From Adaptive Linear to Information Filtering

Jose C. Principe and Deniz Erdogmus

Computational Neuro-Engineering Laboratory
Department of Electrical and Computer Engineering
University of Florida, Gainesville, FL 32611
{principe, deniz}@cnel.ufl.edu

ABSTRACT

Adaptive signal processing theory was born and has lived by exclusively exploiting the mean square error criterion. When we think of the goal of least squares without restrictions of Gaussianity, one has to wonder why an information theoretic error criterion is not utilized instead. After all, the goal of adaptive filtering should be to find the linear projection that best captures the information in the desired response. In this paper we summarize our efforts to extend adaptive linear filtering to information filtering. We briefly review Renyi's entropy definition, Parzen windows and put them together in a framework to estimate entropy directly from samples (nonparametric). Once this criterion is developed we can train linear or nonlinear adaptive networks for entropy maximization or minimization. We present results on the properties of the Renyi's nonparametric entropy estimator, and show how it performs in chaotic time series prediction.

1. INTRODUCTION

Starting with the early work of Wiener [1] on optimum filtering, mean square error (MSE) has been almost exclusively employed in the training of all adaptive systems including linear filters and artificial neural networks. There were mainly two reasons behind this choice: Analytical simplicity, and the assumption that most real-life random phenomena may be expressed accurately by the Gaussian distribution. The probability density function (pdf) of the Gaussian is characterized by only its first and second order statistics. Hence, under these linearity and Gaussianity assumptions, MSE, which concentrates on second order statistics, would be able to extract all possible information from a data set whose statistics are solely defined by its mean and variance.

However, most real-life problems are governed by nonlinear equations and most random phenomena are far from being normally distributed. Therefore, for the training of adaptive systems, a criterion that not only considers the second order statistics but that also takes into account higher-order statistical behavior is a necessity. Moreover, there are classes of problems such as blind equalization and subspace projection (feature extraction) for which higher order statistical information is crucial to obtain useful solutions [2].

The entropy of a given probability distribution function, introduced by Shannon [3], is a scalar quantity that provides a measure of the average information contained in the distribution. By definition, information is a function of the pdf itself, hence the entropy should be estimated directly from the pdf. Application of the entropy criterion to a system identification framework is conceptually quite straightforward. Given a stationary time series produced by an unknown system, the entropy of the estimation error e_k over the training data set must be minimized [4]. In fact, when the entropy of the error is minimized, the expected information contained in the estimation error is minimized; hence the adaptive system is trained optimally in the sense that the mutual information between the time series and the model output is maximized.

The organization of the paper is as follows. First, the backpropagation training algorithms for Renyi's entropy with parameter 2 is given for the one-dimensional case. Second, an analytical proof shows that the global minimum of the entropy is still a minimum of the Parzen window estimated entropy when Gaussian kernels are employed. Then, results of a case study are presented where entropy-error minimization criterion is applied to the short-term prediction of a chaotic time series with a time-delay neural network (TDNN). The performances of MSE trained and entropy trained TDNNs are compared in terms of error power and higher order statistics.

2. TRAINING A TDNN WITH THE ENTROPY CRITERION

A typical system identification scheme with a time delay neural network (TDNN) is shown in Fig. 1. Once the topology of the network is set, the training criterion is what determines the overall performance of the final model.

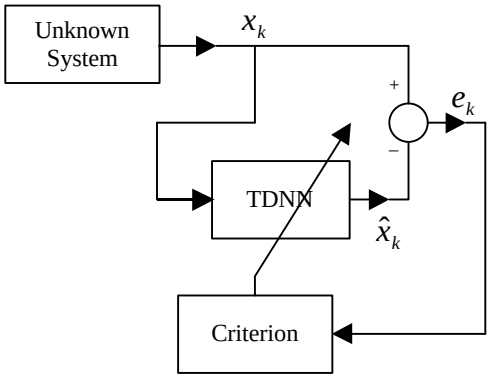

Figure 1. Block diagram for nonlinear adaptive filtering

If the adaptation criterion is chosen to be the minimization of the mean square error, and the optimization procedure is fixed to be steepest descent, then the originating training algorithm is the well-known backpropagation algorithm [6]. However, due to the reasons stated above, if the adaptation criterion is Shannon's entropy of the error, the minimization with steepest descent becomes

$$w(n+1) = w(n) - \eta \frac{\partial H_S(e)}{\partial w} \tag{1}$$

Here, Shannon's entropy [2] is given by (2)

$$H_S(e) = -\int_{-\infty}^{\infty} f_e(x) \log f_e(x) dx \tag{2}$$

In practice, an analytical expression for the stationary random process pdf, which is necessary for the computation of Shannon entropy, is not available. Non-parametric estimators for Shannon entropy are known to be computationally expensive. We have investigated more efficient estimators for entropy based on Renyi's entropy [7], which reads

$$H_{R\alpha}(e) = \frac{1}{1-\alpha} \log \int f_e^{\alpha}(\xi) d\xi \qquad \alpha \geq 0, \alpha \neq 1 \tag{3}$$

Note that, for $\alpha = 2$ (quadratic entropy), this expression becomes

$$H_{R2}(e) = -\log \int_{-\infty}^{\infty} f_e^2(x) dx \tag{4}$$

and for optimization the logarithm can be dropped yielding

$$V(e) = \int_{-\infty}^{\infty} f_e^2(x) dx \tag{5}$$

When the error pdf is approximated from its N samples by the Parzen window estimator [5] with Gaussian kernels of zero mean and variance σ^2

$$\hat{f}_e(x) = \frac{1}{N} \sum_{i=1}^{N} k(x - e_i, \sigma^2) \tag{6}$$

we obtain

$$V(e) = \frac{1}{2N^2\sigma^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (e_j - e_i) \kappa(e_i - e_j, 2\sigma^2) \cdot \tag{7}$$

We called $V(e)$ an *information potential* because it is a decreasing positive function of the distance between the samples [2]. Minimizing quadratic entropy is equivalent to maximizing the information potential. The Gaussian kernel is preferable because it is continuously differentiable, and therefore the sum of Gaussians is continuously differentiable on the space of real vectors of any dimension.

The gradient vector to be used in the steepest ascent algorithm for the maximization of the information potential is found as [2]

$$\frac{\partial \hat{V}(e)}{\partial w} = \frac{1}{2N^2\sigma^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (e_j - e_i) \cdot k(e_i - e_j, 2\sigma^2) \left[\frac{\partial \hat{x}_j}{\partial w} - \frac{\partial \hat{x}_i}{\partial w} \right] \tag{8}$$

This derivation is easily extended to the multi-variable case. The sensitivities in (8) are easily computed with the standard backpropagation algorithm [6]. We will not treat the case of the adaptive linear combiner since it is a special case of this formulation (a linear PE in the TDNN and no hidden layers).

One important point to be mentioned in training with entropy is that the entropy is invariant to the mean of the error over the data set. This can be easily shown by a simple change of variables of the integration, i.e. $z = x - m_e$. Due to this property of entropy, the algorithm will converge, with a probability of one, to a set of optimal weights, which do not yield zero error-mean. However, this can be corrected by modifying the bias weight of the output neuron properly to yield zero mean error over the training data set just after training

ends. It must also be noted that the entropy is a cost function with many local minima.

3. EQUIVALENCE OF LOG-LIKELIHOOD AND NONPARAMETRIC ENTROPY TRAINING

It is worthwhile to remember that nonparametric entropy training is equivalent to maximum likelihood estimation [10] with the big advantage that we do not require a parametric family of densities.

Assume that we have a set of i.i.d. random samples x_i distributed as $p(x_i)$, and we estimate the differential entropy as $\hat{p}(x)$ using Parzen windows

$$H(p_x(x)) \approx -\frac{1}{N}\sum_i \log \hat{p}(x_i) \rightarrow H(p_x(x)) + D(p_x \| \hat{p})$$

where $D(.\|.)$ is the Kullback-Leibler divergence.

Now if we apply the asymptotic equipartition theorem [11], it can be shown that the log-likelihood estimator applied to the same samples assuming a probabilistic model with parameter vector θ yields

$$L_\theta(\{x\}) = \sum_i \log p_\theta(x_i;\theta) \rightarrow -N(H(p_x(x)) + D(p_x \| p_\theta))$$

As we can see the two methods provide a similar result, with the difference that we do not require a family of distributions when using the entropy minimization approach. Clearly, this is preferable in many real world problems where we do not know the data distributions. On the other hand this result also shows that if we assume a distribution and use information theoretic methods, the procedure should be called more precisely maximum likelihood.

4. PROOF OF MINIMA EQUIVALENCE

In an optimization framework, we should further check if the nonparametric estimator preserves the extreme points of the theoretical criterion. We proceed to prove that the global minimum of the entropy is still a minimum of the non-parametrically estimated entropy for the Renyi's entropy, when Parzen windowing with Gaussian kernels is utilized.

The global minimum of Renyi's entropy is achieved when the pdf of error is the Dirac-delta function. Since entropy is independent of the mean of the error, we can concentrate on the case where mean of e is zero without loss of generality. The gradient of the entropy for Gaussian kernels is given as

$$\frac{\partial H_{R\alpha}}{\partial e_j} = \frac{\alpha}{N\sigma^2(1-\alpha)\int_{-\infty}^{\infty}\kappa^\alpha d\xi}\int_{-\infty}^{\infty}\xi\kappa^\alpha d\xi \qquad (9).$$

Evaluating this gradient at $e = [e_1 \quad \dots \quad e_N] = 0$, we obtain a zero value since we are integrating an odd function. Hence $e = 0$ is a stationary point of $H_{R\alpha}(e)$. Now we continue with the computation of the Hessian to see if it is furthermore a minimum. The diagonal and off-diagonal entries of the Hessian are found to be

$$\left.\frac{\partial^2 H_{Ra}}{\partial e_j^2}\right|_{e=0} = \frac{N-1}{N^2 s^2} \qquad (10)$$

$$\left.\frac{\partial^2 H_{Ra}}{\partial e_k \partial e_j}\right|_{e=0} = \frac{-1}{N^2 s^2} \qquad (11)$$

The eigenvalues of the Hessian matrix then can be computed to be $l_0 = 0$ with multiplicity 1 and $l_i = 1/(Ns^2)$ with multiplicity (N-1), hence the Hessian is positive semi definite. The value of the entropy may decrease in the direction of the eigenvector that corresponds to the 0 eigenvalue, which is found as

$$\vec{e}_0 = [1 \quad 1 \quad \dots \quad 1]^T \qquad (12)$$

This means, if the error changes along this direction it will still have constant entries. However, we have shown above that, the mean of e does not change the value of the entropy. Therefore, when e changes along the direction of $\vec{e}_0$, the value of the entropy remains constant. So we conclude that Renyi's entropy approximated by Parzen windowing with Gaussian kernels has a minimum at the point where error is completely constant over the whole data set.

Note that the Hessian matrix for the Renyi's entropy is independent of α, and it is identical to the Shannon's entropy [12]. Hence, it has the same eigenvalues with the Hessian matrix of Shannon's entropy.

5. SHORT TERM PREDICTION OF CHAOTIC TIME SERIES

When a TDNN is trained with the entropy criterion, the expected value of the error over the training data set will not converge to zero. We have mentioned that, this

problem can easily be solved by adjusting the bias weight of the last layer to make the mean of estimation error zero over the training set after the learning finishes. If the output PE in the TDNN is chosen to be linear, this modification is a simple addition of the mean of the current error to the bias weight of the output PE.

As a case study, the short-term prediction of the Mackey-Glass chaotic time series [8] with parameter $t = 30$ using both MSE trained and (Renyi's) entropy trained TDNNs is presented here. The time-delay TDNN has a 3-tap input, 5 neurons in the hidden layer and a single linear output neuron. The embedding dimension is chosen to be 3 here. This is less than the embedding dimension suggested by Taken's Embedding Theorem, namely 5, for the Mackey-Glass series [8]. For this size of the reconstruction space the difficulty level of the prediction problem increases. Increased difficulty is desired since we know that even the MSE criterion performs quite well for this time series when a 5-tap input is employed.

The TDNNs are trained over a training data set of length 200 starting from 100 randomly chosen initial weights, so that hopefully the global optimal solution is one of the solutions suggested by this Monte Carlo type training approach. In this sequence, the weights of the MSE trained TDNN is iterated 100 times for each initial set of weights whereas those of the entropy trained TDNN are iterated for 30 times according to their corresponding backpropogation algorithms using the conjugate gradient algorithm [9]. At the end of the mentioned Monte Carlo training approach, the best set of weights obtained by each of the criteria are taken and checked for further improvement by employing a small constant step size. For this specific case, these extra iterations did not improve the cost function for either of the criteria. Finally, the bias weight of the output neuron of both artificial neural networks are adjusted to give zero error mean over the training data set after training had finished.

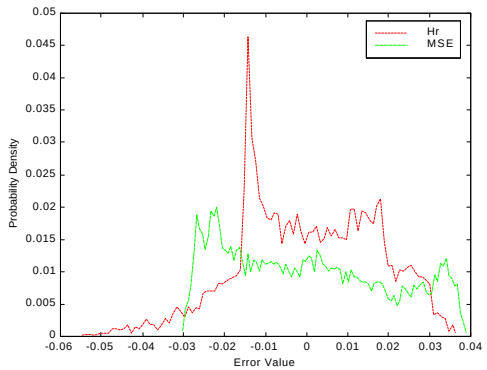

Figure 2: The probability distribution functions for the estimation errors of MSE and entropy trained TDNNs.

The trained networks are tested on an independently created test data set of length 10000. We do not present the error plots of the two TDNNs over the test data set here because they do not bear any information when presented in that form. Rather, we will concentrate on the statistical properties of the error signals. The pdfs of these two error distributions approximated by a histogram of equally spaced 100 bins are shown in Fig. 2. As observed from this plot, the error distribution of the entropy trained TDNN is more concentrated around zero whereas the error distribution of the MSE trained one is more uniformly distributed over its support.

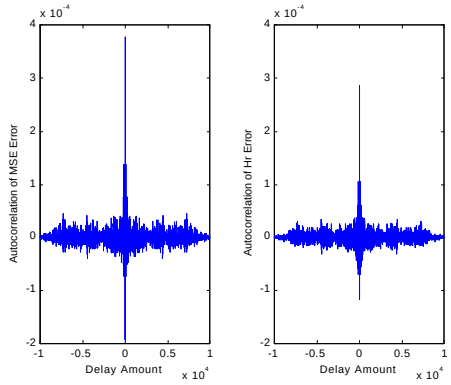

Figure 3: The autocorrelation functions of the estimation errors; MSE trained and entropy trained TDNNs respectively.

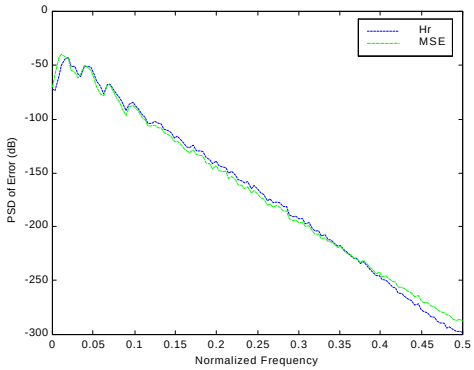

Figure 4: The PSD estimates of the estimation errors.

It is observed from Fig. 3 that, the autocorrelation function of entropy trained TDNN's error contains higher frequency components compared to that of the MSE trained TDNN. This is also evident from the PSD estimates of the two random processes (Fig. 4). The PSD of the error of entropy trained TDNN decays slower than that of the MSE trained one in the lower frequencies.

For completeness, we present here the data with its estimations superimposed (Fig 5). The extreme values of

the horizontal axis are chosen to include the point where the entropy trained TDNN makes its maximum error.

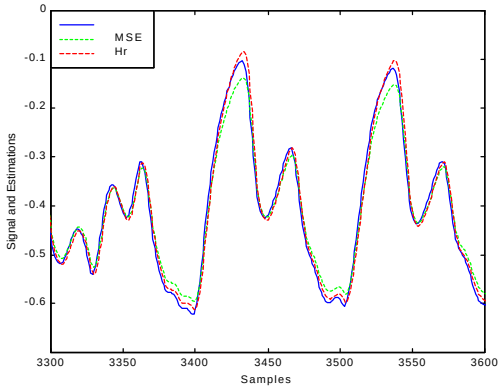

Figure 5: The short-term prediction of Mackey-Glass time series by MSE and entropy trained TDNNs.

We investigate the central moments of the two error distributions for a more quantitative comparison. The obvious ideal solution should have a Dirac-delta distributed error. All central moments of this desired error distribution are zero. In other words, we would like the central moments of the error distribution as close to zero as possible. Table 1 lists the first 5 central distributions of the error for the two training criteria.

n	$E[(e^{MSE} - m_e^{MSE})^n]$	$E[(e^{Ent} - m_e^{Ent})^n]$
1	0	0
2	$0.377\text{x}10^{-3}$	$0.285\text{x}10^{-3}$
3	$0.227\text{x}10^{-5}$	$-0.859\text{x}10^{-6}$
4	$0.271\text{x}10^{-6}$	$0.208\text{x}10^{-6}$
5	$0.372\text{x}10^{-8}$	$-0.27\ \text{x}10^{-8}$

Table 1: Central Moments of the error distributions.

Another way of looking at the same problem is to compare the central moments of the data distribution and the predicted time series distributions. What we are essentially trying to do is to match the probability distribution of the estimated time series to that of the data samples. In other words, we would like the central moments of the predictions to be as close to those of the test data set as possible. The central moments of the test data set and the predictions by the MSE and entropy trained TDNNs are given in Table 2.

n	$E[(x - m_x)^n]$	$E[(\hat{x}^{MSE} - m_{\hat{x}}^{MSE})^n]$	$E[(\hat{x}^{Ent} - m_{\hat{x}}^{Ent})^n]$
1	0	0	0
2	$1.585\text{x}10^{-2}$	$1.248\text{x}10^{-2}$	$1.6\text{x}10^{-2}$
3	$1.085\text{x}10^{-3}$	$0.785\text{x}10^{-3}$	$1.383\text{x}10^{-3}$
4	$6.186\text{x}10^{-4}$	$3.845\text{x}10^{-4}$	$6.838\text{x}10^{-4}$
5	$8.998\text{x}10^{-5}$	$5.125\text{x}10^{-5}$	$12.197\text{x}10^{-5}$

Table 2: Central Moments of the desired data samples and the prediction samples.

Note that the even moments of the entropy-trained predictor are very well matched to the original time series, while the odd moments are still in error, but above the true values. The moments of the MSE trained predictor always are smaller than the original tending to a uniform error distribution.

These results point out clearly that, the TDNN predictions approximate the statistical behavior of the Mackey-Glass chaotic attractor better when it is trained with the entropy criterion compared to the MSE criterion. The pdf estimations of the test data and the TDNN predictions for these data are presented in the following figure with a histogram of equally spaced 100 bins.

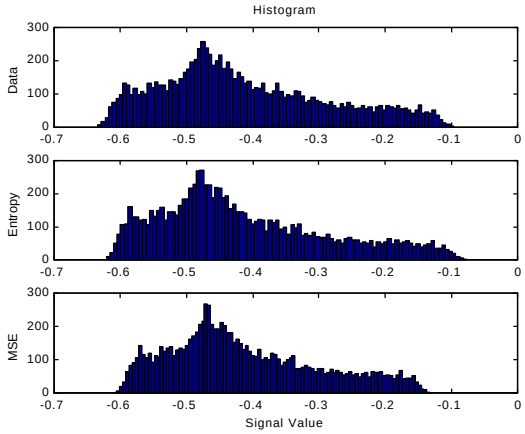

Figure 6: The histogram of the MG30 and its predictions by entropy-trained and MSE-trained TDNNs respectively.

6. OTHER APPLICATIONS OF INFORMATION THEORETIC CRITERIA

Blind source separation is a problem that can be solved by exploiting the higher order statistical information of a set of input signals. Bell and Sejnowski proposed a very efficient algorithm for BSS that maximizes the entropy at the output of a nonlinear system [13]. However, their approach requires knowledge of the kurtosis of the inputs, and can only be applied to mappers that have full rank Jacobians. Alternatively, Renyi's entropy can be used to solve the same problem without the above mentioned limitations. In [14] we compare the performance of our method with Bell and Sejnowski's algorithm and show that our approach is more robust.

We have also extended the entropy criterion to estimate approximations of mutual information [2]. Manipulation of mutual information creates a general framework to train adaptive systems that subsumes supervised and unsupervised learning (see Fig 7).

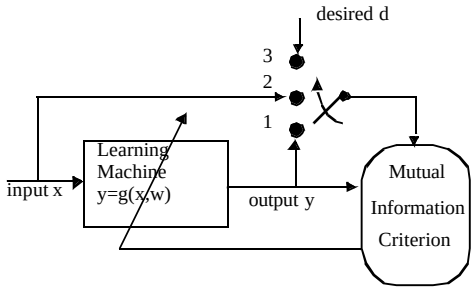

Figure 7. Learning with Information Theoretic Criteria.

When the switch is in position 1 we minimize the mutual information among the outputs of the mapper (used in BSS [15]). When the switch is in position 2 we truly implement Linsker's Infomax principle. Notice that both of these belong to the unsupervised learning case. When the switch is in position 3, we implement information filtering. When the desired signal is made up of class lables or continuous variables, and we maximize the criterion we have feature extraction that preserves the maximum information in the features. These features can be used for classification (indicator variable) [16] or simply sub space mappings (as pose estimation in imagery [2]).

7. CONCLUSIONS

In this paper, an information theoretical learning criterion for adaptive systems, namely estimation-error-entropy-minimization, has been investigated. An analytical proof, which shows that Parzen windowing approximation with Gaussian kernels of the probability distribution function to be used in entropy computation preserves the original global minimum of the entropy as one of the minima, possibly the global minimum, of the estimated entropy function. This proof enables us to use Parzen windowing with Gaussian kernels in entropy estimation safely from the entropy-minimization point of view.

A time-delay neural network has been trained for the short-term prediction of the Mackey-Glass chaotic time series using both the entropy and mean-square-error criteria. A Monte Carlo approach has been taken in terms of the initial weights of the time-delay neural network in order to avoid the local minimum solutions. However, it became evident that even with this approach, the global minimum of the MSE has not been achieved, since the entropy solution had smaller error power. The best solutions obtained by the mean-square-error and the entropy criteria were compared in terms of the statistical behavior of the estimation errors and the prediction values themselves. The comparison of central moments of the error distributions revealed the fact that, the error of the entropy-trained time-delay neural network is closer to the ideal solution. The comparison of the central moments of test data samples and the prediction samples lead to the same conclusion. The predictions of the time-delay neural network trained with the entropy criterion approximate the statistical behavior of the actual output of the unknown system better than the one trained with the men-square-error criterion.

Acknowledgement: This work was partially supported by NSF grant ECS-9900394.

REFERENCES

[1] Haykin, S., *Introduction to Adaptive Filters*, MacMillan, NY, 1984.
[2] Principe, J., Xu, D., Fisher, J., Information Theoretic Learning, in *Unsupervised Adaptive Filtering*, Simon Haykin Editor, 265-319, Wiley, 2000.
[3] Shannon
[4] Fisher, J.W., *Nonlinear Extensions to the Minimum Average Correlation Energy Filter*, Ph.D. Dissertation, University of Florida, 1997.
[5] Parzen, E., "On Estimation of a Probability Density Function and Mode", in *Time Series Analysis Papers*, Holden-Day, Inc., CA, 1967.
[6] Rumelhart, D.E., Chauvin, Y., *Backpropogation: Theory, Architectures, and Applications*, Lawrence Erlbaum Associates, NJ, 1995.
[7] Renyi, A., *A Diary on Information Theory*, Wiley, NY, 1987.
[8] Kaplan, D., Glass, L., *Understanding Nonlinear Dynamics*, Springer-Verlag, NY, 1995.
[9] Luenberger, D.G., *Linear and Nonlinear Programming*, Addison-Wesley Pub. Co., MA, 1973.
[10] Cardoso J., "Infomax and maximum likelihood for blind source separation", IEEE Signal Proc. Letters, vol 4, 112-114, 1997.
[11] Cover T. and J. Thomas, *Elements of Information Theory*, New York, Wiley, 1991.
[12] Erdogmus D., and J. Principe, "Comparison of entropy and mean square error criteria in adaptive system training" accepted in *ICA 2000*, Helsinki, Finland.
[13] Bell T., T. Sejnowski, An information maximization approach to blind source separation and blind deconvolution", *Neural Computation*, vol 6, 1129-1159, 1995.
[14] Fisher J., and J. Principe, "Blind source separation by local interaction of output signals", 1998 *IEEE DSP Workshop*, Bryce Canyon, UT, 1998.
[15] Xu D., J. Principe, J. Fisher H-C Wu, "A novel measure for independent component analysis", in Proc *ICASSP'98*, vol II, 1161-1164, 1998.
[16] Principe, J., D. Xu, Q. Zhao, J Fisher, "Learning from examples with information theoretic criteria", *VLSI Signal Processing Systems*, Special issue on neural networks, 2000 (in press).