MINIMIZING THE FISHER INFORMATION OF THE ERROR IN SUPERVISED ADATIVE FILTER TRAINING

Jian-Wu Xu, Deniz Erdogmus, Jose C. Principe

CNEL, Electrical and Computer Engineering Department
University of Florida, Gainesville, FL 32611, USA

ABSTRACT

In this paper, we propose minimizing the Fisher information of the error in supervised training of linear and nonlinear adaptive filters. Fisher information considers the local structure of the error probability distribution and therefore, it is expected to result in more robust solutions compared to other statistics such as minimum mean-square-error or minimum-error-entropy. A gradient-based training algorithm based on a nonparametric estimator of Fisher information is presented and the performances of the three mentioned optimization criteria is compared using Monte Carlo simulations.

1. INTRODUCTION

Traditionally, supervised training of adaptive filters is peformed using the mean-square-error (MSE) as the optimality criterion. The main reason for the wide use of MSE lies in the fact that quadratic criteria combined with linear systems result in analytically tractable mathematics and lead to solutions like the Wiener-Hopf equation [1]. In the case of linear systems and Gaussian distributed signals, second-order statistics are able to extract all the information present in the data, thus yield optimal training solutions in an information theoretic perspective.

However, many contemporary signal processing problems extend beyond the linearity and Gaussianity assumptions, therefore to achieve optimality in an information theoretic framework, one has to go beyond second-order statistics as optimality criteria in adaptation. In order to achieve these extensions to information theoretically optimal adaptation rules, we need to consider the higher-order statistics of the signals since arbitrary distributions, unlike the Gaussian, are not only characterized by their second-order statistics.

Information theoretic criteria provide natural and intuitive means of dealing with higher-order statistics of the signals, since they are derived based on particular postulates such as additivity [2]. Entropy, which measures the average information content in a random variable with a particular probability distribution was previously proposed as a criterion for supervised adaptive filter training by the authors and it was shown to provide better neural network generalization compared to MSE [3].

As can be intuitively understood from the nature of entropy and from the experimental results in previous publications, minimizing the error entropy (MEE) tends to result in *spikier* optimal error distributions compared to MSE. In certain situations, this *spikiness* of the error distribution might be undesirable, especially when we aim for smooth error distributions. While the minimum error entropy criterion will maximize the information transfer from the training data to the weights of the adaptive system by minimizing the expected information content of the residual error, it does not explicitly act to improve the robustness of the solution in terms of variance in estimated optimal weights.

A criterion that will consider the local smoothness and structure of the error distribution is Fisher information [2]. Minimizing the Fisher information of the error will result in an optimal solution such that small perturbations in the error distribution due to variances in the weights will cause minimal fluctuations in the criterion. The connection between Fisher information and the Cramer-Rao bound [4] also provides another motivation for using this quantity as an optimality criterion. Nevertheless, at this time, there is no rigorous mathematical link between minimizing the error Fisher information and minimizing the variance of the optimal weight estimates.

Since, in general, the signal distributions are either unknown or difficult to *guess*, a nonparametric approach to the estimation of the distributions involved in the problem will provide a more generic solution to the adaptation problem. Parzen windowing is a simple and data efficient density estimation technique, which results in smooth estimates, whose bias and variance can be controlled by the width of the kernel function used (called the kernel size). The smoothness is particularly important in adaptation, since first or second order gradient methods

This work was supported by NSF grant ECS-0300340.

are often used in learning algorithms. In addition, we have observed that there is a functional similarity between the kernel size of the Parzen window density estimate and the width of the smoothing functional in convolution smoothing method for global optimization [5]. Thus, this free parameter might also serve as a tool to achieve global optimization; however, this aspect of the proposed algorithm will be investigated in a future paper.

This paper is organized as follows: first, a brief introduction of Fisher information is presented; second, an analytical proof that shows the preservation of the global minimum of Fisher information when it is estimated by Parzen windowing is given; next, the gradient training algorithm for an MLP using the minimum error Fisher information criterion is derived; and finally, the performance of the proposed training algorithm is compared with that of MSE and MEE on single-step chaotic time-series prediction.

2. FISHER INFORMATION AND ITS ESTIMATION

In the parameter estimation context, the Fisher information matrix is defined as the expected value of the Hessian of the log-likelihood of the data with respect to the parameter vector [4].

$$\begin{aligned} F(\boldsymbol{\theta}) &= -\int p(\mathbf{x};\boldsymbol{\theta}) \frac{\partial^2 Iogp(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} d\mathbf{x} \\ &= -E_X\left[\frac{\partial^2 Iogp(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right] \end{aligned} \tag{1}$$

In (1), $p(\mathbf{x};\boldsymbol{\theta})$ is the data likelihood function parameterized in terms of the vector $\boldsymbol{\theta}$. The well-known Cramer-Rao bound is expressed in terms of this matrix as $Cov(\hat{\boldsymbol{\theta}}) \geq F^{-1}(\boldsymbol{\theta})$, where $\hat{\boldsymbol{\theta}}$ is any unbiased estimator of the underlying *true* parameter vector.

In the context of supervised learning, we will use a different definition of Fisher information, however the latter is still related to the definition in (1) when the parameter vector $\boldsymbol{\theta}$ is assumed to be simply the mean of the data distribution. In this case, the Fisher information for a random variable X with distribution $p(x)$ becomes

$$F(X) = \int \frac{p'^2(x)}{p(x)} dx = E_X\left[\left(\frac{p'(X)}{p(X)}\right)^2\right] \tag{2}$$

It can be shown that this version of Fisher information is affectively measuring the Kullback-Leibler divergence between $p(x)$ and $p(x+\delta x)$ [6]. Hence, in a supervised learning situation, where (2) is evaluated and minimized for the error signal, we expect the optimal solution to result in a set of weights such that small perturbations in the weights will result in minimal localized perturbations in the error distribution. The measure of minimality for these perturbations is the Kullback-Leibler divergence and the error distribution is expected to be smooth and closer to uniform compared to MSE and MEE.

In practice, the Fisher information of the error signal must be estimated from its samples. This requires a smooth (i.e., continuous and differentiable) estimate of its distribution. Parzen windowing is a suitable method [7]. Given independent and identically distributed (iid) samples $\{e_1,\ldots,e_N\}$, the error distribution can be approximated by

$$p_e(\xi) = \frac{1}{N}\sum_{i=1}^{N} \kappa(\xi - e_i, \sigma^2) \tag{3}$$

where $\kappa(x;\sigma^2)$ is typically a zero-mean Gaussian kernel with standard deviation σ. The Fisher information can then be estimated using

$$\begin{aligned} F(e) &= E_e\left[\left(\frac{p'(e)}{p(e)}\right)^2\right] = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{p'(e_j)}{p(e_j)}\right)^2 \\ &= \frac{1}{N}\sum_{j=1}^{N}\left(\frac{\sum_{i=1}^{N}\kappa'(e_j - e_i;\sigma^2)}{\sum_{i=1}^{N}\kappa(e_j - e_i;\sigma^2)}\right)^2 \end{aligned} \tag{4}$$

Now consider the error's Fisher information (EFI) in the first form given in (2): $F(e) = \int p_e'^2(\xi)/p_e(\xi)d\xi$. Note that this expression is invariant to changes in the mean of the probability distribution, therefore we can assume the mean is zero whenever necessary, without loss of generality. The gradient of this quantity with respect to a particular error sample e_k is (assuming Gaussian kernels)

$$\begin{aligned} \frac{\partial F(e)}{\partial e_k} &= \frac{\partial}{\partial e_k}\int p_e(\xi)\frac{p_e'^2(\xi)}{p_e^2(\xi)}d\xi \\ &= \frac{1}{N^2\sigma^4}\int \frac{\kappa(\xi - e_i)}{p^2{}_e(\xi)}\sum_{i=1}^{N}\kappa(\xi - e_i)(e_i - \xi)\cdot \\ &\qquad \left[\begin{array}{l} 2\left(1 - \dfrac{(\xi - e_i)^2}{\sigma^2}\right)p_e(\xi) \\ -\dfrac{(\xi - e_i)}{N\sigma^2}\displaystyle\sum_{i=1}^{N}\kappa(\xi - e_i)(e_i - \xi) \end{array}\right] \end{aligned} \tag{5}$$

This gradient evaluates to zero for the sample set $\mathbf{e}=[e_1,\ldots,e_N]^T=\mathbf{0}$. Thus, this point in the error space is a stationary point of the cost function (if achievable). Evaluating the eigenvalues of the Hessian matrix of the criterion at this point, we also observe that it is a local minimum (with a zero eigenvalue along the direction where only the mean of the error changes, as expected from the mean-invariance property of Fisher information).

Specifically, the diagonal and off-diagonal entries of the Hessian at this point are found as

$$\left.\frac{\partial^2 F(e)}{\partial e_i^{\,2}}\right|_{e=0} = \frac{4N-4}{N^2\sigma^4}$$
$$\left.\frac{\partial^2 F(e)}{\partial e_j \partial e_i}\right|_{e=0} = \frac{-4}{N^2\sigma^4} \tag{6}$$

This matrix has the following eigenvalues: 0 with multiplicity 1 corresponding to the eigenvector $[1,\ldots,1]^T$ and $4/N\sigma^4>0$ with multiplicity N-1 corresponding to the eigenvectors spanning the remaining orthogonal subspace.

In summary, minimizing the Parzen window estimate of Fisher information will try to minimize the error in the vicinity of this small-error solution.

3. GRADIENT LEARNING USING THE FISHER INFORMATION CRITERION

Suppose we are given a training set in the form of input vectors $\mathbf{x}_k$ and desired outputs d_k. Consider the optimization of the parameters of a general class of nonlinear adaptive systems denoted by $y_k=g(\mathbf{x}_k;\mathbf{w})$. The weights are updated according to the steepest descent rule

$$\mathbf{w} \leftarrow \mathbf{w} - \mu \frac{\partial F}{\partial \mathbf{w}} \tag{7}$$

where the gradient is evaluated from

$$\frac{\partial F}{\partial \mathbf{w}} = \frac{2}{N}\sum_j \left[\frac{p_e'(e_j)}{p_e^2(e_j)} \frac{\partial p_e'(e_j)}{\partial \mathbf{w}} - \frac{p_e'^2(e_j)}{p_e^3(e_j)} \frac{\partial p_e(e_j)}{\partial \mathbf{w}} \right]$$
$$p_e(e_j) = \frac{1}{N}\sum_i \kappa(e_j - e_i)$$
$$p_e'(e_j) = \frac{1}{N}\sum_i \kappa'(e_j - e_i) \tag{8}$$
$$\frac{\partial p_e(e_j)}{\partial \mathbf{w}} = \frac{1}{N}\sum_i \kappa'(e_j - e_i)\frac{\partial (e_j - e_i)}{\partial \mathbf{w}}$$
$$\frac{\partial p_e'(e_j)}{\partial \mathbf{w}} = \frac{1}{N}\sum_i \kappa''(e_j - e_i)\frac{\partial (e_j - e_i)}{\partial \mathbf{w}}$$

This learning rule can be interpreted in a particle interaction framework, where the error samples e_k are *physical particles* interacting with each other through the kernel function according to the rules defined by the gradient update expression in (8). The particles exert forces on each other; hence they move in *space*. However, the nonlinear filter topology imposes a constraint on the particles such that their movements are restricted to a manifold that is defined by this topology. A similar analogy was formed for the entropy criteria as well [8].

4. CHAOTIC TIME-SERIES PREDICTION

In this section, we compare the optimal solutions offered by three criteria: MSE, MEE and Fisher information. The example problem selected is the single-step prediction of the chaotic laser time-series [9]. This time-series is particularly difficult to predict at the transition points where the signal value collapses suddenly after a slow regime of expansion. Three structurally identical 14:4(tanh):1(linear) TDNNs [10] are trained using these three criteria and the mean value of the error is set to zero by adjusting the bias of the linear output processing element of the TDNNs for all three criteria.

The training set consists of 1000 samples from the time-series and in order to avoid local optima to some extent, training is started from 100 randomly selected initial conditions and the optimal weights by all three criteria are selected as the weights that perform best according to each individual cost function.

The final TDNNs are then subjected to a test set consisting of 4000 samples in the single-step prediction framework. Fig. 1 shows the histograms of the error distributions of the three TDNNs on this test set. The dynamic ranges of the error samples are [-0.5788,0.6459], [-0.7859,0.6584], and [-0.4512,0.5329] for MSE, MEE, and EFI, respectively. While the EFI criterion results in the smallest dynamic range for the test error, the error distribution at smaller values (around zero) are relatively more spread to approach the targeted uniformity, as we observe in Fig. 1. Upon investigation of the predictions of MSE- and MEE-trained TDNNs, we find out that the large errors occur at the points of collapse. EFI performs better at these locations at the cost of slightly increased error in the expansion regime of the time-series. A sample collapse point in the test sep is shown in Fig. 2 with the predictions made by the three TDNNs. Notice that the EFI-trained network performs better in this region compared to the MSE- and MEE-trained networks.

A particularly difficult problem is to design a closed-loop system that can continue to iteratively predict the chaotic time-series using its previous predictions. Clearly any network will diverge while performing in this closed-loop structure due to the very nature of chaotic signals. Even the smallest error will propagate through the system to create a divergence in the prediction error. Nevertheless, a good indicator of quality of a model for chaotic systems is how long their prediction accuracy survives in this closed-loop prediction scheme. Therefore, we subject the three TDNNs trained by the optimization criteria (on the same training data set) to this test and let them iteratively predict the laser time-series by feeding their own outputs back as inputs. In this procedure, we test the TDNN models on 1000 test sets (each of length 60) starting from randomly selected initial points to obtain a Monte Carlo evaluation. The normalized MSE

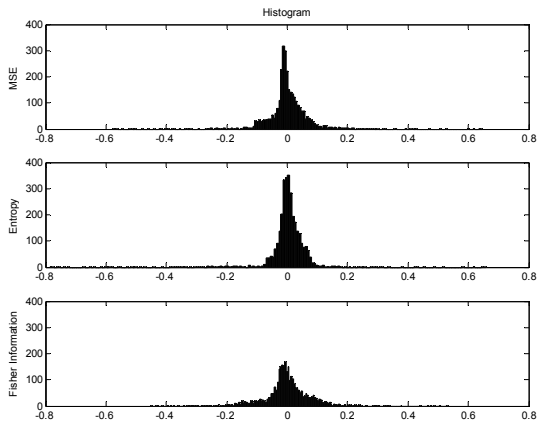

Figure 1. Test error histograms for MSE, MEE, and EFI trained TDNNs.

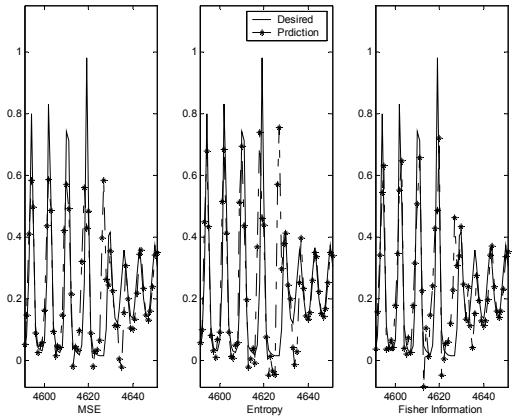

Figure 2. A sample collapse point in the test set from the laser time-series. Actual and predicted values for MSE, MEE, and EFI trained TDNNs in open-loop testing.

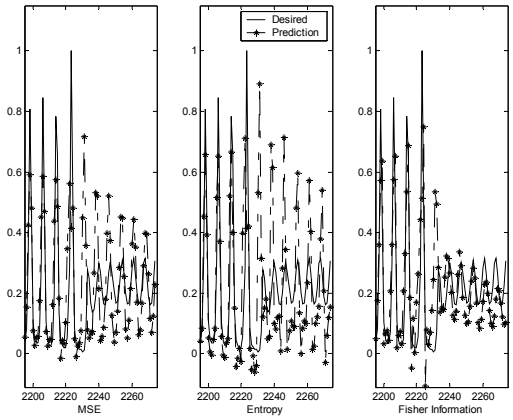

Figure 3. Closed-loop prediction at a sample collapse point in the test set using the MSE, MEE, and EFI trained TDNNs.

and the standard deviation values for the three predictions over the MC test are 1.0306 ± 1.0291 (MSE), 1.4562 ± 1.5448 (MEE) and 1.0690 ± 0.4743 (EFI). A representative prediction output is shown in Fig. 3. Clearly, the MSE and MEE trained networks fail to follow the signal at this abrupt nonstationarity, while the EFI-trained TDNN maintains a relatively high accuracy in its predictions of the signal.

5. CONCLUSIONS

Supervised training of nonlinear adaptive filters using non-Gaussian data requires considering more information than what is present in only the second-order statistics. Therefore, in this paper, we proposed using the Fisher information as an optimality criterion. A nonparametric estimator based on Parzen windowing is presented and the performance of the resulting training methodology is compared with mean-square-error and error-entropy approaches on the laser time-series prediction example. The Fisher information approach yielded a more robust optimal solution that was able to cope with the abrupt nonstationarities in the data more effectively. However, this came at a cost of increased error at the relatively stationary regimes of the signal.

6. REFERENCES

[1] S. Haykin, *Adaptive Filter Theory*, 4th ed., Prentice Hall, NJ, 2002.

[2] T. Cover, J. Thomas, *Elements of Information Theory*, Wiley, NY, 1991.

[3] D. Erdogmus, J.C. Principe, "An Error-Entropy Minimization Algorithm for Supervised Training of Nonlinear Adaptive Systems," IEEE Trans. Signal Processing, vol. 50, no. 7, pp. 1780-1786, 2002.

[4] S.M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, NJ, 1993.

[5] D. Erdogmus, J.C. Principe, "Generalized Information Potential Criterion for Adaptive System Training," IEEE Trans. Neural Networks, vol. 13, no. 5, pp. 1035-1044, 2002.

[6] B.R. Frieden, *Physics From Fisher Information: A Unification*, Cambridge University Press, UK, 1998.

[7] E. Parzen, "On Estimation of a Probability Density Function and Mode", in *Time Series Analysis Papers*, Holden-Day, CA, 1967.

[8] J.C. Principe, D. Xu, J. Fisher, "Information Theoretic Learning," in *Unsupervised Adaptive Filtering*, (Ed. S. Haykin), Wiley, NY, 2000.

[9] A.S. Weigend, N.A. Gershenfeld, *Time Series Prediction: Forecasting the Future and Understanding the Past,* Addison-Wesley, UK, 1994.

[10] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, NJ, 1999.