE) prefers a wide-distributed error on a smaller range. In fact, this can be deduced by the following reasoning. Suppose it is possible to obtain many error distributions with the same variance. Since the Gaussian has the maximum entropy among fixed variance densities, this error distribution would be the least desirable for the entropy criterion. In addition, the uniform would be another undesirable distribution for the error. The entropy would prefer rather spiky distributions, i.e., a number of δ-like concentrated spikes having the same variance. This is observed in Fig. 5. A comparison of the desired output signal and the actual MLP outputs using entropy-trained weights and MSE-trained weights is depicted in Fig. 6 to illustrate the statistical matching property of MEE. Clearly, the entropy-trained TDNN approximates the pdf of the desired output much better around the most probable regions of the domain when compared with the MSE-trained TDNN.

VI. Conclusions

In this paper, an information-theoretic supervised learning criterion for adaptive systems, namely, minimum error entropy (MEE), has been proposed. It is shown that minimizing Renyi's error entropy is equivalent to minimizing a Csiszar distance between the joint densities of system input–output and the desired input–output pairs. It was also proved that the Csiszar distance measure reduces to the well-known Kullback–Leibler divergence when Shannon's entropy is utilized. Furthermore, it is known that there is equivalence between entropy manipulation and maximum likelihood solutions [3], [15].

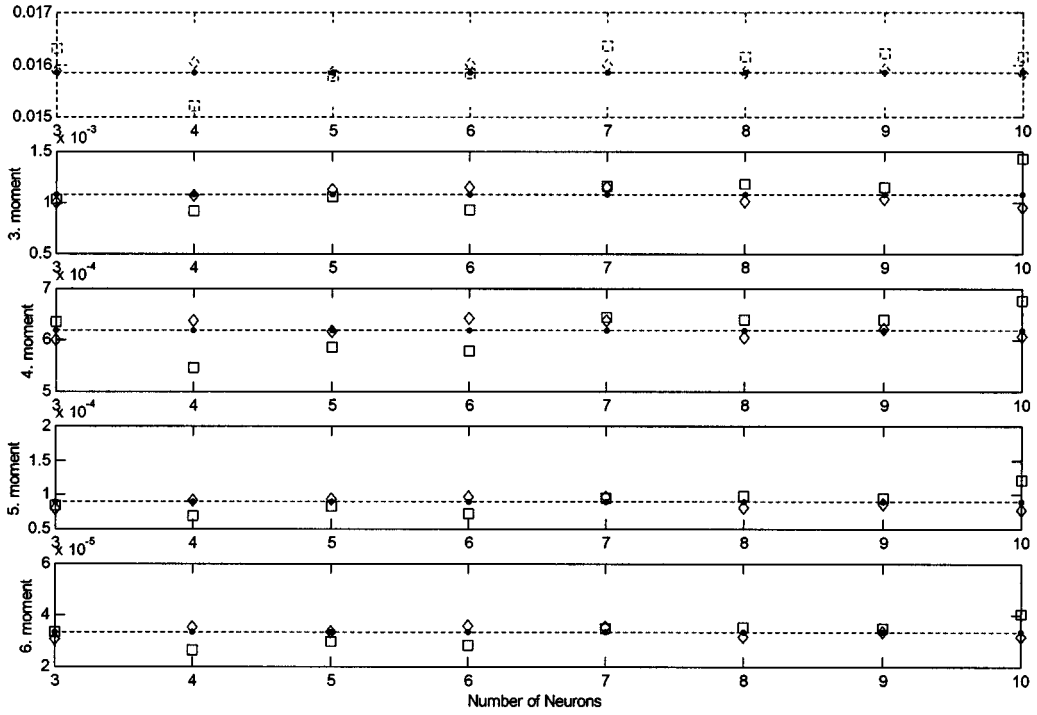

Fig. 4. Desired (dashed lines) central moments and those of the predicted series up to order 6 for entropy-trained (diamonds) and MSE-trained (squares) TDNNs versus number of hidden neurons.

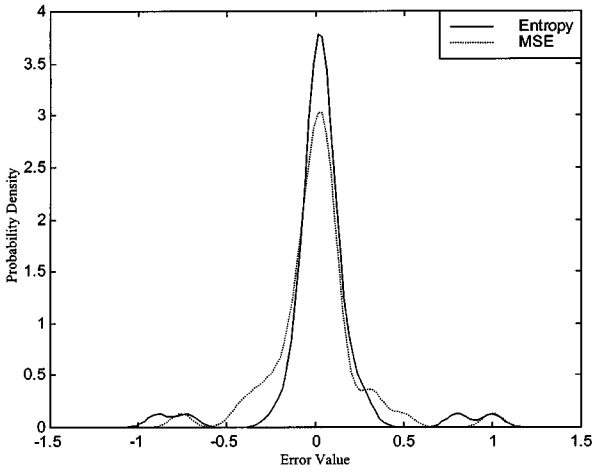

Fig. 5. Distribution of errors of entropy-trained (solid) and MSE-trained (dotted) MLP's.

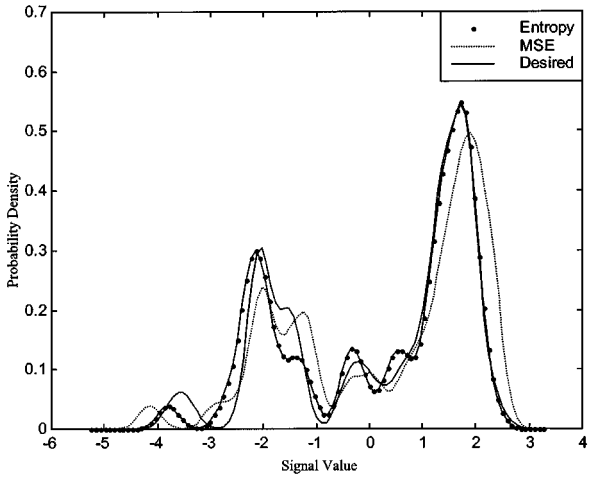

Fig. 6. Distributions of desired (solid) and TDNN outputs. Entropy (dots). MSE (dotted).

A nonparametric entropy estimator based on Parzen windows and Gaussian kernels is presented, and it is proved that the global minimum of the entropy estimator is the same as the global minimum of the actual entropy. This enables us to use the nonparametric entropy estimator for entropy minimization and opens the door to the use of entropy minimization for any type of supervised training applications, such as system identification and time series prediction. Renyi's quadratic entropy is preferred in practice due to the computational efficiency of its nonparametric estimator. It becomes possible to define the *information potential*, which then facilitates an analogy between the presented approach for the information potential computation and the sample test statistics based on kernels [24]. The latter is the basis of the recently proposed Diks test for the equivalence of multidimensional vector distributions [25]. However, unlike Diks' work, we present here an information-theoretic framework, and we use the information potential to adapt directly the parameters of a nonlinear adaptive system.

Two case studies are also presented. The first one investigated the performance of the MEE criterion on the adaptation of time-delay neural networks of various sizes for the short-term prediction of Mackey–Glass chaotic time series. The second one was a nonlinear system identification problem using TDNNs. The optimal solutions obtained by MSE and MEE criteria were compared in terms of the error distributions and their performance in matching the probability density function of the desired output. These analyses demonstrated that the error samples of the entropy-trained TDNNs exhibit a more concentrated density func-

tion, and the distribution of the produced outputs are also closer to that of the desired signals in both case studies. These results indicate the potential advantage of entropy training versus MSE training. Especially since the entropy criterion allows a wider range for the error in favor of a more concentrated distribution for small error values, it can be useful in disregarding outliers in the desired signal if they do not fit the underlying density well. Consequently, this study prompts a new line of research that appears to be very promising by offering a feasible alternative to MSE, which is the workhorse of supervised training.

Further work is needed to study the properties of the entropy cost function for optimization and to find more robust ways to set the kernel size for the information potential estimation. The effect of noise in information-theoretic cost functions must also be addressed. The issue of scalability of the information potential method with the size of the space will also be studied. Finally, we have been applying the information potential method in many other problems (from blind source separation to pattern recognition) with very interesting and promising results [12], [17]. These also provide an encouraging indication for the need to further study this entropy estimator and its applications in related problems.

REFERENCES

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

[2] S. Haykin, *Neural Networks: A Comprehensive Foundation*. New York: MacMillan, 1994.

[3] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed. New York: McGraw-Hill, 1991.

[4] X. Feng, K. Loparo, and Y. Fang, "Optimal state estimation for stochastic systems: An information theoretic approach," *IEEE Trans. Automat. Contr.*, vol. 42, pp. 771–785, June 1997.

[5] J. Casals, C. Jutten, and A. Taleb, "Source separation techniques applied to linear prediction," in *Proc. ICA Conf.*, Helsinki, Finland, 2000, pp. 193–204.

[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.

[7] J. Fisher, A. Ihler, and P. Viola, "Learning informative statistics: A nonparameteric approach," in *Proc NIPS*, vol. 12, 2000, pp. 900–906.

[8] C. E. Shannon, "A mathematical theory of communications," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.

[9] S. Haykin and J. C. Principe, "Dynamic modeling with neural networks," *IEEE Signal Processing Mag.*, vol. 15, p. 66, Mar. 1998.

[10] S. Amari, *Differential—Geometrical Methods in Statistics*. Berlin, Germany: Springer-Verlag, 1985.

[11] E. Parzen, "On estimation of a probability density function and mode," in *Time Series Analysis Papers*. San Francisco, CA: Holden-Day, 1967.

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

[13] A. Renyi, *Probability Theory*. New York: Elsevier, 1970.

[14] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.

[15] S. Kullback, *Information Theory and Statistics*. New York: Dover, 1968.

[16] A. Renyi, "Some fundamental questions of information theory," in *Selected Papers of Alfred Renyi*. Budapest, Hungary: Akademia Kiado, 1976, vol. 2, pp. 526–552.

[17] J. C. Principe, D. Xu, Q. Zhao, and J. Fisher, "Learning from examples with information theoretic criteria," VLSI Signal Process. Syst., 2000, to be published.

[18] A. Waibel, T. Hananzawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time delay neural networks," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 328–339, Feb. 1989.

[19] D. Rumelhart, G. Hinton, and R. Williams, "Learning internal representations by error backpropagation," *Nature*, vol. 323, pp. 533–536, 1986.

[20] D. Kaplan and L. Glass, *Understanding Nonlinear Dynamics*. New York: Springer-Verlag, 1995.

[21] J. M. Kuo, "Nonlinear dynamic modeling with artificial neural networks," Ph.D. dissertation, Univ. Florida, Gainesville, 1993.

[22] D. G. Luenberger, *Linear and Nonlinear Programming*. Reading, MA: Addison-Wesley, 1973.

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

[24] N. Anderson, P. Hall, and D. Titterington, "Two sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimators," *J. Multivariate Anal.*, vol. 50, pp. 41–54, 1994.

[25] C. Diks, J. Houwelingen, F. Takens, and J. deGoede, "Detecting differences between delay vector distributions," *Phys. Rev. E*, vol. 53, pp. 2169–2176, 1996.

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 at 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, 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 (M'83–SM'90–F'00) is Professor of electrical and computer engineering and biomedical engineering at 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 Laboratory has been studying signal and pattern recognition principles based on information-theoretic criteria (entropy and mutual information). He has more than 70 publications in refereed journals, ten book chapters, and 160 conference papers. He has directed 35 Ph.D. dissertations and 45 Master's theses. He recently wrote an interactive electronic book entitled *Neural and Adaptive Systems: Fundamentals Through Simulation* (New York: Wiley).

Dr. Principe is the Chair of the Technical Committee on Neural Networks 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.