Entropy Minimization for Supervised Digital Communications Channel Equalization

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

***Abstract*—This paper investigates the application of error-entropy minimization algorithms to digital communications channel equalization. The pdf of the error between the training sequence and the output of the equalizer is estimated using the Parzen windowing method with a Gaussian kernel, and then, the Renyi's quadratic entropy is minimized using a gradient descent algorithm. By estimating the Renyis entropy over a short sliding window, an online training algorithm is also introduced. Moreover, for a linear equalizer, an orthogonality condition for the minimum entropy solution that leads to an alternative fixed-point iterative minimization method is derived.**

The performance of linear and nonlinear equalizers trained with entropy and mean square error (MSE) is compared. As expected, the results of training a linear equalizer are very similar for both criteria since, even if the input noise is non-Gaussian, the output filtered noise tends to be Gaussian. On the other hand, for nonlinear channels and using a multilayer perceptron (MLP) as the equalizer, differences between both criteria appear. Specifically, it is shown that the additional information used by the entropy criterion yields a faster convergence in comparison with the MSE.

***Index Terms*—Adaptive equalization, equalizers, neural networks, nonlinear equalization, Renyi's entropy.**

I. Introduction

MODERN digital communications systems demand high-speed efficient transmission over bandwidth-limited channels, which distort the signal causing intersymbol interference (ISI). In addition, the digital signal is subject to other impairments such as noise, nonlinear distortion, time-variant channels, etc. At the receiver, an equalizer is used to mitigate these effects and restore the transmitted symbols.

An equalizer is characterized by its structure, the optimization criterion, the adaptive algorithm used to train it, and the availability or not of a training sequence (supervised or blind equalization, respectively). In particular, this paper is focused on supervised equalization, using linear or nonlinear structures trained with backpropagation-like algorithms to minimize a cost function based on the entropy of the error.

Manuscript received December 22, 2000; revised January 28, 2002. This work was supported in part by the National Science Foundation under Grant ECS-9900394, the Spanish Education Ministry under Grant PR2000-0183, and the European Community and CICYT through Project 1FD97-1863-C02-01. The associate editor coordinating the review of this paper and approving it for publication was Dr. Naofal M. W. Al-Dhahir.

I. Santamaría is with the Communications Engineering Department (DICOM), University of Cantabria, Santander, Spain (e-mail: nacho@gtas.dicom.unican.es).

D. Erdogmus and J. C. Principe are with the Computational Neuroengineering Laboratory (CNEL), University of Florida, Gainesville, FL 32611 USA (e-mail: deniz@cnel.ufl.edu; principe@cnel.ufl.edu).

Publisher Item Identifier S 1053-587X(02)03284-1.

The most popular equalizer is the linear transversal equalizer (LTE), trained to minimize the MSE between its output and the desired sequence by means of the LMS or the RLS algorithm [1]–[3]. An interesting and powerful alternative to the LTE is the decision feedback equalizer (DFE). In this case, the past decisions are included in the equalization process to improve the margin against noise and the performance, mainly in channels with deep nulls. Although the DFE structure is nonlinear, it can only cope with very moderate nonlinear distortion. Moreover, it suffers from error propagation due to the feedback part.

When the channel is nonlinear, a nonlinear equalizer is required to eliminate the ISI. Traditionally, Volterra filters were applied for this purpose [4]. However, they require a large number of parameters, and their training is computationally involved. More recently, artificial neural networks have been proven to be attractive alternatives for nonlinear equalization. In particular, the multilayer perceptron (MLP) [5], [6] and the radial basis function (RBF) [7], [8] have demonstrated good performance in several nonlinear equalization problems.

Regarding the cost function or optimization criteria, most of the conventional linear or nonlinear equalizers are trained using an MSE criterion. The error is defined as the difference between the desired training sequence and the output of the equalizer. Some recent attempts have been made to explore other cost functions for this problem such as the structural risk minimization (SRM) principle [9]–[11].

In this paper, we consider an alternative criterion that consists of minimizing the entropy of the error sequence. Information-theoretic criteria have been widely applied to blind equalization and deconvolution. In particular, it is well known that the Shannon entropy provides a measure that can be used to push the probability density function of the equalizer's output away from that of a Gaussian, thus deconvolving the output [13]. Typically, the lack of efficient estimators for Shannon's entropy was circumvented by minimizing a cost function related to entropy but easier to estimate (such as the normalized kurtosis) [13]–[15]. Other approaches aim at forcing a given probability density at the output of the equalizer. As a measure of distance between densities, the Kullback–Leibler distance or relative entropy is used in [16]–[18]. In particular, in [16], it was proven that unlike the MSE, the relative entropy is a well-formed cost function in the sense of Wittner and Denker [19], showing a better capability to track time-varying channels.

Although some of these ideas can be also applied when a training sequence is available, the use of information theoretic criteria for nonblind equalization is not so common. As a new contribution in this line, in this paper, we consider equalization techniques that seek a direct minimization of the error en-

tropy. Entropy is a function of the pdf of the error, and therefore, by minimizing it, we are using much more information than by minimizing just its variance (i.e., the MSE). Hopefully, this extra information in the error sequence can provide some advantage either in terms of performance or in terms of requiring shorter training sequences. Instead of using the widely known Shannon's entropy, which is difficult to estimate directly from samples without a model, we use the quadratic Renyi's entropy [21], [22]. This alternative entropy measure can be more easily estimated from data, and it has shown improved performance over the MSE criterion in other problems [23], [24].

The rest of this paper is organized as follows. In Section II, the problem of linear and nonlinear equalization is briefly introduced. In Section III, we present quadratic Renyi's entropy and describe its estimation from samples using a Parzen windowing method. In Section IV, we analyze the solutions provided by the new criterion and show the equivalence between the MSE and the minimum entropy solutions for an LTE in a small error case. The training of linear equalizers by means of a new fixed-point iterative algorithm, as well as batch and online training algorithms for nonlinear equalizers, are considered in Section V. Some simulation results are presented in Section VI, comparing the performance of linear and nonlinear equalizers trained with MSE and entropy. Finally, Section VII presents the conclusions and points out some lines for further research.

II. Linear and Nonlinear Equalizers

The received signal at the input of the equalizer can be expressed as

$$x_i = \sum_{k=0}^{n_c} h_k s_{i-k} + e_i \tag{1}$$

where the transmitted symbol sequence s_i is assumed to be an equiprobable binary sequence, $\{+1, -1\}$, h_i are the channel coefficients (we assume here an FIR channel), and the measurement noise e_i can be modeled as zero-mean Gaussian with variance σ_n^2.

The equalization problem reduces to correctly classify the transmitted symbols based on the observation vector. For instance, an LTE estimates the value of a transmitted symbol as

$$\hat{s}_{i-d} = \text{sgn}(y_i) = \text{sgn}(\mathbf{w}^T \mathbf{x}_i) \tag{2}$$

where

$y_i = \mathbf{w}^T \mathbf{x}_i$	output of the equalizer;
$\mathbf{w} = [w_0, \ldots, w_{m-1}]^T$	equalizer coefficients;
$\mathbf{x}_i = [x_i, \ldots, x_{i-m+1}]^T$	vector of observations;
d	equalizer delay.

The LTE implements a linear decision border; however, it is well known that even if the channel is linear, the optimal (Bayesian) decision border is nonlinear [7]. As long as the noise increases, the nonlinear character of the optimal border becomes more important.

On the other hand, when the channel is nonlinear, in order to eliminate the ISI, it is necessary to consider a nonlinear equalizer. In this case, the output of the equalizer is given by

$$y_i = g(\mathbf{W}, \mathbf{x}_i) \tag{3}$$

where $g(\cdot)$ is a nonlinear mapping, and $\mathbf{W}$ denotes the parameters of the equalizer. After the mapping, a hard threshold is still needed in order to decide the symbols; in this way, (3) can be viewed as a mapping from the input space to an output space, where the classification becomes possible and hopefully easier.

In this paper, an MLP is considered to be the nonlinear structure to perform that mapping. Assuming an MLP with one hidden layer with n neurons, (3) reduces to

$$y_i = \mathbf{w}_2^T \tanh(\mathbf{W}_1 \mathbf{x}_i + \mathbf{b}_1) + b_2 \tag{4}$$

where

$\mathbf{W}_1$	$n \times m$ matrix connecting the input layer with the hidden layer;
$\mathbf{b}_1$	$n \times 1$ vector of biases for the hidden neurons;
$\mathbf{w}_2$	$n \times 1$ vector of weights connecting the hidden layer to the output neuron;
b_2	bias for the output neuron.

The training of this structure to minimize the MSE criterion can be done using the backpropagation algorithm [12].

III. Quadratic Error-Entropy for Equalization

Conventional equalizers are trained to minimize the MSE between the desired output and the output of the equalizer. For a linear equalizer, for instance, this criterion yields the following cost function

$$J(\mathbf{w}) = \sum_{i=1}^{N} (s_{i-d} - \mathbf{w}^T \mathbf{x}_i)^2. \tag{5}$$

The MSE criterion, which uses only second-order statistics, is adequate under the assumptions of linearity and Gaussianity. When the noise is not Gaussian or the distorting channel (and, therefore, the required equalizer) is not linear, a criterion considering all the higher order statistics of the error signal would be more appropriate.

The entropy of the error sequence is a quantity that takes into account its probability distribution function (pdf). Then, by minimizing the entropy instead of the MSE, all higher order moments (not only the second one) are minimized. Other arguments supporting the minimization of the error entropy as a useful criterion in equalization will be given later.

The most known definition for entropy (Shannon's entropy) is, in general, hard to estimate and minimize since it involves the integral of the logarithm of the pdf. Recently, some efficient procedures for estimating Shannon's entropy have been proposed (see [20] and references therein).

On the other hand, Shannon's entropy is not the only useful definition of entropy. Other alternative definitions have been proposed; in particular, Renyi's entropy with parameter α [21] is defined as

$$H(e) = \frac{1}{1-\alpha} \log\left(\int_{-\infty}^{\infty} p_e(\xi)^\alpha \, d\xi \right) \tag{6}$$

where $p_e(\xi)$ is the pdf for the error. In this paper, only quadratic Renyi's entropy ($\alpha = 2$) will be considered since it can be easily estimated from data. In this case, (6) reduces to

$$H(e) = -\log\left(\int_{-\infty}^{\infty} p_e(\xi)^2 \, d\xi \right). \tag{7}$$

Recently, a nonparametric estimator for quadratic Renyi's entropy has been developed [22]. It allows the maximization or minimization of the entropy criteria using simple gradient descent techniques. Furthermore, this technique has been successfully applied to short-term prediction of chaotic time series [23] and blind source separation [24].

As it is shown in [22], given a set of N error samples $e_i, i = 1, \ldots, N$, the error pdf can be estimated by the Parzen window method using a Gaussian kernel of variance σ^2

$$p_e(\xi) = \frac{1}{N}\sum_{i=1}^{N} G(\xi - e_i, \sigma^2). \tag{8}$$

Then, substituting (8) into (7), the entropy is given by

$$H(e) = -\log V(e) \tag{9}$$

where

$$V(e) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G(e_i - e_j, 2\sigma^2) \tag{10}$$

is called the information potential. In (10), $G(e, 2\sigma^2)$ denotes the Gaussian kernel with variance $2\sigma^2$. To simplify the notation, in the sequel, the kernel size will be omitted.

From (9), it is clear that minimizing the error entropy reduces to maximize the information potential $V(e)$. On the other hand, as it is shown in [23], when Parzen windowing with Gaussian kernels is used to estimate Renyi's entropy, the minima of (7) correspond to points where the error is constant over the data set. Therefore, it can be concluded that the global minimum of Renyi's entropy is preserved as a minimum of the estimated entropy. This allows the use of gradient descent techniques for minimization.

Obviously, the minimum of (7) is obtained when $p_e(\xi) = \delta(\xi - K)$ for any K, i.e., when the error is a constant signal. On the other hand, the pdf of the equalizer's output $y = s + e$, given the training sequence s (considered as deterministic), is

$$p(y \mid s) = p_e(y - s). \tag{11}$$

In this way, the output of the equalizer converges to $y = s + K$ with probability one. In practice, a finite training sequence is used; in this case, by maximizing (10), each error sample interacts with all other errors, pushing the solution toward a constant error signal. The constant term K has no influence since it can be easily eliminated using an additional bias term in the equalizer.

Finally, it has been also proven in [23] that minimizing the error entropy is equivalent to maximizing the mutual information between the output of the equalizer and the training sequence. These arguments support the use of an error-entropy minimization criterion in equalization problems.

IV. Analysis of the Optimal Solutions

To gain some insight into the error entropy criterion, it is interesting to write the information potential as a function of the output of the equalizer (linear or nonlinear) y_i. Considering a binary signal, the set of outputs for the training set y_i can be partitioned according to the desired output $s_{i-d} = \pm 1$ into the following two subsets:

$$R^{(\pm 1)} = \{y_i, s_{i-d} = \pm 1\}. \tag{12}$$

Now, taking into account that

$$e_i - e_j = s_{d-i} - s_{d-j} + y_j - y_i \tag{13}$$

it is easy to show that

$$V(y) = \sum_{i,j}\sum_{\in R^{(+1)}} G(y_i - y_j) + \sum_{i,j}\sum_{\in R^{(-1)}} G(y_i - y_j) + 2\sum_{i\in R^{(+1)}}\sum_{j\in R^{(-1)}} G(y_i - y_j - 2). \tag{14}$$

The first two terms in (14) are maximized when $y_i = y_j$ for $i, j \in R^{(+1)}$ and $i, j \in R^{(-1)}$, respectively. This process can be viewed as minimizing the "intraclass" output entropy, that is, the equalizer tries to cluster the outputs in delta functions for inputs belonging to $R^{(+1)}$ and $R^{(-1)}$.

On the other hand, the third term is maximized when $y_i - y_j = 2$, for $i \in R^{(+1)}$ and $j \in R^{(-1)}$; therefore, it tries to separate the outputs for each class. As a comparison, the MSE criterion in terms of the equalizer outputs is given by

$$\text{MSE}(y) = \sum_{i\in R^{(+1)}} (1 - y_i)^2 + \sum_{i\in R^{(-1)}} (1 + y_i)^2. \tag{15}$$

It can be concluded that the entropy criterion forces additional constraints by exploiting the relationship between each pair of equalizer outputs. Moreover, although the MSE criterion forces a constant modulus for the output signal as in (15), the entropy criterion makes that *the difference* between the outputs for the two classes has a constant value.

This analysis suggests that when the training sequence is short, as occurs, for instance, in packet data transmission (in GSM, for instance, only 21 training bits per packet are used), the additional constraints used in (14) can be very helpful in achieving a faster convergence of the algorithm. This result will be confirmed later by some simulation examples. The analysis performed in this section can be easily extended to M-ary and complex modulations.

Considering now a linear equalizer $y_i = \mathbf{w}^T\mathbf{x}_i$, it is interesting to compare the optimal solutions obtained for an LTE with MSE and entropy criteria. The derivatives of (10) with respect to the linear equalizer coefficients are given by

$$\frac{\partial V(e)}{\partial \mathbf{w}} = \frac{1}{2N^2\sigma^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G(e_i - e_j)(\mathbf{x}_i - \mathbf{x}_j)(e_i - e_j). \tag{16}$$

Equating (16) to zero and taking into account that the Gaussian kernel is a symmetric and positive function that fulfills

$$\sum_{i,j} e_i\mathbf{x}_j G(e_i - e_j) = \sum_{i,j} e_j\mathbf{x}_i G(e_i - e_j) \tag{17}$$

we finally find that the following equation holds at any minima of the cost function

$$\sum_{i=1}^{N} e_i \left(\sum_{j=1}^{N} (\mathbf{x}_i - \mathbf{x}_j) G(e_i - e_j) \right) = 0 \tag{18}$$

which is the entropy counterpart of the orthogonality condition obtained for the MSE criterion

$$\sum_{i=1}^{N} e_i \mathbf{x}_i = 0. \tag{19}$$

The difference is that for entropy, the error must be orthogonal to a nonlinear function of the input, which is a function of the differences between inputs and errors for each pair of training data.

An interesting remark is that if the equalizer can equalize the channel in such a way that the final error sequence is small in comparison with $2\sigma^2$, i.e., the kernel size of the Gaussian window used in (18), then, we have a solution for which

$$G(e_i - e_j) \simeq \frac{1}{\sqrt{4\pi}\sigma}; \quad \forall i, j \tag{20}$$

and (18) reduces to the following condition:

$$\sum_{i=1}^{N} e_i \mathbf{x}_i = N \bar{e} \bar{\mathbf{x}} \tag{21}$$

where $\bar{e}$ and $\bar{\mathbf{x}}$ denote the mean of the error sequence and the mean of the input signal, respectively. This result shows that although, in general, the MSE and entropy solutions are different, for an LTE, if the error is small so that (20) holds, then as long as either the error sequence or the equalizer input is zero mean, the minimum entropy solution is equal to the MSE solution.

V. Training the Equalizers

In this section, we describe batch and online training algorithms for linear and nonlinear equalizers. Basically, gradient descent techniques are used to minimize the entropy cost function over the whole set of training data (batch) or over a short sliding window (online). In addition, for a linear equalizer, a fixed-point method, which converges much faster than a batch gradient descent, is also proposed.

A. Batch Training

Given a set of N input–output training samples $(\mathbf{x}_i, s_{i-d}), i = 1, \ldots, N$, the corresponding set of errors is obtained as $e_i = s_{i-d} - \mathbf{w}^T \mathbf{x}_i$ for a linear equalizer or $e_i = s_{i-d} - g(\mathbf{W}, \mathbf{x}_i)$ for an MLP. The gradient to be used for the maximization of the information potential (10) is given by

$$\frac{\partial V(e)}{\partial w} = \frac{1}{2N^2\sigma^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (e_i - e_j) G(e_i - e_j) \left(\frac{\partial y_i}{\partial w} - \frac{\partial y_j}{\partial w} \right) \tag{22}$$

where w denotes any parameter of the equalizer. For an MLP, it can be any weight or bias parameter, and $\partial y_i / \partial w$ can be computed as in standard backpropagation [12].

For an MLP and taking into account that the entropy of the error does not depend on the mean, the algorithm may converge to a non zero-mean error. This can be easily corrected by choosing the bias of the output neuron to give a zero-mean error. That is, after convergence of the information potential, the bias in (4) is selected as

$$b_2 = \frac{1}{N} \sum_{i=1}^{N} e_i. \tag{23}$$

To accelerate the converge of the algorithm, a variable learning step size is used. As long as the information potential increases, the step size for the next iteration is selected as $\mu = 1.05\,\mu$, whereas if the information potential decreases, the learning rate is also decreased as $\mu = 0.7\,\mu$, and the previous parameters of the equalizer are kept. In the simulations, this optimization technique will be denoted as batch gradient descent (BGD).

B. Online Training

The BGD algorithm described in the previous section can be readily extended to an online (sample-by-sample) adaptive algorithm, which is more sounding in an equalization context and from a practical point of view. In comparison with a batch procedure, an online version allows tracking of the time-varying channels, prevents the introduction of long delays into the decision, and enables a low-cost implementation.

At each time instant n, a window of size N_w is constructed using the current and the past $N_w - 1$ error samples, which have been previously stored in memory $\mathbf{e}(n) = (e_n, \ldots, e_{n-N_w+1})$. Then, using this window, one gradient iteration according to (22) is carried out. This stochastic gradient descent approach can be considered to be the entropy counterpart of the LMS. Specifically, the instantaneous error estimate used in the LMS is replaced here by an estimate of the Renyi's entropy obtained from a short sliding window. In both cases, with each new incoming sample, a single step is taken. As will be shown later, this online approach, which will be denoted as stochastic gradient descent (SGD), is rather effective, even when a short window is used.

Finally, the proposed algorithm can be summarized in the following steps.

1) Initialize the parameters of the algorithm: the stepsize μ, the kernel size σ^2, the window size N_w, the equalizer delay d, and the equalizer parameters (with random values).
2) Initialize the window of errors

$$e_i = 0, \quad \text{for } i = 1, \ldots, N_w.$$

3) For $n = 1, 2, \ldots$

3.1) Update the window with the current error

$$e_i = e_{i+1}, \quad \text{for } i = 1, \ldots N_w - 1$$
$$e_{N_w} = s_{n-d} - y_n.$$

3.2) Update the coefficients of the equalizer as

$$w_{n+1} = w_n + \mu \frac{\partial V(e)}{\partial w}$$

where $\partial V(e)/\partial w$ is given by (22).

3.3) Fix the bias of the output neuron for zero-mean error

$$b_2 = \frac{1}{N_w} \sum_{i=1}^{N_w} e_i.$$

4) End.

C. Fixed-Point Algorithm for Linear Equalizers

For a linear equalizer, the output error entropy could be minimized by applying the BGD or the SGD algorithms described in the previous sections. However, the orthogonality condition (18) suggests an alternative algorithm to the BGD approach, which has proven much faster to converge.

Considering again that the whole set of N input–output training samples is available (batch), the orthogonality condition (18) can be rewritten as

$$\sum_{i=1}^{N} (s_i - \mathbf{w}^T \mathbf{x}_i)\mathbf{t}_i = 0 \tag{24}$$

where $\mathbf{t}_i$ is an $m \times 1$ vector (m being the filter length) given by

$$\mathbf{t}_i = \sum_{j=1, j\neq i}^{N} (\mathbf{x}_i - \mathbf{x}_j) G(e_i - e_j). \tag{25}$$

Equation (24) can be written in matrix form as

$$\mathbf{TXw} = \mathbf{Ts} \tag{26}$$

where

$\mathbf{T}$ $m \times N$ matrix with columns $\mathbf{t}_i$ given by (25);

$\mathbf{X}$ $N \times m$ matrix with rows given by $\mathbf{x}_i^T = [x_i, \ldots, x_{i-m+1}]$;

$\mathbf{s}$ $N \times 1$ vector with the desired responses.

To find a solution to the set of nonlinear (26), we can use the following iterative procedure.

1) Initialize $\mathbf{w}_0$.

2) For $k = 1, 2 \ldots$

2.1) Obtain the output error and estimate matrix $\mathbf{T}_k$ as (25).

2.2) Compute the new solution as

$$\mathbf{w}_k = (\mathbf{T}_k \mathbf{X})^{-1} \mathbf{T}_k \mathbf{s}. \tag{27}$$

3) End.

This well-known technique to find the root of an equation is denoted in the mathematical literature as the method of iteration or the method of successive approximations [25]. Conditions for the convergence of this procedure are also given in [25]. Basically, by rewriting the nonlinear transformation (27) as

$$\begin{aligned} w_1 &= \phi_1(\mathbf{w}) \\ &\vdots \quad \vdots \\ w_m &= \phi_m(\mathbf{w}) \end{aligned} \tag{28}$$

the process of successive approximations converges to a fixed point if

$$\sum_{j=1}^{m} \left| \frac{\partial \phi_j}{\partial w_i} \right| < 1, \quad \forall i = 1, \ldots, m. \tag{29}$$

Although we were not able to prove (29) in our particular case, all the simulations carried out for a number of different situations converged to the correct solution much faster than the BGD technique.

This technique requires the inversion of the $m \times m$ matrix $\mathbf{TX}$. Note, however, that as long as N is sufficiently large ($N > m$), the probability for $\mathbf{TX}$ to be rank-deficient (and hence singular) is very low, i.e., matrix $\mathbf{TX}$ will have rank m for most of the cases.

Regarding the computational cost of this procedure in comparison with the BGD approach, note that since, typically, $N \gg m$, most of the computational cost for both algorithms comes from the evaluation of the N^2 Gaussian kernels in (16) and (25). The additional cost of the fixed-point method, due to matrix inversion, is of order $O(m^3)$, and therefore, it only becomes noticeable when the equalizer length is large. We can conclude that the computational cost per iteration of the fixed point and BGD approaches is roughly the same. However, as it will be shown in the next section, the fixed-point algorithm requires very few iterations to converge, whereas the BGD approach usually takes much longer to converge. Therefore, for a linear equalizer the fixed-point method should be preferred from a computational point of view.

VI. Simulation Results

In this section, we present some simulation results considering linear and nonlinear equalizers trained with MSE and minimum error-entropy criteria. In all the examples, the Gaussian kernel size was $\sigma^2 = 1$.

A. Linear Equalizers

In the first example, a BPSK signal is sent through the channel $H(z) = 0.87 + 0.44z^{-1} + 0.23z^{-2}$, and then, white Gaussian noise for SNR $= 10$ dB was added. The aim of this example is twofold: first, to compare the convergence rate and computational cost of the BGD and the fixed-point algorithms and, second, to validate the analysis carried out in Section IV. We train an LTE with $m = 7$ coefficients and an equalization delay of $d = 3$ using MSE and entropy. The equalizer taps were initialized to zero, and it was trained using a known sequence of 100 symbols. We use a least squares (batch) method for the MSE and the fixed-point and BGD algorithms for the entropy. The BGD used an adaptive stepsize to speed up the convergence, the initial learning rate is $\mu = 0.05$.

Fig. 1 shows the normalized information potential (10) versus the number of iterations using the BGD algorithm and the proposed fixed-point technique. Three iterations of the fixed-point procedure provides a solution that requires more than 30 gradient iterations. On the other hand, Table I compares the computational cost per iteration for the fixed-point and the BGD algorithms using different sizes of the training set $N = 25, 50, 100$, and 200: The cost per iteration is roughly the same.

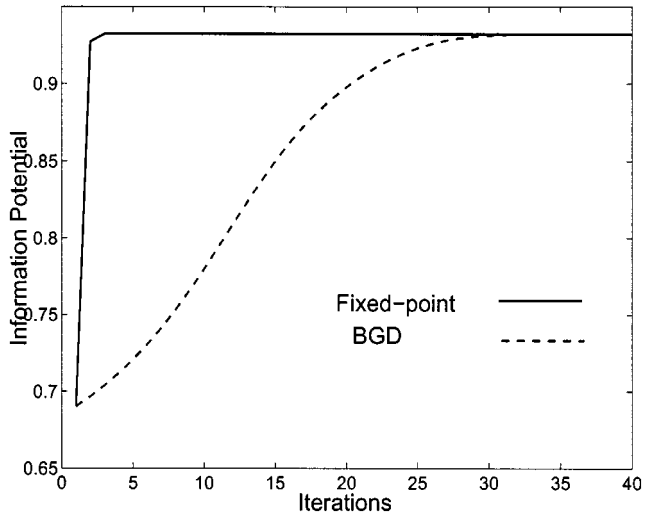

Fig. 1. Information potential $V(e)$ versus iterations for the fixed-point and the batch gradient descent (BGD) algorithms.

TABLE I
COMPARISON OF THE COMPUTATIONAL COST (FLOPS) PER ITERATION FOR THE FIXED-POINT AND THE BGD ALGORITHMS. TRAINING SET SIZE $= N$, EQUALIZER COEFFICIENTS $m = 7$

	$N = 25$	$N = 50$	$N = 100$	$N = 200$
Fixed-point	$1.7\ 10^4$	$5.6\ 10^4$	$2.5\ 10^5$	$8.4\ 10^5$
BGD	$1.4\ 10^4$	$5.4\ 10^4$	$2.3\ 10^5$	$8.0\ 10^5$

Now, we compare the MSE and entropy criteria; for this example, the error at the output of the equalizer is small in comparison with $2\sigma^2$, and therefore, for the minimum entropy solution, (20) and the orthogonality condition (21) hold. Fig. 2 shows the error sequence at the output of the equalizer for the MSE and entropy solutions, respectively. As it was proven by the analysis carried out in Section IV, although the cost functions are very different, the final solution is exactly the same.

In the second example, we consider a situation for which (20) does not hold, and then, the minimum entropy and MSE solutions are different. Now the channel is $H(z) = 0.35 + 0.8z^{-1} + z^{-2} + 0.8z^{-3}$, the LTE has $m = 9$ coefficients, and the equalizer delay is $d = 5$.

As in the previous example, the equalizer was initialized to zero, and then it was trained with 100 known symbols using a least squares method for the MSE and the fixed-point algorithm for the entropy criterion. Finally, the BER was evaluated by counting errors after transmitting 10^5 or 10^6 symbols, depending on the SNR. We run 50 independent simulations. Fig. 3 shows the BER curves for this example. It can be seen that although the solutions are different, from a classification point of view, both criteria provide practically the same results.

It can be concluded that if the structure of the equalizer is linear and the noise is Gaussian, then minimizing the variance is equivalent, from a practical standpoint, to minimizing the error entropy. In fact, if the error is Gaussian, as long as its variance decreases, its entropy decreases as well.

Extensive simulations using other noise distributions (uniform, impulsive, zero-mean Rayleigh, ...) confirm the results

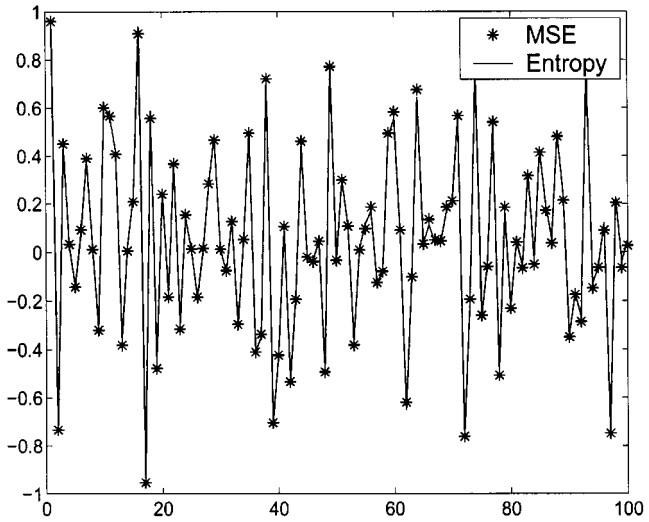

Fig. 2. Error sequence at the output of an LTE for the MSE solution ("$*$"), and the minimum entropy solution (solid line). Channel $H(z) = 0.87 + 0.44z^{-1} + 0.23z^{-2}, m = 7, d = 3$, SNR $= 10$ dB, kernel size $\sigma^2 = 1$, and 100 training samples (batch training).

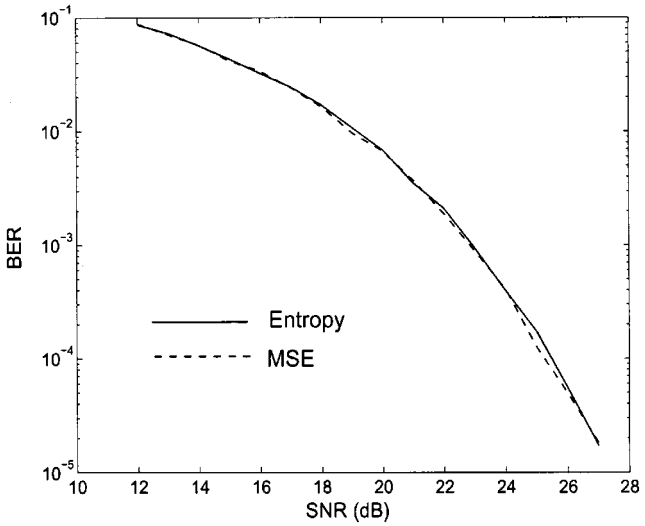

Fig. 3. BER comparison for the minimum entropy (solid) and MSE (dashed) linear equalizers. Channel $H(z) = 0.35 + 0.8z^{-1} + z^{-2} + 0.8z^{-3}, m = 9, d = 5$ and 100 training samples (batch training).

presented for the Gaussian noise. The explanation to this lies in the fact that the noise filtered after the linear equalizer and added to the residual ISI tends to be practically Gaussian, regardless of the input noise distribution.

B. Nonlinear Equalizers

In this example, we consider a nonlinear channel composed of a linear channel followed by a memoryless nonlinearity. The transmitted binary sequence is passed through a linear channel, and the output of the channel is added to some static nonlinear function. Such a nonlinear model can be encountered in digital satellite communications [26] and as nonlinear channel models for digital magnetic recording [27], [28]. The linear channel considered is $H(z) = 0.3482 + 0.8704z^{-1} + 0.3482z^{-2}$, and the nonlinear function applied is $z = x + 0.2x^2 - 0.1x^3$, where x is the linear channel output. Finally, white Gaussian noise for SNR $= 16$ dB was added. The nonlinear equalizer structure is an MLP with

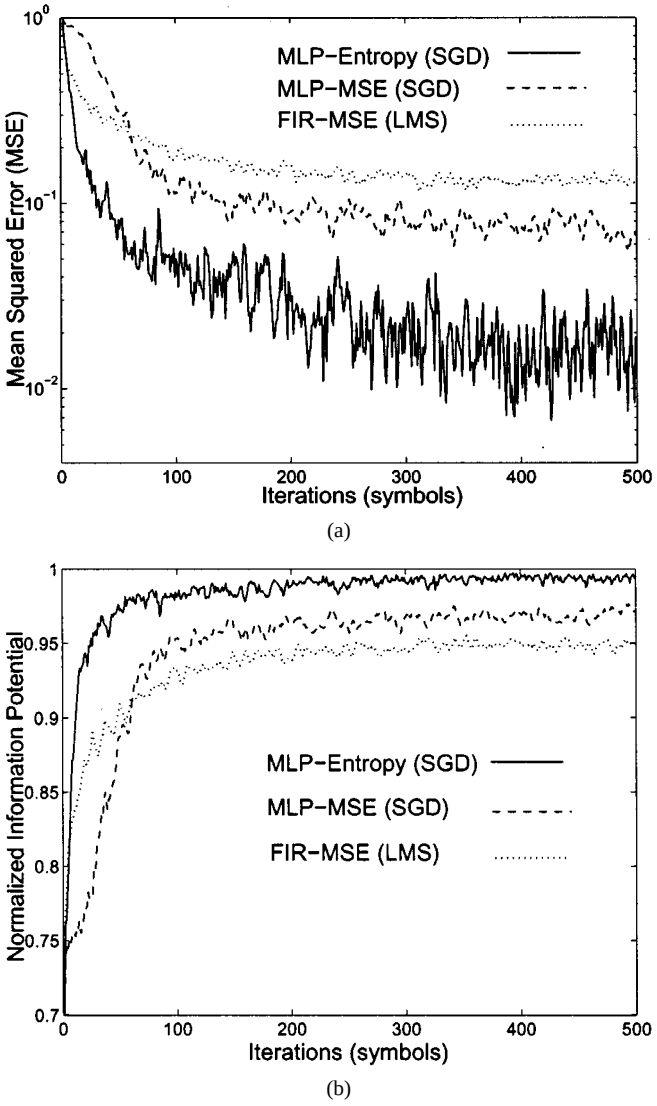

Fig. 4. Convergence characteristics of online adaptive algorithms for a linear equalizer with $m = 7$ coefficients (dotted line) and an MLP(7-4-1) trained with MSE (dashed line) and minimum entropy criteria (solid line) using a window size of $N_w = 5$ samples. (a) MSE. (b) Normalized information potential.

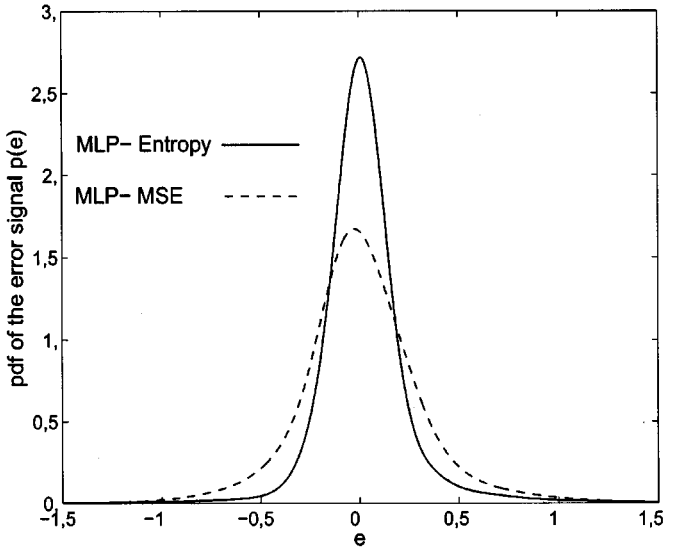

Fig. 5. Estimated pdf of the error at the equalizer's output for the MSE (dashed line) and the minimum entroy (solid line).

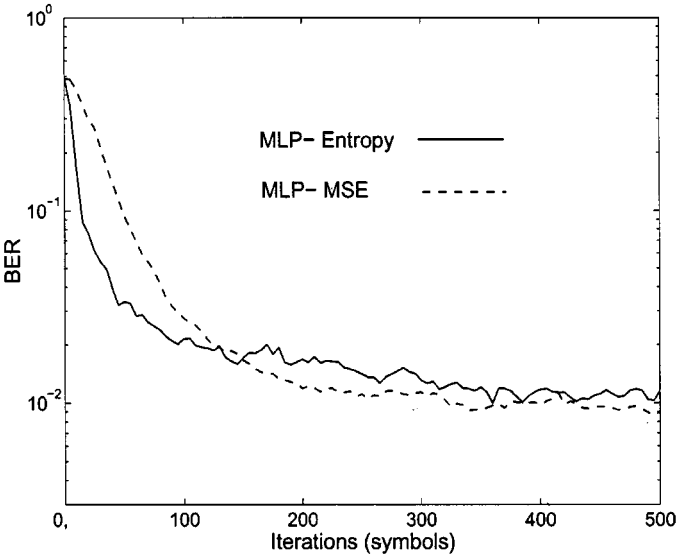

Fig. 6. Convergence of the BER with the number of iterations for the MSE (dashed line) and the minimum entropy (solid line).

seven neurons in the input layer and three neurons in the hidden layer [MLP(7, 3, 1)], and the equalization delay is $d = 4$.

For this example, the online adaptive algorithm (SGD) described in Section V-B is applied. A short sliding window of just $N_w = 5$ error samples is used to minimize the MSE or the error entropy using a backpropagation-like algorithm. At each iteration, a single step was taken. For both criteria, a fixed stepsize $\mu = 0.01$ was used, which is the largest stepsize for which the algorithms converged in all trials. The results provided by a linear (FIR) equalizer with $m = 7$ coefficients and trained with an MSE criterion were also obtained. In this case, a conventional LMS algorithm with a fixed stepsize $\mu = 0.05$ was used.

Fig. 4 shows the convergence of the normalized information potential and the MSE evaluated over the sliding window for the three algorithms. These results were obtained by averaging 100 independent simulations. Each method is characterized by an structure (MLP/FIR), an optimization criterion (Entropy/MSE), and the adaptive algorithm used. It can be seen that the MLP trained with the entropy criterion achieves the best results, and it also provides the fastest convergence, whereas the linear equalizer is not able to remove the nonlinear part of the ISI. It is interesting to point out that even though the entropy criterion does not directly minimize the MSE, surprisingly, it achieves a lower MSE than a direct minimization of this criterion. The explanation of this fact is that, in comparison to the MSE, the entropy criterion yields a spiky error with more abrupt changes (higher kurtosis) but with a lower MSE.

In Fig. 5, the pdf of the error sequence (estimated using Parzen windowing method with $\sigma^2 = 0.01$) for the MLP trained with both criteria are depicted. As it was discussed in Section III, it can be seen that the minimization of the error entropy tries to push the pdf of the error closer to a delta function.

The convergence of the BER with the number of training symbols is shown in Fig. 6; the entropy criterion achieves a very fast convergence, but the final BER is slightly worse than

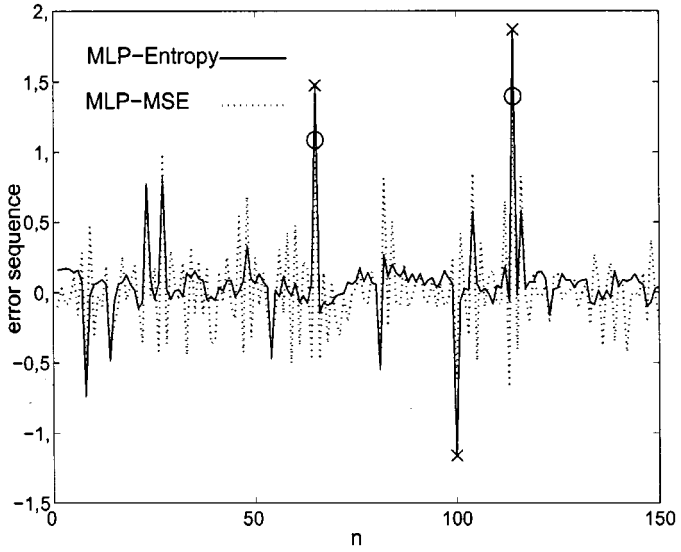

Fig. 7. Error sequence at the output of nonlinear equalizer for the MSE solution (dashed line), and the minimum entropy solution (solid line). Circles and crosses indicate classification errors for MSE and entropy, respectively.

the BER obtained by the MSE criterion. This apparent discrepancy between MSE and BER can be explained by examining the error sequence obtained after the convergence of the equalizer in Fig. 7. As we previously pointed out, the entropy criterion yields a spiky signal with a low MSE and low entropy (remember Fig. 4). However, only the error samples larger than one (in absolute value) cause a classification error. For instance, in Fig. 7, the entropy sequence generates three errors (depicted with crosses), whereas the MSE sequence only causes two errors (depicted with circles). To switch from an initial entropy criterion during the first training symbols (thus achieving a fast convergence) to an MSE criterion during the last symbols of the training sequence seems to be the best strategy to exploit the benefits of the entropy criterion. As a final comment, the results of Figs. 6 and 7 suggest that for this particular application, a criterion that directly minimizes the number of errors (instead of the MSE or entropy) might be a better choice. Recent works on the use of support vector machines [9]–[11] or a direct minimization of the BER [29] point in this direction.

VII. Conclusion

In this paper, we have considered the use of a new optimization criterion based on the Renyi's error entropy for supervised equalization of digital communications channels. For an LTE, it has been shown that this criterion tends to minimize the intraclass entropy, whereas, at the same time, it tries to separate the classes. Moreover, it was also shown that if the LTE structure allows an effective equalization in the sense that the output error is small, the minimum entropy solution becomes the MSE solution. In general, for linear filtering problems and regardless of the noise distribution, the entropy and MSE criteria provide similar results. The differences appear when a nonlinear equalizer is considered; in this case, some simulation results indicate that the minimization of all higher order moments of the error through entropy yields a faster convergence in comparison with the MSE. This suggests that this new criterion can be useful for packet-based data transmission when the training sequences can be short.

This preliminary work using the entropy criterion in nonblind equalization problems has provided interesting results. However, it is our belief that the main interest of this technique will appear when applying this criterion to blind equalization problems. This is, undoubtedly, an interesting line for further research.

Acknowledgment

The authors would like to thank the referees for providing us with valuable comments and insightful suggestions that have greatly improved this paper.

References

[1] J. Proakis, *Digital Communications*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
[2] B. Mulgrew and C. F. N. Cowan, *Adaptive Filters and Equalisers*. Boston, MA: Kluwer, 1988.
[3] S. U. H. Qureshi, "Adaptive equalization," *Proc. IEEE*, vol. 73, pp. 1349–1387, Sept. 1985.
[4] S. Benedetto and E. Biglieri, "Nonlinear equalization of digital satellite channels," *IEEE J. Select. Areas Commun.*, vol. SAC-1, pp. 57–62, Jan. 1983.
[5] S. Chen, G. J. Gibson, C. F. N. Cowan, and P. M. Grant, "Adaptive equalization of finite nonlinear channels using multilayer perceptrons," *Signal Process.*, vol. 10, pp. 107–119, 1990.
[6] P. R. Chang and B. C. Wang, "Adaptive decision feedback equalization for digital channels using multilayer neural networks," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 316–324, Feb. 1995.
[7] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," *IEEE Trans. Neural Networks*, vol. 4, pp. 570–578, July 1993.
[8] J. Cid-Sueiro, A. Artés-Rodríguez, and A. R. Figueiras-Vidal, "Recurrent radial basis function networks for optimal symbol-by-symbol equalization," *Signal Process.*, vol. 40, pp. 53–63, 1994.
[9] S. Chen and C. J. Harris, "Design of the optimal separating hyperplane for the decision feedback equalizer using support vector machines," in *Proc. ICASSP*, Istanbul, Turkey, 2000.
[10] I. Santamaría, C. Pantaleón, and J. C. Principe, "Minimizing BER in DFE's with the Adatron algorithm," in *IEEE Neural Netw. Signal Process. Workshop*, Falmouth, MA, Sept. 2001, pp. 423–432.
[11] D. J. Sebald and J. A. Bucklew, "Support vector machines for nonlinear equalization," *IEEE Trans. Signal Processing*, vol. 48, pp. 3217–3226, Nov. 2000.
[12] S. Haykin, *Neural Networks, A Comprehensive Foundation*. New York: Macmillan, 1998.
[13] D. Donoho, "On minimum entropy deconvolution," in *Applied Time Series Analysis II*. New York: Academic, 1981, pp. 565–609.
[14] A. T. Walden, "Non-Gaussian reflectivity, entropy and deconvolution," *Geophys.*, vol. 50, pp. 2862–2888, Dec. 1985.
[15] O. Shalvi and E. Weinstein, "Super-exponential methods for blind deconvolution," *IEEE Trans. Inform. Theory*, vol. 39, pp. 504–519, Mar. 1993.
[16] T. Adali, X. Liu, and M. K. Sönmez, "Conditional distribution learning with neural networks and its application to channel equalization," *IEEE Trans. Signal Processing*, vol. 45, pp. 1051–1064, Apr. 1997.
[17] J. Sala-Alvarez and G. Vazquez-Grau, "Statistical reference criteria for adaptive signal processing in digital communications," *IEEE Trans. Signal Processing*, vol. 45, pp. 14–31, Jan. 1997.
[18] J. Montalvao, B. Dorizzi, and J. C. Mota, "Identification des coefficients du model MA du canal par sous-estimation de la densité de probabilité du signal recu," in *Proc. 17th Colloque GRETSI*, France, Sept. 1999, pp. 1105–1108.
[19] B. S. Wittner and J. S. Denker, "Strategies for teaching layered networks classification tasks," in *Neural Inform. Proc. Syst.*, Denver, CO, 1989, pp. 850–859.
[20] J. F. Bercher and C. Vignat, "Estimating the entropy of a signal with applications," *IEEE Trans. Signal Processing*, vol. 48, pp. 1687–1694, June 2000.

[21] A. Renyi, "Some fundamental questions of information theory," in *Selected Papers of Alfred Renyi*. Budapest, Hungary: Akademia Kiado, 1976, vol. 2, pp. 526–552.
[22] J. C. Principe, D. Xu, Q. Zhao, and J. W. Fisher III, "Learning from examples with information theoretic criteria," *J. VLSI Signal Process.*, vol. 26, pp. 61–77, 2000.
[23] D. Erdogmus and J. C. Principe, "An entropy-minimization algorithm for short-term prediction of chaotic time series," *IEEE Trans. Signal Processing*, 2000, submitted for publication.
[24] K. E. Hild, 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.
[25] B. P. Demidovich and I. A. Maron, *Computational Mathematics*. Moscow, USSR: MIR, 1987.
[26] G. Kechriotis, E. Zervas, and E. S. Manolakos, "Using recurrent neural network for adaptive communication channel equalization," *IEEE Trans. Neural Networks*, vol. 5, pp. 267–278, Mar. 1994.
[27] N. P. Sands and J. M. Cioffi, "Nonlinear channel models for digital magnetic recording," *IEEE Trans. Magn.*, vol. 29, pp. 3996–3998, Nov. 1993.
[28] ——, "An improved detector for channels with nonlinear intersymbol interference," in *Proc. Int. Conf. Commun.*, vol. 2, 1994, pp. 1226–1230.
[29] S. Chen, B. Mulgrew, and L. Hanzo, "Adaptive least error rate algorithm for neural network classifiers," in *Proc. IEEE Neural Netw. Signal Process. Workshop*, Falmouth, MA, Sept. 2001, pp. 223–232.

Ignacio Santamaría was born in Vitoria, Spain, in 1967. He received the Telecommunication Engineer and Ph.D. degrees in electrical engineering from the Polytechnic University of Madrid, Madrid, Spain in 1991 and 1995, respectively.

In 1992, he joined the Departamento de Ingeniería de Comunicaciones, Universidad de Cantabria, Santander, Spain, where he is currently an Associate Professor. In 2000, he was a Visiting Professor with the Computational NeuroEngineering Laboratory (CNEL), University of Florida, Gainesville. His current research interests include nonlinear modeling techniques and adaptive systems and their application to digital communication systems.

Deniz Erdogmus (S'99) received the B.S. degree in electrical and electronics engineering and the B.S. degree in mathematics in 1997 and the M.S. degree in electrical and electronics engineering, with emphasis on systems and control, in 1999, all from the Middle East Technical University, Ankara, Turkey. Since 1999, he has been pursuing the Ph.D. degree at the Electrical and Computer Engineering Department, University of Florida, Gainesville, under the supervision of J. C. Principe.

From 1997 to 1999, he worked as a Research Engineer at the Defense Industries Research and Development Institute (SAGE) under The Scientific and Technical Research Council of Turkey (TUBITAK). His current research interests include information theory and its applications to adaptive systems and adaptive systems for signal processing, communications, and control.

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

Jose C. Principe (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 (ANN's) 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 70 publications in refereed journals, ten book chapters, and 160 conference papers. He directed 35 Ph.D. dissertations and 45 M.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.