

# A MAXIMUM PARTIAL LIKELIHOOD FRAMEWORK FOR CHANNEL EQUALIZATION BY DISTRIBUTION LEARNING <sup>1</sup>

*Tülay Adalı<sup>†</sup>, Xiao Liu<sup>†</sup>, Ning Li<sup>†</sup>, and M. Kemal Sönmez<sup>‡</sup>*

<sup>†</sup>Information Technology Laboratory, Dept. of Electrical Engineering  
University of Maryland Baltimore County, Baltimore, MD 21228-5398

<sup>‡</sup>Institute for Systems Research, University of Maryland  
College Park, MD 20742

{adali,xliu,nli2}@engr.umbc.edu    kemal@isr.umd.edu

**Abstract-** We present the general formulation for adaptive equalization by distribution learning [1] in which conditional probability mass function (pmf) of the transmitted signal given the received is parametrized by a general neural network structure. The parameters of the pmf are computed by minimization of the accumulated relative entropy (ARE) cost function. The equivalence of ARE minimization to maximum partial log-likelihood (MPLL) estimation is established under certain regularity conditions which enables us to bypass the requirement that the true conditionals be known. The large sample properties of MPLL estimator are obtained under further regularity conditions, and the binary case with sigmoidal perceptron as the conditional pmf model [1, 2] is shown to be a special case of the new framework. Results are presented which show that the multilayer perceptron (MLP) equalizer based on ARE minimization can always recover from convergence at the wrong extreme whereas the mean square error (MSE) based MLP can not.

## INTRODUCTION

As more complex channels are required to carry increasing amounts of data in today's demanding communications applications, the need to develop more sophisticated equalization schemes has become more evident. To overcome the inherent limitation of linear equalizers, a number of neural network adaptive equalizers have been introduced (see e.g. [5, 6, 9]), and it is shown that these equalizers can successfully equalize nonlinear channels where linear equalizers might fail. The neural network equalizer also offers the advantage of low-power low-complexity analog hardware implementation which is particularly important in portable applications. These neural network equalizers view channel equalization as a classification problem and are based on the traditional mean square error (MSE) performance criterion. Recently, we have introduced a new approach to channel equalization which is based on probability distribution learning [1], and uses relative entropy (RE) between the true and estimated conditional probability density functions as the performance measure to be minimized. The conditional probability mass function

---

<sup>1</sup>Research supported in part by Engineering Foundation grant RI-A-94-08.

(pmf) of the transmitted signal given the received signal is parametrized by a general neural network architecture. It is shown that when multilayer perceptrons (MLP) are chosen as the parametrized model, the equalizer can successfully combat multi-path [1] and nonlinear distortions [2], and can always recover from convergence at the wrong extreme as opposed to the MSE based MLP's [1, 2].

In this paper, we extend our distribution learning formulation to finite symbol alphabets by working in the *partial likelihood* framework. Partial likelihood, a relatively new method in estimation theory, allows for inference as the time unfolds. Hence it bypasses the problem with maximum likelihood or quasi-maximum likelihood estimation which require that the auxiliary information be known in full throughout the period of observation when they are extended to dependent observations [4]. The general formulation we present for channel equalization here encompasses both supervised and unsupervised (blind) mode of operation for a general neural network structure. In this framework, adaptive channel equalization can be considered as a conditional pmf estimation problem by accumulated relative entropy (ARE) minimization which we show to be equivalent to *maximum partial log-likelihood* (MPLL) statistical estimation problem. Unlike ARE minimization, MPLL estimation does not require that we know the true conditionals which in general are never available, hence the parameters of the conditional distribution model can be directly learned on the chosen neural network model. We show that the consistency and asymptotic normality of MPLL estimator can be obtained under further regularity conditions. In [2], we consider a simple binary communication channel equalization problem and use the sigmoidal perceptron to parametrize the conditional pmf. We then employ first order stochastic approximation of the true conditionals to write the stochastic variant of the RE cost function, and then note its equivalence to MPLL estimation. Here, we consider the binary case with the sigmoidal model as an example and show that it is a special case of the new framework. Also, for the perceptron model, we present simulation studies which show that the ARE based MLP equalizer can always recover from convergence at the wrong extreme whereas the MSE based MLP can not. This property of the RE based equalizer is discussed in [2] within an extension of the *well-formed* cost functions framework of Wittner and Denker [11].

## CHANNEL EQUALIZATION BY DISTRIBUTION LEARNING

We formulate adaptive equalization problem as follows: A sequence of symbols  $x(n)$ , taking values from a finite alphabet  $\mathcal{S} = \{a_0, a_1, \dots, a_M\}$ , is transmitted through a channel  $h$  which acts as a nonlinear operator on the incoming signal. Let  $\mathcal{F}_n$  be the  $\sigma$ -field generated by the events of the form  $\mathbf{z}(n) = [x(n-1), \dots, x(1), x(0); y(n), \dots, y(1), y(0)]$ , where,  $y(n)$ 's are the time dependent covariates of  $x(n)$ . Typically,  $y(n)$  is the noise corrupted channel output. A

common model for the channel output would be  $y(n) = h(\mathbf{x}_K(n)) + v(n)$  where  $v(n)$  is the additive noise component,  $\mathbf{x}_K(n) = [x(n), x(n-1), \dots, x(n-K+1)]$ , and  $h : \mathbf{R}^K \rightarrow \mathbf{R}$ . If  $\mathcal{F}_n$  does not include the transmitted sequence  $x(n)$  but only its covariates, this results in unsupervised (blind) mode of operation for the equalizer. Thus  $\mathcal{F}_n = \sigma\{1, \mathbf{z}(n)\}$  represents all that is known to the observer at time  $n$ , and  $\mathcal{F}_{n-1} \subset \mathcal{F}_n$ . Note that since  $\mathcal{F}_n$  includes the entire history  $p_\theta(x|\mathcal{F}_n)$  can have a recurrent structure as well.

Our aim is to estimate the conditional pmf  $p(x|\mathcal{F}_n)$ ,  $\forall x \in \mathcal{S}$ . We parametrize the conditional probability by a neural network as follows:

$$p_\theta(x|\mathcal{F}_n) = f(x, c(\theta), g(z_N(n), \theta)). \quad (1)$$

Here,  $\theta$  is the vector of network weights,  $\theta \in \Theta$  where  $\Theta$  is a compact parameter set and  $\mathbf{z}_N(n)$  is a subset of  $\mathbf{z}(n)$  containing the most recent  $N$  values of  $\mathbf{z}(n)$ . The term  $g(\mathbf{z}_N(n), \theta)$  is the output of the neural network,  $f(\cdot)$  and  $g(\cdot)$  are continuous differentiable functions, and  $c(\theta)$  and  $f(\cdot)$  are chosen such that  $\sum_{x \in \mathcal{S}} p_\theta(x|\mathcal{F}_n) = 1$ .

The relative entropy (RE), or the Kullback-Leibler distance, [8] a fundamental information theoretic measure of how accurate the estimated conditional pmf is an approximation to the true conditional pmf,

$$D_n(p_{\theta_0}||p_\theta) = \sum_{x \in \mathcal{S}} p_{\theta_0}(x|\mathcal{F}_n) \ln \frac{p_{\theta_0}(x|\mathcal{F}_n)}{p_\theta(x|\mathcal{F}_n)} \quad (2)$$

arises as the natural cost function for this formulation. Note that it is non-negative, and is equal to zero only when  $p_{\theta_0} = p_\theta$ . In (2) we assumed that  $\theta_0$  is the weight vector for which  $f(\cdot)$  achieves the true conditional pmf. The goal is then to learn  $\theta$  which minimizes the *accumulated* relative entropy (ARE) given by

$$\mathcal{I}_n = \sum_{i=1}^n D_i(p_{\theta_0}||p_\theta) \quad (3)$$

in the sequence of observations  $\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_n$ . However, note that the minimization of this cost function requires that the true conditionals, or that  $\theta_0$  be known. In the next section we show that the ARE minimization problem is equivalent to MPLL estimation which allows us to overcome this problem.

## MAXIMUM PARTIAL LIKELIHOOD ESTIMATION

The optimal network parameters  $\theta_0$  have the fundamental information theoretic interpretation that they minimize the Kullback-Leibler information given the chosen architecture and the ARE performance measure (3). Thus viewing learning as related to Kullback-Leibler information minimization in this way implies that learning is a *maximum likelihood* statistical estimation procedure for independent observations [10]. Though this may be extended

to dependent data by discounting the dependence structure in some sense, this still requires that the auxiliary information be known in full throughout the period of observation [7]. It is obvious that this requirement can not be satisfied in data communications. *Partial likelihood* (PL) can bypass this problem by allowing inference from the available information.

The distribution learning problem posed in the previous section can be cast as a MPLL estimation problem. To show this, we define

$$r_i = \ln \frac{p_{\theta_0}(x|\mathcal{F}_i)}{p_{\theta}(x|\mathcal{F}_i)} \quad \text{and} \quad \mathcal{J}_n = \sum_{i=1}^n \text{Var}_{\theta_0}(r_i|\mathcal{F}_i)$$

and based on the theory of partial likelihood [12], show the following:

**Theorem 1:** If there exist a constant  $\delta > 0$ ,  $\alpha_n \uparrow \infty$ , continuous functions  $f(\cdot)$  and  $g(\cdot)$  such that

$$P(\mathcal{I}_n/\alpha_n > \delta) \longrightarrow 1 \quad \text{and} \quad \mathcal{J}_n/\alpha_n \longrightarrow_p 0 \quad (4)$$

then ARE minimization is equivalent to MPLL estimation, i.e.,

$$\arg \left( \min_{\theta} \mathcal{I}_n \right) = \arg \left( \max_{\theta} \bar{\mathcal{L}}_n \right)$$

where  $\mathcal{L}_n = \prod_{i=1}^n p_{\theta}(x|\mathcal{F}_i)$  is the partial likelihood function and  $\bar{\mathcal{L}}_n = \ln \mathcal{L}_n$  is the partial log-likelihood. (Proof is given in the appendix.)

It then suffices to maximize  $\bar{\mathcal{L}}_n$  to estimate the conditional distribution, and the value  $\hat{\theta}$ , maximizing  $\bar{\mathcal{L}}_n$ , provides an estimate of the true parameter  $\theta_0$ . 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 parametrized model of (1) we show the following large sample properties of the MPLL estimator:

**Theorem 2:** For  $f(\cdot)$  and  $g(\cdot)$  as given in Theorem 1, assume conditions given in (4) hold and that the first and second order derivatives of  $f(\cdot)$  and  $g(\cdot)$  exist and are continuous. Then if there exist a  $\beta_n \uparrow \infty$  and positive definite matrices  $Q$  and  $Q_1$  such that

$$\beta_n^{-1} U_n(\theta_0) \longrightarrow_p Q_1 \quad \text{and} \quad \beta_n^{-1} V_n(\theta_0) \longrightarrow_p Q \quad (5)$$

where

$$U_n(\theta) = \sum_{i=1}^n E(\nabla u_i \nabla u_i^T), \quad V_n(\theta) = \sum_{i=1}^n \nabla u_i \quad \text{and} \quad u_i = \nabla \ln p_{\theta}(x|\mathcal{F}_i) \quad (6)$$

then, we can guarantee that  $\hat{\theta}$  is almost surely unique for all sufficiently large  $n$  and as  $n \rightarrow \infty$ ,

- (i)  $\hat{\boldsymbol{\theta}} \rightarrow \boldsymbol{\theta}_0$  in probability,  
(ii)  $\sqrt{\beta_n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \rightarrow \mathcal{N}[\mathbf{0}, \Lambda]$  in distribution,  
where  $\Lambda = Q^{-1}Q_1Q^{-1}$  is the information matrix per observation for estimating the true parameter  $\boldsymbol{\theta}_0$ . (Proof is given in the appendix.)

### EXAMPLE: THE BINARY ALPHABET

Consider the adaptive channel equalization problem where the probability that the transmitted signal  $x(n) = 1$  from the alphabet  $\{0, 1\}$  is to be determined from a training sequence, given the finite past of the received signal:  $\mathbf{y}_N(n) = [y(n), y(n-1), \dots, y(n-N+1)]$ , i.e.,  $\mathbf{z}_{N+1} = [x(n), \mathbf{y}_N(n)]$ . The conditional pmf  $p_\theta : \mathbf{R}^N \rightarrow [0, 1]$  is parametrized such that

$$p_\theta(x(n) = 1 | \mathcal{F}_n) = g(\boldsymbol{\theta}^T \mathbf{y}_N(n))$$

where  $g(\cdot)$  is a differentiable non-linearity such that  $g'(s) > 0$  for all  $s$  and can be chosen as  $g(s) = 1/(1 + e^{-s})$ . The pmf is then written as

$$f(\cdot) = g(\cdot)^{x(n)}(1 - g(\cdot))^{1-x(n)}.$$

This is the sigmoidal perceptron model we used in [1].

We can reformulate  $f(\cdot)$  as

$$f(\cdot) = \exp(x(n)\gamma_n(\boldsymbol{\theta}) - b_n(\boldsymbol{\theta})) \quad (7)$$

where  $\gamma_n(\boldsymbol{\theta}) = \boldsymbol{\theta}^T \mathbf{z}_N(n)$  and

$$b_n = \boldsymbol{\theta}^T \mathbf{z}_N(n) - \ln \frac{1}{1 + \exp(-\boldsymbol{\theta}^T \mathbf{y}_N(n))}.$$

For this exponential model (7), we have

$$\begin{aligned} r_n(\boldsymbol{\theta}) &= -x(n)(\gamma_n(\boldsymbol{\theta}) - \gamma_n(\boldsymbol{\theta}_0)) + b_n(\boldsymbol{\theta}) - b_n(\boldsymbol{\theta}_0) \\ E(r_n | \mathcal{F}_n) &= b_n(\boldsymbol{\theta}) - b_n(\boldsymbol{\theta}_0) - b'(\bar{\boldsymbol{\theta}})(\gamma_n(\boldsymbol{\theta}) - \gamma_n(\boldsymbol{\theta}_0)) \\ Var(r_n | \mathcal{F}_n) &= b''(\bar{\boldsymbol{\theta}})(\gamma_n(\boldsymbol{\theta}) - \gamma_n(\boldsymbol{\theta}_0))^2 \end{aligned}$$

where each  $\bar{\theta}_i$  is a value between  $\theta_{0i}$  and  $\theta_i$ , and the prime denotes the derivative with respect to  $\theta$ .

By Lemma 3A [12]:

$$\alpha_n^{-2} \sum_{i=1}^n (\gamma_i(\boldsymbol{\theta}) - \gamma_i(\boldsymbol{\theta}_0))^2 \rightarrow_p 0.$$

Since  $\alpha_n^{-1} \sum_{i=1}^n (\gamma_i(\boldsymbol{\theta}) - \gamma_i(\boldsymbol{\theta}_0))^2$  is locally uniformly bounded away from zero conditions given in (4) hold and ARE minimization for this problem is

equivalent to MPLL estimation. For this model, the consistency conditions are also satisfied [7].

### Dynamics of the LRE algorithm

In [1], we consider the binary case and show that  $\tilde{\mathcal{L}}_n(\boldsymbol{\theta}) = -\bar{\mathcal{L}}_n(\boldsymbol{\theta})$  where  $\tilde{\mathcal{L}}_n$  is the stochastic relative entropy (SRE) cost function which results when we employ first order stochastic approximations for the true conditionals. In this paper, we consider the sigmoidal perceptron as a special case of the new formulation and establish the equivalence of ARE minimization to MPLL estimation under the conditions given in (4). The parameters of the conditional pmf model can be estimated by minimizing the SRE or by maximizing PLL in a number of ways, gradient descent (ascent) learning is one popular alternative. We derive the least relative entropy (LRE) algorithm by gradient descent minimization of the SRE cost function for the single layer perceptron and show that it successfully equalizes multipath channels in [1]. The general formulation for distribution learning with the MLP model is presented in [2] and is applied to the equalization of nonlinear channels.

The properties of gradient descent learning on the SRE cost function is considered in [1, 2]. Particularly, it is shown that the SRE cost function for single layer perceptron is a *well-formed* cost function in the sense of Wittner and Denker [11] and hence gradient descent learning on this cost function is guaranteed to find a solution. As is well known, there is no such guarantee with the MSE cost function when used on MLP's, even on those without any hidden units. The dynamics of gradient descent learning on the SRE cost function is also studied by considering its parameter updates [2] and it is shown that for LRE updates the backpropagated output error is always a non-vanishing control signal and hence the algorithm can recover from convergence at the wrong extreme while the MSE based MLP can not. In this paper, we present a simulation study to demonstrate this fact.

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  $0.5(\cdot)^3$ , and the PAM communication system has 8 bits per sample with Nyquist pulse shaping. We implement the LRE algorithm for binary alphabet given in [2] 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. In Figure 1(a), 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 notable is that when we introduce an abrupt change (an exact sign change) in the channel characteristics after 150 iterations, causing the decision region to rotate suddenly the LRE can very rapidly adapt

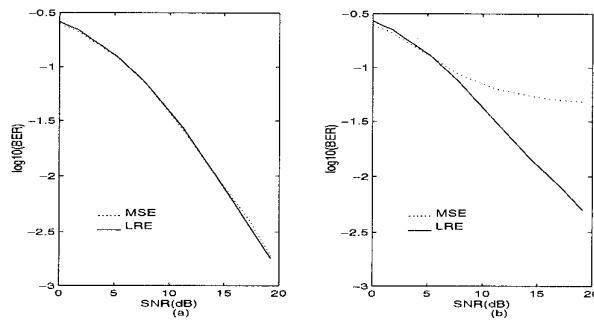


Figure 1: BER Comparison for MSE and LRE MLP Equalizers  
 (a) without (b) with an abrupt change at 150 iterations

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 1(b), MSE produces many wrong decisions before it can adapt to this new operating condition. In Figure 2(a), we show the transient characteristics of both algorithms with the abrupt change at 150 iterations at a signal to noise ratio (SNR) of 19 dB. As seen in the figure, LRE can recover from convergence at the wrong extreme very effectively whereas MSE based MLP needs a considerable amount of time for the same task. 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 Figure

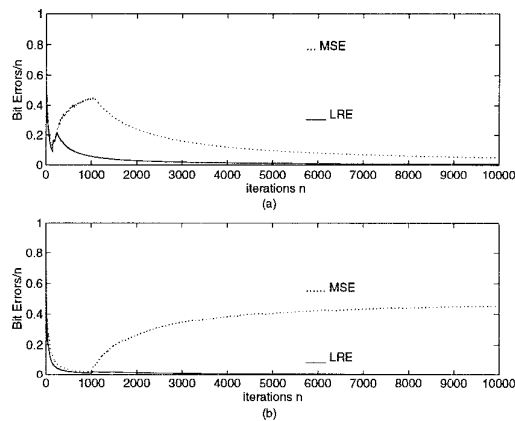


Figure 2: Recovery Characteristics for MSE and LRE MLP Equalizers  
 with an abrupt change at (a) 150 (b) 1000 iterations (SNR = 19 dB)

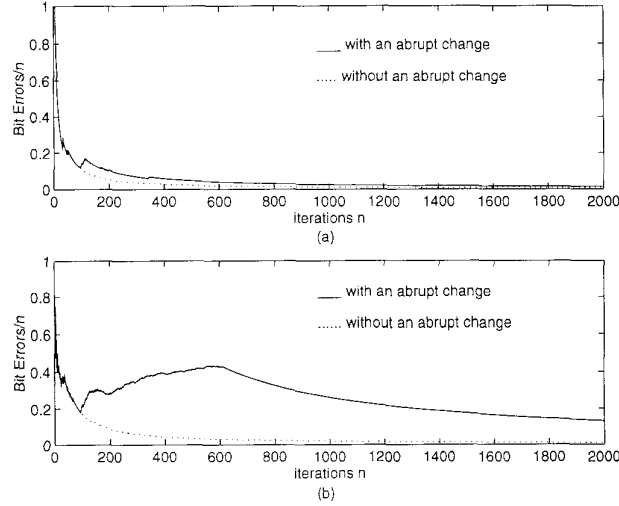


Figure 3: Recovery and Convergence Characteristics for (a) LRE (b) MSE MLP Equalizers (with abrupt change at  $n = 100$ , SNR = 19 dB)

2(b) by introducing the sudden change at iteration 1000. Again LRE can very rapidly adapt to the new operating condition, rapidly recovering from convergence at the wrong extreme. Figures 3 and 4 show the convergence and recovery characteristics of both MLP equalizers (ARE and MSE based) with and without the abrupt change when the change occurs at 100 and 1000 iterations respectively.

## APPENDIX

*Proof of Theorem 1:* Let  $\mathcal{R}_n = \sum_{i=1}^n r_i$ , by Theorem 2A [12]

$$\mathcal{R}_n / \mathcal{I}_n \xrightarrow{p} 1.$$

Therefore,  $\forall \epsilon > 0$ ,  $\exists N$  for  $n \geq N$ , for any  $\theta \in \Theta$ , we have

$$\mathcal{I}_n(\theta_n^*)(1 - \epsilon) < \mathcal{R}_n(\bar{\theta}_n^*) < \mathcal{I}_n(\theta_n^*)(1 + \epsilon)$$

$$\mathcal{I}_n(\theta_n^*)(1 - \epsilon) < \mathcal{R}_n(\theta_n^*) < \mathcal{I}_n(\theta_n^*)(1 + \epsilon)$$

where  $\theta_n^*$  and  $\bar{\theta}_n^*$  are the values which minimize  $\mathcal{I}_n$  and  $\mathcal{R}_n$  respectively. Since  $\epsilon$  is arbitrary, for sufficiently large  $N$ , we have

$$\mathcal{I}_n(\theta_n^*) = \mathcal{R}_n(\bar{\theta}_n^*) = \mathcal{R}_n(\theta_n^*)$$

almost surely on  $\Theta$ . In addition, we can express  $\mathcal{R}_n(\theta_n^*)$  as

$$\mathcal{R}_n(\theta_n^*) = \bar{\mathcal{L}}_n(\theta_0) - \bar{\mathcal{L}}_n(\theta_n^*)$$

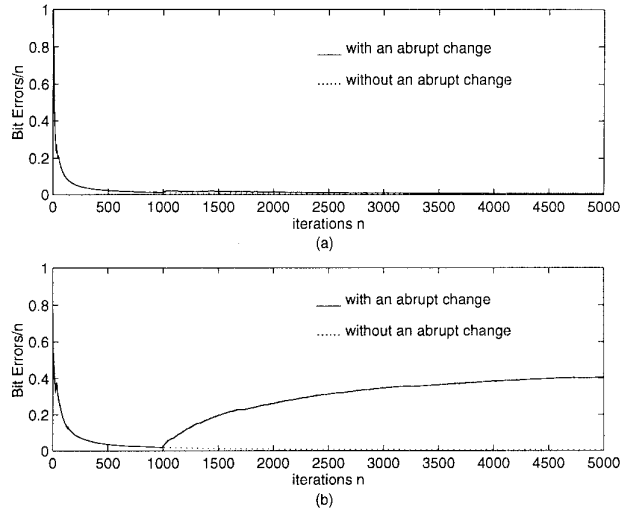


Figure 4: Recovery and Convergence Characteristics for (a) LRE (b) MSE MLP Equalizers (with abrupt change at  $n = 1000$ , SNR = 19 dB)

which implies:  $\arg(\min_{\theta} \mathcal{I}_n) = \arg(\max_{\theta} \bar{\mathcal{L}}_n)$ .

*Proof of Theorem 2:* We have  $V_n(\theta) = \nabla \nabla^T \mathcal{L}_n(\theta)$  (6). Since conditions given in (5) hold,  $\bar{\mathcal{L}}_n(\theta)$  is concave with respect to  $\theta$  (Theorem 2E [12]). Therefore  $\hat{\theta} \rightarrow \theta_0$  in probability.

Let  $\mathcal{S}_n(\theta) = \sum_{i=1}^n u_i(\theta)$ , then, we have

$$E(u_i | \mathcal{F}_i) = 0 \quad \text{and} \quad U_n(\theta) = - \sum_{i=1}^n E(v_i | \mathcal{F}_i)$$

therefore  $\mathcal{S}_n(\theta)$  is a martingale. By considering the first two terms in the Taylor expansion of  $\nabla \bar{\mathcal{L}}_n(\theta_0)$  we can write

$$0 = \nabla \bar{\mathcal{L}}_n(\theta_0) \approx \mathcal{S}_n(\theta_0) + (\hat{\theta} - \theta_0) V_n(\theta_0)$$

then

$$\sqrt{\beta_n} (\hat{\theta} - \theta_0) \approx \left( \beta_n^{-\frac{1}{2}} \mathcal{S}_n(\theta_0) \right) \left( -\beta_n V_n(\theta_0)^{-1} \right).$$

Since the conditions given in (5) are satisfied and that  $\mathcal{S}_n(\theta)$  is a martingale, by invoking Martingale central limit theorem [3], we have

$$\beta_n^{-\frac{1}{2}} \mathcal{S}_n(\theta_0) \rightarrow_D \mathcal{N}[0, Q_1]$$

therefore

$$\sqrt{\beta_n}(\hat{\theta} - \theta_0) \rightarrow_D \mathcal{N}[0, Q^{-1}Q_1Q^{-1}].$$

## REFERENCES

- [1] 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.
- [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* (Detroit, Michigan), May 1995, pp. 929-932.
- [3] B. M. Brown, "Martingale central limit theorems," *Ann. Math. Statist.*, vol. 44, pp. 59-66, 1971.
- [4] D.R. Cox, "Partial likelihood," *Biometrika*, vol. 62, pp. 69-72, 1975.
- [5] 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.
- [6] 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.
- [7] B. Kedem, "Time series analysis by higher order crossings", IEEE press, New York, N.Y., 1994.
- [8] L. Kullback, and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics* 22, pp. 79-86, 1951.
- [9] M. Meyer and G. Pfeiffer, "Multilayer perceptron based decision feedback equalisers for channels with intersymbol interference," *IEE Proceedings*, Vol. 140, No.6, pp. 420-424, 1993.
- [10] H. White, "Learning in artificial neural networks: A statistical perspective," *Neural Computation*, vol. 1, pp. 425-464, 1989.
- [11] B.S. Wittner and J.S. Denker, "Strategies for teaching layered networks classification tasks," *Neural Info. Proc. Systems*, (Denver, CO), p.850-859, 1988.
- [12] W. H. Wong, "Theory of partial likelihood," *Ann. Statist.*, 14, pp. 88-123, 1986.