A LIKELIHOOD FRAMEWORK FOR NONLINEAR SIGNAL PROCESSING WITH FINITE NORMAL MIXTURES

Tülay Adalı, Bo Wang, Xiao Liu and Jianhua Xuan
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County, Baltimore, MD 21250
Tel: +1 410 4551000/3521; fax: +1 410 4553969
e-mail: adali@engr.umbc.edu

ABSTRACT

We introduce a likelihood framework for nonlinear signal processing using partial likelihood and use the result to derive the information geometric *em* algorithm for distribution learning through information-theoretic projections. We demonstrate the superior convergence of the *em* algorithm as compared to least relative entropy (LRE) algorithm by simulations. The performance of finite normal mixtures (FNM) based equalizers with different number of mixtures and different dimension observation vectors is also discussed.

1 INTRODUCTION

The increasing demand for digital communication systems to operate at higher data and lower bit error rates has emphasized the need for sophisticated signal processing schemes which can function in non-linear, non-stationary environments. To overcome the inherent limitation of linear filters, among other nonlinear techniques, a number of neural network based signal processing systems have been introduced (for a recent collection of these applications see e.g. [9],[10]). These systems have provided significant performance improvements especially when the underlying process involves nonlinearities and/or the signal-to-noise ratio (SNR) is poor. Among these approaches, radial basis functions (RBF) have found unique application in communications, (e.g. in interference rejection [6], [8], and channel equalization [7]) and have been noted for their ability to approximate the optimal Bayesian decision boundary. In this paper, based on a FNM probability model, which is closely related to the RBF, we introduce a likelihood framework for nonlinear signal processing. We use a recent extension of maximum likelihood (ML), *partial likelihood* (PL), as the cost function, which allows for sequential processing of dependent observations to develop the probabilistic framework for signal processing with FNM. We have shown that the two conditions given in [1] for the equivalence of *accumulated* relative entropy (ARE) and *maximum partial likelihood* (MPL) are satisfied for the FNM [2]. In [12], FNM are applied to channel equalization. However, for estimating the FNM parameters, the batch expectation-maximization (EM) scheme is used which is not suitable for an application such as channel equalization which has to be ideally on-line. In this paper, based on the FNM, we derive the *on-line* information geometric *em* algorithm such that PL is maximized (or relative entropy is minimized). We demonstrate the superior performance of the *em* algorithm as compared to gradient descent based LRE algorithm by simulations. We also discuss the performance of the FNM based equalizer with different number of mixtures and different dimension observation vectors.

2 MPL FOR SIGNAL PROCESSING WITH FNMS

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 use a conditional distribution learning framework for real-time signal processing based on the partial likelihood theory. Obtained as a partial factorization of the full likelihood, PL possesses nice large sample properties of ML, and more importantly, it can easily be characterized for dependent data and easily used for 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 observations be known for the whole period (i.e., including future observations). In these cases, the learning algorithm for conditional likelihood must be in batch mode. PL, thus provides us with a particularly suitable formation for real-time signal processing, which most of the time requires on-line processing of dependent observations.

We can introduce the partial likelihood as follows: Given a time series $\{x_n\}$, $n = 0, 1, 2, \cdots$ that takes values from a finite alphabet $\mathcal{S} = \{a_0, a_1, ..., a_M\}$, and its time-dependent covariates (observations) $\{\mathbf{y}_n\}$, esti-

mate the probability that x_n takes a value from the given alphabet $\mathcal{S}$. We assume $\mathcal{F}_n = \sigma\{1, [x_{n-1}, \cdots, x_1, x_0], [\mathbf{y}_n, \cdots, \mathbf{y}_1, \mathbf{y}_0]\}$. Our aim is to estimate the conditional probabilities: $P(x_n = a_i|\mathcal{F}_n)\ \forall a_i \in \mathcal{S}$. The *partial likelihood* can then be written as

$$\mathcal{L}^p(\mathbf{x}_n; \boldsymbol{\theta}) \equiv \mathcal{L}^p_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} p_\theta(x_i|\mathcal{F}_i) \tag{1}$$

where $\mathbf{x}_n = [x_n, \cdots, x_1]$.

The relative entropy (RE), or the Kullback-Leibler distance $D_n(p\|p_\theta)$ [11], is a fundamental information-theoretic measure of how accurate the estimated conditional pmf $p_\theta(x_n|\mathcal{F}_n)$ is an approximation to the true conditional pmf $p(x_n|\mathcal{F}_n)$. The ARE can be defined as $\mathcal{I}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} D_i(p_{\theta_0}\|p_\theta)$. It is relatively easy to demonstrate the equivalence of ML estimation to ARE minimization when the observations are i.i.d. For the neural network model defined in [1], we established the equivalence of PL estimation to ARE minimization for the general case of dependent observations. These two conditions are: (1) the asymptotical stability of variance and (2) the condition on the rate by which information accumulates. In [2], we show that the equivalence of MPL to ARE minimization is also valid for the FNM model. Therefore, we can estimate/learn the parameters of the FNM model directly by PL maximization, which minimizes the ARE distance between the true and estimated conditional probabilities.

3 INFORMATION GEOMETRY OF MAXIMUM PARTIAL LIKELIHOOD ESTIMATION

To construct the information geometry of PL estimation such that the FNM parameters can be learned by sequential updates, we proceed as follows: Given an information source from a certain environment, the set of all related probability distributions form the manifold $\mathcal{S}$. The set of distributions which are realizable by a *selected neural network structure* is embedded, as a submanifold $\mathcal{M}$, in $\mathcal{S}$. On the other hand, the distributions suggested by the observed *partial data* form a submanifold $\mathcal{D}$ in $\mathcal{S}$. The problem can then be posed as finding a conditional probability model (neural network) that minimizes the distance between the *realizable* $\mathcal{M}$ and the *observed* $\mathcal{D}$. A suitable distance measure in this framework is relative entropy. This minimization problem can be solved by *em* algorithm, an alternating minimization of the RE, which is proposed by Csiszár and Tusnády [5]. The network in $\mathcal{M}$ that minimizes the distance is selected as the desired one. Then, the point in $\mathcal{D}$ that minimizes the divergence gives the estimated data completing the partial observed data. Repeatedly applying above procedure produces a sequence of neural networks, each with the same parametrized structure but with different parameter values. It can be shown that this procedure will converge to the infimum distance between $\mathcal{M}$ and $\mathcal{D}$ if $\mathcal{M}$ and $\mathcal{D}$ are convex sets with finite measures [5].

Assume that true distribution of channel output vectors are included in a curved exponential family $\mathcal{M}$ and the observed data is in manifold $\mathcal{D}$. It can be shown that for a given $Q \in \mathcal{D}$, the point $\hat{P} \in \mathcal{M}$ that maximizes the partial likelihood is given by the m-projection of Q onto $\mathcal{M}$. Dual to the above statement, for a given $P \in \mathcal{M}$, the point $\hat{Q} \in \mathcal{D}$ that maximizes the partial likelihood is given by the e-projection of P onto $\mathcal{D}$. Hence we can formulate the geometric em-algorithm [3] (e- and m-projection algorithm) for maximum partial likelihood estimation as follows:

Consider a FNM with hidden variable z (the index within the mixture pdf) written as:

$$p(\mathbf{y}, z) = \sum_{i=0}^{N} \frac{\delta_i(z)\pi_i}{(\sqrt{2\pi})^d |\Sigma_i|^{d/2}} exp\{-\frac{1}{2}(\mathbf{y} - \mu_i)^T \Sigma_i^{-1}(\mathbf{y} - \mu_i)\}, \tag{2}$$

where d is the dimension of the observation vector $\mathbf{y}$ and $\delta_i(z)$ is the component index of the mixture model. We proceed by writing the logarithm of the probability distribution as:

$$\begin{aligned}
\mathcal{P}(\mathbf{y}, z) &= \log p(\mathbf{y}, z) \\
&= \mu_0{}^T \Sigma_0^{-1}\mathbf{y} - \frac{1}{2}\mathbf{y}^T\Sigma_0^{-1}\mathbf{y} \\
&\quad + \sum_{i=1}^{N} \delta_i(z)(\log\frac{\pi_i}{\pi_0} - \log\frac{|\Sigma_i|^{1/2}}{|\Sigma_0|^{1/2}} \\
&\quad - \frac{1}{2}\mu_i^T\Sigma_i^{-1}\mu_i + \frac{1}{2}\mu_0^T\Sigma_0^{-1}\mu_0) \\
&\quad + \sum_{i=1}^{N} \delta_i(z)\mathbf{y}^T(\Sigma_i^{-1}\mu_i - \Sigma_0^{-1}\mu_0) \\
&\quad - \sum_{i=1}^{N} \delta_i(z)\mathbf{y}^T(\frac{1}{2}\Sigma_i^{-1} - \frac{1}{2}\Sigma_0^{-1})\mathbf{y} \\
&\quad + \log\frac{\pi_0}{|\Sigma_0^{-1}|^{1/2}} \\
&\quad - \frac{1}{2}\mu_0^T\Sigma_0^{-1}\mu_0 - \log(\sqrt{2\pi})^d.
\end{aligned} \tag{3}$$

This is a generalization of the construction given in [3] to the multidimensional case. Note that in the channel equalization application, all components of $\mathbf{y}_i$ are conditionally independent given the input sequence $\{x_i\}$ as a consequence of the statistical independence of the additive noise. So, the covariance matrix Σ_i is a diagonal matrix, $\Sigma_i = diag[\sigma_{i0}^2, \cdots, \sigma_{i(d-1)}^2], i = 1, \cdots, N$. Let $\zeta_i = [\sigma_{i0}^{-2}, \cdots, \sigma_{i(d-1)}^{-2}]^T, i = 1, \cdots, N$, $\xi_i = [\sigma_{i0}^2, \cdots, \sigma_{i(d-1)}^2]^T, i = 1, \cdots, N$, $\mathbf{y} = [y_0, \cdots, y_{d-1}]^T$ and $\mathbf{Y} = [y_0^2, \cdots, y_{d-1}^2]^T$ to write

$$
\begin{array}{ll}
\mathbf{r}_{11} = \mathbf{y}, & \theta_{11} = {\mu_0}^T \Sigma_0^{-1}, \\
\mathbf{r}_{12} = \mathbf{Y}, & \theta_{12} = \frac{1}{2}\zeta_0, \\
\mathbf{r}_{2i} = \delta_i(z), & \theta_{2i} = \log \frac{\pi_i}{\pi_0} - \log \frac{|\Sigma_i|^{1/2}}{|\Sigma_0|^{1/2}} \\
 & -\frac{1}{2}\mu_i^T \Sigma_i^{-1} \mu_i + \frac{1}{2}\mu_0^T \Sigma_0^{-1} \mu_0, \\
\mathbf{r}_{3i} = \delta_i(z)\mathbf{y}, & \theta_{3i} = \Sigma_i^{-1}\mu_i - \Sigma_0^{-1}\mu_0, \\
\mathbf{r}_{4i} = \delta_i(z)\mathbf{Y}, & \theta_{4i} = \frac{1}{2}\zeta_i - \frac{1}{2}\zeta_0,
\end{array}
\tag{4}
$$

where $i = 1, \cdots, N$.

The expectation parameter, $\eta = E_\theta(\mathbf{r})$, called the η-coordinates of $\mathcal{M}$, can be represented as

$$
\begin{array}{l}
\eta_{11} = \sum_{i=1}^N \pi_i \mu_i, \\
\eta_{12} = \sum_{i=1}^N \pi_i (\omega_i + \xi_i), \\
\eta_{2i} = \pi_i, \\
\eta_{3i} = \pi_i \mu_i, \\
\eta_{4i} = \pi_i (\omega_i + \xi_i),
\end{array}
\tag{5}
$$

where $\mu_i = [\mu_{i0}, ..., \mu_{i(d-1)}]^T$, and $\omega_i = [\mu_{i0}^2, ..., \mu_{i(d-1)}^2]^T$.

With the above θ- and η-coordinates, we can represent the information geometric *em*-algorithm as follows:

- Select an arbitrary initial vector $\hat{\mathbf{u}}_0$, which gives the initial distribution $\hat{P}_0 \in \mathcal{M}$. Set $t = 0$. Repeat the following two steps:
  - *e*-step: Calculate the *e*-projection of the present $\hat{P}_t$ onto D_{t+1}. Because at time $t+1$, $\mathbf{y}_{t+1}$ is observed but z_{t+1} is not observed, the observed data sub-manifold D_{t+1} is given by

$$
\begin{array}{rcl}
D_{t+1} & = & \{\hat{\mathbf{r}}_{t+1} | \hat{\mathbf{r}}_{11}^{t+1} = \mathbf{y}_{t+1}, \hat{\mathbf{r}}_{12}^{t+1} = \mathbf{Y}_{t+1}, \\
 & & \hat{\mathbf{r}}_{2i}^{t+1} = \alpha_i, \hat{\mathbf{r}}_{3i}^{t+1} = \alpha_i \mathbf{y}_{t+1}, \\
 & & \hat{\mathbf{r}}_{4i}^{t+1} = \alpha_i \mathbf{Y}_{t+1}\},
\end{array}
$$

    where α_is are the free parameters corresponding to the unobserved $\delta_i(z_{t+1})$, which can be estimated as

$$
\begin{array}{rcl}
\alpha_i & = & E_{\hat{P}_t}(\delta_i(z_{t+1}) | \mathbf{y}_{t+1}) \\
 & = & \dfrac{\hat{\pi}_i \exp\{-\frac{1}{2}(\mathbf{y}_{t+1} - \hat{\mu}_i)^T \Sigma^{-1} (\mathbf{y}_{t+1} - \hat{\mu}_i)\}}{\sum_j \hat{\pi}_j \exp\{-\frac{1}{2}(\mathbf{y}_{t+1} - \hat{\mu}_j)^T \Sigma^{-1} (\mathbf{y}_{t+1} - \hat{\mu}_j)\}}
\end{array}
$$

    Then, using $\hat{\mathbf{r}}_{t+1}$ calculated above, we modify the η-coordinates by

$$
\hat{\eta}_{t+1} = (1 - \epsilon_t)\hat{\eta}_t + \epsilon_t \hat{\mathbf{r}}_{t+1}, \tag{6}
$$

    where ϵ_t is the learning rate selected as a decreasing sequence.
  - *m*-step: Use the gradient method to get the next $\hat{\mathbf{u}}_{t+1}$ by

$$
\hat{\mathbf{u}}_{t+1} = \hat{\mathbf{u}}_t + \epsilon_t \mathbf{B}[\hat{\eta}_{t+1} - \eta(\hat{\mathbf{u}}_t)], \tag{7}
$$

    where $\mathbf{B} = \frac{\partial}{\partial \mathbf{u}} \theta(\mathbf{u})$ is the gradient matrix.
- Increment the iteration index, $t = t + 1$.

The given *em*-algorithm for MPL provides good estimates of the optimum behavior through its *e*-projections on the set of desirable distributions $\mathcal{D}$. Each of these *e*-projections is then used by *m*-projection to find the corresponding best neural network. The algorithm can be thought of consisting of two parts [4]: One part provides estimate of the best network behavior and the other part finds a neural network whose behavior closely approximates this estimation. Hence, information geometric *em* algorithm provides us not only with a new learning algorithm, but also a method to understand the learning process.

4 APPLICATION TO CHANNEL EQUALIZATION

In this section, we present application of the likelihood framework for nonlinear signal processing with FNM to adaptive channel equalization. We consider transmission of simple binary pulse amplitude modulated data $x(n) \in \{-1, 1\}$ through a nonlinear nonminimum phase channel such that the received signal is given by $y_l(n) - 0.2y_l^2(n)$ where $y_l(n) = 0.3482x(n) + 0.8704x(n-1) + 0.3482x(n-2)$. In the first experiment, the observation vector at time n consists of $y(n)$ and $y(n-1)$, and 120 training samples are used to train the FNM equalizer with 16 normal distributions using the *em* algorithm such that the partial likelihood given by (1) is maximized. The former part of the training data are used to initialize the 16 normal distributions. The average values of the first 3 observed vectors which belong to the same normal distribution are assigned as the means of the 16 normal distributions. The remaining 72 training data are used to train the FNM equalizer sequentially.

The performance of the FNM equalizer is compared with that of multilayer perceptron (MLP) equalizer of similar complexity. A 2-18-1 perceptron equalizer is trained by the traditional mean square error (MSE) and the partial likelihood costs with 1000 training samples. Number of training samples is chosen such that the algorithm converges. There is also one delay in the MLP equalizer. Fig. 1 shows the bit error rate (BER) curves for the three cases which are averaged over 50 independent runs. The information geometric *em* algorithm based on FNM probability model can achieve much faster convergence compared to those of the backpropagation (MLP with MSE cost) and the LRE [1] (MLP with MPL cost) algorithms while it still can achieve better BERs.

In the second simulation example, we address the problem of correct network complexity determination (order selection for the FNM model and the observation vector). We consider FNM models with 8, 16, and 32 normal components respectively. When the FNM model has 8 mixtures and the observation vector is two dimenstional, after 100 training samples, the equalizer

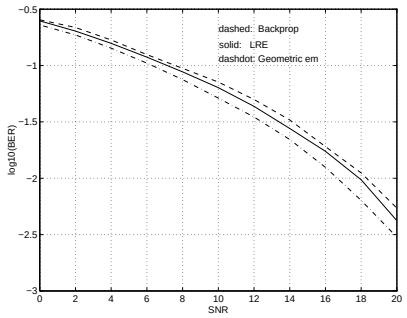

Figure 1: BER curves for the geometric em (FNM with 16 normal mixtures) algorithm, the backpropgation (2-18-1 MLP) and the LRE (2-18-1 MLP) algorithms

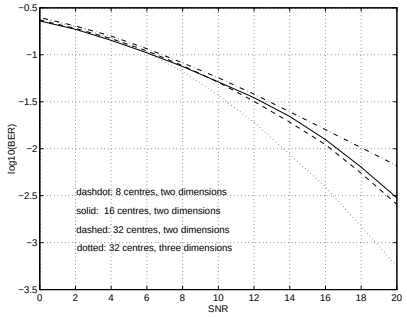

Figure 2: BER curves for the geometric em for FNM with different number of mixtures and different dimensions of observation vector

converges. For the FNM model with 32 mixtures and two dimenstional observation vectors, 180 training samples are used for training. In Fig. 2, the BER curve for FNM equalizer with 8 mixtures and two-dimension observation vectors is quite close to that with 16 mixtures and the same dimension observation vectors, especially at low SNR values, as expected. The BER curve obtained by the FNM equalizer with 32 mixtures and two-dimension observation vectors is slightly better than the 16 mixture FNM equalizer with same dimension observation vectors at high SNR values but performs slightly worse at low SNRs as overparametrization is likely to generate problems for generalization at increased noise levels. When we increase the dimension of observation vectors from 2 to 3 for FNM equalizer with 32 mixtures, but using same 180 training samples, the BER improves a lot at high SNR values. This is due to the increase of the minimum distance between noise-free centres of FNM when the dimension of observation vector increases.

References

[1] T. Adalı, X. Liu, and M. K. Sönmez, "Conditional distribution learning with neural networks and its application to channel equalization," *IEEE Trans. Signal Processing*, vol. 45, no. 4, pp. 1051-1064, Apr. 1997.

[2] T. Adalı, B. Wang, X. Liu, and J. Xuan, "A likelihood framework for nonlinear signal processing with finite normal mixtures," submitted to *IEEE Trans. Signal Processing*.

[3] S. Amari, "Information Geometry of the EM and em Algorithms for Neural Networks," *Neural Networks*, vol. 8, No. 9, pp. 1379-1408, 1995.

[4] W. Byrne, "Alternating minimization and Boltzmann machine learning," *IEEE Trans. Neural Networks*, vol. 3, no. 4, pp. 612-620, 1992.

[5] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedure," in *Statistics and decisions, Supplementary issue, No. 1,* (E. Dedewicz *et. al.*,eds.), pp. 205-237, Munich, Oldenburg Verlag, 1984.

[6] I. Cha and S.A. Kassam, "Interference Cancellation Using Radial Basis Function Networks," *Signal Processing*, vol. 47, no. 3, pp. 247-268, Dec, 1995.

[7] S. Chen, G.J. Gibson, C.F.N. Cowan, and P.M. Grant, "Reconstruction of Binary Signals Using an Adaptive Radial Basis Function Equalizer," *Signal Processing*, vol 22, no. 2, pp. 77-93, 1991.

[8] S. Chen and B. Mulgrew, "Overcoming Co-channel Interference Using an Adaptive Radial Basis Function Equalizer," *Signal Processing*, vol. 28, no. 1, pp. 77-93, Jul., 1995.

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

[10] J. Principe, L. Giles, N. Morgan, and E. Wilson, *Neural Networks for Signal Processing VII,* Proc. IEEE Workshop, Amelia Island, FL, 1997.

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

[12] L. Xu, "Channel Equalization by Finite Mixtures and the EM Algorithm," *Proc. IEEE Workshop on Neural Networks for Signal Processing*, pp. 603-612, Boston, MA, 1995.