# Conditional Distribution Learning with Neural Networks and Its Application to Channel Equalization

Tülay Adalı, *Member, IEEE*, Xiao Liu, and M. Kemal Sönmez

*Abstract*— **We present a conditional distribution learning formulation for real-time signal processing with neural networks based on a recent extension of maximum likelihood theory—partial likelihood (PL) estimation—which allows for i) dependent observations and ii) sequential processing. For a general neural network conditional distribution model, we establish a fundamental information-theoretic connection, the equivalence of maximum PL estimation, and accumulated relative entropy (ARE) minimization, and obtain large sample properties of PL for the general case of dependent observations. As an example, the binary case with the sigmoidal perceptron as the probability model is presented. It is shown that the single and multilayer perceptron (MLP) models satisfy conditions for the equivalence of the two cost functions: ARE and negative log partial likelihood. The practical issue of their gradient descent minimization is then studied within the well-formed cost functions framework. It is shown that these are well-formed cost functions for networks without hidden units; hence, their gradient descent minimization is guaranteed to converge to a solution if one exists on such networks. The formulation is applied to adaptive channel equalization, and simulation results are presented to show the ability of the least relative entropy equalizer to realize complex decision boundaries and to recover during training from convergence at the wrong extreme in cases where the mean square error-based MLP equalizer cannot.**

## I. Introduction

RECENTLY, neural networks have been applied to a wide range of signal processing applications because of the growing need for alternatives to the linear structure that is typically assumed. Nonlinear signal processing with neural networks has provided significant performance improvement in a variety of applications (for a recent collection of these applications see, e.g., [16] and [20]) when the underlying process involves nonlinearities and/or the signal-to-noise ratio (SNR) is poor. Neural networks are also deemed attractive because they lend themselves to low-complexity, low-power, analog hardware implementations that are likely to be widely available in the near future; this is a feature that is particularly important in portable applications such as equalization for cellular communications, where there are strict restrictions on power and space. A certain view of learning in neural networks, regarding the mechanism as a statistical estimation of a parameterized probability model, has proven quite useful [21], [22], [30]. This view offers advantages both in understanding properties of neural network learning paradigms and in developing new approaches for learning (e.g., [5], [19], [30]).

Statistical parameter estimation theory has as its fundamental support maximum likelihood (ML) estimation that provides estimators with nice large sample optimality properties and invariant with respect to functions of the parameters. However, ML theory is traditionally developed for independent observations, and a majority of signal processing applications require processing of dependent observations. In this paper, we introduce a conditional distribution learning framework for real-time signal processing with neural networks based on partial likelihood (PL) theory [11], [33]. Obtained as a partial factorization of the full likelihood, PL also possesses nice large sample properties of ML, and more importantly, it can easily be characterized for dependent data and sequential processing. Hence, it overcomes the difficulties with other extensions of ML for dependent data, such as conditional likelihood, which, for easy specification, requires that the auxiliary information be known for the whole period (i.e., including future observations) [27]. Some of the other problems with other factorizations of likelihood for dependent data are initial state specification requirements (e.g., when using Markovian representations for the data [1]) and the problems when dealing with missing data. Therefore, PL provides us with a particularly suitable formulation for real-time signal processing, which most of the time requires on-line processing of dependent observations.

We introduce a general neural network conditional probability model, and for this model, we establish a key information-theoretic connection, namely, the equivalence of maximum PL estimation and accumulated relative entropy (ARE) minimization. Hence, distribution learning using relative entropy between the true and estimated probability mass functions can be achieved by maximum PL estimation, which does not require that the true conditionals be known (which, in general,

Manuscript received September 2, 1995; revised October 23, 1996. The work of T. Adalı and X. Liu was supported in part by Engineering Foundation Grant RI-A-94-08 and the University of Maryland Graduate School Research Initiation Award. The associate editor coordinating the review of this paper and approving it for publication was Dr. Shigeru Katagiri.

T. Adalı is with the Department of Computer Science and Electrical Engineering, University of Maryland-Baltimore County, Baltimore, MD 21250 USA (e-mail: adali@engr.umbc.edu).

X. Liu is with the Department of Computer Science and Electrical Engineering, University of Maryland-Baltimore County, Baltimore, MD 21250 USA.

M. K. Sönmez was with the Institute for Systems Research, University of Maryland, College Park, MD 20742 USA. He is now with the Speech Technology and Research Laboratory, SRI International, Menlo Park, CA 94025 USA.

Publisher Item Identifier S 1053-587X(97)02567-1.

are not available). This result can be regarded as the extension of the ML and minimum ARE equivalence for independent and identically distributed (i.i.d.) data [30] to the general case of dependent observations. While providing the theoretical foundation for statistical analysis of maximum PL estimation, this connection can also be used to derive a new class of real-time signal processing algorithms based on information-theoretic alternating projections [5], [12]. We establish the consistency and asymptotic normality of the PL estimator for the general neural network model under some regularity conditions.

In the second part of the paper, we consider a perceptron probability model for binary distribution learning and its application to adaptive channel equalization. We show that the multilayer perceptron (MLP) probability model satisfies the conditions for ARE-PL equivalence. For the MLP model, we derive the least relative entropy (LRE) algorithm by gradient optimization and show that it possesses nice dynamical properties that can be beneficial to the channel equalization problem. Particularly, it is shown that, for networks without hidden units, LRE can always recover from convergence at the wrong extreme, whereas the mean-squared-error (MSE)-based gradient descent learning on these networks cannot. This property of the algorithm is discussed within the *well-formed* cost functions framework of Wittner and Denker [32], stating that gradient descent learning on such cost functions is always guaranteed to find a solution if one exists. In a gradient descent dynamics framework, it has been shown that for networks without hidden units, MSE cost function is not a *well-formed* cost function [32]; therefore, finding a solution cannot always be guaranteed, and the MSE-based learning algorithms may not be able to recover from convergence at a wrong extreme. We present simulation results for the adaptive equalization application that we consider, which demonstrate that the advantages of using a well-formed cost function also carry over to networks with hidden units.

Adaptive channel equalization is an area in which neural networks have been found to be quite valuable as they provide a unique computational structure to compensate for the nonlinear distortion and to achieve reliable communication at increasing data rates (e.g., [8]–[10], [15], [18], [23], [25], [28]). The neural network equalizers such as those employing MLP's [10], [15], [25], radial basis functions [8], [9], [18], [28], and recurrent networks [18], [23] have been shown to equalize nonlinear channels successfully in situations where linear equalizers may fail. They also outperform other nonlinear approaches in which nonlinearity is incorporated in the way the transmitted sequence is recovered, such as decision feedback equalizers and maximum likelihood sequence detectors [14], [23]. The main difference between the approaches taken in the neural network equalizers mentioned above and ours is in the cost function that is being utilized. The algorithms employed in the above-mentioned equalizers use the traditional MSE as the cost to be minimized. There are, however, some important drawbacks of MSE minimization related to the dynamics of gradient descent learning with nonlinear structures resulting from the fact that MSE is not a *well-formed* cost function, as mentioned above. We present simulation studies to show that the learning algorithm we derive within the PL framework using a MLP model (the least relative entropy (LRE) algorithm) is just as effective as the MSE-based backpropagation on a MLP for compensating for various channel distortions. However, when there is an abrupt change in the channel characteristics during training, LRE algorithm can very rapidly adapt to the new operating condition while the MSE-based MLP equalizer either responds very slowly or cannot adapt to the change at all if the algorithm has fully converged.

The organization of the paper is as follows: In Section II, we present the formulation for distribution learning with neural networks and introduce partial likelihood. We define ARE and establish the equivalence of ARE minimization to maximum PL estimation as well as the large sample properties of the maximum PL estimator under some regularity conditions. In Section III, we consider the binary case with sigmoidal perceptron as the conditional probability mass function model and show that the conditions for equivalence of ARE minimization to maximum PL estimation are satisfied for the single-layer perceptron model. We then extend this equivalence to MLP's. Finally, we discuss the dynamics of PL (or ARE) learning for the binary equalization problem and present simulation results to demonstrate the capability of our MLP equalizer to realize complex decision boundaries and to recover from convergence at the wrong extreme. Section IV presents the conclusions.

## II. Distribution Learning by Neural Networks

### A. Problem Formulation

The distribution learning problem is posed as follows: Given a time series $\{x_n\}, n = 0, 1, 2, \cdots$ that takes values from a finite alphabet $\mathcal{S} = \{a_0, a_1, \cdots, a_M\}$, and its time-dependent covariates $\{y_n\}$, estimate the probability that $x_n$ takes a value from the given finite alphabet $\mathcal{S}$. We define $\mathcal{F}_n$ as the $\sigma$ field generated by events of the form $[x_{n-1}, \cdots, x_1, x_0], [y_n, y_{n-1}, \cdots, y_1, y_0]$

$$\mathcal{F}_n = \sigma\{1, [x_{n-1}, \cdots, x_1, x_0], [y_n, \cdots, y_1, y_0]\}.$$

The $\sigma$ field $\mathcal{F}_n$ represents all that is known to the observer at time instant $n$; hence, $\mathcal{F}_{n-1} \subset \mathcal{F}_n$. In its definition, 1 is included to account for the constant bias term in the probability model. If we consider the example of channel equalization, $x_n$ is the transmitted sequence of symbols, and the covariates of $x_n, y_n$ are the noise-corrupted outputs of the channel. If multiple covariates are available for the problem or can be computed/defined by using some *a priori* information, they can also be incorporated into the formulation.

Hence, our goal is to estimate the conditional probabilities

$$p(x_n = a_i | \mathcal{F}_n) \quad \forall a_i \in \mathcal{S}. \tag{1}$$

The total distribution information—the conditional probability mass function (pmf) $p_\theta(x_n|\mathcal{F}_n)$—can be used in a variety of ways depending on the application. The simplest decision rule would be to select $a_k$ such that $k = \arg\max_i p(x_n = a_i|\mathcal{F}_n)$ as the most likely symbol, given the history $\mathcal{F}_n$.

The conditional probability mass function $p_\theta(x_n|\mathcal{F}_n)$ is parameterized by a neural network as follows:

$$p_\theta(x_n|\mathcal{F}_n) = f(x_n, g(\boldsymbol{y}_n, \boldsymbol{\theta})). \tag{2}$$

Here, $\boldsymbol{\theta}$ is the vector of network weights $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is a compact parameter set, and $\boldsymbol{y}_n = [y_n, y_{n-1}, \cdots, y_{n-N+1}]$. The term $g(\boldsymbol{y}_n, \boldsymbol{\theta})$ is the output of the neural network, $f(\cdot)$ and $g(\cdot)$ are continuous and differentiable functions, and $f(\cdot)$ is chosen such that

$$\sum_{a_j \in \mathcal{S}} p_\theta(x_n = a_j|\mathcal{F}_n) = 1. \tag{3}$$

Note that since $\mathcal{F}_n$ includes the entire history, $p_\theta(x_n|\mathcal{F}_n)$ can have a *recurrent* structure as well, and formulation for the unsupervised case is obtained if $[x_{n-1}, \cdots, x_1, x_0]$ is excluded from the definition of $\mathcal{F}_n$ and $x_n$ from the definition of $p_\theta(x_n|\mathcal{F}_n)$ in (2).

Once we select a particular neural network structure for the conditional probability model, we can compute the parameters of the network such that the likelihood is maximized. The theory of maximum likelihood (ML) is historically developed for independent observations, but the independence condition is very restrictive for almost all practical applications. In a typical real-time communications application such as channel equalization, the observations will be highly correlated because of the memory of the channel (intersymbol interference). Partial likelihood is a relatively recent extension of ML estimation introduced by Cox [11], and its theoretical justification, including large sample properties, is given in [33]. It processes data *as they become available*, requiring only the information present at a given time. Again, it is important to note that while doing this processing, we do not make any assumptions on the *dependence* structure of the data for characterizing the PL. This is critical for many practical applications since the observations most often have strong dependencies in time. Partial likelihood is defined as [27], [33]

$$\mathcal{L}_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} p_\theta(x_i|\mathcal{F}_i). \tag{4}$$

If, at each point in time, the history generated by the inputs $x_i$ and its covariates, i.e., $\mathcal{F}_i$, is known for the whole sequence of observations $i = 1, \cdots, n$, then the PL we defined in (4) becomes equal to the conditional likelihood [27]. Given the selected probability model, which is (2) in our case, PL estimate $\hat{\boldsymbol{\theta}}$ will approximate the true distribution such that PL is maximized in the sequence of observations $\mathcal{F}_1, \mathcal{F}_2, \cdots, \mathcal{F}_n$. In the next section, we provide an information-theoretic connection for PL estimation, which shows that the distribution for $\boldsymbol{\theta}$ also minimizes the information-theoretic distance with the true distribution.

### *B. Information-Theoretic View and Large Sample Properties of PL*

The relative entropy (RE), or the Kullback–Leibler distance [27], is a fundamental information-theoretic measure of how accurate the estimated conditional pmf $p_\theta(x_n|\mathcal{F}_n)$ is an approximation of the true conditional pmf $p(x_n|\mathcal{F}_n)$ and is given by

$$D_n(p||p_\theta) = \sum_{a_j \in \mathcal{S}} p(x_n = a_j|\mathcal{F}_n) \ln \frac{p(x_n = a_j|\mathcal{F}_n)}{p_\theta(x_n = a_j|\mathcal{F}_n)}. \tag{5}$$

It is an information-theoretic measure of the average surprise experienced when we believe that the conditional pmf is $p_\theta(x_n|\mathcal{F}_n)$ and are informed that it actually is $p(x_n|\mathcal{F}_n)$. In addition, it is important to note that $D_n(p||p_\theta)$ is nonnegative and is equal to zero only when $p = p_\theta$. We can define the ARE as $\mathcal{I}_n(\boldsymbol{\theta}) = \Sigma_{i=1}^{n} D_i(p||p_\theta)$ and rewrite it as

$$\mathcal{I}_n(\boldsymbol{\theta}) = \sum_{k=1}^{n} \sum_{a_j \in \mathcal{S}} p_{\theta_0}(x_k = a_j|\mathcal{F}_k) \ln \frac{p_{\theta_0}(x_k = a_j|\mathcal{F}_k)}{p_\theta(x_k = a_j|\mathcal{F}_k)} \tag{6}$$

if we assume that $\boldsymbol{\theta}_0$ is the weight vector for which $f(\cdot)$ defined in (2) achieves the true conditional pmf. The optimal network parameters $\boldsymbol{\theta}_0$ thus have the fundamental information-theoretic interpretation that they minimize the Kullback–Leibler information given a certain network architecture [31]. Thus, viewing learning as related to Kullback–Leibler information minimization in this way implies that learning is a ML statistical estimation procedure for independent observations [30]. As we have noted before, its extension to dependent data requires that the auxiliary information be known in full throughout the period of observation [27], which is a condition most often not satisfied in real-time applications. Partial likelihood, on the other hand, allows for sequential inference from the available data.

It is relatively easy to demonstrate the equivalence of ML estimation to ARE minimization when the observations are i.i.d. [29], [30]. As one of the key results of our development, we establish the equivalence of PL estimation to ARE minimization for the neural network model defined in (2) for the general case of dependent observations. In [34], we exploit this relationship to derive a new learning rule based on information-theoretic alternating projections for maximum PL estimation of the parameters of the conditional pmf model of (2).

Note that the ARE defined in (6) can also be written as $\mathcal{I}_n(\boldsymbol{\theta}) = \sum_{k=1}^{n} i_k(\boldsymbol{\theta})$, where

$$i_k(\boldsymbol{\theta}) = E\{r_k(\boldsymbol{\theta})|\mathcal{F}_k\} \tag{7}$$

and

$$r_k(\boldsymbol{\theta}) = \ln \frac{p_{\theta_0}(x_k|\mathcal{F}_k)}{p_\theta(x_k|\mathcal{F}_k)}. \tag{8}$$

To state the equivalence result for the parameterized pmf model of (2), we also define $\mathcal{J}_n(\boldsymbol{\theta}) = \Sigma_{k=1}^{n} j_k(\boldsymbol{\theta})$, where

$$j_k(\boldsymbol{\theta}) = \mathrm{Var}\,\{r_k(\boldsymbol{\theta})|\mathcal{F}_k\}. \tag{9}$$

In both definitions (7) and (9), the expectations are with respect to the true distribution $p_{\theta_0}(x_k|\mathcal{F}_k)$. Based on the theory of PL [33], we establish the relationship between PL and ARE by the following theorem and the corollary to Theorem 2.

*Theorem 1:* Given continuous functions $f(\cdot)$ and $g(\cdot)$, if, for each $\boldsymbol{\theta} \neq \boldsymbol{\theta}_o$, there exists a constant $\delta > 0$ such that, as $n \to \infty$

$$P(\mathcal{I}_n(\boldsymbol{\theta})/n > \delta) \to 1 \tag{10}$$

and

$$\mathcal{J}_n(\boldsymbol{\theta})/n^2 \to 0 \text{ in probability} \tag{11}$$

then at least one $\arg\min_\theta \mathcal{I}_n(\boldsymbol{\theta})$ tends to one $\arg\max_\theta \overline{\mathcal{L}}_n(\boldsymbol{\theta})$ almost surely on $\Omega = \{\boldsymbol{\theta} | \mathcal{I}_n(\boldsymbol{\theta}) \uparrow \infty, \Sigma_{i=1}^n j_i(\boldsymbol{\theta})/\mathcal{I}_i^2(\boldsymbol{\theta}) < \infty\}$, where $\overline{\mathcal{L}}_n(\boldsymbol{\theta}) \equiv \ln \mathcal{L}_n(\boldsymbol{\theta})$.

Proof of the theorem is given in Appendix A. Note that the first condition of the theorem [see (10)] represents the rate by which the Kullback–Leibler information accumulates with $n$ and guarantees that for each $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0, \mathcal{I}_n(\boldsymbol{\theta}) \to \infty$ as $n \to \infty$, i.e., the information continues to accumulate. The second condition [see (11)], on the other hand, implies asymptotical stability of variance.

Consistency and asymptotic normality are essential properties to ensure that as the network experience grows, the probability of the network approximation error exceeding any specified level tends to zero. For the parameterized model of (2), we show the following large sample properties of the maximum PL estimator.

*Theorem 2:* Assume $f(\cdot)$ and $g(\cdot)$ are continuously differentiable and satisfy (10) and (11) of Theorem 1. Let $\boldsymbol{\Theta}$ be a convex set and $\boldsymbol{\theta}_0$ be in the interior of $\boldsymbol{\Theta}$. Then, if there exist positive definite matrices $\boldsymbol{Q}_\theta$ and $\boldsymbol{Q}'_\theta$ such that

$$n^{-1}\boldsymbol{U}_n(\boldsymbol{\theta}) \to \boldsymbol{Q}'_\theta \quad \text{in probability} \tag{12}$$

and

$$n^{-1}\boldsymbol{V}_n(\boldsymbol{\theta}) \to \boldsymbol{Q}_\theta \quad \text{in probability} \tag{13}$$

where

$$\boldsymbol{U}_n(\boldsymbol{\theta}) = \sum_{i=1}^n E\{\boldsymbol{u}_i(\boldsymbol{\theta})\{\boldsymbol{u}_i^T(\boldsymbol{\theta})\}, \tag{14}$$

$$\boldsymbol{V}_n(\boldsymbol{\theta}) = \sum_{i=1}^n \nabla \boldsymbol{u}_i(\boldsymbol{\theta}) \tag{15}$$

and

$$\boldsymbol{u}_i(\boldsymbol{\theta}) = \nabla \ln p_\theta(x_i|\mathcal{F}_i) \tag{16}$$

and if $||\boldsymbol{u}_i(\boldsymbol{\theta})||$ is finite, then $\hat{\boldsymbol{\theta}}_n = \arg\max_\theta \mathcal{L}_n(\boldsymbol{\theta})$ is almost surely unique for all sufficiently large $n$, and as $n \to \infty$

1) $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$ almost surely,
2) $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \to \mathcal{N}[0, \boldsymbol{Q}_{\theta_0}^{-1}\boldsymbol{Q}'_{\theta_0}\boldsymbol{Q}_{\theta_0}^{-1}]$ in distribution.

Hence, the maximum PL estimate for the neural network model of (2) is consistent and asymptotically normal. Proof of the theorem is given in Appendix B.

*Corollary 1:* If the conditions of Theorem 2 are satisfied, then

$$\arg\min_\theta \mathcal{I}_n \to \arg\max_\theta \overline{\mathcal{L}}_n \tag{17}$$

almost surely on $\Omega$.

Proof of the corollary directly follows from the conclusion of Theorem 1 and the almost sure uniqueness of $\hat{\boldsymbol{\theta}}_n$ given by Theorem 2. Thus, the maximum PL estimate $\hat{\boldsymbol{\theta}}_n$ also minimizes ARE distance between the true and estimated conditional distributions asymptotically providing an estimate of the true parameter $\boldsymbol{\theta}_0$. We emphasize the fact that the result holds for the general case of dependent observations and, hence, provides a generalization of the ML and ARE equivalence, which is shown for independent observations [29], [30].

## III. Perceptron Conditional Probability Model and Its Application to Channel Equalization

### A. Binary Distribution Learning

If we consider the special case where the sequence $x_n$ takes values from the binary alphabet $\mathcal{S} = \{0, 1\}$, the problem reduces to estimation of the conditional probability $p(x_n = 1|\mathcal{F}_n)$. We can write the ARE cost function for the binary alphabet as

$$\mathcal{I}_n(\boldsymbol{\theta}) = \sum_{i=1}^n p(x_i = 1|\mathcal{F}_i) \ln \frac{p(x_i = 1|\mathcal{F}_i)}{p_\theta(x_i = 1|\mathcal{F}_i)} + \sum_{i=1}^n (1 - p(x_i = 1|\mathcal{F}_i)) \ln \frac{1 - p(x_i = 1|\mathcal{F}_i)}{1 - p_\theta(x_i = 1|\mathcal{F}_i)}. \tag{18}$$

In Theorem 1, we establish the equivalence of maximum PL estimation and ARE minimization for a finite alphabet $\mathcal{S}$ under two regularity conditions: (10) and (11). In the following two subsections, we show that these regularity conditions are satisfied for the single layer and multilayer perceptron probability models for a binary alphabet. However, here, we note an interesting observation for the binary alphabet, which we have originally presented as the connection between PL estimation and ARE minimization in [3] and [26]. If we use first-order stochastic approximations

$$p(x_n = 1|\mathcal{F}_n) = E\{x_n|\mathcal{F}_n\} \approx x_n \tag{19}$$

we can write the stochastic variant of the ARE cost function

$$\hat{\mathcal{I}}_n(\boldsymbol{\theta}) = -\left(\sum_{i=1}^n x_i \ln p_\theta(x_i = 1|\mathcal{F}_i) + \sum_{i=1}^n (1 - x_i) \ln (1 - p_\theta(x_i = 1|\mathcal{F}_i))\right) \tag{20}$$

after we substitute (19) in (18) and use the fact that $x_n$ is binary and that continuity requires $0 \log 0 = 0$.

For the binary alphabet, the PL function is written as

$$\mathcal{L}_n(\boldsymbol{\theta}) = \prod_{i=1}^n p_\theta(x_i = 1|\mathcal{F}_i)^{x_i}(1 - p_\theta(x_i = 1|\mathcal{F}_i))^{1-x_i}. \tag{21}$$

It can then be observed that

$$\hat{\mathcal{I}}_n(\boldsymbol{\theta}) = -\overline{\mathcal{L}}_n(\boldsymbol{\theta}) = -\ln \mathcal{L}_n(\boldsymbol{\theta}) \tag{22}$$

i.e., for binary $x_n$, under first-order stochastic approximation for the true conditionals, the ARE cost function is exactly equal to the negative log PL.

In the next two subsections, we consider the two basic perceptron models for $p_{\boldsymbol{\theta}}(x_n = 1|\mathcal{F}_n)$: The single and the multilayer perceptron model show that the conditions of Theorem 1 are satisfied for both models, and hence, we can estimate/learn the parameters of the perceptron model directly by PL maximization, which will also minimize the ARE distance between the true and estimated conditional pmf. We then derive the least relative entropy algorithm for the binary alphabet and study its dynamical properties in the following subsections.

### *B. Single-Layer Perceptron Model*

Consider the problem of determining the probability that the binary series $x_n$ takes the value 1 from a training sequence, given the finite past of the covariates $\boldsymbol{y}_n = [y_n, y_{n-1}, \cdots, y_{n-N+1}]^T$. Note the limitation of the single layer perceptron model for problems that are not linearly separable. The conditional pmf $p_\theta: \boldsymbol{R}^N \rightarrow [0, 1]$ is parameterized such that

$$p_{\boldsymbol{\theta}}(x_n = 1|\mathcal{F}_n) = g(\boldsymbol{\theta}^T \boldsymbol{y}_n) \tag{23}$$

where $g(\cdot)$ is the sigmoidal nonlinearity

$$g(s) = \frac{1}{1+e^{-s}}. \tag{24}$$

The pmf $p_\theta(x_n|\mathcal{F}_n)$ defined in (2) is then written as

$$f(\cdot) = g(\cdot)^{x_n}(1-g(\cdot))^{1-x_n}. \tag{25}$$

We can rewrite $f(\cdot)$ in the exponential form as

$$f(\cdot) = \exp\left(x_n \gamma_n(\boldsymbol{\theta}) - b(\gamma_n(\boldsymbol{\theta}))\right) \tag{26}$$

where

$$\gamma_n(\boldsymbol{\theta}) = \boldsymbol{\theta}^T \boldsymbol{y}_n \tag{27}$$

and

$$b(\gamma_n(\boldsymbol{\theta})) = \gamma_n(\boldsymbol{\theta}) - \ln \frac{1}{1+\exp\left(-\gamma_n(\boldsymbol{\theta})\right)}. \tag{28}$$

To simplify the expressions in what follows, we make the definition

$$\gamma_n^* \equiv \gamma_n(\boldsymbol{\theta}_*) \tag{29}$$

while keeping in mind that $\gamma_n^*$ is simply a linear function of $\boldsymbol{\theta}$.

Using the exponential representation in (26), we can evaluate the following:

$$\begin{aligned} r_n(\boldsymbol{\theta}) &= -x_n(\gamma_n - \gamma_n^0) + b(\gamma_n) - b(\gamma_n^0) && (30)\\ i_n(\boldsymbol{\theta}) &= E\{r_n(\boldsymbol{\theta})|\mathcal{F}_n\} \\ &= b''(\gamma_n^\beta)(\gamma_n^\alpha - \gamma_n^0)(\gamma_n - \gamma_n^0) && (31) \end{aligned}$$

and

$$j_n(\boldsymbol{\theta}) = \operatorname{Var}\{r_n(\boldsymbol{\theta})|\mathcal{F}_n\} = b''(\gamma_n^0)(\gamma_n - \gamma_n^0)^2 \tag{32}$$

where $r_n(\boldsymbol{\theta})$ is defined in (8), $\gamma_n^\alpha \in (\min(\gamma_n, \gamma_n^0), \max(\gamma_n, \gamma_n^0))$, and $\gamma_n^\beta \in (\min(\gamma_n^\alpha, \gamma_n^0), \max(\gamma_n^\alpha, \gamma_n^0))$. The derivations for (30)–(32) are given in Appendix C.

In [33, Lemma 3A], it is shown that for a stationary process $\{y_n\}$, if $b''(\cdot)$ is uniformly bounded away from 0 and $\infty$ for all possible values of $\boldsymbol{\theta}$, then

$$n^{-2}\sum_{i=1}^{n}(\gamma_i - \gamma_i^0)^2 \rightarrow 0 \quad \text{in probability} \tag{33}$$

but

$$n^{-1}\sum_{i=1}^{n}(\gamma_i - \gamma_i^0)^2 \quad \text{is locally uniformly bounded away from zero.} \tag{34}$$

To show that the conditions of Theorem 1, (10), and (11) are satisfied, we need to consider the asymptotic behavior of $\mathcal{I}_n(\boldsymbol{\theta})$ and $\mathcal{J}_n(\boldsymbol{\theta})$, which are defined as the sum of $i_k(\boldsymbol{\theta})$ and $j_k(\boldsymbol{\theta}), k = 1, \cdots, n$, respectively. Consider the expressions given for $i_n(\boldsymbol{\theta})$ and $j_n(\boldsymbol{\theta})$ in (31) and (32), and observe that by its definition in (28) $b''(\cdot)$ is finite for all $\boldsymbol{\theta}$. In addition, since we assume that $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is the compact parameter space, $b''(\cdot)$ is also bounded away from 0. Therefore, the asymptotic behavior of $i_k(\boldsymbol{\theta})$ and $j_k(\boldsymbol{\theta}), k = 1, \cdots, n$ depend only on that of $\Sigma_{i=1}^n(\gamma_i - \gamma_i^0)^2$. Then, the result given in (33) together with the expression derived for $j_n(\boldsymbol{\theta})$ in (32) implies the second condition of Theorem 1 (11). Since the product $(\gamma_n^\alpha - \gamma_n^0)(\gamma_n - \gamma_n^0)$ is bounded by $(\gamma_n^\alpha - \gamma_n^0)^2$ and $(\gamma_n - \gamma_n^0)^2$ and (34) holds for all $\boldsymbol{\theta}$, the first condition of the theorem [see (10)] is also satisfied. Hence, maximum PL estimation for the single-layer sigmoidal perceptron model (23) is equivalent to ARE minimization.

### *C. Multilayer Perceptron Model*

In what follows, we present extension of the single-layer perceptron model of previous section to multilayer perceptrons (MLP). For simplicity in notation, we consider a single hidden layer MLP structure. Extension to multi hidden layer MLP's is immediate.

Consider the following single hidden layer MLP structure as the conditional pmf model:

$$p_\theta(x_n = 1|\mathcal{F}_n) = g\left(\sum_{i=1}^{q} h(\boldsymbol{y}_n^T \boldsymbol{w}^i) v^i\right) \tag{35}$$

where $\boldsymbol{w}^i \in \boldsymbol{R}^{N\times 1}$ is the weight vector between the input layer and the hidden node $i$, $(i = 1, \cdots, q$, where $q$ is the number of hidden nodes), $\boldsymbol{y}_n \in \boldsymbol{R}^{N\times 1}$ is the observation vector, and $v^i$ is the weight between the hidden node $i$ and the output node. We represent the entire set of weights by

$$\boldsymbol{\theta} = [\boldsymbol{W}, \boldsymbol{v}] \in \boldsymbol{R}^{q\times(N+1)}$$

where

$$\boldsymbol{W} = [\boldsymbol{w}^1, \boldsymbol{w}^2, \cdots, \boldsymbol{w}^q]^T \in \boldsymbol{R}^{q\times N} \quad \text{and}$$
$$\boldsymbol{v} = [v^1, v^2, \cdots, v^q]^T \in \boldsymbol{R}^{q\times 1}.$$

The hidden node activation function $h(\cdot)$ is chosen to ensure network approximation capabilities [30], e.g., it can be chosen as the familiar logistic or the radial basis function. However, for learning parameters by gradient descent minimization, note that $g(\cdot)$ has to be chosen such that $g'(\cdot) > 0$.

If we choose both $g(\cdot)$ and $h(\cdot)$ as a sigmoidal function

$$g(\boldsymbol{s}_n^T \boldsymbol{v}) = \frac{1}{1+\exp(-\boldsymbol{s}_n^T \boldsymbol{v})} \tag{36}$$

where

$$s_n^i = \frac{1}{1+\exp(-\boldsymbol{y}_n^T \boldsymbol{w}^i)}$$

for $i = 1, \cdots, q, \boldsymbol{s}_n = [s_n^1, s_n^2, \cdots, s_n^q]^T$. Then, the exponential formulation for the binary pmf (25) can again be written as in (26)

$$f(\cdot) = \exp(x_n \gamma_n(\boldsymbol{\theta}) - b_n(\gamma_n(\boldsymbol{\theta}))) \tag{37}$$

with $b_n(\gamma_n(\boldsymbol{\theta}))$ defined as in (28), but $\gamma_n(\boldsymbol{\theta})$ this time defined as

$$\gamma_n(\boldsymbol{\theta}) = \boldsymbol{s}_n^T \boldsymbol{v}. \tag{38}$$

Hence, we can derive the exact same expressions as in (30)–(32) with this new definition of $\gamma_n$. Again, the asymptotic behavior of $\mathcal{I}_n(\boldsymbol{\theta})$ and $\mathcal{J}_n(\boldsymbol{\theta})$ depends solely on that of $\Sigma_{i=1}^n (\gamma_i - \gamma_i^0)^2$. To satisfy both condition (10) and (11), we have to show that $n^{-2} \Sigma_{i=1}^n (\gamma_i - \gamma_i^0)^2 \to 0$ in probability while $n^{-1} \sum_{i=1}^n (\gamma_i - \gamma_i^0)^2$ is locally uniformly bounded away from zero. We write

$$\begin{aligned}(\gamma_k - \gamma_k^0)^2 &= \left( \sum_{i=1}^M \left( v^i \frac{1}{1+\exp(-\boldsymbol{y}_k^T \boldsymbol{w})} \right.\right. \\ &\qquad \left.\left. - \; v_0^i \frac{1}{1+\exp(-\boldsymbol{y}_k^T \boldsymbol{w}_0)} \right) \right)^2 \\ &\le \left( \sum_{i=1}^M (|v^i| + |v_0^i|) \right)^2. \end{aligned} \tag{39}$$

Hence, $E\{(\gamma_k - \gamma_k^0)^2\}$ is bounded and for stationary $\{y_n\}$, by Birkhoff–Khinchin theorem [17]

$$n^{-1} \sum_{k=1}^n (\gamma_k - \gamma_k^0)^2 \to \hat{h} \quad \text{almost everywhere} \tag{40}$$

where $\hat{h}$ is a random variable. Hence, the conditions of Theorem 1, (10), and (11) are satisfied. If, in addition to being stationary, $\{y_n\}$ is also ergodic, then $\hat{h} = E\{(\gamma_k - \gamma_k^0)^2\}$. Note that extension to multi-hidden layer, single-output MLP's proceeds exactly the same with proper definition of $\gamma_n$.

### D. Dynamics of Relative Entropy Minimization

Thus far, we have shown the statistical equivalence of ARE minimization and partial likelihood maximization and obtained large sample properties for conditional distribution learning. In this section, we are concerned with the actual dynamics of sequential minimization of relative entropy (or maximization of partial likelihood) by stochastic gradient descent. The mathematical framework for the dynamics has been provided by Wittner and Denker in [32]. We simply show that the negative log PL cost function meets the criteria to be a *well-formed* function in the sense of the definition in [32], and therefore, convergence to a solution, if one exists, can be assured.

Gradient descent minimization of the negative log PL cost function [which is given by (20)–(22)] for the single-layer perceptron conditional probability model of (23) results in the least relative entropy (LRE) algorithm given by the following update:

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n + \mu(x_n - g(\boldsymbol{y}_n^T \boldsymbol{\theta}_n))\boldsymbol{y}_n \tag{41}$$

where $\mu$ is the adaptation (step) size. Note that gradient descent learning of the parameter vector $\boldsymbol{\theta}$ imposes the additional constraint that $g'(\cdot) > 0$ on the model given in (2).

If we choose sigmoidal nonlinearity for both $h(\cdot)$ and $g(\cdot)$ for the MLP model in (35), gradient descent minimization of the negative log PL cost function results in the following updates:

$$v_{n+1}^i = v_n^i + \mu_1 s_n^i e_n \tag{42}$$

$$\boldsymbol{w}_{n+1}^i = \boldsymbol{w}_n^i + \mu_2 \boldsymbol{y}_n g(s_n^i)(1 - g(s_n^i)) v_n^i e_n \tag{43}$$

for $i = 0, \cdots, q,$ where

$$s_n^i = g(\boldsymbol{y}_n^T \boldsymbol{w}_n^i) \tag{44}$$

and

$$\begin{aligned} e_n &= x_n - g(\boldsymbol{s}_n^T \boldsymbol{v}_n) \\ &= x_n - p_\theta(x_n = 1 | \mathcal{F}_n) \end{aligned} \tag{45}$$

where $\mu_1$ and $\mu_2$ are the step sizes at the output and hidden layers, respectively. The derivations for the update equations (41)–(43) are given in Appendix D.

The binary equalization problem can be rephrased as follows in order to comply with the development in [32]. For the remainder of the section, assume that the nonlinearity is the hyperbolic tangent, which is an odd function, without loss of generality. Therefore, $p_\theta(x_n = 1 | \mathcal{F}_n) \in (-1, 1)$. (Note that the transformation to the probability measure is immediate by the application of transformation $\frac{1}{2}[(\cdot) + 1]$).

Divide the training set $\mathcal{Y} = \{\boldsymbol{y}_n \in \boldsymbol{R}^{N \times 1}, n = 1, \cdots, M\}$ into two disjoint subsets specified by the desired output

$$\begin{aligned} \mathcal{Y} = &\underbrace{\{\boldsymbol{y}_n | p_\theta(x_n = 1 | \mathcal{F}_n) \ge 0\}}_{B_1} \\ &\cup \underbrace{\{\boldsymbol{y}_n | p_\theta(x_n = 1 | \mathcal{F}_n) < 0\}}_{B_2}. \end{aligned} \tag{46}$$

Now, define

$$B \equiv B_1 \cup \{-\boldsymbol{y}_n | \boldsymbol{y}_n \in B_2\} \tag{47}$$

so that the solution set can be defined as

$$\boldsymbol{\Theta} \equiv \{\boldsymbol{\theta} | p_\theta(x_n = 1 | \mathcal{F}_n) > 0, \quad \forall \boldsymbol{y}_n \in B\}. \tag{48}$$

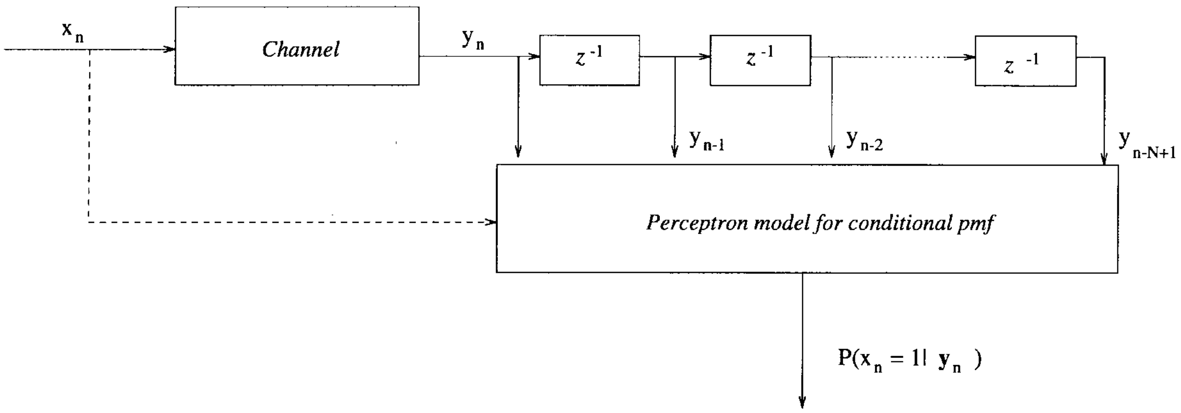


Fig. 1. Conditional distribution learning for the binary communications channel.

Next, we state the definition of a well-formed cost function in the sense of Wittner and Denker given for networks without hidden units [32]. Consider cost functions of the form

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{n} \nu(\boldsymbol{y}_i^T \boldsymbol{\theta}). \tag{49}$$

*Definition 1:* The cost function $J(\cdot)$ is *well-formed* if $\nu(\cdot)$ is differentiable and satisfies the following:

1) For all $s, -\nu'(s) \geq 0$ ($\nu(\cdot)$ does not push in the wrong direction).
2) There exists some $\epsilon > 0$ such that $-\nu'(s) \geq \epsilon$ for all $s \leq 0$ ($\nu(\cdot)$ keeps pushing if there is a misclassification).
3) $\nu(\cdot)$ is bounded below.

*Proposition 1:* If the cost function is well-formed, then gradient descent is guaranteed to enter $\boldsymbol{\Theta}$, provided $\boldsymbol{\Theta}$ is not empty.

*Proof:* See [32].

*Proposition 2:* The negative log PL cost function

$$-\mathcal{L}_n = -\sum_{i=1}^{n} \left[ \frac{1+x_i}{2} \ln \left( \frac{1+p_\theta(x_i=1|\mathcal{F}_i)}{2} \right) + \frac{1-x_i}{2} \ln \left( \frac{1-p_\theta(x_i=1|\mathcal{F}_i)}{2} \right) \right] \tag{50}$$

is well formed.

*Proof:* With the hyperbolic tangent as the nonlinearity and for the target $x$, $\nu$ becomes

$$\nu(s) = -\frac{1+x}{2} \ln \left( \frac{1+\tanh(s)}{2} \right) - \frac{1-x}{2} \cdot \ln \left( \frac{1-\tanh(s)}{2} \right) \tag{51}$$

with

$$-\nu'(s) = x - \tanh(s). \tag{52}$$

In the rephrased version of the binary equalization problem for the development in this section, the target $x = 1$, and therefore

$$-\nu'(s) = 1 - \tanh(s) \tag{53}$$

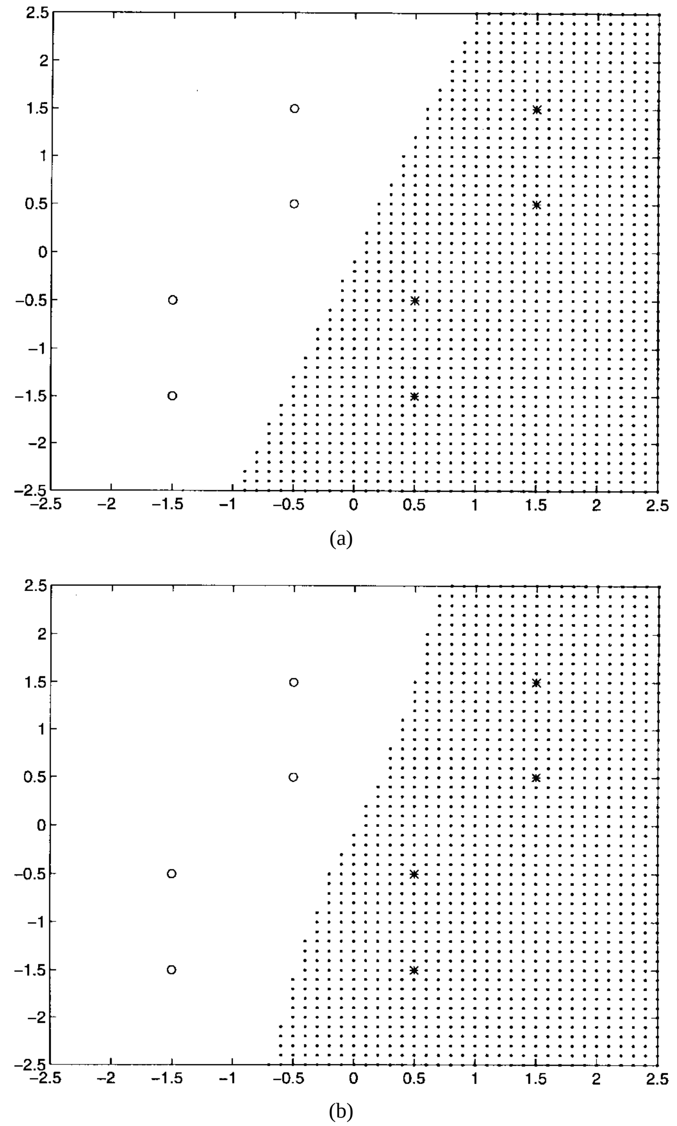
and

1) $-\nu'(s) = 1 - \tanh(s) \geq 0$,
2) $-\nu'(s) = 1 - \tanh(s) \geq 1$ for $s \leq 0$,
3) $\nu(s) \geq \ln 2$.

Therefore, gradient descent on the negative log PL cost function is guaranteed to find a solution, provided that one exists for networks without any hidden unit. As is well known, there is no such guarantee with the MSE cost function.

Note that the above analysis is for cost functions that can be represented in the form in (49), i.e., it is for networks without any hidden units. As pointed out in [32], it is probably impossible to extend Proposition 1 to networks with hidden units, as even though property 2) in Definition 1 assures that the top layer of weights gets a nonvanishing error control signal for misclassified inputs, the lower layers might be still receiving a vanishingly weak signal. In the next section, we present simulation results for a single hidden layer MLP model demonstrating that the advantages in using a well-formed cost function also carry over to networks with hidden units. Some further aspects of gradient descent learning on the ARE cost function and generalizations to MLP's are considered in [2] and [3]. In particular, the dynamics is studied by considering its parameter updates [2], and simulation results are presented that show that LRE can recover from convergence at the wrong extreme in cases where the MSE-based MLP may not.

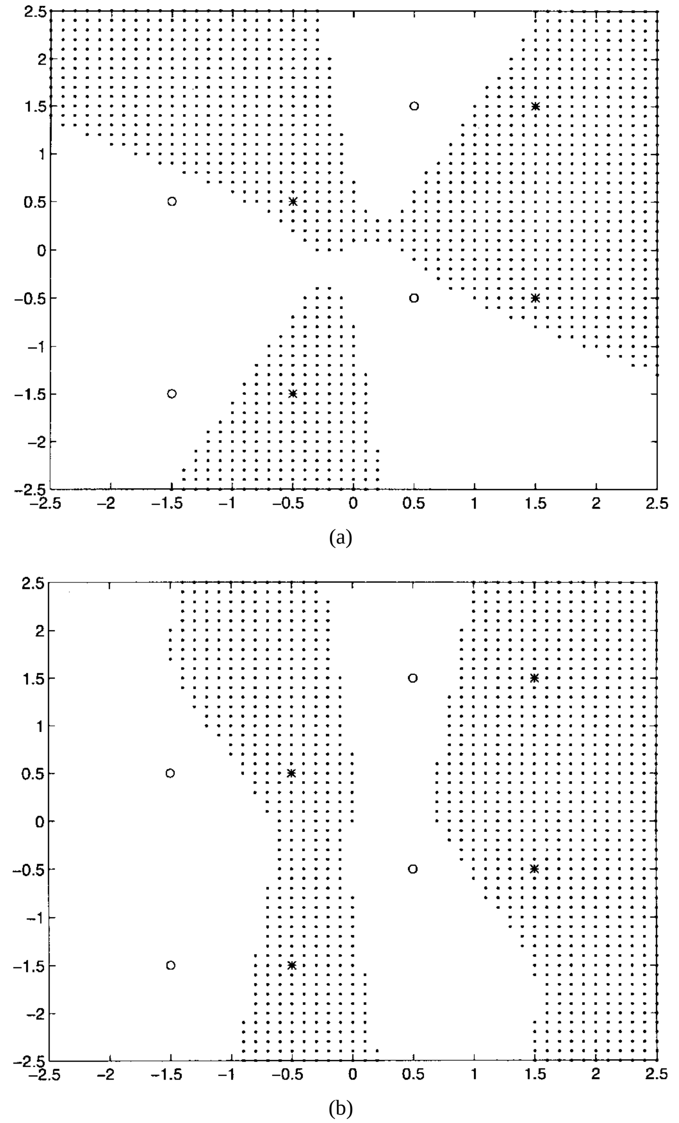
The ability of the algorithm to track large variations during training can be quite beneficial for channel equalization. For example, in low earth orbit satellite (LEOS) communication systems, these abrupt changes occur quite frequently. Due to the Doppler shift, combined with multipath reflections, the channel characteristics undergo an abrupt change as the channel is switched from one satellite (usually receding with a negative carrier shift) to the next successive satellite (usually approaching with a positive carrier shift). Another typical case is in land mobile communications, where multiple cells are transmitting the same information (usually with a small frequency offset) to cover an entire area. In this case, the channel variation occurs when the mobile unit switches reception from one antenna to another one having a stronger signal at that particular point. In the next section, we present



(a)

(b)

Fig. 2. Decision regions formed by the LRE equalizer for $H(z) = 1 + 0.5z^{-1}$ for (a) 21 dB and (b) 11 dB SNR ("*" represents $x_n = 1$ and "o" $x_n = -1$).



(a)

(b)

Fig. 3. Decision regions formed by the LRE equalizer for $H(z) = 0.5 + z^{-1}$ for (a) 21 dB and (b) 11 dB SNR ("*" represents $x_n = 1$ and "o" $x_n = -1$).

simulation results to demonstrate these dynamics for LRE and MSE minimizations in practical channel equalization schemes.

### *E. Application to Adaptive Channel Equalization*

In this section, we present application of the conditional distribution learning framework to adaptive channel equalization. We consider a simple binary pulse amplitude modulation (PAM) data transmission system transmitting ±1 and pose the supervised adaptive channel equalization problem as follows: Learn the probability that the transmitted signal $x_n$ takes the value 1 from the binary alphabet from a training sequence given the finite past of the received signal $\boldsymbol{y}_n = [y_n, y_{n-1}, \cdots, y_{n-N+1}]$. The equalizer structure is shown in Fig. 1. Note that the activation functions for the binary alphabet $\mathcal{S} = \{-1, 1\}$ are chosen as the hyperbolic tangent, and the probability $p_\theta(x_n = 1|\mathcal{F}_n) \in (-1, 1)$ can be transformed into a probability measure by applying the transformation $\frac{1}{2}[(\cdot) + 1]$. The negative PL cost function then takes the form given in (50) and in the update equation (43), $g(s_n^i)(1 - g(s_n^i))$ is replaced by $(1 - g^2(s_n^i))$.

We study the performance of the LRE algorithm derived in Section III-D for the MLP probability model as follows: First, we present test results to demonstrate the capability of the structure to realize complex decision boundaries and of the algorithm to learn parameters to achieve these boundaries. This is done for minimum and nonminimum phase channels at different SNR levels. We then present simulation results to demonstrate the ability of the algorithm to track abrupt changes during training: a property we discussed in Section III-D. The performance of LRE is compared with that of the steepest descent learning (backpropagation) based on the MSE criterion for the same structure, i.e., for the perceptron model, since this is the structure we have considered and analyzed in this section.

For the first simulation study, we consider two simple multipath channels: $H(z) = 1 + 0.5z^{-1}$ and $H(z) = 0.5 +$

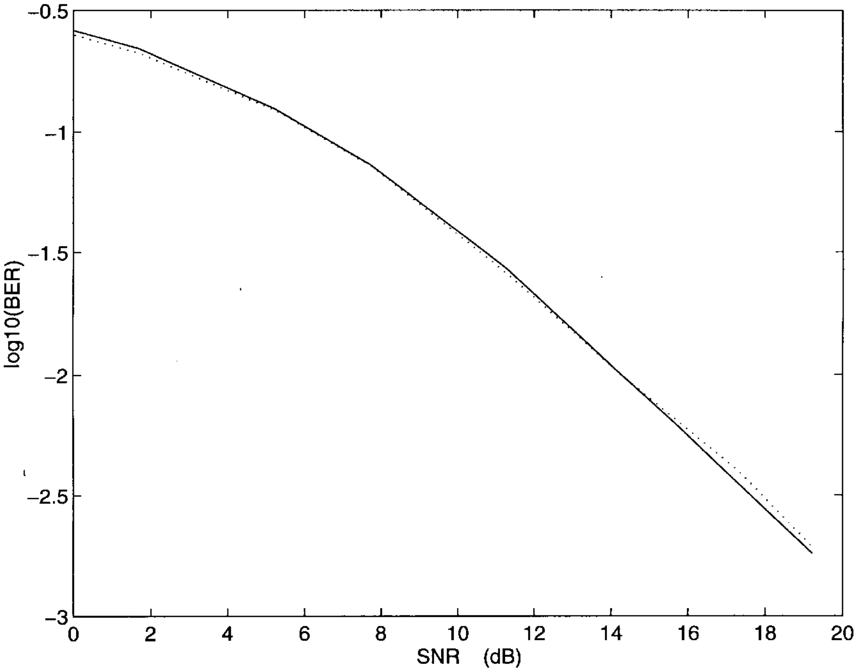


Fig. 4. BER comparison for MSE (dotted) and LRE (solid) MLP equalizers.

$z^{-1}$, i.e., a minimum- and a nonminimum-phase channel, respectively. Fig. 2 shows the decision regions for the first, and Fig. 3 shows the regions for the second channel, for approximately 21 and 11 dB SNR, respectively. The figures are drawn as $y_{n-1}$ versus $y_n$. The MLP structure used in these figures is 2-7-1, and the step sizes used are $\mu_1 = 0.2$ and $\mu_2 = 0.1$. As observed in both cases, LRE successfully learns the coefficients for achieving the given partitions. The test results are presented for convergence after 100 samples for Fig. 2(a), 300 for Fig. 2(b), 1000 for Fig. 3(a), and 5000 for 3(b). The results we show in Figs. 2 and 3 with the given convergence times compare favorably with those presented in [15] for the MLP equalizer based on the MSE criterion. Reference [15] also includes examples for these cases that show that the MSE-based MLP equalizer outperforms linear equalizers, especially at low SNR's.

Next, we consider a nonlinear channel example and compare the learning characteristics of LRE with those of MSE based backpropagation. We model the nonlinear channel as a multipath channel $(H(z) = 1 + 0.5z^{-6} + 0.25z^{-16})$ followed by a nonlinearity $0.5(\cdot)^3$, and the binary ±1 PAM communication system has 8 bits per sample with Nyquist pulse shaping. Note that since 8 bit pulse shaping is used, the multipath structure corresponds to fractional previous symbol interference and full interference of second previous symbol. We implement the LRE algorithm for binary alphabet given in (42) and (43) with the hyperbolic tangent activation function as explained above and the gradient descent minimization of the MSE on the same MLP structure for equalization of the given channel. Both algorithms have a 3-8-1 MLP structure, and the step sizes are chosen such that both algorithms have the average best performance. They are chosen as $\mu_1 = 0.1$ and $\mu_2 = 0.4$. The figures shown (Figs. 4–7) are averaged over 25 independent runs.

In Fig. 4, we show the bit error rate (BER) curves for the equalization of this channel. It can be observed that both algorithms do an equally good job of learning the parameters to compensate for the channel distortion.

To show the recovery property of LRE discussed in the previous section, we introduce an abrupt change (an exact sign change) in the channel characteristics after 150 iterations, effectively causing the current parameter estimates to be at the wrong extreme. In Fig. 5(a), we show the transient characteristics of both algorithms with the abrupt change at 150 iterations at a SNR of 19 dB. As observed in the figure, LRE can recover from convergence at the wrong extreme very effectively. Starting from the very first iteration after the change, it can follow the changes by adapting both its hidden and output layer weights in a few iterations. As we can observe in Fig. 5(a), MSE-based MLP reacts slower to the same change. Note that both algorithms have not fully converged at 150 iterations, and if the sudden change causing misclassifications occurs later, MSE-based MLP might not be able to recover. This is shown in Fig. 5(b), by introducing the sudden change at iteration 1500. Again, LRE can very rapidly adapt to the new operating condition, rapidly recovering from convergence at the wrong extreme. MSE-based MLP, on the other hand, can not recover in the next 8500 iterations after the abrupt change. Figs. 6 and 7 show the convergence and recovery characteristics of both MLP equalizers (LRE and MSE based backpropagation) with and without the abrupt change when the change occurs at 150 and 1500 iterations, respectively.

In Section III-D, we study dynamics of the algorithm that relates to its convergence properties, and here, we present simulation results to support the analysis results of Section III-D and show its extension to multilayer networks. Another important property for equalizers is their tracking ability, i.e.,

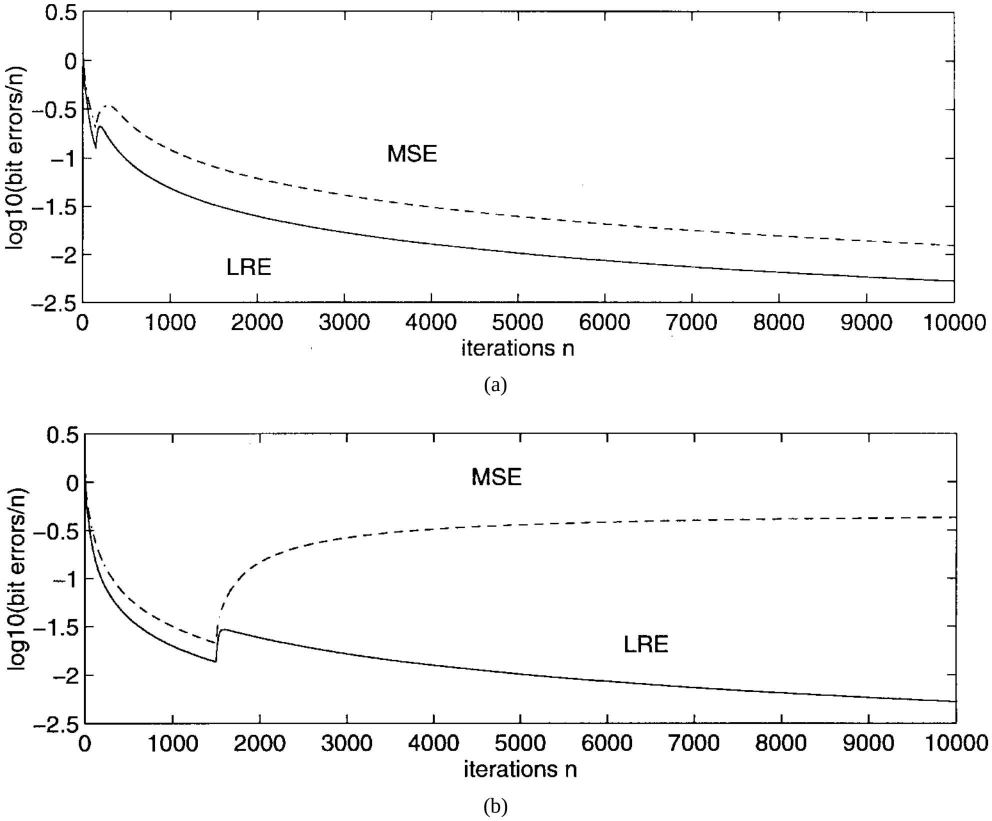


Fig. 5. Recovery characteristics for MSE (dotted) and LRE (solid) MLP equalizers with an abrupt change at (a) 150 (b) 1500 iterations.

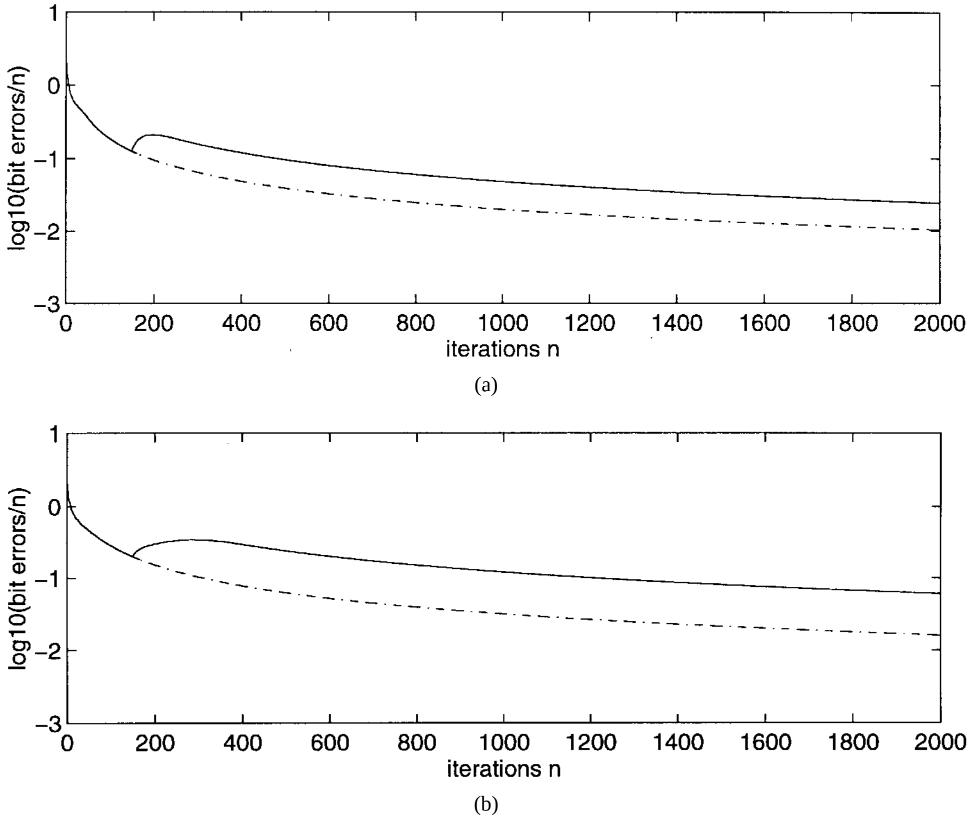
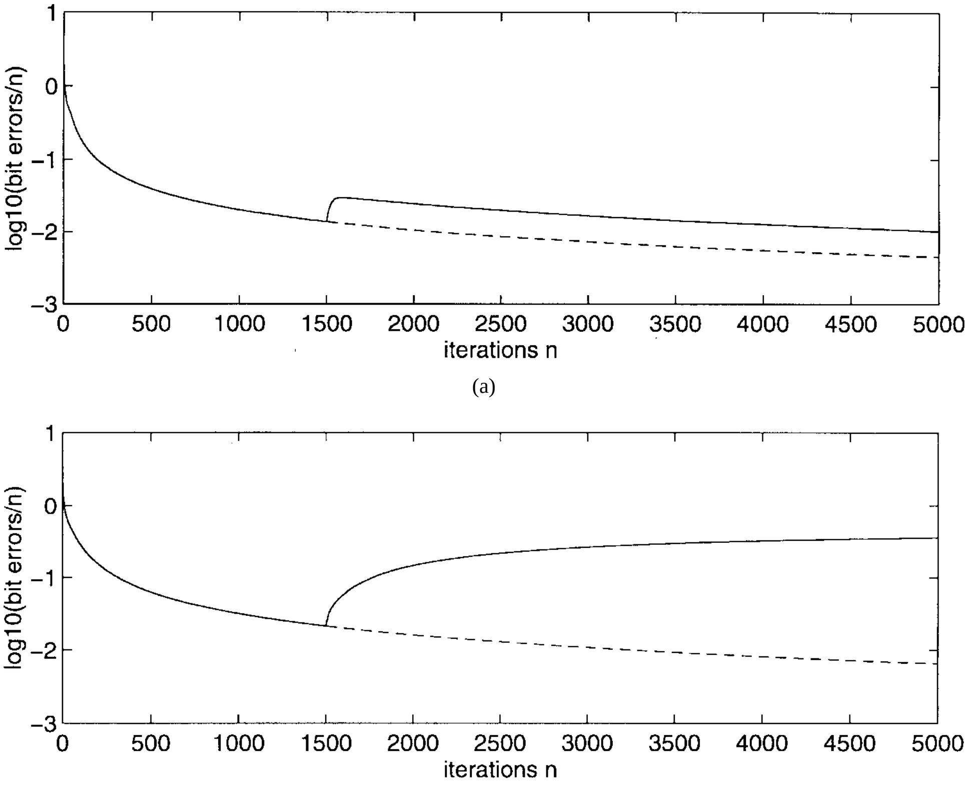


Fig. 6. Recovery and convergence characteristics for (a) LRE (b) MSE MLP equalizers [with abrupt change at $n = 150$ (solid) and without abrupt change (dotted)].



(a)

(b)

Fig. 7. Recovery and convergence characteristics for (a) LRE (b) MSE MLP equalizers [with abrupt change at $n = 1500$ (solid) and without abrupt change (dotted)].

the ability of the algorithm to track changes in the channel response after convergence, in the absence of a training sequence, in a decision-directed mode. A recent work [13] considers this property and application of our framework to tracking in a decision directed mode. It is shown that the PL or relative entropy cost function also provides advantages for this case. For an abrupt change in a linear channel response, LRE provides the most robust performance in that, among the 1000 realizations simulated, it can track the changes in a greater number of the cases than the linear least mean square and MSE-based backpropagation algorithms.

## IV. Conclusions

We have presented a distribution learning formulation for real-time signal processing applications in which conditional probabilities are parameterized by a general neural network structure. We have shown that this formulation can be exploited to derive efficient learning rules and lends itself to statistical analysis via introduction of an extended likelihood framework. The central result that renders the framework fruitful in terms of analysis is the equivalence of ARE minimization and partial likelihood maximization under some technical conditions. Partial likelihood is a relatively recently developed extension of likelihood, which, by design, allows us to obtain the consistency and asymptotic normality of conditional distribution learning for a general neural network structure without making the assumption of independent observations. The notion of maximum partial likelihood estimation should prove to be a very useful tool in developing real-time signal processing theories, where the historic development of likelihood-based frameworks have come to force independence or some further simplifying assumptions.

We have used the well-formed cost function framework to study the dynamics of on-line gradient descent minimization of ARE (or negative log PL). We point out the behavior differences in adapting to abrupt parameter changes during training between the MSE and ARE. We apply the formulation to adaptive channel equalization and demonstrate the capability of the least relative entropy equalizer to achieve complex decision boundaries and, as a consequence of the aforementioned property, to recover from abrupt changes in the channel response during training. Thus, the speedy (and guaranteed) recovery of ARE minimization algorithms from convergence at wrong extremes should be taken into consideration while comparing cost functions in channel equalization problems.

## Appendix A
## Proof of Theorem 1

Let

$$\mathcal{R}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} r_i(\boldsymbol{\theta}) \tag{54}$$

where $r_i$ is defined in (8). Lemma 2B in [33] states that if (10) and (11) are satisfied, then for each $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, as $n \to \infty$

$$(\mathcal{R}_n(\boldsymbol{\theta}) - \mathcal{I}_n(\boldsymbol{\theta}))/\mathcal{I}_n(\boldsymbol{\theta}) \to 0 \tag{55}$$

almost surely on the set

$$\Omega = \left\{ \boldsymbol{\theta} | \mathcal{I}_n(\boldsymbol{\theta}) \uparrow \infty, \sum_{i=1}^{n} j_i(\boldsymbol{\theta}) / \mathcal{I}_i^2(\boldsymbol{\theta}) < \infty \right\}. \tag{56}$$

Therefore, for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\forall \epsilon > 0$, there exists an $N$ such that

$$\mathcal{I}_n(\boldsymbol{\theta})(1-\epsilon) < \mathcal{R}_n(\boldsymbol{\theta}) < \mathcal{I}_n(\boldsymbol{\theta})(1+\epsilon) \tag{57}$$

if $n > N$.

If we assume that $\mathcal{I}_n(\boldsymbol{\theta})$ achieves its minimum on $\Omega_{\mathcal{I}}^{(n)} \subset \Omega$ and $\mathcal{R}_n(\boldsymbol{\theta})$ on $\Omega_{\mathcal{R}}^{(n)} \subset \Omega$, then for $\boldsymbol{\theta}_n^* \in \Omega_{\mathcal{I}}^{(n)}$ and $\overline{\boldsymbol{\theta}}_n^* \in \Omega_{\mathcal{R}}^{(n)}$, we can write

$$\mathcal{I}_n(\boldsymbol{\theta}_n^*)(1-\epsilon) \le \mathcal{I}_n(\overline{\boldsymbol{\theta}}_n^*)(1-\epsilon) < \mathcal{R}_n(\overline{\boldsymbol{\theta}}_n^*) \tag{58}$$

and

$$\mathcal{R}_n(\overline{\boldsymbol{\theta}}_n^*) \le \mathcal{R}_n(\boldsymbol{\theta}_n^*) < \mathcal{I}_n(\boldsymbol{\theta}_n^*)(1+\epsilon) \tag{59}$$

or equivalently

$$\mathcal{I}_n(\boldsymbol{\theta}_n^*)(1-\epsilon) < \mathcal{R}_n(\overline{\boldsymbol{\theta}}_n^*) < \mathcal{I}_n(\boldsymbol{\theta}_n^*)(1+\epsilon). \tag{60}$$

We can also write (57) for $\boldsymbol{\theta}_n^*$ as

$$\mathcal{I}_n(\boldsymbol{\theta}_n^*)(1-\epsilon) < \mathcal{R}_n(\boldsymbol{\theta}_n^*) < \mathcal{I}_n(\boldsymbol{\theta}_n^*)(1+\epsilon). \tag{61}$$

Hence, for sufficiently large $n$, we have

$$\mathcal{R}_n(\overline{\boldsymbol{\theta}}_n^*) / \mathcal{I}_n(\boldsymbol{\theta}_n^*) \to 1 \tag{62}$$

by (60), and

$$\mathcal{R}_n(\boldsymbol{\theta}_n^*) / \mathcal{I}_n(\boldsymbol{\theta}_n^*) \to 1 \tag{63}$$

by (61) almost surely on $\Omega$. Hence, by (62) and (63), at least a point in $\Omega_{\mathcal{I}}^{(n)}$ tends to a point in $\Omega_{\mathcal{R}}^{(n)}$ almost surely.

Since we can express $\mathcal{R}_n(\boldsymbol{\theta}_n^*)$ in terms of the log PL $\overline{\mathcal{L}}_n(\boldsymbol{\theta}_n)$ as

$$\mathcal{R}_n(\boldsymbol{\theta}_n^*) = \overline{\mathcal{L}}_n(\boldsymbol{\theta}_0) - \overline{\mathcal{L}}_n(\boldsymbol{\theta}_n^*) \tag{64}$$

where the first term $\overline{\mathcal{L}}_n(\boldsymbol{\theta}_0)$ is constant, the conclusion of the theorem follows.

## Appendix B
## Proof of Theorem 2

Since conditions given in (12) and (13) hold, then $\overline{\mathcal{L}}_n(\boldsymbol{\theta})$ is concave with respect to $\boldsymbol{\theta}$, and (10) and (11) are also satisfied. Therefore, as $n \to \infty$

$$\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$$

almost surely by [33, Th. 2E].

Define the score vector process

$$\boldsymbol{S}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{u}_i(\boldsymbol{\theta}) \tag{65}$$

where $\boldsymbol{u}_i(\boldsymbol{\theta})$ is defined in (16). For $\boldsymbol{u}_i(\boldsymbol{\theta})$, it is also easy to show that

$$E\{\boldsymbol{u}_i(\boldsymbol{\theta}) | \mathcal{F}_i\} = 0. \tag{66}$$

Hence, $\boldsymbol{u}_i(\boldsymbol{\theta})$ is a Martingale difference, and $\boldsymbol{S}_n(\boldsymbol{\theta})$ is also a martingale.

By considering the first two terms in the Taylor expansion of $\nabla \overline{\mathcal{L}}_n(\boldsymbol{\theta})$ around $\boldsymbol{\theta}_0$, we can write

$$\nabla \overline{\mathcal{L}}_n(\hat{\boldsymbol{\theta}}_n) \approx \boldsymbol{S}_n(\boldsymbol{\theta}_0) + (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \boldsymbol{V}_n(\boldsymbol{\theta}_0) \tag{67}$$

and for $\nabla \overline{\mathcal{L}}_n(\hat{\boldsymbol{\theta}}_n) = 0$, we then have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \approx n^{-(1/2)} \boldsymbol{S}_n(\boldsymbol{\theta}_0)(-n\boldsymbol{V}_n^{-1}(\boldsymbol{\theta}_0)). \tag{68}$$

By (66) and the definition of $\boldsymbol{U}_n(\boldsymbol{\theta})$ (14), it is easily observed that

$$\boldsymbol{U}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathrm{Var}\,\{\boldsymbol{u}_i(\boldsymbol{\theta}) | \mathcal{F}_i\}. \tag{69}$$

We define $t_i \equiv \boldsymbol{e}^T \boldsymbol{u}_i(\boldsymbol{\theta})$, where $\boldsymbol{e}$ is a constant unit vector. Hence, $t_i$ is also a martingale with respect to $\mathcal{F}_i$, and we can write

$$n^{-1} \sum_{i=1}^{n} \mathrm{Var}\,\{t_i | \mathcal{F}_i\} = n^{-1} \boldsymbol{e}^T \boldsymbol{U}_n(\boldsymbol{\theta}) \boldsymbol{e} \to \boldsymbol{e}^T \boldsymbol{Q}'_\theta \boldsymbol{e} \quad \text{in probability} \tag{70}$$

by condition (12). We can then use the Markov inequality to write, for any $\epsilon > 0$

$$\sum_{i=1}^{n} P(|t_i| > n\epsilon) < \sum_{i=1}^{n} \frac{E\{t_i^2\}}{(n\epsilon)^2} = \boldsymbol{e}^T \frac{\boldsymbol{U}_n(\boldsymbol{\theta})}{(n\epsilon)^2} \boldsymbol{e} \to 0 \quad \text{in probability.} \tag{71}$$

Now, using the indicator function $I(\cdot)$, we write

$$\begin{aligned} n^{-1} &\sum_{i=1}^{n} E\{t_i^2 I(|t_i| > n\epsilon)\} \\ &\le n^{-1} \max_{i,\theta} t_i^2 \sum_{i=1}^{n} E\{I(|t_i| > n\epsilon)\} \\ &= n^{-1} \max_{i,\theta} ||\boldsymbol{u}_i(\boldsymbol{\theta})||^2 \sum_{i=1}^{n} P(|t_i| > n\epsilon) \to 0 \\ &\text{in probability} \end{aligned} \tag{72}$$

by (71) and the assumption that $||\boldsymbol{u}_i(\boldsymbol{\theta})||$ is bounded.

Thus, all conditions for the martingale central limit theorem [6] are verified for the martingale $n^{-(1/2)} \Sigma_{i=1}^n t_i$, and it follows that

$$\boldsymbol{e}^T (n^{-(1/2)} \boldsymbol{S}_n(\boldsymbol{\theta}_0)) \boldsymbol{e} = n^{-(1/2)} \sum_{i=1}^{n} t_i \to \mathcal{N}[\boldsymbol{0}, \boldsymbol{e}^T \boldsymbol{Q}'_{\theta_0} e] \quad \text{in distribution.} \tag{73}$$

Since this is true for any unit vector $\boldsymbol{e}$

$$n^{-(1/2)} \sum_{i=1}^{n} \boldsymbol{S}_n(\boldsymbol{\theta}_0) \to \mathcal{N}[\boldsymbol{0}, \boldsymbol{Q}'_{\theta_0}] \quad \text{in distribution.} \tag{74}$$

We can then use (68) to write

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \to \mathcal{N}[\boldsymbol{0}, \boldsymbol{Q}_{\theta_0}^{-1} \boldsymbol{Q}'_{\theta_0} \boldsymbol{Q}_{\theta_0}^{-1}] \quad \text{in distribution.} \tag{75}$$

## APPENDIX C
## DERIVATIONS FOR (30)–(32)

Using the exponential formulation for $f(\cdot)$ in (26) and the definition for $\gamma_n^x$ (29), we write

$$\begin{aligned} r_n(\boldsymbol{\theta}) &= \ln \frac{p_{\theta_0}(x_n|\mathcal{F}_n)}{p_\theta(x_n|\mathcal{F}_n)} \\ &= \ln\left(\exp\left[x_n\gamma_n^0 - b(\gamma_n^0) - (x_n\gamma_n - b(\gamma_n))\right]\right) \\ &= -x_n(\gamma_n - \gamma_n^0) + b(\gamma_n) - b(\gamma_n^0). \end{aligned} \tag{76}$$

The same definition (26) can also be used to write

$$\begin{aligned} E\{x_n|\mathcal{F}_n\} &= \sum_{x_n\in\{0,1\}} x_n p_{\theta_0}(x_n|\mathcal{F}_n) \\ &= \exp\left(\gamma_n^0 - b(\gamma_n^0)\right) \\ &= \frac{1}{1+\exp\left(-\gamma_n^0\right)} \end{aligned} \tag{77}$$

where the last equality follows from definition of $b_n(\gamma_n^0)$. Then, since $b'(s) = 1/(1+\exp(-s))$, we have

$$E\{x_n|\mathcal{F}_n\} = b'(\gamma_n^0). \tag{78}$$

We can now use (76) and (78) to write

$$\begin{aligned} i_n(\boldsymbol{\theta}) &= E\{r_n(\boldsymbol{\theta})|\mathcal{F}_n\} \\ &= -E\{x_n|\mathcal{F}_n\}(\gamma_n - \gamma_n^0) + b(\gamma_n) - b(\gamma_n^0) \\ &= b(\gamma_n) - b(\gamma_n^0) - b'(\gamma_n^0)(\gamma_n - \gamma_n^0). \end{aligned} \tag{79}$$

Since $b(\cdot)$ is continuous at each point in $[\min(\gamma_n, \gamma_n^0), \max(\gamma_n, \gamma_n^0)]$ and differentiable in $(\min(\gamma_n, \gamma_n^0), \max(\gamma_n, \gamma_n^0))$, we can apply the mean value theorem and substitute $b(\gamma_n) - b(\gamma_n^0) = b'(\gamma_n^\alpha)(\gamma_n - \gamma_n^0)$ in the above expression to write

$$i_n(\boldsymbol{\theta}) = (b'(\gamma_n^\alpha) - b'(\gamma_n^0))(\gamma_n - \gamma_n^0) \tag{80}$$

where $\gamma_n^\alpha \in (\min(\gamma_n, \gamma_n^0), \max(\gamma_n, \gamma_n^0))$. In addition, by applying the mean value theorem for the second time, we get

$$i_n(\boldsymbol{\theta}) = b''(\gamma_n^\beta)(\gamma_n^\alpha - \gamma_n^0)(\gamma_n - \gamma_n^0) \tag{81}$$

where $\gamma_n^\beta \in (\min(\gamma_n^\alpha, \gamma_n^0), \max(\gamma_n^\alpha, \gamma_n^0))$.

By using (79) and (76), we get

$$\begin{aligned} &j_n(\boldsymbol{\theta}) \\ &\quad= \operatorname{Var}\{r_n(\boldsymbol{\theta})|\mathcal{F}_n\} \\ &\quad= E\{(r_n(\boldsymbol{\theta}) - E\{r_n(\boldsymbol{\theta})|\mathcal{F}_n\})^2|\mathcal{F}_n\} \\ &\quad= E\{(-x_n(\gamma_n - \gamma_n^0) + b'(\gamma_n^0)(\gamma_n - \gamma_n^0))^2|\mathcal{F}_n\} \\ &\quad= E\{(x_n - b'(\gamma_n^0))^2|\mathcal{F}_n\}(\gamma_n - \gamma_n^0)^2. \end{aligned} \tag{82}$$

Using (78) and the fact that $b_n''(s) = b_n'(s) - b_n'^2(s)$, we finally write

$$\begin{aligned} j_n(\boldsymbol{\theta}) &= (E\{x_n^2|\mathcal{F}_n\} - b'^2(\gamma_n^0))(\gamma_n - \gamma_n^0)^2 \\ &= b''(\gamma_n^0)(\gamma_n - \gamma_n^0)^2. \end{aligned} \tag{83}$$

## APPENDIX D
## DERIVATIONS FOR (41)–(43)

For the the single-layer perceptron, conditional probability model of (23), the negative log PL cost function (21), or the stochastic variant of ARE (22) is given by

$$\begin{aligned} \hat{\mathcal{I}}_n(\boldsymbol{\theta}) &= -\ln \mathcal{L}_n(\boldsymbol{\theta}) = -\overline{\mathcal{L}}_n(\boldsymbol{\theta}) \\ &= -\sum_{i=1}^{n} l_i(\boldsymbol{\theta}) \end{aligned} \tag{84}$$

where

$$l_n(\boldsymbol{\theta}) = -(1 - x_n)\boldsymbol{y}_n^T\boldsymbol{\theta}_n - \ln(1 + e^{-\boldsymbol{y}_n^T\boldsymbol{\theta}_n}).$$

The gradient of $l_n(\boldsymbol{\theta})$ can be written as

$$\begin{aligned} \nabla_{\boldsymbol{\theta}_n} l_n(\boldsymbol{\theta}) &= -(1 - x_n)\boldsymbol{y}_n + \frac{e^{-\boldsymbol{y}_n^T\boldsymbol{\theta}_n}}{1 + e^{-\boldsymbol{y}_n^T\boldsymbol{\theta}_n}}\boldsymbol{y}_n \\ &= (x_n - g(\boldsymbol{y}_n^T\boldsymbol{\theta}_n))\boldsymbol{y}_n. \end{aligned} \tag{85}$$

The least relative entropy (LRE) update for the single layer perceptron probability model (23) is then given by the update

$$\begin{aligned} \boldsymbol{\theta}_{n+1} &= \boldsymbol{\theta}_n + \mu\nabla_{\boldsymbol{\theta}_n} l_n(\boldsymbol{\theta}) \\ &= \boldsymbol{\theta}_n + \mu(x_n - g(\boldsymbol{y}_n^T\boldsymbol{\theta}_n))\boldsymbol{y}_n. \end{aligned} \tag{86}$$

If we choose sigmoidal nonlinearity for both $h(\cdot)$ and $g(\cdot)$ for the MLP model in (35), $l_n(\boldsymbol{\theta})$ is given by

$$l_n(\boldsymbol{\theta}) = -(1 - x_n)\boldsymbol{s}_n^T\boldsymbol{v}_n - \ln(1 + e^{-\boldsymbol{s}_n^T\boldsymbol{v}_n}) \tag{87}$$

where

$$s_n^i = g(\boldsymbol{y}_n^T\boldsymbol{w}_n^i) \tag{88}$$

for $i = 1, \cdots, q, \boldsymbol{s}_n = [s_n^1, s_n^2, \cdots, s_n^q]^T$, and $\boldsymbol{v}_n = [v_n^1, v_n^2, \cdots, v_n^q]^T$. Then, we have

$$\begin{aligned} \frac{\partial l_n(\boldsymbol{\theta})}{\partial v_n^i} &= -(1 - x_n)s_n^i + \frac{e^{-\boldsymbol{s}_n^T\boldsymbol{v}_n}}{1 + e^{-\boldsymbol{s}_n^T\boldsymbol{v}_n}}s_n^i \\ &= s_n^i e_n \end{aligned} \tag{89}$$

and

$$\begin{aligned} \nabla_{\boldsymbol{w}_n^i} l_n(\boldsymbol{\theta}) &= -(1 - x_n)v_n^i\nabla_{\boldsymbol{w}_n^i} s_n^i + \frac{e^{-\boldsymbol{s}_n^T\boldsymbol{v}_n}}{1 + e^{-\boldsymbol{s}_n^T\boldsymbol{v}_n}}v_n^i\nabla_{\boldsymbol{w}_n^i} s_n^i \\ &= v_n^i e_n \nabla_{\boldsymbol{w}_n^i} s_n^i \\ &= \boldsymbol{y}_n g(s_n^i)(1 - g(s_n^i))v_n^i e_n \end{aligned} \tag{90}$$

where $e_n = (x_n - g(\boldsymbol{s}_n^T\boldsymbol{v}_n))$, and we have used the fact that $g'(s) = g(s)(1 - g(s))$ for the sigmoidal nonlinearity $g(s) = 1/(1 + e^{-s})$. Gradient descent minimization of $\hat{\mathcal{I}}_n(\boldsymbol{\theta})$ [or $-\overline{\mathcal{L}}_n(\boldsymbol{\theta})$] hence results in the updates given in (42) and (43).

## REFERENCES

[1] T. Adalı, M. K. Sönmez, and X. Liu, "Partial likelihood estimation for real-time signal processing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Atlanta, GA, 1996, vol. 6, pp. 3562–3565.

[2] T. Adalı, M. K. Sönmez, and K. Patel, "On the dynamics of the LRE algorithm: A distribution learning approach to adaptive equalization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Detroit, MI, 1995, pp. 929–932.

[3] T. Adalı and M. K. Sönmez, "Channel equalization with perceptrons: An information-theoretic approach," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Adelaide, Australia, 1994, pp. 297–300.

[4] T. Adalı, X. Liu, N. Li, and M. K. Sönmez, "A maximum partial likelihood framework for channel equalization by distribution learning," in *Proc. IEEE Workshop Neural Networks Signal Processing*, Boston, MA, 1995, pp. 541–550.

[5] S. Amari, "The EM algorithms and information geometry in neural network learning," *Neural Computat.*, vol. 7, pp. 13–18, 1994.

[6] B. M. Brown, "Martingale central limit theorems," *Ann. Math. Statist.*, vol. 44, pp. 59–66, 1971.

[7] T. X. Brown, "Neural networks for adaptive equalization," in *Applications Neural Networks Telecommun.*, J. Alspector, R. Goodman, and T. X. Brown Eds. Hillsdale, NJ: Erlbaum, 1993, pp. 27–33.

[8] I. Cha and S. A. Kassam, "Channel equalization using adaptive complex radial basis function networks," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 122–131, 1995.

[9] S. Chen, S. McLaughlin, and B. Mulgrew, "Complex-valued radial basis function network, Part II: Application to digital communications channel equalization," *Signal Processing*, vol. 35, pp. 19–31, 1994.

[10] S. Chen, G. J. Gibson, C. F. N. Cowan, and P. M. Grant, "Adaptive equalization of finite nonlinear channels using multilayer perceptrons," *Signal Processing*, vol. 10, pp. 107–119, 1990.

[11] D. R. Cox, "Partial likelihood," *Biometrika*, vol. 62, pp. 69–72, 1975.

[12] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedure," *Statistics and decisions, Supplementary issue, no. 1*, E. Dedewicz *et al.*, Eds. Munich: Oldenburg Verlag, 1984, pp. 205–237.

[13] A. R. Figueiras-Vidal, A. Artés-Rodríguez, J. Sid-Sueiro, and M. Martinez-Ramon, "Adaptive signal processing: A discussion of tradeoffs from the perspective of artificial learning," in *Proc. 1996 European Conf. Signal Processing*, Trieste, Italy, 1996, pp. 1635–1640.

[14] G. D. Forney, Jr., "Maximum likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 378–383, 1972.

[15] G. J. Gibson, S. Siu, and C. F. N. Cowan, "The application of nonlinear structures to the reconstruction of binary signals," *IEEE Trans. Signal Processing*, vol. 39, pp. 1877–1884, 1991.

[16] F. Girosi *et al.*, Eds., *Neural Networks for Signal Processing V, Proc. IEEE Workshop, Boston, MA*. New York: IEEE, 1995.

[17] R. M. Gray and L. D. Davisson, *Random Processes: A Mathematical Approach for Engineers*. Englewood Cliffs, NJ: Prentice-Hall, 1986.

[18] J. Cid-Sueiro, A. Artès-Rodriguez, and A. R. Figueiras-Vidal, "Recurrent radial basis function networks for optimal symbol-by-symbol equalization," *Signal Processing*, vol. 40, pp. 53–63, 1994.

[19] M. I. Jordan and R. A. Jacobs, "Hierarchical mixture of experts and the EM algorithm," *Neural Comput.*, no. 6, pp. 181–214, 1994.

[20] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Englewood Cliffs, NJ: Macmillan, 1994.

[21] G. E. Hinton, "Connectionist learning procedures," *Artif. Intell.*, vol. 40, pp. 185–234, 1989.

[22] J. J. Hopfield, "Learning algorithms and probability distributions in feed-forward and feed-back networks," in *Proc. Nat. Acad. Sci. USA*, 1987, vol. 84, pp. 8429–8433.

[23] G. Kechriotis, E. Zervas, and E. S. Manolakos, "Using recurrent neural networks for adaptive communication channel equalization," *IEEE Trans. Neural Networks*, vol. 5, pp. 267–278, 1994.

[24] L. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.

[25] M. Meyer and G. Pfeiffer, "Multilayer perceptron based decision feedback equalisers for channels with intersymbol interference," in *Proc. Inst. Elec. Eng.*, vol. 140, no. 6, pp. 420–424, 1993.

[26] M. K. Sönmez and J. S. Baras, "Time series modeling by perceptrons: A likelihood approach," in *Proc. World Congr. Neural Networks*, Portland, OR, 1993, pp. 601–604.

[27] E. Slud and B. Kedem, "Partial likelihood analysis of logistic regression and autoregression," *Statistica Sinica*, vol. 4, no. 1, pp. 89–106, 1994.

[28] S. Theodoridis, C. F. N. Cowan, and C. M. S. See, "Schemes for equalization of communication channels with nonlinear impairments," in *Proc. Inst. Elec. Commun.*, vol. 142, no. 3, pp. 1350–2425, 1995.

[29] Y. Wang and T. Adalı, "Efficient learning of finite normal mixtures for image quantification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Atlanta, GA, vol. 6, 1996, pp. 3423–3426.

[30] H. White, *Estimation, Inference, and Specification Analysis*. New York: Cambridge Univ. Press, 1994.

[31] ——, "Learning in artificial neural networks: A statistical perspective," *Neural Comput.*, vol. 1, pp. 425–464, 1989.

[32] B. S. Wittner and J. S. Denker, "Strategies for teaching layered networks classification tasks," in *Neural Inform. Proc. Syst.*, Denver, CO, 1988, pp. 850–859.

[33] W. H. Wong,"Theory of partial likelihood," *Ann. Statist.*, vol. 14, pp. 88–123, 1986.

[34] J. Xuan, T. Adalı, and X. Liu, "Information geometry of maximum partial likelihood estimation for channel equalization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Atlanta, GA, vol. 6, 1996, pp. 3534–3537.

**Tülay Adalı** (S'88–M'92) received the B.S. degree from Middle East Technical University, Ankara, Turkey, in 1987 and the M.S. and Ph.D. degrees from North Carolina State University, Raleigh, in 1988 and 1992, respectively, all in electrical engineering.

In 1992, she joined the Department of Electrical Engineering, University of Maryland-Baltimore County, Baltimore, as an assistant professor. Her research interests include statistical signal processing, neural computation, adaptive signal processing, and their applications in channel equalization, system identification, time-series prediction, and image analysis.

Dr. Adalı is the recipient of a 1997 National Science Foundation Career Award.

**Xiao Liu** received the B.S. and M.S. degrees in control theory from Shandong University, Jinan, China, in 1984 and 1987, respectively.

From 1987 to 1990, he worked with the Flight Control Laboratory in the Department of Automatic Control, Nanjing University of Aeronautics and Astronautics, Nanjing, China. From 1990 to 1994, he was an assistant professor in the same department. Presently, he is working toward the Ph.D. degree in the Department of Computer Science and Electrical Engineering, University of Maryland-Baltimore County, Baltimore. His current research interests include digital signal processing and communications, neural networks, and nonlinear system modeling and identification.

**M. Kemal Sönmez** received the B.S. degree from Middle East Technical University, Ankara, Turkey, in 1989 and the M.S. degree from North Carolina State University, Raleigh, in 1991, both in electrical engineering. He is currently completing the Ph.D. degree in electrical engineering at the University of Maryland, College Park.

While doing his Ph.D. research, he was a research assistant at the Institute for Systems Research, University of Maryland. He joined the Speech Technology and Research Laboratory at SRI, Menlo Park, CA, in November 1996. His research interests include robust speech and speaker recognition, adaptive channel equalization, nonlinear/multiresolution signal models for recognition and compression, statistical signal processing, and artificial neural networks.