

---

# PARTIAL LIKELIHOOD FOR REAL-TIME SIGNAL PROCESSING WITH FINITE NORMAL MIXTURES

Bo Wang, Tülay Adalı, Xiao Liu, and Jianhua Xuan

Department of Computer Science and Electrical Engineering  
University of Maryland Baltimore County, Baltimore, MD 21250  
{bwang1, adali}@enr.umbc.edu

**Abstract-** We introduce a unified framework for nonlinear signal processing with finite normal mixtures (FNM) by using *maximum partial likelihood* (MPL) theory. We show that the equivalence of MPL to *accumulated relative entropy* (ARE) minimization is valid for the FNM. Then, we define the information geometry of MPL and use the result to derive the *em* algorithm for distribution learning based on the FNM model. The superior convergence of the *em* algorithm as compared to the least relative entropy (LRE) and the backpropagation algorithms is demonstrated by simulations. We also discuss the performance of the FNM based equalizers with different number of mixtures and observation vector sizes.

## I. 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, and 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. [8],[9]). 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 [5], [7], and channel equalization [6]) and have been noted for their ability to approximate the optimal Bayesian decision boundary.

In this paper, based on the FNM model, which is closely related to the RBF, we introduce an information geometric 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 unified framework for signal processing with FNM. We show that the two conditions given in [1] for the equivalence of ARE and MPL are satisfied for the FNM. In [11], 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. 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 [1] by simulations. We also discuss the performance of the FNM based equalizer with different number of mixtures and different dimension observation vectors.

## II. MPL FOR SIGNAL PROCESSING WITH FNM

Statistical parameter estimation theory has as its fundamental support 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  $\{x_n\}$ ,  $n = 0, 1, 2, \dots$ , that takes values from a finite alphabet  $\mathcal{S} = \{a_0, a_1, \dots, a_M\}$ , and its time-dependent covariates (observations)  $\{y_n\}$ , estimate the probability that  $x_n$  takes a value from the given alphabet  $\mathcal{S}$ . We assume  $\mathcal{F}_n = \sigma\{1, [x_{n-1}, \dots, x_1, x_0], [y_n, \dots, y_1, y_0]\}$ . The *partial likelihood* can then be written as

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

where  $\mathbf{x}_n = [x_n, \dots, x_1]$ , and  $\theta$  is the parameter set for the selected probability mass function (pmf) model  $p_\theta(x_i | \mathcal{F}_i)$ . Our goal is to estimate the conditional probabilities  $p_\theta(x_i | \mathcal{F}_i)$ , which can be used in a number of ways depending on the application.

The relative entropy (RE), or the Kullback-Leibler distance  $D_n(p_{\theta_0} \| p_\theta)$  [10], 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  $p_{\theta_0}(x_n | \mathcal{F}_n)$ . The accumulated RE can be defined as

$$\mathcal{I}_n(\theta) = \sum_{k=1}^n D_k(p_{\theta_0} \| p_\theta) = \sum_{k=1}^n i_k(\theta)$$

$$= \sum_{k=1}^n \sum_{a_j \in \mathcal{S}} P_{\theta_0}(x_k = a_j | \mathcal{F}_k) \ln \frac{P_{\theta_0}(x_k = a_j | \mathcal{F}_k)}{P_{\theta}(x_k = a_j | \mathcal{F}_k)} \quad (2)$$

It is relatively easy to demonstrate the equivalence of ML estimation to ARE minimization when the observations are i.i.d. However, the independence condition is very restrictive for almost all practical applications. Also, PL can process data as they become available, requiring only the information present at a given time. The PL estimation is equivalent to ARE minimization for a selected pmf model if the two conditions in the following theorem [1] are satisfied.

*Theorem:* If there exists a constant  $\delta > 0$  and a continuous function  $p_{\theta}$  such that for each  $\theta \neq \theta_0$ , as  $n \rightarrow \infty$ ,

$$P(\mathcal{I}_n(\theta)/n > \delta) \rightarrow 1 \quad (3)$$

and

$$\mathcal{J}_n(\theta)/n^2 \rightarrow 0 \text{ in probability} \quad (4)$$

then at least one  $\arg \min_{\theta} \mathcal{I}_n(\theta)$  tends to one  $\arg \max_{\theta} \bar{\mathcal{L}}_n(\theta)$  almost surely on  $\Omega = \{\theta \mid \mathcal{I}_n(\theta) \uparrow \infty, \sum_{i=1}^n j_i(\theta)/\mathcal{I}_i^2(\theta) < \infty\}$  where  $\bar{\mathcal{L}}_n(\theta) \equiv \ln \mathcal{L}_n(\theta)$ ,  $\mathcal{J}_n(\theta) = \sum_{k=1}^n j_k(\theta)$ ,  $j_k(\theta) = \text{Var}\{r_k(\theta) | \mathcal{F}_k\}$ , and  $r_k(\theta) = \ln \frac{p_{\theta_0}(x_k | \mathcal{F}_k)}{p_{\theta}(x_k | \mathcal{F}_k)}$ .

The two conditions defined in (3) and (4) are the asymptotical stability of variance and the condition on the rate by which information accumulates. In [1], we prove that these two ARE-PL equivalence conditions are satisfied for the multi-layer perceptron (MLP) probability model. In this paper, we show that they also hold for the FNM model. Let us consider a binary distribution which can be expressed as

$$p_{\theta}(x_n | \mathbf{y}_n) = P_{\theta}(x_n = 1 | \mathbf{y}_n)^{x_n} P_{\theta}(x_n = 0 | \mathbf{y}_n)^{1-x_n} \quad (5)$$

where  $x_n = \{0, 1\}$ . Note that since a feedforward network structure is assumed,  $p_{\theta}(x_n | \mathcal{F}_n) = p_{\theta}(x_n | \mathbf{y}_n)$ . Using Bayes' theorem, we can write the conditional probabilities as

$$P_{\theta}(x_n = i | \mathbf{y}_n) = \frac{p_{\theta}(\mathbf{y}_n | x_n = i) P(x_n = i)}{p_{\theta}(\mathbf{y}_n)} \quad i = 0, 1 \quad (6)$$

We assume a FNM distribution model for  $p_{\theta}(\mathbf{y}_n | x_n = i)$  in (6), i.e., we let

$$p_{\theta}(\mathbf{y}_n | x_n = 1) = \sum_{i=1}^{N_1} \frac{\pi_i}{(\sqrt{2\pi})^d |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}_n - \mu_i)^T \Sigma_i^{-1} (\mathbf{y}_n - \mu_i)\right\} \quad (7)$$

$$p_{\theta}(\mathbf{y}_n | x_n = 0) = \sum_{j=N_1+1}^{N_1+N_2} \frac{\pi_j}{(\sqrt{2\pi})^d |\Sigma_j|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}_n - \mu_j)^T \Sigma_j^{-1} (\mathbf{y}_n - \mu_j)\right\} \quad (8)$$

where  $i = 1, \dots, N_1$  and  $j = N_1 + 1, \dots, N_1 + N_2$ . We proceed by writing the logarithm of  $p_\theta(x_n|\mathbf{y}_n)$  in (5) and using (6) to obtain

$$\begin{aligned} \ln p_\theta(x_n|\mathbf{y}_n) &= x_n \ln P_\theta(x_n = 1|\mathbf{y}_n) + (1 - x_n) \ln P_\theta(x_n = 0|\mathbf{y}_n) \\ &= x_n \left[ \ln \frac{P(x_n = 1)}{P(x_n = 0)} + \ln \frac{p_\theta(\mathbf{y}_n|x_n = 1)}{p_\theta(\mathbf{y}_n|x_n = 0)} \right] \\ &\quad + [\ln p_\theta(\mathbf{y}_n|x_n = 0) + \ln P(x_n = 0) - \ln p_\theta(\mathbf{y}_n)] \\ &= x_n a(\theta) + b(\theta) \end{aligned} \quad (9)$$

where  $a(\theta) = \ln \frac{P(x_n=1)}{P(x_n=0)} + \ln \frac{p_\theta(\mathbf{y}_n|x_n=1)}{p_\theta(\mathbf{y}_n|x_n=0)}$  and  $b(\theta) = \ln p_\theta(\mathbf{y}_n|x_n = 0) + \ln P(x_n = 0) - \ln p_\theta(\mathbf{y}_n)$ . Using the definition in the above theorem and Eqn. (9), we can get

$$\begin{aligned} r_n(\theta) &\equiv \ln \frac{p_{\theta_0}(x_n|\mathcal{F}_n)}{p_\theta(x_n|\mathcal{F}_n)} = \ln \frac{p_{\theta_0}(x_n|\mathbf{y}_n)}{p_\theta(x_n|\mathbf{y}_n)} \\ &= x_n(a(\theta_0) - a(\theta)) + (b(\theta_0) - b(\theta)) \end{aligned} \quad (10)$$

and for the binary case, we have

$$\begin{aligned} E[x_n|\mathcal{F}_n] &= E[x_n|\mathbf{y}_n] = \sum_{x_n \in \{0,1\}} x_n p_{\theta_0}(x_n|\mathbf{y}_n) \\ &= P_{\theta_0}(x_n = 1|\mathbf{y}_n) \equiv C(\theta_0). \end{aligned} \quad (11)$$

Using the result in (9), we can also write

$$\begin{aligned} i_n(\theta) &= E[r_n(\theta)|\mathcal{F}_n] = E\left[\ln \frac{p_{\theta_0}(x_n|\mathcal{F}_n)}{p_\theta(x_n|\mathcal{F}_n)}|\mathcal{F}_n\right] \\ &= E[x_n|\mathcal{F}_n][a(\theta_0) - a(\theta)] + [b(\theta_0) - b(\theta)] \\ &= C(\theta_0)[a(\theta_0) - a(\theta)] + [b(\theta_0) - b(\theta)] \end{aligned} \quad (12)$$

and

$$\begin{aligned} j_n(\theta) &= \text{Var}(r_n(\theta)|\mathcal{F}_n) = E[(r_n(\theta) - i_n(\theta))^2|\mathcal{F}_n] \\ &= E[(x_n - C(\theta_0))^2|\mathcal{F}_n](a(\theta_0) - a(\theta))^2 \\ &= (C(\theta_0) - C^2(\theta_0))(a(\theta_0) - a(\theta))^2 \end{aligned} \quad (13)$$

In both definitions, (12) and (13), the expectations are with respect to the true distribution  $p_{\theta_0}(x_n|\mathbf{y}_n)$ . From the definition of  $i_n(\theta)$ , we know  $i_n(\theta) > 0$ , for  $\theta \neq \theta_0$ . In addition, we assume  $\theta \in \Theta$ , where  $\Theta$  is a compact parameter set. Thus  $i_n(\theta)$  is finite. So there exists a constant  $\delta > 0$  such that, as  $n \rightarrow \infty$ ,

$$P(\mathcal{I}_n(\theta)/n > \delta) \rightarrow 1 \quad (14)$$

We also have  $j_n(\theta) \geq 0$  from its definition. Also  $\theta \in \Theta$ .  $C(\theta_0)$ ,  $a(\theta_0)$  and  $a(\theta)$  are finite, so  $j_n(\theta)$  is finite. Then we can get

$$\mathcal{J}_n(\theta)/n^2 = \sum_{k=1}^n j_k(\theta)/n^2 \rightarrow 0 \text{ in probability} \quad (15)$$

The two conditions in the above Theorem are satisfied for the FNM model, then at least one  $\arg \min_{\theta} \mathcal{I}_n(\theta)$  tends to one  $\arg \max_{\theta} \hat{\mathcal{L}}_n(\theta)$  almost surely on  $\Omega = \{\theta \mid \mathcal{I}_n(\theta) \uparrow \infty, \sum_{i=1}^n j_i(\theta)/\mathcal{I}_i^2(\theta) < \infty\}$ . 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.

### III. 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 the *em* algorithm, an alternating minimization of the RE, which is proposed by Csiszár and Tusnády [4]. 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 application of these two updates 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 [4].

Assume that the true distribution of the channel output vectors is included in a curved exponential family  $\mathcal{M}$  and the observed data are 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 [2] (*e*- and *m*- projection algorithm) for maximum partial likelihood estimation as follows:

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

$$p(\mathbf{y}, z) = \sum_{i=0}^N \frac{\delta_i(z)\pi_i}{(\sqrt{2\pi})^d |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mu_i)^T \Sigma_i^{-1} (\mathbf{y} - \mu_i)\right\} \quad (16)$$

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

$$\begin{aligned}
\mathcal{P}(\mathbf{y}, z) &= \ln p(\mathbf{y}, z) \\
&= \delta_0(z) \ln \left\{ \frac{\pi_0}{(\sqrt{2\pi})^d |\Sigma_0|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mu_0)^T \Sigma_0^{-1} (\mathbf{y} - \mu_0)\right) \right\} \\
&+ \sum_{i=1}^N \delta_i(z) \ln \left\{ \frac{\pi_i}{(\sqrt{2\pi})^d |\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mu_i)^T \Sigma_i^{-1} (\mathbf{y} - \mu_i)\right) \right\} \\
&= \ln \left\{ \frac{\pi_0}{(\sqrt{2\pi})^d |\Sigma_0|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mu_0)^T \Sigma_0^{-1} (\mathbf{y} - \mu_0)\right) \right\} \\
&+ \sum_{i=1}^N \delta_i(z) \left[ \ln \left\{ \frac{\pi_i}{(\sqrt{2\pi})^d |\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mu_i)^T \Sigma_i^{-1} (\mathbf{y} - \mu_i)\right) \right\} \right. \\
&\left. - \ln \left\{ \frac{\pi_0}{(\sqrt{2\pi})^d |\Sigma_0|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mu_0)^T \Sigma_0^{-1} (\mathbf{y} - \mu_0)\right) \right\} \right] \quad (17)
\end{aligned}$$

where we have used the condition that  $\sum_{i=0}^N \delta_i(z) = 1$ .  $\mathcal{P}(\mathbf{y}, z)$  can be further expressed as

$$\begin{aligned}
\mathcal{P}(\mathbf{y}, z) &= \mu_0^T \Sigma_0^{-1} \mathbf{y} - \frac{1}{2} \mathbf{y}^T \Sigma_0^{-1} \mathbf{y} + \sum_{i=1}^N \delta_i(z) \left( \ln \frac{\pi_i}{\pi_0} - \ln \frac{|\Sigma_i|^{1/2}}{|\Sigma_0|^{1/2}} \right. \\
&\quad \left. - \frac{1}{2} \mu_i^T \Sigma_i^{-1} \mu_i + \frac{1}{2} \mu_0^T \Sigma_0^{-1} \mu_0 \right) + \sum_{i=1}^N \delta_i(z) \mathbf{y}^T (\Sigma_i^{-1} \mu_i - \Sigma_0^{-1} \mu_0) \\
&\quad - \sum_{i=1}^N \delta_i(z) \mathbf{y}^T \left( \frac{1}{2} \Sigma_i^{-1} - \frac{1}{2} \Sigma_0^{-1} \right) \mathbf{y} + \ln \frac{\pi_0}{|\Sigma_0^{-1}|^{1/2}} \\
&\quad \left. - \frac{1}{2} \mu_0^T \Sigma_0^{-1} \mu_0 - \ln (\sqrt{2\pi})^d \right). \quad (18)
\end{aligned}$$

This is a generalization of the construction given in [2] to the multidimensional case. We assume that all components of  $\mathbf{y}_i$  are conditionally independent given the input sequence  $\{x_i\}$ . In channel equalization, for example, although the channel output vectors are highly correlated because of intersymbol interference (ISI), the output vector  $\mathbf{y}_i$  given the transmitted symbol  $x_i$  is independent. So, the covariance matrix  $\Sigma_i$  is a diagonal matrix,  $\Sigma_i = \text{diag}[\sigma_{i0}^2, \dots, \sigma_{i(d-1)}^2]$ ,  $i = 1, \dots, N$ . Let  $\zeta_i = [\sigma_{i0}^{-2}, \dots, \sigma_{i(d-1)}^{-2}]^T$ ,  $i = 1, \dots, N$ ,  $\xi_i = [\sigma_{i0}^2, \dots, \sigma_{i(d-1)}^2]^T$ ,  $i = 1, \dots, N$ ,  $\mathbf{y} = [y_0, \dots, y_{d-1}]^T$  and

$\mathbf{Y} = [y_0^2, \dots, y_{d-1}^2]^T$  to write

$$\begin{aligned}
\mathbf{r}_{11} &= \mathbf{y}, & \theta_{11} &= \mu_0^T \Sigma_0^{-1}, \\
\mathbf{r}_{12} &= \mathbf{Y}, & \theta_{12} &= -\frac{1}{2} \zeta_0, \\
\mathbf{r}_{2i} &= \delta_i(z), & \theta_{2i} &= \ln \frac{\pi_i}{\pi_0} - \ln \frac{|\Sigma_i|^{1/2}}{|\Sigma_0|^{1/2}} \\
& & & - \frac{1}{2} \mu_i^T \Sigma_i^{-1} \mu_i + \frac{1}{2} \mu_0^T \Sigma_0^{-1} \mu_0, \\
\mathbf{r}_{3i} &= \delta_i(z) \mathbf{y}, & \theta_{3i} &= \Sigma_i^{-1} \mu_i - \Sigma_0^{-1} \mu_0, \\
\mathbf{r}_{4i} &= \delta_i(z) \mathbf{Y}, & \theta_{4i} &= -\frac{1}{2} \zeta_i + \frac{1}{2} \zeta_0,
\end{aligned} \tag{19}$$

where  $i = 1, \dots, N$ . We can use the above representation to show that the FNM model given in (2) belongs to an exponential family:

$$p(\mathbf{y}, z) = \exp\{\boldsymbol{\theta}(\mathbf{u}) \cdot \mathbf{r} - \psi(\boldsymbol{\theta}(\mathbf{u}))\} \tag{20}$$

where  $\mathbf{u}^T = (\pi_0, \dots, \pi_N, \mu_0^T, \dots, \mu_N^T, \xi_0^T, \dots, \xi_N^T)$  is a vector of all the parameters of the model to be estimated,  $\boldsymbol{\theta} \cdot \mathbf{r}$  is the dot product, and

$$\psi(\boldsymbol{\theta}) = -\ln \frac{\pi_0}{|\Sigma_0^{-1}|^{1/2}} + \frac{1}{2} \mu_0^T \Sigma_0^{-1} \mu_0 + \log(\sqrt{2\pi})^d \tag{21}$$

We can now obtain the new observation vector defined as  $\mathbf{r}^T = (\mathbf{r}_{11}^T, \mathbf{r}_{12}^T, \mathbf{r}_{2i}^T, \mathbf{r}_{3i}^T, \mathbf{r}_{4i}^T)$ , and the new coordinate system (the  $\theta$ -coordinates)  $\boldsymbol{\theta}(\mathbf{u}) = (\theta_{11}, \theta_{12}, \theta_{2i}, \theta_{3i}, \theta_{4i})$  in the manifold  $\mathcal{M}$  of  $p(\mathbf{y}, z)$ . The expectation parameters,  $\boldsymbol{\eta} = E_{\boldsymbol{\theta}}(\mathbf{r})$ , called the  $\eta$ -coordinates of  $\mathcal{M}$ , can be represented as

$$\begin{aligned}
\eta_{11} &= \sum_{i=0}^N \pi_i \mu_i, & \eta_{12} &= \sum_{i=0}^N \pi_i (\omega_i + \xi_i), \\
\eta_{2i} &= \pi_i, & \eta_{3i} &= \pi_i \mu_i, \\
\eta_{4i} &= \pi_i (\omega_i + \xi_i),
\end{aligned} \tag{22}$$

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

With the above  $\theta$ - and  $\eta$ -coordinates, we can get the information geometric *em*-algorithm for the FNM based on the alternating minimization of the divergence  $\mathcal{K}$  between submanifold  $\mathcal{M}$  which is realizable by the FNM model and the submanifold  $\mathcal{D}$  suggested by the observed partial data.  $\mathcal{K}$  can be e.g. selected as relative entropy.

1. Select an arbitrary initial vector  $\hat{\mathbf{u}}_0$ , which gives the initial distribution  $\hat{P}_0 \in \mathcal{M}$ . Set  $t = 0$ .
2. Perform the em update
  - *e*-step: Calculate the *e*-projection of the present  $\hat{P}_t$  onto  $\mathcal{D}$ . This gives  $Q_{t+1} \in \mathcal{D}$  that minimizes  $K(Q||P_t)$ ,  $Q \in \mathcal{D}$ .

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

where  $\alpha_i$ s are the free parameters, which can be estimated as

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

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

$$\hat{\eta}_{t+1} = (1 - \epsilon_t)\hat{\eta}_t + \epsilon_t \hat{\mathbf{r}}_{t+1} \quad (23)$$

where  $\epsilon_t$  is the learning rate selected as a decreasing sequence.

- $m$ -step: Use the gradient method to get the next  $\hat{\mathbf{u}}_{t+1}$ , which gives  $P_{t+1}$  that minimizes  $K(Q_{t+1}||P)$ ,  $P \in \mathcal{M}$ .

$$\hat{\mathbf{u}}_{t+1} = \hat{\mathbf{u}}_t + \epsilon_t \mathbf{B}[\hat{\eta}_{t+1} - \eta(\hat{\mathbf{u}}_t)] \quad (24)$$

where  $\mathbf{B} = \frac{\partial}{\partial \mathbf{u}}\theta(\mathbf{u})$  is the gradient matrix.

3. Increment the iteration index,  $t = t + 1$ .
4. Repeat step 2 until convergence (using a suitable convergence criterion).

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 the  $m$ -projection to find the corresponding best neural network. The algorithm can be thought of consisting of two parts [3]: 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.

#### IV. APPLICATION TO CHANNEL EQUALIZATION

In this section, we present application of the information geometric 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 multi-layer 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 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 dimensional, after 100 training samples, the equalizer converges. For the FNM model with 32 mixtures and two dimensional 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 the 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 the FNM equalizer with 32 mixtures, but using same number of training samples, the BER improves considerably at high SNR values. This is due to the increase of the minimum distance between noise-free centers of the 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] S. Amari, "Information Geometry of the EM and em Algorithms for Neural Networks," *Neural Networks*, vol. 8, No. 9, pp. 1379-1408, 1995.
- [3] W. Byrne, "Alternating minimization and Boltzmann machine learning," *IEEE Trans. Neural Networks*, vol. 3, no. 4, pp. 612-620, 1992.
- [4] 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.

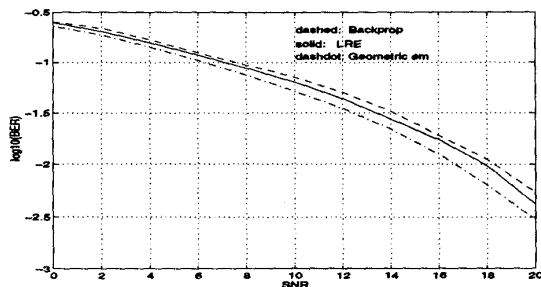


Figure 1: BER curves for the geometric  $em$  (FNM with 16 normal mixtures), the backpropagation (2-18-1 MLP), and the LRE (2-18-1 MLP) algorithms

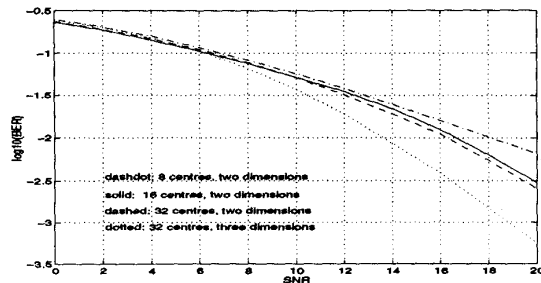


Figure 2: BER curves for the geometric  $em$  for FNM with different number of mixtures and observation vector sizes

- [5] I. Cha and S.A. Kassam, "Interference Cancellation Using Radial Basis Function Networks," *Signal Processing*, vol. 47, no. 3, pp. 247-268, Dec, 1995.
- [6] 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.
- [7] 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.
- [8] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Englewood Cliffs, NJ: Macmillan, 1994.
- [9] J. Principe, L. Giles, N. Morgan, and E. Wilson, *Neural Networks for Signal Processing VII*, Proc. IEEE Workshop, Amelia Island, FL, 1997.
- [10] L. Kullback, and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics* 22, pp. 79-86, 1951.
- [11] 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.