ON THE DYNAMICS OF THE LRE ALGORITHM: A DISTRIBUTION LEARNING APPROACH TO ADAPTIVE EQUALIZATION

Tülay Adalı[1] *M. Kemal Sönmez*[2] *Kartik Patel*[1]

[1]Information Technology Laboratory, Department of Electrical Engineering
University of Maryland Baltimore County, Baltimore, MD 21228-5398
[2]Institute for Systems Research, University of Maryland, College Park, MD 20742

ABSTRACT

We present the general formulation for the adaptive equalization by distribution learning introduced in [Adalı 94]. In this framework, adaptive equalization can be viewed as a parametrized conditional distribution estimation problem where the parameter estimation is achieved by learning on a multilayer perceptron (MLP). Depending on the definition of the conditioning event set either supervised or unsupervised (blind) algorithms in either recurrent or feedforward networks result. We derive the least relative entropy (LRE) algorithm for binary data communications and analyze its statistical and dynamical properties. Particularly, we show that LRE learning is consistent and asymptotically normal by working in the partial likelihood estimation framework, and that the algorithm can always recover from convergence at the wrong extreme as opposed to the MSE based MLP's by working within an extension of the *well-formed* cost functions framework of Wittner and Denker [Wittner 88]. We present simulation examples to demonstrate this fact.

1. INTRODUCTION

As the volume of data traffic gets ever more demanding on communication systems, the need for sophisticated signal processing techniques to increase achievable data communication rates becomes more evident. Adaptive equalization techniques developed during the last two decades have been successfully implemented in high speed data transmission for correcting the distorting effect of channels having substantial amplitude and delay distortion. These techniques, based on linear models, have inherent limitations in combating nonlinear time varying distortion which is the major factor hindering further increase in the attainable data rate today.

Recently, a number of adaptive equalizers that use nonlinear structures are introduced (see e.g. [Gibson 91, Kechriotis 94]), and it is shown that these neural network based equalizers can successfully equalize nonlinear channels where linear equalizers might fail. These adaptive equalizers view channel equalization as a classification problem and are based on the traditional mean square error (MSE) performance criterion. We have recently introduced a new approach to the problem of adaptive channel equalization [Adalı 94] in which conditional probability distribution of the transmitted signal given the received signal is modeled by a sigmoidal perceptron. In this framework, channel equalization can be viewed as a distribution learning problem on an artificial neural network. This new approach results in a perceptron based algorithm, which can successfully combat nonlinear channel distortions [Adalı 94]. In this paper, we extend our formulation to finite symbol alphabets, and present general formulation for this framework which encompasses both supervised and unsupervised (blind) mode of operation for either feedforward or recurrent networks, and analyze the statistical and dynamical properties of the LRE algorithm which we derive within this framework. The paper is structured as follows: in section 2, we give the general problem formulation, derive the LRE algorithm for binary data communications in section 3, present its analysis in section 4, and give a simulation example comparing its performance with the MSE based equalizer in section 5.

2. CHANNEL EQUALIZATION BY DISTRIBUTION LEARNING

We formulate adaptive equalization problem as follows: A sequence of symbols $x(n)$, taking values from a finite alphabet $\mathcal{S} = \{a_0, a_1, ..., a_M\}$ is transmitted through a channel $\mathbf{h}$ which acts as a nonlinear operator on the incoming signal. Usually, the channel is assumed to have finite memory ($\mathbf{h} : \mathbf{R}^K \rightarrow \mathbf{R}$). The output of the channel is corrupted by additive noise $v(n)$, hence the input to the equalizer can be written as $y(n) = \mathbf{h}(\mathbf{x}_K(n)) +$

0-7803-2431-5/95 $4.00

$v(n)$ where $\mathbf{x}_K(n) = [x(n), x(n-1), ..., x(n-K+1)]$.

Let $\mathcal{F}_n$ be the σ-field generated by events of the form $\mathbf{x}(n) = [x(n), \cdots, x(1), x(0)]$ and $\mathbf{z}(n) = [z(n), \cdots, z(1), z(0)]$; the vector of time dependent covariates of $x(n)$. Typically the covariates $z(n)$ are the noise corrupted channel outputs $y(n)$, however, they can also be defined as functions of $y(n)$ and $x(n)$ depending on the channel model or the application. If $\mathcal{F}_n$ is defined as being created by the covariates which are functions of the channel output $y(n)$ only (and possibly some other auxiliary information other than the transmitted data sequence $x(n)$) this results in unsupervised (blind) mode of operation for the equalizer. Since $\mathcal{F}_n = \sigma\{1, \mathbf{x}(n), \mathbf{z}(n)\}$ represents all that is known to the observer at the equalizer end at time n, $\mathcal{F}_{n-1} \subset \mathcal{F}_n$.

Our aim is then to estimate the conditional probabilities:

$$P(X = x(n)|\mathcal{F}_n)$$

where the the conditional probability mass function (pmf) $p(x|\mathcal{F}_n)$ corresponding to the distribution P can be written as:

$$p(x|\mathcal{F}_n) = P(X = x|\mathcal{F}_n) \qquad \forall x \in \mathcal{S}.$$

We parametrize the conditional probabilities by a multilayer perceptron p_θ and achieve learning by adjusting the parameter vector θ based on the information represented by $\mathcal{F}_n$. Note that since $\mathcal{F}_n$ includes the entire history, $p_\theta(x|\mathcal{F}_n)$ can have a recurrent structure as well. In [White 93], it is shown that a probability density can be associated with the output of any network indexed by weights θ when we select a continuous, nonconstant, and bounded hidden layer activation function, and a nonnegative, continuous, increasing output activation function. The normalization condition required by the probabilities can be achieved by imposing constraints on the output bias weight.

The relative entropy (RE), or the Kullback-Leibler (KL) distance [Kullback 51], a fundamental information theoretic measure of how accurate the estimated conditional pmf is an approximation to the true conditional pmf

$$D_n(p\|p_\theta) = \sum_{x \in \mathcal{S}} p(x|\mathcal{F}_n) \ln \frac{p(x|\mathcal{F}_n)}{p_\theta(x|\mathcal{F}_n)} \qquad (1)$$

arises as the natural cost function for this formulation. Note that it is nonnegative, and is equal to zero only when $p = p_\theta$. Within this framework, the goal is then to learn the parameters θ which minimize the *accumulated* KL information (total RE)

$$D_{total_n}(p\|p_\theta) = \sum_{i=0}^{n} D_i(p\|p_\theta) \qquad (2)$$

in the sequence of conditional observations $x(0)|\mathcal{F}_0$, $x(1)|\mathcal{F}_1, \cdots, x(n)|\mathcal{F}_n$ on a multilayer perceptron.

3. MULTILAYER LRE ALGORITHM

In what follows, we assume that $x(n)$ is an independent sequence taking values from the binary alphabet $\{0, 1\}$, and derive supervised version of the LRE algorithm for a single hidden layer feedforward perceptron probability model. In [Adalı 95], we consider an arbitrary finite alphabet $\mathcal{S}$ and show that under rather mild regularity conditions, the accumulated relative entropy error minimization problem is equivalent to maximum partial likelihood estimation. We then derive the supervised and unsupervised versions of the LRE algorithm and the large sample properties of the partial likelihood estimator.

The accumulated RE distance for binary data can be written as

$$D_{total_n}(p\|p_\theta) = \sum_{i=0}^{n} (p(1|\mathcal{F}_i) \ln \frac{p(1|\mathcal{F}_i)}{p_\theta(1|\mathcal{F}_i)} + (1 - p(1|\mathcal{F}_i)) \ln \frac{1 - p(1|\mathcal{F}_i)}{p_\theta(1|\mathcal{F}_i)}). \qquad (3)$$

Since the true conditionals are not available we use first order stochastic approximations:

$$p(1|\mathcal{F}_n) = E\{x(n)|\mathcal{F}_n\} \approx x(n)$$

to write the stochastic relative entropy (SRE) cost function as:

$$D_{total_n}(\hat{p}\|p_\theta) = -\sum_{i=0}^{n} [x(i) \ln p_\theta(1|\mathcal{F}_i) + (1 - x(i)) \ln (1 - p_\theta(1|\mathcal{F}_i))] \qquad (4)$$

after some simplifications [Adalı 94].

For the binary case, the pmf can be represented by a single output multilayer perceptron. Assume the single hidden layer feedforward structure:

$$p_\theta(1|\mathcal{F}_n) = f\left(v_0(n) + \sum_{i=1}^{q} g(\overline{\mathbf{y}}_{N+1}^T(n)\mathbf{w}_i(n))v_i(n)\right) \qquad (5)$$

which is a leading example of the networks under consideration for the conditional probability model where $\mathcal{F}_n = \sigma\{1, x(n), \mathbf{y}_N(n)\}$. In (5), $\overline{\mathbf{y}}_{N+1}(n)$ is the augmented input vector containing the last N samples of $y(n)$ (vector $\mathbf{y}_N(n)$) together with the bias input 1, $\mathbf{w}_i(n)$ is the $N+1$ dimensional weight vector from the input layer to hidden node i, ($i = 1, \cdots, q$ where q is the number of hidden nodes) and $v_i(n)$'s are the hidden to output layer weights $v_0(n)$ being the bias

weight. We represent the entire vector of weights by θ, and choose the hidden node activation function g to ensure network approximation capabilities [White 93]. It can be chosen as any of the familiar logistic, hyperbolic tangent, or the radial basis activation functions. Assume choosing both as the sigmoidal or hyperbolic tangent followed by $\frac{1}{2}[(\cdot)+1]$ transformation (to ensure $f(\cdot) \in [0,1]$) functions. Gradient descent minimization of the SRE given in (4) by using the probability model of (5) then results in the least relative entropy (LRE) algorithm with the parameter updates:

$$v_i(n+1) = v_i(n) + \mu_1 s_i(n) e(n) \tag{6}$$

$$\mathbf{w}_i(n+1) = \mathbf{w}_i(n) + \mu_2 \overline{\mathbf{y}}_{N+1}(n) f'(s_i(n)) v_i(n) e(n) \tag{7}$$

$i = 0, \cdots, q$ where $s_i(n) = f(\overline{\mathbf{y}}^T_{N+1}(n)\mathbf{w}_i(n))$ and $e(n) = x(n) - p_\theta(1|\mathcal{F}_n)$. Note that we can also use variants such as momentum and batching.

4. PROPERTIES OF LRE

4.1. Statistics:

To study large sample properties of learning with the SRE cost function, we cast the problem as a maximum partial likelihood estimation problem. Partial likelihood (PL) is a generalization of both likelihood and conditional likelihood and is introduced in [Cox 75]. We show that learning on the SRE cost function is consistent and asymptotically normal by working in the PL estimation framework.

The cost function $D(\hat{p}\|p_\theta)$ may also be viewed as the negative of the *partial log-likelihood* relative to parameters θ as follows: Since $x(n)$ is binary, the pmf can be written as

$$p_\theta(1|\mathcal{F}_n)^{x(n)}(1 - p_\theta(1|\mathcal{F}_n))^{1-x(n)}. \tag{8}$$

Then the corresponding partial likelihood is given by the product

$$PL(\theta) = \prod_{i=1}^{n} p_\theta(1|\mathcal{F}_i)^{x(i)}(1 - p_\theta(1|\mathcal{F}_i))^{1-x(i)}. \tag{9}$$

Now, it is easy to see that $D(\hat{p}\|p_\theta) = -\ln(PL(\theta))$. Thus, the maximization of the PL function is equal to minimization of the SRE cost function, i.e., $\theta_0 = \arg\max PL(\theta) = \arg\min D(\hat{p}\|p_\theta)$.

We assume that a single layer sigmoidal perceptron is used as the probability distribution function model, i.e. $p_\theta(1|\mathcal{F}_n) = 1/(1 + exp(-\theta^T(n)\mathbf{y}_N(n))$. Define the score vector process

$$\mathbf{S}_\theta(n) \equiv \sum_{i=1}^{n} \mathbf{y}_N(i)\,(x(i) - p_\theta(x|\mathcal{F}_i))$$

as the gradient ($\nabla \equiv (\partial/\partial\theta_0, \partial/\partial\theta_1, \cdots, \partial/\partial\theta_N)^T$) of $PL(\theta)$ which is a martingale with respect to the filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$. The sample observation matrix is defined as $\mathbf{I}_\theta(n) \equiv \nabla\nabla'(-logPL(\theta))$ and can be written as the sum of conditional covariance matrices

$$\mathbf{I}_\theta(n) = \sum_{i=1}^{n} var_\theta\{\mathbf{y}_N(i)(x(i) - \mathbf{y}_N(i))|\mathcal{F}_i\}.$$

Since $\mathbf{S}_\theta(n)$ is a martingale, $\mathbf{I}_{\theta_0}(n)$ is the cumulative conditional variance matrix for $\mathbf{S}_\theta(n)$ at $\theta = \theta_0$. By following the general development of [Slud 92], the large sample properties of the SRE minimization or the maximum PL estimation are then studied by using $\mathbf{S}_n(\theta)$ and $\mathbf{I}(\theta)$ based on the martingale central limit theorem for $\mathbf{S}_n(\theta)/\sqrt{n}$ and the almost sure concavity of the sample information matrix $\mathbf{I}_\theta(n)/\sqrt{n}$.

Hence, the estimator $\hat{\theta}$ is almost surely unique for all sufficiently large n and as $n \to \infty$,
(i) $\hat{\theta} \to \theta_0$ in probability,
(ii) $\sqrt{n}(\hat{\theta} - \theta_0) \to \mathcal{N}[0, \Lambda^{-1}(\theta_0)]$ in distribution,
where $\Lambda(\theta_0)$ is the information matrix per observation for estimating the true parameter θ_0. Thus, learning by minimizing the stochastic relative entropy is consistent and asymptotically normal. Note that, in this proof, we have not made the common but unrealistic assumption of independent observation vectors since PL estimation bypasses the problem of data dependency in the maximum likelihood estimation by permitting sequential conditional inference. In [Adalı 95], the large sample properties are derived for the general probability model.

4.2. Dynamics:

In [Adalı 94], we have shown that SRE cost function is *well-formed* in the sense of Wittner and Denker [Wittner 88], and hence gradient descent on this cost function is guaranteed to find a solution. As is well known, there is no such guarantee with the mean square error (MSE) cost function when used on neural networks, even on those without any hidden units. We can also study the dynamics of gradient descent learning on the SRE cost function by considering its parameter updates. The parameter updates of the LRE algorithm for gradient descent learning for the single hidden layer feedforward model of (5) are given in (6) and (7). If we compare these updates by those of the commonly used form of backpropagation algorithm based on MSE minimization (e.g. [Haykin 94]), it is easy to see that the main difference between the two minimization approaches is the absence of the term proportional to

output activation function gradient $F'(s)$ in the updates. This term directly affects the response of the algorithms in tracking changes in the unknown (true) system parameters. Assume that after initial convergence, output (or output of one unit in the multi-output case) saturates at the wrong extreme, then since SRE is a well-formed cost function, for LRE, there always exists an $\epsilon > 0$ for which $F'(s) > \epsilon$, [Adalı 94]. By considering the LRE updates given in (6) and (7), it is easy to see that when the backpropagated output error will be a non-vanishing control signal even for misclassified inputs, and the parameters will easily recover from convergence at the wrong extreme. This fact can also be observed by considering the two cost functions: in the MSE cost function

$$E_{total_n} = \sum_{i=0}^{n} (x(i) - \hat{x}_\theta(i))^2$$

the total error would float around on a relatively flat plateau when the network output $\hat{x}_\theta(n)$ saturates at the wrong extreme whereas the RE cost function (2) diverges.

5. SIMULATION RESULTS

We demonstrate the situation discussed in section 4.2 with a simulation example. Consider a binary pulse amplitude modulation (PAM) data transmission system. An abrupt change in the channel response happens during training of the equalizer and causes misclassifications after initial convergence. We model the nonlinear channel as a multipath channel ($H(z) = 1 + 0.5z^{-6} + 0.25z^{-16}$) followed by a nonlinearity $0.5(.)^3$, and the PAM communication system has 8 bits per sample with Nyquist pulse shaping.

We implement the LRE algorithm given in (6,7) 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 perceptron structure. In Figure 1, we show the bit error rate (BER) curves for the equalization of this channel which show that both algorithms do an equally good job of partitioning the decision region. What is remarkable is however, when we introduce an abrupt change (an exact sign change) in the channel characteristics after convergence, causing the decision region to rotate suddenly the LRE can very rapidly adapt to this new operating condition. 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 Figure 2, MSE produces many wrong decisions before it can adapt to this new operating condition.

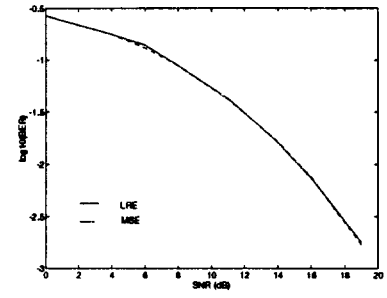

Figure 1: BER Comparison for MSE and LRE multilayer perceptrons

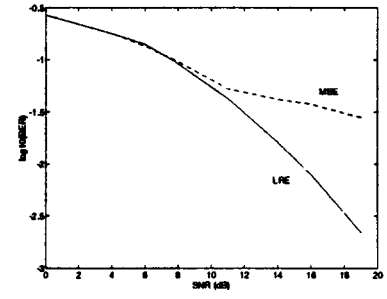

Figure 2: BER Comparison for MSE and LRE multilayer perceptrons (with an abrupt change in the channel response)

REFERENCES

[Adalı 94] T. Adalı and M. K. Sönmez, "Channel equalization with perceptrons: an information theoretic approach," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (Adelaide, Australia), April 1994, vol 3, pp. 297-300.

[Adalı 95] T. Adalı, M. K. Sönmez, and X. Liu, "Channel equalization by distribution learning," submitted to *IEEE Trans. Signal Processing.*

[Cox 75] D.R. Cox, "Partial likelihood", *Biometrika* vol.62, no. 2, pp. 269-276, 1975.

[Haykin 94] S. Haykin, "Neural networks: a comprehensive foundation," Macmillan: NY, 1994.

[Kullback 51] L. Kullback, and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics 22*, pp. 79-86, 1951.

[Gibson 91] G. J. Gibson, S. Siu, and C. F. N. Cowan, "The application of nonlinear structures to the reconstruction of binary signals," in *IEEE Trans. Signal Processing*, pp. 1877-1884, vol. 39, no. 8, Aug. 1991.

[Kechriotis 94] G. Kechriotis, E. Zervas, and E. S. Manolakos, "Using recurrent neural networks for adaptive communication channel equalization," *IEEE Trans. on Neural Networks*, vol. 5, no. 2, pp. 267-278, March 1994.

[Slud 92] E. Slud, "Martingale methods in statistics," 1992.

[White 93] H. White, "Estimation, inference and specification testing" 1993.

[Wittner 88] B.S. Wittner and J.S. Denker, "Strategies for teaching layered networks classification tasks", *Neural Info. Proc. Systems*, (Denver, CO), p.850-859, 1988.