AN ON-LINE ADAPTATION ALGORITHM FOR ADAPTIVE SYSTEM TRAINING WITH MINIMUM ERROR ENTROPY: STOCHASTIC INFORMATION GRADIENT

Deniz Erdogmus, José C. Príncipe

Computational NeuroEngineering Laboratory
Department of Electrical and Computer Engineering
University of Florida, Gainesville, FL 32611
E-mail: [deniz, principe]@cnel.ufl.edu

ABSTRACT

We have recently reported on the use of minimum error entropy criterion as an alternative to minimum square error (MSE) in supervised adaptive system training. A nonparametric estimator for Renyi's entropy was formulated by employing Parzen windowing. This formulation revealed interesting insights about the process of information theoretical learning, namely information potential and information forces. Variants of this criterion were applied to the training of linear and nonlinear adaptive topologies in blind source separation, channel equalization, and chaotic time-series prediction with superior results. In this paper, we propose an on-line version of the error entropy minimization algorithm, which can be used to train linear or nonlinear topologies in a supervised fashion. The algorithms used for blind source separation and deconvolution can be modified in a similar fashion. For the sake of simplicity, we present preliminary experimental results for FIR filter adaptation using this on-line algorithm and compare the performance with LMS.

1. INTRODUCTION

Mean square error (MSE) has been the focus of optimal filtering and function approximation research since Wiener and Kolmogorov established the perspective of regarding adaptive filters as statistical function approximators [1]. When applied to the FIR filter training, an analytical solution to MSE is given by the Wiener-Hopf equation [2]. In many real-time applications, however, this solution was not practical, hence simple variants of the steepest descent algorithm emerged, LMS by Widrow being the most popular [3]. In contrast, we, among others [4] have proposed the use of error entropy minimization as the performance criterion in adaptation, since manipulation of information is better suited to adaptation rather than merely second order statistics.

Originally, the error entropy minimization algorithm we proposed relied on the use of quadratic Renyi's entropy and the Gaussian kernels in Parzen windowing due to analytical simplicities [5]. Recently, we have formulated a new nonparametric entropy estimator, again with the use of Parzen windowing, which made possible the use of any order of entropy and any suitable kernel function [6]. We have also proved that minimizing the error entropy is equivalent to minimizing the divergence, as defined by Amari in [7], between the output probability density function (pdf) and the desired signal's pdf [8]. In addition, the generalized estimator reduced to the previously utilized estimator for quadratic entropy, for the specific choices of entropy order $\alpha=2$, and Gaussian kernels [6]. With this generalized estimator, it also became possible to generalize the concepts of information potential and information force, to any order α, which were previously defined in the context of blind source separation for the quadratic case [9].

Batch version of the steepest descent algorithm was the basis of our work in minimizing Renyi's entropy. This contrasts with the formulation of InfoMax [10] and FastICA [11], which lead to on-line algorithms that are practical in real-time applications. However, batch learning brought the ability to conduct theoretic work on the adaptation properties of the algorithm. We have recently analyzed the structure of the performance surface in the neighborhood of the optimal solution. We have derived the difference equations, which govern the dynamics of the weights for an FIR filter. This analysis also led to a theoretical upper bound for the step size for stability, as well as an understanding of the effect of the two free parameters α and σ on this structure and the dynamics [12].

In this paper, we propose an approximation to the cost function that yields a stochastic instantaneous estimation

for the gradient. This stochastic gradient can be used both for linear and nonlinear topologies, leading to what we like to call information filtering..

The organization of this paper is as follows. First, we present a brief overview of the generalized information potential criterion for batch adaptation. Next, we derive the instantaneous gradient estimator for information potential and define the stochastic information gradient algorithm. In Section 4, we analyze the convergence of the stochastic information gradient algorithm. Section 5 investigates the connection between the stochastic information gradient and LMS motivated by the previously established link between information potential and MSE. Finally, we present two FIR training examples and conclusions in Sections 6 and 7.

2. ENTROPY CRITERION

Consider the supervised training scheme depicted in Fig. 1. We have previously showed that minimizing Renyi's entropy of the error results in minimization of the α-divergence of Amari [7] between the joint pdfs of the input-desired and input-output signal pairs [8]. In the special case of Shannon's entropy, this is equivalent to minimizing the Kullback-Leibler divergence. From a statistical function approximation point of view, this is exactly what we seek for estimating the unknown system parameters.

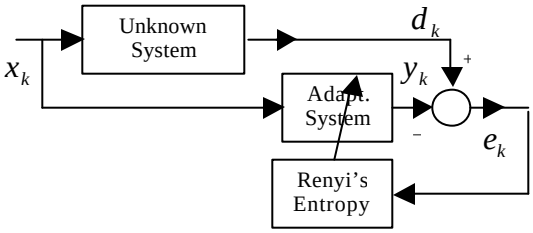

Figure 1. Supervised adaptive system training

Renyi's entropy for a random variable e is defined as [13]

$$H_a(e) = \frac{1}{1-a}\log\int_{-\infty}^{\infty} f_e^a(e)de \qquad (1)$$

where $\alpha>0$ is the order of entropy. In the limit, Renyi's entropy approaches Shannon's entropy as $\alpha\to 1$. Employing L'Hopital's rule easily proves this. In [6], we defined the argument of the log to be the (order-α) information potential. Writing the information potential as an expected value, we get

$$V_a(e) = \int f_e^a(e)de = E\left[f_e^{a-1}(e)\right] \approx \frac{1}{N}\sum_i f_e^{a-1}(e_i) \qquad (2)$$

and then substituting the Parzen window estimation for the pdf [14], we obtain the nonparametric estimator of information potential

$$V_a(e) = \frac{1}{N^a}\sum_j\left(\sum_i k_s(e_j - e_i)\right)^{a-1} \qquad (3)$$

Here, k_s is the kernel function in Parzen windowing with σ denoting the width of the window. In terms of a unit-width kernel κ, the σ-wide kernel is written as

$$k_s(x) = \frac{1}{s}k\left(\frac{x}{s}\right) \qquad (4)$$

The information potential is a crucial quantity because making use of the fact that the log is a monotonic function, we can reduce the minimization of Renyi's entropy to the maximization of the information potential for $\alpha>1$. Therefore, the information potential can replace the entropy criterion in adaptation with computational savings.

We have investigated the mathematical properties of this information potential criterion. One of the most significant properties is its relationship with the convolution smoothing method in global optimization. We have demonstrated that this cost function, in the limit as the number of samples go to infinity, approaches to a convolution smoothed version of the actual information potential, thus eliminating all local optima [6]. Together with the dilation property of this cost function in the weight space for a large kernel size, this link motivated a new approach to steepest descent adaptation. We have proposed the use of a variable kernel size, starting from a large value and decreasing to a preset small value. This way, it is possible to avoid any local minima in the entropy and achieve global optimization even when steepest descent is utilized [6]. Another interesting equivalence materializes when the first order Taylor approximation to the kernel function satisfies specific conditions and the kernel size is very large. In this case, as we have shown, the information potential criterion reduces to the MSE criterion, thus MSE is a limit case of our error entropy [6]. One final remark we would like to make at this point is that, after extensive simulations we have not observed any local minima for the entropy in FIR filter training, thus we conjecture, based upon this observation that the information potential given in (3) has a unique maximum, which in turn is the global optimum solution.

3. STOCHASTIC GRADIENT

Suppose the adaptive system under consideration in Fig. 1 is a linear (e.g. FIR) or nonlinear (e.g. TDNN) topology with a weight vector w. The error samples can be

represented by $e_k = d_k - y_k$. Then the gradient of the information potential estimator given in (3) with respect to the weight vector is simply

$$\frac{\partial V_a}{\partial w} = \frac{(a-1)}{N^a} \sum_j \left(\sum_i k_s (e_j - e_i) \right)^{a-2} \cdot \left(\sum_i k_s' (e_j - e_i)(\frac{\partial y_i}{\partial w} - \frac{\partial y_j}{\partial w}) \right) \quad (5)$$

The obvious approach to obtain a stochastic gradient would be to drop the expected value operator in the information potential expression in (2), which would eliminate in (5) the summation over the index j, and use a window of samples for i, extending back in time to estimate the pdf and hence the gradient associated with this stochastic version of the criterion.

Since all data are considered in pairs in the adopted formulation of the entropy, it is convenient to regard the training data set for this problem, which consists of N input-desired output pairs of the form $\{(x_1, d_1), ..., (x_N, d_N)\}$, in an alternative indexation. This results in a new training data set in the form $\{(x_{ij}, d_{ij})\}_{i,j=1,...,N}$, where

$$x_{ij} = x_i - x_j$$
$$d_{ij} = d_i - d_j \quad (6)$$

We can define new variables for the error samples and the gradients of the output with respect to the weights for different input vectors similarly, where the double-index denotes subtraction in the appropriate order. With this new notation, our gradient expression in (5) simplifies to

$$\frac{\partial V_a}{\partial w} = \frac{(1-a)}{N^a} \sum_{ji} C_j(a, s) k_s'(e_{ji}) \frac{\partial y_{ji}}{\partial w} \quad (7)$$

where the summation over ji is performed over the index of the new training set. When writing (7), we defined

$$C_j(a, s) = \left(\sum_i k_s (e_j - e_{j-i}) \right)^{a-2} \quad (8)$$

This new formulation of the gradient in terms of the samples in an alternative expression of the training set allows us to see clearly now how to obtain a stochastic gradient estimate for the information potential using the most recent sample only. The stochastic gradient at step k obtained by this approximation, considering only the most recent sample for k and a window of samples extending back in time from k for i, is now written as

$$\frac{\partial \hat{V}_a}{\partial w} = \frac{(1-a)}{N^{a-1}} C_k(a, s) \left(\sum_{i=0}^{N-1} k_s' (e_{k,k-i}) \frac{\partial y_{k,k-i}}{\partial w} \right) \quad (9)$$

where N is the window length. Notice that in (9), the term corresponding to i=0 is zero. The expected value of this stochastic gradient expression is equal to the actual gradient in (5), as will be proven in the next section.

We can further simplify the stochastic gradient expression. As far as the new training data set, which consists of differences of pairs, is concerned, the most recent time index is (k,k-1). As in LMS, we now consider this instantaneous gradient as a stochastic estimation to the actual gradient in (8), and write the instantaneous gradient at time instant k as follows.

$$\left(\frac{\partial V_a}{\partial w} \right)_k = \frac{(1-a)}{2^{a-1}} C_k(a, s) k_s'(e_{k,k-1}) \frac{\partial y_{k,k-1}}{\partial w} \quad (10)$$

Notice that this stochastic information gradient (SIG) expression corresponds to the drastic choice of N=2 in (9). Now, using this instantaneous gradient estimator, we can maximize the information potential using a steepest ascent algorithm with the following update.

$$w_{k+1} = w_k + h \left(\frac{\partial V_a}{\partial w} \right)_k \quad (11)$$

The evaluation of this instantaneous gradient given in (10) requires a single kernel evaluation (no summation) for $C_k(a, s)$ at $e_{k,k-1} = e_k - e_{k-1}$ if $a \neq 2$, and a single evaluation of the derivative of the kernel function at the same point $e_{k,k-1}$. If the kernel function is chosen to be a Gaussian, its derivative is simply a product of its argument and the Gaussian evaluated at the same point again, hence one needs only a single kernel evaluation to compute both $C_k(a, s)$ and k_s' terms.

For FIR filter adaptation, further simplification is possible. In this case, the gradient of the output with respect to the weight vector is simply the input vector, thus we substitute

$$\frac{\partial y_{k,k-1}}{\partial w} = \frac{\partial y_k}{\partial w} - \frac{\partial y_{k-1}}{\partial w} = x_k - x_{k-1} \quad (12)$$

In this section, we have derived an instantaneous gradient estimator for the information potential following a methodology similar to the LMS algorithm. Only this time we have used a modified training data set that consists of pair-wise differences of the samples in contrast to using the actual input-output pairs as done in LMS. Hence, this viewpoint involving an alternative training data set enabled us to discover the similarities between our stochastic information gradient algorithm and LMS, as

well as it clarifies the differences. We have observed that the stochastic information gradient is similar to Widrow's instantaneous gradient estimator in LMS in structure when we introduce the modified training set, except for the scaling factors that depend on the kernel evaluation and the entropy order. These extra scale factors, however, will lead to a smoother convergence both in terms of the cost function and the weight trajectories as we will see later.

Obviously, when the drastically reduced SIG given in (10) is used in adaptation, the weight tracks will be noisy. One way to get a smoother convergence of the weights is to use an average of the stochastic gradient in (10) over a window of samples and update once at the end of every window by this average gradient. This is given by

$$\overline{\left(\frac{\partial V_a}{\partial w}\right)}_k = \frac{1}{L}\sum_{j=k-L+1}^{k} \frac{(1-a)}{2^{a-1}} C_j(a,s)k'_s(e_{j,j-1})\frac{\partial y_{j,j-1}}{\partial w} \qquad (13)$$

Note that this is different from the gradient expression suggested in (9) as the averaging is done over the gradients of a window of consecutive sample pairs, whereas in (9), the gradient is evaluated over pairs of samples formed by the instantaneous sample and a window of samples that preceded it.

4. CONVERGENCE IN THE MEAN FOR FIR FILTERS TRAINED WITH SIG

One compelling property of LMS is its convergence in the mean to the Wiener-Hopf solution. This property is mathematically stated as

$$\lim_{k\to\infty} E[w_{k+1} - w_*] = 0 \qquad (14)$$

Here, w_* is the optimal MSE solution. In addition to this, it is also known and easy to prove that for a quadratic cost surface, the expected value of Widrow's instantaneous gradient is equal to the actual gradient vector [2]. This fact will become useful also in proving the convergence in the mean for the stochastic information gradient later.

Now, consider the exact information gradient given in (5) and the stochastic gradient given in (6) or (10). We will denote the former by $\nabla V(w_k)$ and the latter by $G(w_k)$ evaluated at a given weight vector w_k. Since the stochastic gradient is derived as an approximation to the exact gradient simply by dropping out terms in the summation, we can write

$$\nabla V(w_k) = G(w_k) + H(w_k) \qquad (15)$$

where $H(w_k)$ contains all the left over terms. The weight updates are done using the stochastic gradient as given in (11). Substituting (15),

$$\begin{aligned} w_{k+1} &= w_k + h\ G(w_k) \\ &= w_k + h\ \left(\nabla V(w_k) - H(w_k)\right) \\ &= w_k + h\ \nabla V(w_k) - h\ H(w_k) \end{aligned} \qquad (16)$$

We have mentioned that the expected value of Widrow's stochastic gradient estimator for MSE is equal to the actual gradient. This is due to the quadratic nature of the performance surface with respect to the weights. We have shown in [12] that the information potential becomes quadratic in the vicinity of the optimal solution. Furthermore, we proved that, we are able to control the volume of the region, where the performance surface is approximately quadratic, simply by adjusting the size of the kernel function used in the Parzen window estimator [6]. Combining this fact and the fact that the weights will converge to the vicinity of w_* as the number of iterations increases, due to the existence of a single maximum, we conclude that in the limit the expected value of the stochastic information gradient is equal to the exact information gradient.

$$\lim_{k\to\infty} E[G(w_k)] = \nabla V(w_k) \qquad (17)$$

Employing the identity given in (15), we have

$$\lim_{k\to\infty} E[H(w_k)] = 0 \qquad (18)$$

Consequently, the update equation given in (16) reduces to the steepest ascent algorithm in the mean, that is

$$E[w_{k+1}] = E[w_k] + h\ \nabla V(w_k) \qquad (19)$$

In the limit, the two expected values in (19) converge to the same value hence the exact gradient of the information potential converges to zero. Thus the weight vector converges to the optimal solution, w_*.

5. RELATIONSHIP OF SIG WITH LMS

In [6], we have shown that for the quadratic entropy, if the kernel function satisfies certain conditions, the minimization of error entropy approaches to minimizing MSE in the limit when the kernel size is increased. This equivalence in the limit encourages us to investigate any possible connection between the SIG and LMS. Consider the instantaneous gradient estimator for information potential given in (10). When the kernel size is very large (or when the weights are very close to the optimal solution such that the entropy of the error is small), we can

approximate the kernel's derivative around zero by a line (with negative slope). Also, choosing $a = 2$ eliminates the coefficient $C_k(a, s)$ due to the power $(a-2)$.

$$G(w_k) = \bar{c} \cdot e_{k,k-1} \cdot x_{k,k-1} \tag{20}$$

where the coefficient $\bar{c}$ gets smaller with increasing kernel size. Clearly seen from (20), for the choice of quadratic entropy and Gaussian kernels, SIG is nothing more than the LMS applied to the instantaneous increments of the error and the input vector.

6. SIMULATIONS

In order to demonstrate the SIG algorithm at work, we present two simulation results in this section, both using the quadratic entropy and the gradient given in (10). The first example considered here is a time-series prediction problem. The time-series to be predicted is a sum of three sinusoids given by

$$x(t) = \sin 20t + 2\sin 40t + 3\sin 60t \tag{21}$$

sampled at 100Hz. The training sample consists of 32 samples, which corresponds to approximately one period of the signal. In this example, we use different step sizes for SIG (0.1) and LMS (0.001). These choices guarantee that the convergence of both algorithms occurs at around 150 epochs. The results for this case are presented in Fig. 2. Notice that, since we have decreased the step size of LMS, the weight trajectories are now much smoother, but still slightly jagged compared to those of the SIG.

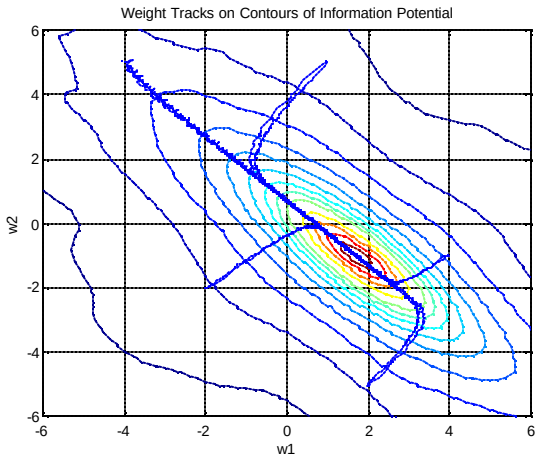

Figure 2. Weight Tracks for SIG (solid) and LMS (dotted) on the Contour Plot of the Information Potential Function – Sinusoidal Signal Prediction

For the fixed convergence time, SIG is observed to converge smoother than LMS. The reason for this is, the information potential relies on pair-wise differences of the error samples and in addition, the variance of these differences are made even smaller since they are passed through the kernel function. This constriction acts like a lowpass filter on the gradient vector and prevents the gradient vectors of consecutive iterations from fluctuating drastically. It also enables SIG to use larger learning rates, values at which LMS would become unstable. For example, LMS would become unstable in this problem if it used the step size assumed by SIG. In general, one would start with a very small step size to be on the safe side, but this will reduce the convergence speed. Clearly, using the stochastic gradient avoids this problem of determining a suitable step size, because the upper limit on the step size for stability of the algorithm is higher compared to LMS.

The second example is the frequency doubling problem, where a 2-tap FIR filter is trained to double the frequency of the sinusoid presented at its input. The motivation for this example is to investigate the behavior of the SIG algorithm and the shape of the entropy as a cost function in a case where we know the optimal solution has large approximation errors. Clearly, the performance of a 2-tap FIR filter will be limited in this problem. Shown in Fig. 3 below, there is a unique maximum and two local minima.

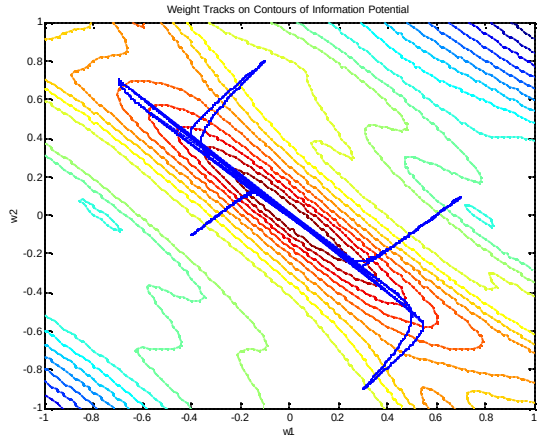

Figure 3. Weight Tracks for SIG (solid) and LMS (dotted) on the Contour Plot of the Information Potential Function – Frequency Doubling

Since it is time consuming to evaluate the information potential on a grid in the weight space for a large number of training samples, we have restricted the number of training samples to 20 and performed the weight updates for both the SIG and LMS over the data 1000 epochs to be consistent with the contour plots. However, the algorithm works perfectly well using only a single pass over a larger training set. The kernel size used to obtain this contour plot is a relatively small value (σ=0.2). When a larger kernel size is utilized, the area of the region where the quadratic approximation is accurate increases, as expected.

Notice that the contours of the information potential for weight vectors that are far from the optimal solution have changed shape compared to the previous example; however, near w_* both performance surfaces are accurately represented by a quadratic function.

7. CONCLUSIONS

We have recently proposed the minimization of Renyi's error entropy as an alternative adaptation criterion, to perform information theoretical learning. Initially, we have employed Renyi's quadratic entropy, due to analytical complexities involved with other orders of entropy. as the criterion with success in blind source separation, equalization, and chaotic time-series prediction. Motivated by good results, we sought for extensions to any order of Renyi's entropy. The extended α-order nonparametric estimator reduced to the previously used estimator for quadratic entropy, and performed equally well in the same problems for various orders of entropy. In fact, we believe there is an optimal choice of the entropy order and the kernel function for each specific adaptation problem, however, at this time it is not clear how to achieve this and we pose it as an open question.

One drawback of the entropy algorithms was that they had to work in batch mode. With the work presented in this paper, we overcame this shortcoming and proposed a successful on-line version through an approximation of the gradient vector, which we named the stochastic information gradient (SIG). We have proved that SIG converges to the optimal solution of the actual entropy minimization problem in the mean. We have investigated the relationship between the LMS and the stochastic information gradient in the limiting case of a large kernel size, and found out that the latter incorporates some momentum terms into the update algorithm compared to that of LMS. Also presented here were two comparative examples of FIR training, where the advantages of SIG over the classical LMS have been demonstrated.

Although not reported here, we have successfully applied the stochastic information gradient to blind source separation, blind deconvolution, and projection pursuit problems. Those results will be presented elsewhere.

Further studies must be conducted on the properties of entropy minimization criterion and the convergence properties of the stochastic information gradient algorithm to set up a solid information-theoretic learning theory, similar to that behind MSE criterion and the LMS algorithm. Items of primary interest are the counterparts of the misadjustment and excess MSE. In addition, a rigorous proof of the non-existence of local optimum solutions will be useful.

Acknowledgments: This work was supported by the NSF grant ECS-9900394.

REFERENCES

[1] Wiener N., *Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications*, MIT Press, Cambridge, MA, 1949.

[2] Haykin, S., *Adaptive Filter Theory*, 3rd ed., Prentice Hall, Inc., NJ, 1996.

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

[4] Casals, J.S., Jutten, C., Taeb, A., "Source Separation Techniques Applied to Linear Prediction," in Proceedings of Independent Component Analysis 2000, pp.193-198, Helsinki, Finland, June 2000.

[5] Erdogmus D., J.C.Principe, "Comparison of Entropy and Mean Square Error Criteria in Adaptive System Training Using Higher Order Statistics", in Proceedings of Independent Component Analysis 2000, pp. 75-80, Helsinki, Finland, June 2000.

[6] Erdogmus, D., J.C.Principe, "Generalized Information Potential Criterion for Adaptive System Training," submitted to IEEE Transactions on Neural Networks, Feb. 2001.

[7] Amari, S., *Differential–Geometrical Methods in Statistics*, Springer-Verlag, Berlin, 1985.

[8] Erdogmus, D., J.C.Principe, "An Entropy Minimization Algorithm for Short-Term Prediction of Chaotic Time Series," submitted to IEEE Transactions on Signal Processing, Sept. 2000.

[9] Principe, J.C., D.Xu, J.Fisher, Information Theoretic Learning, in *Unsupervised Adaptive Filtering*, vol I, Simon Haykin Editor, 265-319, Wiley, 2000.

[10] Bell, A., Sejnowski, T., "An Information-Maximization Approach to Blind Separation and Blind Deconvolution," *Neural Computation*, vol. 7, pp. 1129-1159, 1995.

[11]Hyvarinen, A., "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, pp. 626-634, 1999.

[12] Erdogmus, D., J.C.Principe, "Convergence Analysis of the Information Potential Criterion in ADALINE Training,", accepted to Neural Networks in Signal Processing 2001, Falmouth, Massachusetts, USA, Dec 2001.

[13] Renyi, A., *Probability Theory*, American Elsevier Publishing Company Inc., New York, 1970.

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