SUPERVISED TRAINING OF ADAPTIVE SYSTEMS WITH PARTIALLY LABELED DATA

Deniz Erdogmus[1], *Yadunandana N. Rao*[2], *Jose C. Principe*[3]

[1] Oregon Graduate Institute, OHSU, Portland, OR 97006, USA
[2] Motorola Inc., Plantation, FL 33324, USA
[3] Computational NeuroEngineering Laboratory, University of Florida, Gainesville, FL 32611, USA
derdogmus@ieee.org, yadu@motorola.com, principe@cnel.ufl.edu

ABSTRACT

Supervised adaptive system training is traditionally performed with available pairs of input-output data and the system weights are fixed following this training procedure. Recently, in the context of machine learning, where the desired outputs are discrete-valued, the idea of exploiting unlabeled samples for improving classification performance has been proposed. In this paper, we introduce an information theoretic framework based on density divergence minimization to obtain extended training algorithms. Our goal is to provide a theoretical framework upon which we can build efficient algorithms to this end.

1. INTRODUCTION

Traditionally system identification and nonlinear regression have been approached in a supervised learning framework where different optimality criteria are utilized based on the statistics of the error between the adaptive system output and the desired output values [1-3]. All labeled pairs of available data are used in determining the *optimal* weights (typically by splitting this data to training and testing sets) and no further optimization is carried out over the unlabeled data samples in the actual application phase.[1]

This approach has been deemed quite natural by everyone; after all, how could you use an input sample for training further an adaptive system if you did not know what output it should produce?[2] Only recently, in the machine learning literature the concept of *making use of unlabeled data* in supervised learning to enhance classifier performance has been addressed. This is due to the fact that in pattern recognition, labeled samples are much more expensive to collect compared to unlabeled feature vectors. The most prominent approach is to use the well-known EM algorithm in a maximum likelihood framework [4,5]. Another interesting approach utilizes the representer theorem in the context of regularization to exploit the unlabeled data for *smoother* function approximation [6].

The purpose of this paper is to create a theoretical framework that allows the training of adaptive systems in supervised learning settings, using both labeled and unlabeled data. Under this framework, one can continue to train a system even after supervised training is completed. To this end, information theoretic approaches will be considered at a theoretical level and for some possible criteria, connections to existing methods will be pointed out. For illustration purposes, a special case of the proposed framework will also be studied.

2. PROBLEM DEFINITION

Consider the function approximation problem. For convenience assume that independent and identically distributed (iid) input-output data $\{(\mathbf{x}_1,d_1),\ldots,(\mathbf{x}_T,d_T)\}$ are available from an unknown nonlinear function as follows:[3]

$$d = f(\mathbf{x}) + n \qquad (1)$$

The observed output (desired response) d is called the label of the input $\mathbf{x}$, borrowing the terminology from pattern recognition. In function approximation, the labels are continuous-valued and corrupted by noise.[4]

An adaptive system with input $\mathbf{x}$, output y, and weights $\mathbf{w}$ is used to approximate f:

$$y = g(\mathbf{x}, \mathbf{w}) \qquad (2)$$

The adaptive system could be linear filter ($y=\mathbf{w}^T\mathbf{x}$), a neural network, or any other topology whose coefficients need to be optimized for a specific task. In supervised learning, the optimization is carried out by minimizing or maximizing an optimality criterion. The usual choice of this criterion is MSE [1,2], however alternative selections such as minimum error entropy (MEE) [7] or the ε-insensitive loss function [8] are also possible and equally valid. The error is defined as the difference

[1] Throughout this paper, labeled data pairs are those that are available in the form (x,y) where x denotes the input sample and y is the corresponding desired output value. All unlabeled data consists only of x values for which the corresponding output values are not specified or unknown.

[2] We are concerned about *supervised learning*, where a function approximator is optimized. Unsupervised learning, which is based on training from only input samples is not being addressed here.

[3] In some cases, the input and noise samples could be correlated violating the independence assumption. For the sake of argument, we assume independence at this point.

[4] For convenience a single output system is considered here, but the ideas generalize to multidimensional systems.

0-7803-8874-7/05/$20.00 ©2005 IEEE

between the available desired output and the output generated by the adaptive system for a specific input: $e=d-y$. For future use we define the following variables:

$$\tilde{d} = f(\mathbf{x}) \qquad \tilde{e} = f(\mathbf{x}) - g(\mathbf{x},\mathbf{w}) \tag{3}$$

Furthermore, we assume that the input vector is a random variable $\mathbf{X}$ with an unknown probability density function (pdf) $p_{\mathbf{X}}(.)$, the measurement noise N has an unknown pdf $p_N(.)$. The variables $\mathbf{X}$ and N are independent from each other.

In the training phase, an approximation to f could be obtained by minimizing the error (using any viable criterion). In the application phase, where the *trained* network is utilized on novel input data $\{\mathbf{x}_{T+1},\ldots,\mathbf{x}_K\}$, the weights determined in the training phase are fixed. The following question is to be answered: *How do we continue to update the weights of the adaptive network in the application phase?*

3. AN INFORMATION THEORETIC APPROACH

Suppose that we have a data set $\{(\mathbf{x}_1,d_1),\ldots,(\mathbf{x}_T,d_T)\}\cup\{\mathbf{x}_{T+1},\ldots,\mathbf{x}_K\}$. The data is collected as described in the previous section and an adaptive topology $g(.,\mathbf{w})$ is to be optimized.

3.1. Joint Distribution Based Criteria

Consider the joint distribution of the input-desired data:

$$p_{D\mathbf{X}}(d,\mathbf{x}) = p_{D|\mathbf{X}}(d\mid\mathbf{x})p_{\mathbf{X}}(\mathbf{x}) = p_N(d-f(\mathbf{x}))p_{\mathbf{X}}(\mathbf{x}) \tag{4}$$

The substitution for the conditional density is based on (1). The distributions in (4) will be estimated from available data whenever appropriate. Specifically, we can use $\{(\mathbf{x}_1,d_1),\ldots,(\mathbf{x}_T,d_T)\}$ to estimate $p_{D\mathbf{X}}(d,\mathbf{x})$ and $\{\mathbf{x}_1,\ldots,\mathbf{x}_K\}$ to estimate $p_{\mathbf{X}}(\mathbf{x})$. In other words, the unlabeled data that is acquired in the application phase can be used to continually update our estimate for the input distribution.

A natural information theoretic approach is to minimize the Kullback-Leibler (KL) divergence [9] between the estimates of both sides of (4) based on the available data. We propose minimizing $D_{KL}(p_N(d-f(\mathbf{x}))p_{\mathbf{X}}(\mathbf{x})\parallel p_{D\mathbf{X}}(d,\mathbf{x}))$ with respect to $\mathbf{w}$. With a change of variables, it can be shown that this corresponds to minimizing the following expression:

$$\begin{aligned} &D_{KL}(p_N(d-f(\mathbf{x}))p_{\mathbf{X}}(\mathbf{x})\parallel p_{D\mathbf{X}}(d,\mathbf{x})) \\ &\quad= E_{\mathbf{X}N}\left[\log\frac{p_N(N)p_{\mathbf{X}}(\mathbf{X})}{p_{D\mathbf{X}}(N+f(\mathbf{X}),\mathbf{X})}\right] \\ &\quad= E_{\mathbf{X}N}[\log p_N(N)] + E_{\mathbf{X}N}[\log p_{\mathbf{X}}(\mathbf{X})] \\ &\qquad - E_{\mathbf{X}N}[\log p_{D\mathbf{X}}(N+f(\mathbf{X}),\mathbf{X})] \end{aligned} \tag{5}$$

The first and second terms are the entropies of the noise and the input. Therefore, they are constant (under the iid assumption), hence are independent of $\mathbf{w}$. Consequently, it suffices to solve the following to minimize the KL divergence between the two distributions:

$$\min_{\mathbf{w}} -E_{\mathbf{X}}\left[E_{N|\mathbf{X}}[\log p_{D\mathbf{X}}(N+f(\mathbf{X}),\mathbf{X})]\right] \tag{6}$$

It is important to note that the final form of the optimization problem in (6) has been targeted from the beginning, in order to eventually facilitate a stochastic-gradient approach to the training of adaptive systems in application phase. By arriving at a criterion that is expressed as the expectation over $\mathbf{X}$, we can design an extended-learning algorithm that utilizes the incoming samples $\{\mathbf{x}_{T+1},\ldots,\mathbf{x}_K\}$ in an on-line fashion for sample-by-sample updates that still converge in-the-mean to the desired solution.

In a more general framework, one can minimize various definitions of the divergence/distance between the estimates of the distributions on two sides of the equality in (4). Possibilities include Csiszar divergence, Renyi's α-divergence, Euclidean distance, and angular distance (also called Cauchy-Schwartz distance) [10].[5] For example, Renyi's α-divergence leads to the following optimization criterion:

$$\begin{aligned} &D_\alpha(p_N(d-f(\mathbf{x}))p_{\mathbf{X}}(\mathbf{x})\parallel p_{D\mathbf{X}}(d,\mathbf{x})) \\ &\quad= \frac{1}{1-\alpha}\log E_{\mathbf{X}N}\left[\frac{p_N^{\alpha-1}(N)p_{\mathbf{X}}^{\alpha-1}(\mathbf{X})}{p_{D\mathbf{X}}^{\alpha-1}(N+f(\mathbf{X}),\mathbf{X})}\right] \end{aligned} \tag{7}$$

Due to L'Hopital's rule, in the limit as $\alpha\rightarrow1$, (7) approaches (5).

Note that if the noise and input distributions are *known*, then (6) corresponds to a maximum likelihood solution. The generalized version in (7) allows us to manipulate the free parameter α to select how to emphasize denser and sparser regions in the joint $\mathbf{X}N$ probability space in the optimization. We have seen that $\alpha=1$ corresponds to maximum likelihood. Larger α will emphasize dense regions more while smaller α will emphasize sparse regions. This plays a critical role when the input distribution is not uniform over the domain of the function.

Unfortunately, in most realistic situations, the noise distribution is unknown, therefore it needs to be estimated or approximated. We propose to use the error distribution as an approximation to that of the noise. Due to the independence of $\mathbf{X}$ and N, we have

$$p_E(\xi,\mathbf{w}) = p_{\tilde{E}}(\xi,\mathbf{w}) * p_N(\xi) \tag{8}$$

For a general divergence measure (such as (5) and (7), this approximation leads to the following problem:

$$\min_{\mathbf{w}} D(p_E(e,\mathbf{w})p_{\mathbf{X}}(\mathbf{x})\parallel p_{D\mathbf{X}}(e+g(\mathbf{x},\mathbf{w}),\mathbf{x})) \tag{9}$$

Fact 1. The divergence measure in (9) becomes zero if and only if $f(\mathbf{x})=g(\mathbf{x},\mathbf{w})$ for all $\mathbf{x}$ in the support of $p_{\mathbf{X}}(.)$.[6]

Proof. If $f(\mathbf{x})=g(\mathbf{x},\mathbf{w})$ for all $\mathbf{x}$ in the support of $p_{\mathbf{X}}(.)$, then $p_E(.)$ becomes identical to $p_N(.)$, since $p_{\tilde{E}}(.)$ reduces to a Dirac-δ distribution. Due to this and (4), the divergence in (9) becomes zero. Conversely, if the divergence in (9) is zero, then the two distributions in the argument are identical over the support of the distribution $p_{D\mathbf{X}}$. Therefore, it is easy to see that for all possible $(d,\mathbf{x})$ values $p_E(d-g(\mathbf{x},\mathbf{w}))=p_N(d-f(\mathbf{x}))$. Substituting the right hand side of (8) for the error distribution, we have

$$\begin{aligned} &p_{\tilde{E}}(d-g(\mathbf{x},\mathbf{w})) * p_N(d-g(\mathbf{x},\mathbf{w})) \\ &\qquad= p_N(d-g(\mathbf{x},\mathbf{w})-f(\mathbf{x})) \\ &\qquad= p_N(d-g(\mathbf{x},\mathbf{w})) * \delta(d-f(\mathbf{x})) \end{aligned} \tag{10}$$

This means $p_{\tilde{E}}(d-g(\mathbf{x},\mathbf{w})) = \delta(d-f(\mathbf{x}))$. Thus, $f(\mathbf{x})=g(\mathbf{x},\mathbf{w})$ for all $\mathbf{x}$ in the support of $p_{\mathbf{X}}(.)$. □

[5] Definitions of these measures are in the appendix.

[6] In general, it is unlikely that there exists a $\mathbf{w}^*$ such that $f(.)=g(.,\mathbf{w}^*)$. However, the point of this fact is to validate the use of a divergence criterion if it was possible to access the true underlying function, and not to address what happens due to the shortcomings of the approximation topology.

Fact 1 validates the minimization of a divergence measure as shown in (9) for function approximation. The minimization of such a divergence to zero is necessary and sufficient for *exact* function approximation. Unfortunately, in practice, achieving a zero divergence is not always possible. In general, the unknown function $f(.)$ may not be a member of the parametric family of functions spanned by $g(.,\mathbf{w})$. In that case, minimizing a divergence will yield the projection of f onto the manifold where the family of functions $g(.,\mathbf{w})$ lie, wherein the projection itself is determined by the divergence measure utilized.

3.2. Marginal Distribution Based Criteria

The approach presented above, which is based on the joint distribution of the input and desired output variables, will typically be prone to the difficulties associated with the *curse of dimensionality*. For problems with large input dimensionality, estimating the joint distribution $p_{D\mathbf{X}}(d,\mathbf{x})$ from the available labeled training data will become exponentially harder. In order to avoid such difficulties, we can resort to measures that do not consider the input distribution explicitly.

Before continuing further, consider the distribution of the desired output and the adaptive system output error. Due to the independence assumptions, the following identities hold.

$$p_D(\xi) = p_{\tilde{D}}(\xi) * p_N(\xi) \quad p_E(\xi,\mathbf{w}) = p_{\tilde{E}}(\xi,\mathbf{w}) * p_N(\xi) \quad (11)$$

If the adaptive system function matches the true function f perfectly, then the output distribution $p_Y(.)$ becomes identical to $p_{\tilde{D}}(.)$. However, the latter distribution is not available in practice. Consequently, similar to the previous section, we can employ various divergence measures to match $p_Y(.)*p_N(.)$ to $p_D(.)$. Once again, assuming that the error distribution approximates the noise distribution, the following divergence must be minimized:

$$\min_{\mathbf{w}} D(p_Y(\xi,\mathbf{w}) * p_E(\xi,\mathbf{w}) \,\|\, p_D(\xi)) \quad (12)$$

Similar to the reasoning in Fact 1, we can show that (12) becomes zero if and only if $g(\mathbf{x},\mathbf{w})=f(\mathbf{x})$ for all $\mathbf{x}$ in the support of $p_{\mathbf{X}}(.)$. Hence, minimizing (12) is necessary and sufficient for exact function matching. The same argument about f being a member of the function family g can be reiterated here.

4. ALGORITHMIC POSSIBILITIES

For illustration, we consider the minimization of KL divergence formulated in (6). First, an estimate of the joint distribution $p_{D\mathbf{X}}(d,\mathbf{x})$ must be obtained using the labeled portion of the data set $\{(\mathbf{x}_1,d_1),\ldots,(\mathbf{x}_T,d_T)\}$. This could be achieved by parametric or nonparametric approaches.

The parametric approach involves assuming a specific structure for the distribution in terms of a prespecified family of distributions, such as exponential or mixture models. An exponential distribution assumption leads to a simple overall criterion. For example, if we assume that the distributions of interest are of the form

$$p_{D\mathbf{X}}(d,\mathbf{x}) = \exp\big(q(d,\mathbf{x},\boldsymbol{\lambda})\big) \qquad p_E(e) = \exp(r(e,\boldsymbol{\gamma})) \quad (13)$$

where $q(d,\mathbf{x},\boldsymbol{\lambda})$ is a polynomial function of $(d,\mathbf{x})$ with coefficients $\boldsymbol{\lambda}$, and $r(e,\boldsymbol{\gamma})$ is a polynomial over e, minimization of the KL divergence in accordance with (9) becomes

$$\begin{aligned}
&\min_{\mathbf{w}} D_{KL}(p_E(e)p_{\mathbf{X}}(\mathbf{x}) \,\|\, p_{D\mathbf{X}}(e+g(\mathbf{x},\mathbf{w}),\mathbf{x})) \\
&\equiv \min_{\mathbf{w}} E_{\mathbf{X}}\big[E_{E|\mathbf{X}}[\log p_E(E)]\big] \\
&\qquad - E_{\mathbf{X}}\big[E_{E|\mathbf{X}}[\log p_{D\mathbf{X}}(E+g(\mathbf{X},\mathbf{w}),\mathbf{X})]\big] \\
&\equiv \min_{\mathbf{w}} E_E[r(E,\boldsymbol{\gamma})] - E_{\mathbf{X}}\big[E_{E|\mathbf{X}}[q(E+g(\mathbf{X},\mathbf{w}),\mathbf{X},\boldsymbol{\lambda})]\big]
\end{aligned} \quad (14)$$

The criterion is a combination of error entropy and maximum likelihood terms, and with the exponential density assumptions, it simply consists of the moments of the error and joint moments of the input variables. The coefficients of the polynomials can be estimated using the maximum likelihood principle or alternative analytical solutions such as Jaynes' maximum entropy principle [11,12]. Furthermore, in the application phase, the second term can be approximately optimized using a stochastic gradient approach where each new unlabeled input sample is utilized for a single-sample update. Additional complexity reduction could be achieved by also making the expectation over E stochastic. Just for illustration purposes, if these polynomials are quadratic (leading to the unrealistic Gaussian distribution assumption)

$$\begin{aligned}
q(d,\mathbf{x}) &= \begin{bmatrix} d & \mathbf{x}^T \end{bmatrix} \boldsymbol{\lambda}_2 \begin{bmatrix} d \\ \mathbf{x} \end{bmatrix} + \boldsymbol{\lambda}_1^T \begin{bmatrix} d \\ \mathbf{x} \end{bmatrix} + \lambda_0 \\
r(e,\boldsymbol{\gamma}) &= \gamma_2 e^2 + \gamma_1 e + \gamma_0
\end{aligned} \quad (15)$$

and sample means over available data are used to approximate the expectations, the extended-learning criterion in (14) becomes

$$\begin{aligned}
&\min_{\mathbf{w}} \frac{1}{T}\sum_{t=1}^{T}\Big[\gamma_2 e_t^2(\mathbf{w}) + \gamma_1 e_t(\mathbf{w}) + \gamma_0\Big] \\
&\quad - \frac{1}{KT}\sum_{k=1}^{K}\sum_{t=1}^{T}\left(\begin{aligned}&\begin{bmatrix} e_t(\mathbf{w})+g(\mathbf{x}_k,\mathbf{w}) & \mathbf{x}_k^T \end{bmatrix}\boldsymbol{\lambda}_2 \begin{bmatrix} e_t(\mathbf{w})+g(\mathbf{x}_k,\mathbf{w}) \\ \mathbf{x}_k \end{bmatrix} \\ &+ \boldsymbol{\lambda}_1^T \begin{bmatrix} e_t(\mathbf{w})+g(\mathbf{x}_k,\mathbf{w}) \\ \mathbf{x}_k \end{bmatrix} + \lambda_0 \end{aligned}\right)
\end{aligned} \quad (16)$$

Notice that the error samples are evaluated using the labeled data using the most recent weight values and the expectation over $\mathbf{X}$ is evaluated using all available input data (labeled and unlabeled). In practice, the second term could be approximated stochastically using only the most recent (unlabeled) input sample $\mathbf{x}_k$ to approximate the sample average over K samples $\{\mathbf{x}_1,\ldots,\mathbf{x}_K\}$.

The nonparametric approach can be implemented using a Parzen window estimate for the distributions (also called as kernel density estimates) [13,14]. Using appropriate kernel functions $K(.)$, the densities are approximated by

$$\begin{aligned}
p_{D\mathbf{X}}(d,\mathbf{x}) &= \frac{1}{T}\sum_{t=1}^{T} K_{D\mathbf{X}}(d-d_t,\mathbf{x}-\mathbf{x}_t) \\
p_E(e) &= \frac{1}{T}\sum_{t=1}^{T} K_E(e-e_t)
\end{aligned} \quad (17)$$

where d_t, $\mathbf{x}_t$, and e_t are evaluated over the labeled data pairs (training set). In this case, the KL divergence criterion, in accordance with resubstitution estimates of information theoretic measures [10,15], is given in (18). As in the parametric case, computational complexity could be reduced by resorting to stochastic gradients. The stochastic gradient for information theoretic measures using the kernel resubstitution estimates has been studied earlier [16].

$$\begin{aligned}&\min_{\mathbf{w}} D_{KL}(p_E(e)p_{\mathbf{X}}(\mathbf{x}) \,\|\, p_{D\mathbf{X}}(e+g(\mathbf{x},\mathbf{w}),\mathbf{x})) \\ &\equiv \min_{\mathbf{w}} \frac{1}{T}\sum_{u=1}^{T}\log\frac{1}{T}\sum_{t=1}^{T}K_E(e_u-e_t) \qquad (18) \\ &\quad -E_{\mathbf{X}}\left[\frac{1}{T}\sum_{u=1}^{T}\log\frac{1}{T}\sum_{t=1}^{T}K_{D\mathbf{X}}(e_u+g(\mathbf{X},\mathbf{w})-d_t,\mathbf{X}-\mathbf{x}_t)\right]\end{aligned}$$

5. DISCUSSION

In this paper, we have discussed the feasibility of training adaptive systems using unlabeled input data in the application phase. Traditional adaptive systems theory optimizes the weights of the neural network by minimizing an error function evaluated over a labeled training set, where corresponding output values for specific inputs are available. The *trained* network is then applied to novel data (which we call the application phase) and no training of filter coefficients is performed at this stage. Specifically, we have investigated some information theoretic possibilities as the criterion for *training* in the application phase. This adaptation process during actual application is called extended supervised learning, and is essentially unsupervised. However, the criteria for extended supervised learning have been carefully designed to extract the most possible information from the labeled data (training set), as well as the unlabeled data.

The discussion here has been mostly theoretical, while some hints on how to implement these ideas in a practical learning system have been provided. Both parametric and nonparametric statistical approaches have been considered for a feasible implementation and at this point they both seem to be equally viable. However, the performance of these possible implementations as well as different divergence definitions are not studied here due to lack of space. These details will be visited in a future paper.

Acknowledgment: This work is partially supported by the NSF grant ECS-0300340.

APPENDIX

A wide range of possibilities exists for density divergences that could be used in this framework. We list a few of them.

Renyi $D_\alpha(p(\mathbf{x}) \,\|\, q(\mathbf{x})) = (\alpha-1)^{-1}\int p^\alpha(\mathbf{x})q^{1-\alpha}(\mathbf{x})d\mathbf{x}$

Csiszar $D_h(p(\mathbf{x}) \,\|\, q(\mathbf{x})) = \int p(\mathbf{x})h(q(\mathbf{x})p^{-1}(\mathbf{x}))d\mathbf{x}$

Euclidean $D_E(p(\mathbf{x}) \,\|\, q(\mathbf{x})) = \int (p(\mathbf{x})-q(\mathbf{x}))^2 d\mathbf{x}$

Cauchy $D_C(p(\mathbf{x}) \,\|\, q(\mathbf{x})) = \log\dfrac{\left(\int p^2(\mathbf{x})d\mathbf{x}\int q^2(\mathbf{x})d\mathbf{x}\right)^{1/2}}{\int p(\mathbf{x})q(\mathbf{x})d\mathbf{x}}$

Renyi's divergence becomes KL divergence in the limit as $\alpha \to 1$ [17]. In Csiszar divergence, h is a convex function with $h(1)=0$ [18]. If $h(.)=-log(.)$, then we obtain the KL divergence. Euclidean and Cauchy measures are distances drawn from the linear algebra of function spaces [19].

REFERENCES

[1] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed., Prentice Hall, New Jersey, 1999.

[2] A.H. Sayed, *Fundamentals of Adaptive Filtering,* Wiley, New York, 2003.

[3] B. Widrow, S.D. Stearns, *Adaptive Signal Processing*, Prentice Hall, New Jersey, 1985.

[4] A. Blum, T. Mitchell, "Combining Labeled and Unlabeled Data with Co-Training," Proceedings of the Conference on Computational Learning Theory, pg. 92-100, 1998.

[5] K. Nigram, A. McCallum, S. Thrun, T. Mitchell, "Text Classification from Labeled and Unlabeled Documents sing EM," Machine Learning, vol. 39, no. 2, pp. 103-134, 2000.

[6] M. Belkin, P. Niyogi, V. Sindhwani, "Manifold Regularization: A Geometric Framework for Learning from Examples," Technical Report TR-2004-06, Department of Computer Science, University of Chicago, Chicago, Illinois, 2004.

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

[8] N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.

[9] T. Cover, J. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[10] D. Erdogmus, *Information Theoretic Learning: Renyi's Entropy and its Applications to Adaptive System Training*, PhD Dissertation, University of Florida, Gainesville, Florida, May 2002.

[11] E.T. Jaynes, "Information Theory and Statistical Mechanics," Physical Review, vol. 106, pp. 620-630, 1957.

[12] D. Erdogmus, K.E. Hild II, Y.N. Rao, J.C. Principe, "Minimax Mutual Information Approach for Independent Components Analysis," Neural Computation, vol. 16, no. 6, pp. 1235-1252, 2004.

[13] L. Devroye, G. Lugosi, *Combinatorial Methods in Density Estimation*, Springer, New York, 2001.

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

[15] J. Beirlant, E.J. Dudewicz, L. Gyorfi, E.C. van der Meulen, "Nonparametric Entropy Estimation: An Overview," International Journal of Mathematical and Statistical Sciences, vol. 6, no. 1, pp. 17-39, 1997.

[16] D. Erdogmus, J.C. Principe, K.E. Hild II, "On-Line Entropy Manipulation: Stochastic Information Gradient," IEEE Signal Processing Letters, vol. 10, no. 8, pp. 242-245, 2003.

[17] A. Renyi, *Probability Theory*, North-Holland Publishing Company, Amsterdam, 1970.

[18] I. Csiszár, J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1981.

[19] J.C. Principe, J.W. Fisher, D. Xu, "Information Theoretic Learning," in *Unsupervised Adaptive Filtering*, S. Haykin Editor, Wiley, New York, pp. 265-319, 2000.