
Natural Computing **1:** 85–108, 2002.

Beyond second-order statistics for learning: A pairwise interaction model for entropy estimation

DENIZ ERDOGMUS, JOSE C. PRINCIPE and KENNETH E. HILD II
Computational NeuroEngineering Laboratory, Electrical & Computer Engineering Department, University of Florida, Gainesville, FL 32611, USA

Abstract. Second order statistics have formed the basis of learning and adaptation due to its appeal and analytical simplicity. On the other hand, in many realistic engineering problems requiring adaptive solutions, it is not sufficient to consider only the second order statistics of the underlying distributions. Entropy, being the average information content of a distribution, is a better-suited criterion for adaptation purposes, since it allows the designer to manipulate the information content of the signals rather than merely their power. This paper introduces a nonparametric estimator of Renyi's entropy, which can be utilized in any adaptation scenario where entropy plays a role. This nonparametric estimator leads to an interesting analogy between learning and interacting particles in a potential field. It turns out that learning by second order statistics is a special case of this interaction model for learning. We investigate the mathematical properties of this nonparametric entropy estimator, provide batch and stochastic gradient expressions for off-line and on-line adaptation, and illustrate the performance of the corresponding algorithms in examples of supervised and unsupervised training, including time-series prediction and ICA.

Key words: adaptation, information theory, learning, Renyi's entropy

1. Introduction

The mean square error (MSE) has been the workhorse of optimal data fitting models since the early work of Gauss in the 19th century. Both optimal linear filtering and pattern recognition formulations have utilized extensively MSE for very good reasons. In data fitting with the linear model, MSE yields a solution that is linear in the weights and can be analytically computed (the famous least square method). Under the Gaussian assumption for the error, the MSE provides the maximum likelihood solution, and so it has gained acceptance in parameter estimation (Scharf 1990). The classical work of Wiener on optimal filters in the MSE sense provided the theoretical framework (Wiener 1949) and the stochastic gradient by Widrow (Widrow and Stearns 1985), which gave rise to the LMS algorithm, tremendously decreased the computational complexity of adapting filters, and more importantly opened new horizons for adaptive systems.

The MSE is also very popular in pattern recognition, in spite of the fact that minimizing the power of the output error of a classifier does not guarantee in general minimal classification error (Fukunaga 1972). We know today that minimizing the MSE at the output of a classifier provides an estimate for the a posteriori probability of the class given the sample (Bishop 1995). Although the MSE approach has well-established properties for linear systems in the case of Gaussian distributed signals, the shortcomings of the method become evident in the nonlinear systems, non-Gaussian signals case (Deco and Obradovic 1996). Under these circumstances, one needs an adaptation criterion that utilizes the higher order statistical properties of the signal under consideration.

The concept underlying the use of MSE in optimal system design is very appealing. The error between the output of the system and a desired response is a measure of mismatch; hence one should minimize the error for optimal performance. A productive way is to minimize its variance, or the error energy. The variance is intimately related to correlation, so MSE is really only constraining the second order statistics of the error. Effectively, to transfer all the information from the desired response to the parameters of the mapper, we should constrain *all* the moments of the probability density function of the error, i.e. we would like to make the error approach a delta function distribution. When the error is assumed Gaussian, the minimization of the variance (achieved with MSE) leads to the best possible solution.

When the error is not Gaussian, higher order moments of the probability density function of the error should also be brought into play for optimality. Entropy, defined as the average information content of a probability distribution by Shannon (Shannon 1948), is a measure of uncertainty of the underlying error distribution. Being the expectation of a function of the probability distribution, entropy inherently encompasses higher order statistics of the density. Thus entropy is a suitable candidate for an adaptation criterion to manipulate the *information content* of the signals rather than operating on the second order statistics. Recently, Shannon's entropy and mutual information, Kullback-Leibler divergence, and other forms of higher order statistics (e.g. kurtosis and higher order cumulants) have found their ways into the adaptive systems literature in the context of independent components analysis (ICA) and blind deconvolution (Lee et al. 1997; Hyvarinen 1999; Bell and Sejnowski 1995; Comon 1994; Yang and Amari 1997).

When a Gaussian distribution appropriately models the error, Shannon's entropy still provides a manageable option for design, as the extensive work in communication theory clearly demonstrates (Cover and Thomas 1991). Unfortunately, whenever nonlinear systems are the basis for system modeling, the Gaussian assumption fails the realism test, and it has been

difficult to apply Shannon's definition of entropy as a criterion for adaptive systems training without invoking the Gaussian assumption.

In this paper, we submit that Shannon's definition is not always the most appropriate choice for a given problem. In fact, by utilizing Renyi's parametric definition of entropy (or mutual information), which includes Shannon's definition as a special case, the designer gains extra freedom by choosing the free parameter that is introduced into the scheme. Even more importantly, Renyi's definition of entropy leads to a practical estimator for entropy directly from samples when combined with a Parzen estimator (i.e. nonparametric estimator). Therefore, the MSE can be simply substituted by the new entropy estimator in any practical problem. After working on this problem for more than four years, we dare to say that Shannon's definition of entropy does not provide an intuitive understanding of what is gained when using entropy instead of MSE for training an adaptive system.

Renyi's entropy and mutual information are parametric families described by (Renyi 1970)

$$\begin{aligned} H_\alpha(X) &= \frac{1}{1-\alpha}\log\int_{-\infty}^{\infty} f_X^\alpha(x)dx \\ I_\alpha(X;Y) &= \frac{1}{\alpha-1}\log\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{f_{XY}^\alpha(x,y)}{f_X^{\alpha-1}(x)\cdot f_Y^{\alpha-1}(y)}dxdy \end{aligned} \tag{1}$$

where X and Y are two random variables with the designated marginal and joint probability density functions (pdf), and α is the order parameter. As can be shown, using L'Hopital's rule, the limit of Renyi's definitions for entropy and mutual information as α approaches to one yields Shannon's definitions. Hence, even though Renyi's parametric entropy family has a discontinuity at $\alpha = 1$, one can approximate Shannon's entropy arbitrarily close.

Finally, Renyi's entropy provides a "physical analogy" for learning from samples by means of an interaction model, which is increasing tremendously our understanding about entropic learning. In fact, Renyi's entropy combined with the Parzen estimator in (Parzen 1967), computes interactions among pairs of samples, which has analogies with physical potential fields (and that we called an *information potential* in the context of information theoretic learning (Principe et al. 2000)). Moreover, when further steps are taken in order to obtain a stochastic approximation of the gradient vector for sample-by-sample training purposes, the resulting expressions reveal invaluable insights about the relations between entropy training and the LMS algorithm, and even the biologically plausible Hebbian learning. At this point we can state that entropic learning with the information potential is the natural extension for MSE learning in adaptive systems, because it extracts more information from the data samples during the learning process.

As the reader may already expect, this paper is intended as a review of the fundamental concepts, so it will be light in the mathematics and will emphasize understanding and relationship with more traditional concepts. For a more in depth analysis, we provide the appropriate references.

2. Entropy estimator

In this section, a nonparametric estimator for Renyi's entropy is derived, the information potentials and information forces are defined and their role in adaptation is pointed out. We start by writing the entropy definition in (1) in a different way, using the expectation operator.

$$H_\alpha(X) = \frac{1}{1-\alpha} \log E[f_X^{\alpha-1}(X)] \tag{2}$$

The Parzen window estimator [15] for the pdf $f_X(.)$ is evaluated using a kernel function $\kappa_\sigma(.)$, where σ is a parameter that controls the width of the kernel function.

$$\hat{f}_X(x) = \frac{1}{N} \sum_{i=1}^{N} \kappa_\sigma(x - x_i) \tag{3}$$

In the multidimensional pdf estimation case, this can be a vector or the covariance matrix of the kernel function. In general, we suggest using joint kernels of the type

$$\kappa_\Sigma(x) = \prod_{o=1}^{n} \kappa_{\sigma_o}(x^o) \tag{4}$$

where x^o is the oth component of the input vector. This multi-dimensional kernel used to estimate the joint pdf is equal to the product of single dimensional kernels used for estimating the marginal pdfs. In this way, the joint pdf estimation performed with this multi-dimensional kernel and the marginal density estimates evaluated using the individual single-dimensional kernels are consistent.

We can now replace the expected value in (2) by the sample mean and obtain the following nonparametric estimator for Renyi's entropy (Erdogmus and Principe 2001; Erdogmus et al. 2002).

$$H_\alpha(X) \approx \frac{1}{1-\alpha} \log \frac{1}{N_\alpha} \sum_{j=1}^{N} \left(\sum_{i=1}^{N} \kappa_\sigma(x_j - x_i) \right)^{\alpha-1} \tag{5}$$

This nonparametric estimator allows the designer to choose any entropy order α and any kernel function. For the special choices of quadratic entropy ($\alpha = 2$), and Gaussian kernels, (5) reduces to the estimator defined by Principe except for a change in kernel size (Principe et al. 2000).

$$\begin{aligned} H_2(X) &\approx -\log V_2(X) \\ V_2(X) &= \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} G_{\sigma\sqrt{2}}(x_j - x_i) \end{aligned} \tag{6}$$

It is interesting to point out that this definition is achieved without any sample approximations as in (5) due to the mathematical properties of the Gaussian kernel. We therefore conclude in (Erdogmus et al. 2001) that for Renyi's entropy evaluation the additional variance introduced by the sample mean approximation of the expected value operator can be *exactly* compensated by a change of the shape and size of the kernel in the Parzen window.

The properties of Renyi's entropy for estimation have been studied extensively in the statistical literature (Renyi 1970). However, we are going to use Renyi's entropy in an adaptation framework, therefore the estimator in (5) requires further study. The key concern is if the extrema locations of Renyi's entropy are preserved when (5) is used as an estimator. We have proven that (Erdogmus and Principe 2001).

Theorem 1. If the kernel function κ_σ is a symmetric, continuous and differentiable pdf, then the global minimum of the nonparametric entropy estimator in (5) occurs when all samples x_j are equal. Furthermore, this minimum is smooth.

Proof: In (Erdogmus and Principe 2001), it was shown that the eigenvalues of the Hessian matrix of (5) at the minimum point were strictly positive (except for a single zero eigenvalue due to the fact that entropy is invariant to the mean of the pdf) if the kernel is positive, its derivative is zero at zero, and its second derivative is negative when evaluated at zero. A kernel as described in the theorem satisfies all these conditions. This also proves the smoothness. By comparing the value of (5) for an arbitrary set of samples to that of the case where all the samples are identical, the global nature of the solution is shown. □

3. Information particles and their potential

$V_2(X)$ in the quadratic entropy definition (6) was named the *information potential* (Principe et al. 2000). The reason is the observation that $V_2(X)$

is a positive function monotonically decreasing with the distances between the samples. Using an analogy to physics, the samples can then be thought of as *information particles* in an information field. From (5) we obtain the following nonparametric estimator of the information potential

$$V_\alpha(X) = \int f_X^\alpha(x)dx \approx \frac{1}{N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\kappa_\sigma(x_j - x_I)\right)^{\alpha-1} \tag{7}$$

Having defined the information potential, we can then introduce the information potential created by a single particle x_j. Towards this goal, we will use again the fact that the total potential energy of a system of particles is a summation of the potential energies of the individual particles. Thus, the potential for x_j, parametrically dependent on the entropy order and the kernel size as well as the kernel function itself, becomes

$$\hat{V}_\alpha(x_j) - \frac{1}{N^\alpha}\left(\sum_i \kappa_\sigma(x_j - x_i)\right)^{\alpha-1} \tag{8}$$

From this, the definition of the *information force* acting on this particle follows immediately. Recall that the force acting on a particle due to a potential field is calculated by taking the derivative of that field with respect to the position of the particle. Following the same principle, the α-order information force on x_j is

$$\begin{aligned} F_\alpha(x_j) &= \frac{\partial \hat{V}_\alpha(x_j)}{\partial x_j} = \frac{(\alpha-1)}{N^\alpha}\left(\sum_i \kappa_\sigma(x_j - x_i)\right)^{\alpha-2}\left(\sum_{i\neq j}\kappa'_\alpha(x_j - x_i)\right) \\ &= (\alpha-1)\hat{f}_X^{\alpha-2}(x_j)F_2(x_j) \end{aligned} \tag{9}$$

To obtain the closed form in the second line in (9), the sum of the kernels is collected in the pdf estimate for the particle x_j and the quadratic force is defined as in Principe et al. (2000).

$$F_2(x_j) = \frac{1}{N^2}\left(\sum_{i\neq j}\kappa'_\sigma(x_j - x_i)\right) \tag{10}$$

The choice of the kernel and the Renyi's entropy order affect the interaction among the information particles. Equation (9) is important because it explains how the entropy order choice affects the behavior of the adaptation

algorithm. As will be demonstrated later, the information forces are an essential component of the gradient (or the natural gradient (Amari 1998) if used) and their behavior dominates the behavior of the gradient.

One can then ask what is the force component acting on a particle due to another specific particle. Since forces are also additive, we can decompose the quadratic force in (9) into its components composed of each element in the summation and define the α-order information force acting on x_j due to particle x_i as

$$F_\alpha(x_j; x_i) = (\alpha - 1)\hat{f}_e^{\alpha-2}(x_j)F_2(x_j; x_i) \tag{11}$$

An interpretation of the results obtained here is in order. Consider the α-force in (9). It is clear that for $\alpha > 2$ the quadratic force is scaled up in magnitude for samples in dense regions of the sample space as their density estimates will have values greater than one. Similarly, for $\alpha < 2$ the force acting on samples in the sparse regions will be scaled up. Thus, it is possible to choose the entropy order α to emphasize dense or sparse regions of the sample space, which is linked to the kurtosis of the distributions. Choosing $\alpha = 2$ will put no emphasis on either region. In this respect the entropy order behaves as the p in L_p Euclidean space norms.

The formulation that has been presented above develops entropic training as a pair-wise interaction model among training samples. Notice that the kernel function chosen determines the *potential field* that emanates from a specific *information particle*, thus leading to the information force notion that these particles exert on each other during adaptation. As expected from the analogy formed between this framework and physics, these forces depend on the relative locations of these particles with respect to each other.

Even more interesting, MSE can be shown to be a special case of this framework. It is possible to regard the behavior of the adaptation algorithm arising from the MSE criterion as an interaction between the training samples and their sample mean. The following theorem states the equivalence between MSE and quadratic entropy in the context of supervised training.

Theorem 2. If the kernel function satisfies the conditions in Theorem 1, then minimizing the quadratic entropy expression given in (5) is equivalent to minimizing the variance in the limit as the kernel size tends to infinity.

Proof: Recall that minimizing the quadratic entropy is equivalent to maximizing the quadratic information potential. Consider a second order truncated Taylor series approximation to the kernel function, expanded around the origin.

$$\begin{aligned}\kappa_\sigma(\xi) &\approx \kappa_\sigma(0) + \kappa'_\sigma(0)\xi + \kappa''_\sigma(0)\xi^2/2 \\ &= \kappa_\sigma(0) + \kappa''_\sigma(0)\xi^2/2 \end{aligned} \tag{12}$$

When a symmetric kernel as described in Theorem 1 is used the first order derivative at the origin is zero, hence the corresponding term drops. Now we substitute this second order approximation for the kernel in the quadratic information potential and obtain

$$\begin{aligned}V_\alpha(X) &\approx \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N}(\kappa_\sigma(0) + \kappa''_\sigma(0)(x_j - x_i)/2) \\ &= \kappa_\sigma(0) + \kappa''_\sigma(0)/2\left(\frac{1}{N^2}\sum_j\sum_j(x_j^2 - 2x_jx_i + x_i^2)\right) \\ &= \kappa_\sigma(0) + \kappa''_\sigma(0)\cdot\left(\overline{x^2} - \bar{x}^2\right) \end{aligned} \tag{13}$$

Since the second derivative of the kernel evaluated at zero is negative, maximizing the information potential shown above, estimated using a very large kernel size (much larger than the dynamic range of the data), is equivalent to minimizing the sample variance. In order to force the sample mean to zero, an additional term involving the square of the sample mean could be introduced. Note also that the information potential is a biased estimator for the MSE. □

This understanding highlights the unifying perspective brought by the proposed particle interaction model as a novel model for learning from examples, and leads to our conviction that the proposed entropy criterion indeed exploits more information about the data set than second order statistics.

4. Particle interaction model for learning

The equivalence between the sample entropy and the sample variance in the limit of large kernels given in Theorem 2 brings the question of how potential fields and forces look like when MSE (or rather the variance) is used as the criterion. As we will see now, the interaction in the MSE case will be between each individual sample and the sample mean, as if a collective (quadratic) potential field is generated by a single particle located at the sample mean. Consider the following maximization problem, which is nothing but the minimization of the sample variance.

$$V_{MSE} = \frac{-1}{N}\sum_i(x_i - \bar{x})^2 \tag{14}$$

where $\bar{x}$ denotes the sample mean. Conceiving this quantity as the total potential of the particle system, we can write the contribution of a single particle as follows.

$$V_{MSE}(x_i; \bar{x}) = \frac{-1}{N}(x_i - \bar{x})^2 \tag{15}$$

where the source of the potential field is the sample mean and the potential of a particle depends on its (Euclidean) distance to the source. Generalizing from this, at any point x, the potential is given by

$$V_{MSE}(x; \bar{x}) = \frac{-1}{N}(x - \bar{x})^2 \tag{16}$$

Thus, the potential in the case of second order statistics is quadratic with respect to the distance from the source. Now, let's investigate the corresponding force acting on a particle x. The force is naturally defined as the derivative of the potential with respect to the position of the particle.

$$F_{MSE}(x) = \frac{\partial V_{MSE}(x)}{\partial x} = \frac{-2}{N}(x - \bar{x}) \tag{17}$$

Now let's consider the quadratic force in the case of a large kernel, where the second order approximation in (12) is valid. In that case, the derivative of the kernel can be approximated by a linear expression, explicitly given by $\kappa'_\sigma(\xi) \approx \kappa''_\sigma(0)\xi$, therefore, the quadratic force in (10) approximately becomes

$$F_2(x) \approx \frac{\kappa''_\sigma(0)}{N}(x - \bar{x}) \tag{18}$$

which is in the same form as the force in (17).

Simulations for 4 information particles randomly located in a single dimensional space are illustrated in Figure 1. The forces are plotted in the 1st column and potentials are plotted in the second column. Figure 1a illustrates the force field emanating from a single particle and Figure 1c depicts the total force field at a given point as a superposition of the force fields due to each individual particle. Figure 1b and 1d, similarly illustrate the individual potential field due to a single particle as a function of the relative position and the total potential of a particle at a given position. Note that, as the kernel size increases, the information forces are better and better approximated by a line in the dynamic range of the samples, thus becoming similar to MSE (Figure 1c). Likewise, the total information potential (Figure 1d) becomes better approximated by a quadratic polynomial in this range, becoming more

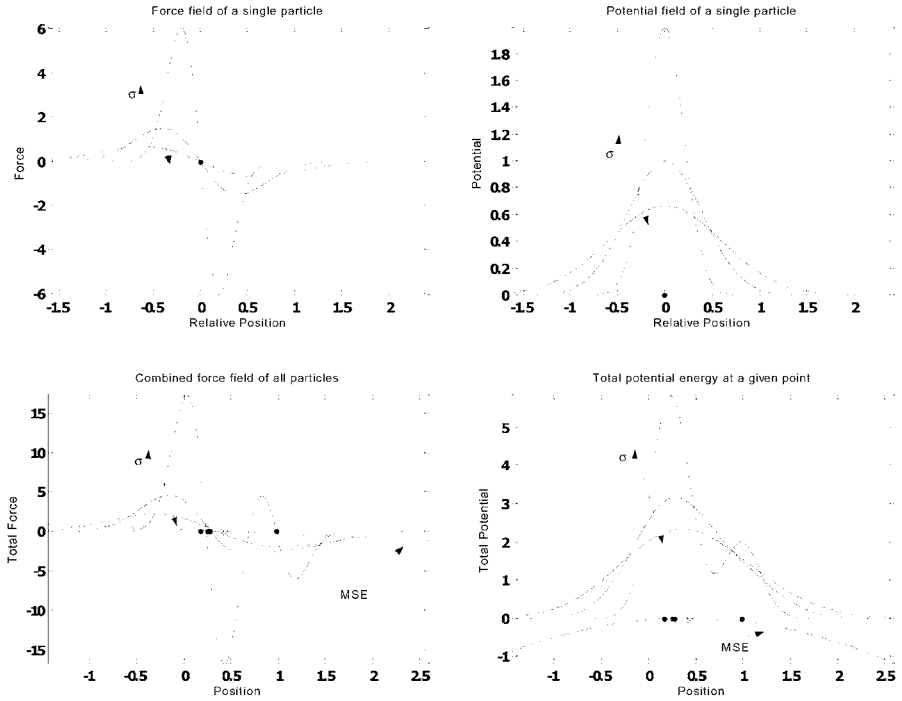

Figure 1. Forces and ptentials as a function of position for different values of kernel size (a) force due to a single particle (b) potential due to a single particle (c) overall force at a given position (d) total potential at a given position.

and more similar to the potential field of MSE, which emanates from the sample mean.

The basic difference between MSE and entropic learning is that, in the second order statistics, the source of the potential (and hence the force) is the sample mean, while in the entropy model each sample is a source itself. For MSE, this corresponds to using an average model of the physical interactions with the sample mean acting as a center of gravity originating an average field, whereas for entropy each sample generates its own potential and force field leading to an accurate and a complete model of the physical interactions between all the particles. For these reasons, MSE algorithms can estimate the forces one sample at a time, while for entropic learning pairwise interactions are required.

Notice that if one aims at minimizing the entropy of the samples, the forces are attractive (with a + sign due to maximization of information potential) and pull the other particles towards the sample, thus minimizing the entropy and increasing the potential energy of the system of particles. On the other hand,

if entropy needs to be maximized, these forces become repulsive (due to the – sign for minimization of potential) and the particles tend to spread to fill the space.

5. Gradient of entropy

There are numerous possible applications where entropy can be used as the optimization criterion to find the parameters w of a linear or nonlinear parametric mapper $y = g(x, w)$. In such cases, the gradient of the cost function will require the evaluation of the gradient of entropy. This can be written in terms of the information forces as

$$\frac{\partial H_\alpha(X)}{\partial w} = \frac{1}{(1-\alpha)V_\alpha(X)} \sum_j F_\alpha(x_j) S_w(x_j) \tag{19}$$

where $S_w(x_j) = dx_j/dw$ is the sensitivity of the sample with respect to the weight vector of the adaptive system under training. In addition, the information potential can normally replace the entropy in the cost function, (as in supervised training with minimum error entropy criterion or in maximization of joint entropy as in the Bell & Sejnowski algorithm), therefore the *log* can be dropped leaving the information potential as the cost function. The gradient in this case, simplifies down to only the summation term given in (19), eliminating the need to evaluate the information potential at every iteration. Of course, one needs to be careful about the switching from a maximization problem to a minimization problem or vice versa for values of $\alpha > 1$ or $\alpha < 1$. In supervised training, the error samples constitute the sample set used to evaluate the entropy and its gradient, whereas in unsupervised learning schemes, the system output is directly used (Principe et al. 2000). These samples can be single dimensional or multi-dimensional depending on the system being trained.

6. Stochastic gradient

The drawback of the gradient algorithm in (19) is that it requires a double summation and kernel evaluation operations which are $O(N^2)$, where N is the number of samples. This complexity demands high computational bandwidth and makes the information potential algorithm unattractive for on-line learning situations involving time series. Motivated by this, and inspired by Widrow's stochastic gradient that led to the well-known LMS algorithm

(Widrow and Stearns 1985; Haykin 1984), we derive a stochastic information gradient (SIG) algorithm for the maximization/minimization of entropy. Once again, we can derive two stochastic gradient algorithms, one for the information potential, and one for the entropy itself. Detailed derivations of these expressions are given in (Erdogmus and Principe 2001; Erdogmus et al. 2002; Hild II et al. 2001a). Here, we simply present the final results.

The SIG for the information potential basically consists of the kernel terms evaluated at consecutive samples in time. Recall that in the batch gradient, all possible pair-wise combinations of samples in the training set are considered in evaluating the gradient. This simplification to concentrate only on the forces between pairs of consecutive samples reduces the required computation effort significantly, and the algorithm becomes O(N). A gradient update may be performed after accumulating and averaging L samples as shown in (20)

$$\overline{\left(\frac{\partial V_\alpha}{\partial w}\right)_k} = \frac{1}{L}\sum_{j=k-L+1}^{k} \frac{(1-\alpha)}{L^{\alpha-1}} C_j(\alpha,\sigma)\kappa'_\sigma(x_j - x_{j-1})(S_w(x_j) - S_w(x_{j-1})) \quad (20)$$

where we define the coefficient

$$C_j(\alpha,\sigma) = \left(\sum_i \kappa_\sigma(e_j - e_{j-i})\right)^{\alpha-2} \quad (21)$$

It can easily be shown that this stochastic gradient is an unbiased estimator of the actual gradient that uses the whole data set (Erdogmus and Principe 2001). This conclusion can also be seen from the trivial fact that the stochastic gradient in (20) is obtained by taking the derivative of the information potential with respect to the weights after dropping the expectation operator in its definition.

Alternatively, a gradient update can be applied to the weights after every sample. This corresponds to choosing $L = 2$ in (20). In that case, the SIG for information potential becomes

$$\left(\frac{\partial V_\alpha}{\partial w}\right)_k = \frac{(1-\alpha)}{2^{\alpha-1}} C_k(\alpha,\sigma)\kappa'_\sigma(x_k, -x_{k-1})(S_w(x_k) - S_w(x_{k-1})) \quad (22)$$

where the coefficient $C_k(\alpha,\sigma)$ now becomes a single kernel evaluation (Erdogmus and Principe 2001), and it is intrinsically sample by sample, just like the LMS algorithm.

The SIG algorithm for entropy, although similar in structure, reveals interesting insights about entropic and Hebbian learning, which challenges our current understanding of Hebbian synapses. In fact we show below that using Donald's Hebb's definition of synaptic learning applied to innovations instead of current values, an Hebbian synapse can estimate entropy instead of correlation. The derivation is similar to that of the information potential SIG, however, the information potential in the denominator of (19) is also approximated by only differences of consecutive samples. This leads to the following SIG for entropy

$$\begin{aligned}\left(\frac{\partial H_\alpha(X)}{\partial w}\right)_k &= \frac{1}{1-\alpha}\frac{(\partial \hat{V}_\alpha(y)/\partial w)_k}{(\hat{V}_\alpha(y))_k} \\ &= \frac{\kappa'_\sigma(x_k - x_{k-1}) \cdot (S_w(x_k) - S_w(x_{k-1}))}{\kappa_\sigma(x_k - x_{k-1})}\end{aligned} \tag{23}$$

This SIG can be compactly written as

$$\left(\frac{\partial H_\alpha(X)}{\partial w}\right)_k = f(x_k - x_{k-1}) \cdot (S_w(x_k) - S_w(x_{k-1})) \tag{24}$$

where $S_w(x_k)$ is the sensitivity of the output to the weight, and the nonlinear function $f(x) = -\kappa'_\sigma(x)/\kappa_\sigma(x)$ regulates the magnitude of the gradient according to the chosen kernel function, i.e. the interaction law between particles. It is noteworthy that $sign(f(x)) = sign(x)$ when kernels satisfying the conditions in Theorem 1 are used. This property allows us to interpret the gradient in (23) to adapt the weight w of a mapper, and therefore the corresponding learning rule, as Hebbian (or anti-Hebbian). However, in entropic learning, Hebb's rule is not applied to the current value of the input and output, but rather to the *instantaneous differences (innovations)* of the related values. Specifically for the choice of Gaussian kernels and an ADALINE structure (Widrow and Stearns 1985), the learning rule in (23) becomes

$$\left(\frac{\partial H_\alpha(X)}{\partial w}\right)_k = \frac{1}{\sigma^2}(x_k - x_{k-1}) \cdot (u_k - u_{k-1}) \tag{25}$$

where u_k is the input vector at time k, and x_k is the corresponding output. This coincides with the generally accepted definition of Hebbian update rule, only applied to the instantaneous differences (the innovations) instead of the instantaneous values. It is well known that when an ADALINE is trained with Hebbian rule, it can identify the direction of maximum variance in the input space with its weights (Oja 1983). It is remarkable that when the same network is trained using the entropy SIG given in (23), it can identify

the direction of maximum entropy! Thus, one can implement information theoretic learning using Hebbian updates (Erdogmus et al. 2002).

The gradient expression in (24) also applies, for example, to the supervised training of FIR filters by replacing the difference of output values with the difference of error values at consecutive time instants. This algorithm compared to the well known LMS, which only utilizes the product of the current error value with the current input vector, takes into account the interactions between the input vectors and error values at different time indices (e.g. consecutive time indices), which encodes correlations at nonzero delays as well as at zero delay. This property of the SIG algorithm can be exploited as a regularization term in parallel with the LMS algorithm. It is known that the step size of the LMS algorithm determines a trade-off between the misadjustment and the tracking ability of the adaptive system (Haykin 1984). While larger step sizes result in increased tracking ability, they also result in a higher misadjustment due to weight updating solely dependent on the current value of the error. At this point, we suggest using the SIG given in (24) along with the regular LMS update, perhaps with a smaller weighting factor, as a regularizer, which will prevent high magnitude fluctuations in the weight space, thus decrease the size of the ripples in the tracking error. For this, the following combined gradient expression must be utilized in adapting the weights for an ADALINE or an FIR structure.

$$\nabla J_k = -2e_k u_k - \lambda(e_k - e_{k-1})(u_k - u_{k-1})/\sigma^2 \tag{26}$$

where e_k is the instantaneous error and $0 < \lambda < 1$ is the regularization coefficient that controls the influence of SIG in the overall adaptation of the weights. This update rule, while tracking the changes in a nonstationary environment by employing fast tracking LMS algorithm, will, in the mean, regulate the fluctuations in the tracking error by controlling the difference between consecutive error values. As a final comment, note that utilizing the hybrid update rule proposed in (25) corresponds to employing a mixed adaptation criterion that consists of a combination of MSE and error entropy, both estimated stochastically from the most recent samples.

7. Applications of the entropy estimator to learning

In this section, we will demonstrate a number of applications where the before-mentioned entropy estimator and the associated batch and stochastic gradient algorithms may be employed. One of these applications is the substitution of MSE with error entropy for prediction and system identification. This is an application of entropy that has not been extensively studied in

adaptive systems literature. Another example is a well-known problem where entropy and other information theoretic measures are extensively utilized, namely independent component analysis (ICA).

7.1 *Error entropy minimization for prediction and system identification*

Although most engineering applications can be reasonably treated using linearity and Gaussianity assumptions, sometimes better models and improved strategies are necessary. In such cases, it is advantageous to manipulate the information content of signals rather than acting on merely the second order or any fixed higher order statistics. Entropy, under these circumstances emerges as the natural choice. Suppose that the input-output mapping of the adaptive system is defined by the parametric function $y = g(x; w)$, where w represents the adjustable weights of the system. In order to approximate the target mapping using a finite number of input-output pairs and as good as possible in the sense that the residual error carries minimal information content, one needs to solve for the following optimization problem for the error entropy where the error is defined as the difference between the desired and actual outputs, i.e. $e = d - y$.

$$w_* = \arg\min_w H_\alpha(e) \tag{27}$$

It can be shown that minimizing Renyi's error entropy of order $\alpha > 1$ is equivalent to minimizing a Csiszar divergence between the joint densities $p_{dx}(.,.)$ and $p_{yx}(.,.)$ as shown in (27) (Erdogmus and Principe 2001). Similarly, equivalence can be established between minimizing Shannon's error entropy and the Kullback-Leibler divergence between these joint densities. The interesting point is that the Csiszar divergence and the Kullback-Leibler divergence are closely associated with the *α-divergence* defined by Amari in the context of Riemannian structure of probability density function (pdf) spaces (Amari 1985).

$$\min_w H_\alpha(e) \equiv \min_w \int\int p_{yx}(y,x)\left(\frac{p_{dx}(y,x)}{p_{yx}(y,x)}\right)^{1-\alpha} dxdy \tag{28}$$

Minimizing the error entropy minimizes the Riemannian distance (i.e. on the geodesic of the nonlinear manifold) between the aforementioned joint (therefore conditional) densities. Thus, the curved structure of this space is fully considered by the information theoretic criterion (Erdogmus and Principe 2001).

To illustrate this, consider the single-step prediction of the Mackey-Glass chaotic time series using a TDNN with 6 inputs, 6 hidden neurons, with

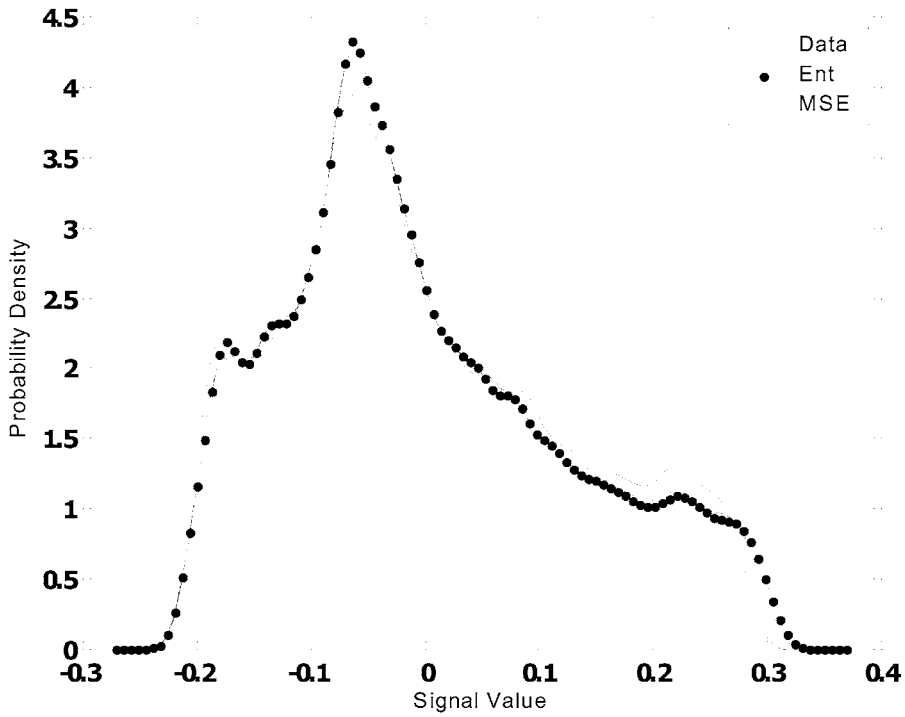

Figure 2. Pdf of desired output compared with pdfs of MLP outputs obtained by entropy and MSE training. Adapted from (Erdogmus and Principe 2001).

tanh activation functions, and a single linear output neuron whose bias term is adjusted to yield zero error mean. All other weights are adapted to yield minimum error entropy over a training set of 200 samples. Training is performed over 1000 random initial sets of weights and the best resulting weights that yield the minimum entropy is selected. The performance of the optimal weights is then tested on an independently generated 10000-sample set. For comparison, the pdf fit obtained by the MSE criterion is also presented (optimal weights selected after a similar training process) (Erdogmus and Principe 2001). As observed in Figure 2, minimum error entropy provides a better fit than MSE to the desired pdf, reminiscent of an L1 norm fit.

As a second example, consider an ADALINE training problem, where one aims to approximate the input-output mapping from a 1000-input vector space to a single desired output time series. The input vector consists of 10 delayed values of 100 measuring probes surgically implanted in the motor cortex of a monkey, whereas the desired output is the time series produced by the current position of the monkey's arm (one dimensional in this example).

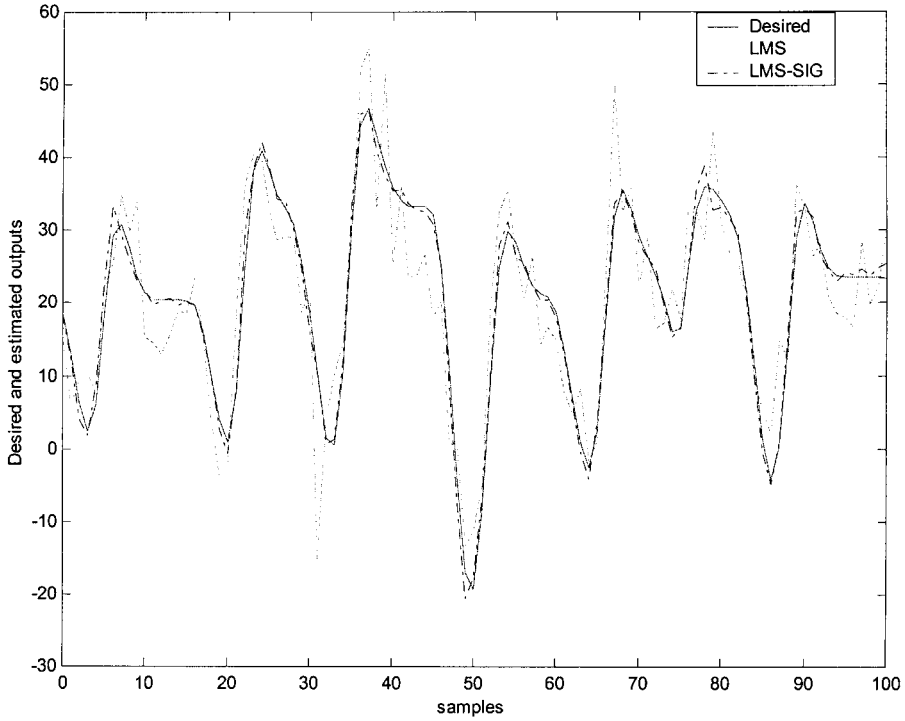

Figure 3. Desired filter output (solid) and the actual filter outputs using LMS (dot) and LMS-SIG (dash).

The data was collected by the group of Dr. Nicolelis at Duke University (Wessberg et al. 2000). In this high-dimensional problem, the compromise between tracking ability and misadjustment has significant implications on the performance of the filter. The same filter is trained using the normalized LMS algorithm with a step size of 0.6, and also trained using the hybrid LMS (normalized)-SIG algorithm, given in (25), using the same step size and the regulation factor $\lambda = 0.5$. In Figure 3 the last 100 samples of a training set of 10000 are presented. Clearly, the introduction of SIG into the update rule leads to the reduction of over and undershoots, which results in a much better estimate.

7.2 *ICA via maximum entropy*

The maximum entropy criterion for independent components analysis (ICA) was first proposed by Bell & Sejnowski, and their algorithm, which is based on Shannon's entropy definition, became one of the benchmarks in ICA and blind source separation (BSS) literature (Hyvarinen et al. 2001). The topology

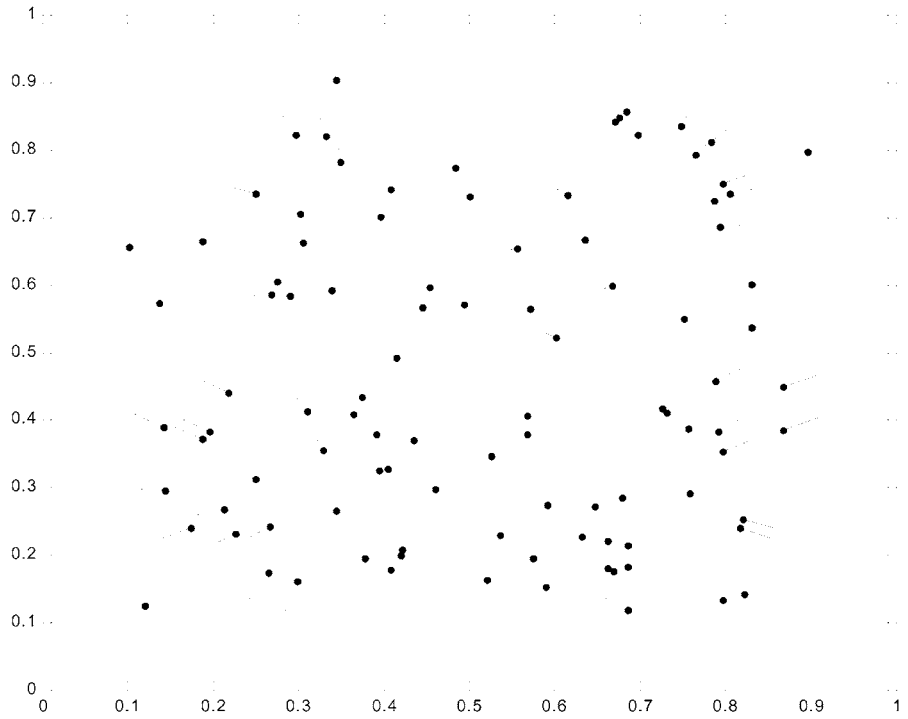

Figure 4. Samples (information particles) with the information forces acting on them (coded as a voctor) in the 2-D output of the mapper (joint space).

used to separate the n independent components (in the square case) from the n measurements consists of a single layer MLP with an $n \times n$ demixing matrix followed by nonlinearities at each output channel, which are tuned to the cumulative distribution functions (cdf) of the sources (independent components). Maximizing the joint entropy after the nonlinearities will result in a uniform distribution and the signals before the nonlinearities will have the desired source densities; furthermore, they will be independent. It is a trivial step to substitute Shannon's joint entropy with Renyi's joint entropy. In the context of ICA/BSS via joint entropy maximization, the relevant optimization problem that needs to be solved is as follows (for $\alpha > 1$).

$$w_* = \arg\max_W H_\alpha(y) = \arg\min_W V_\alpha(y) \tag{29}$$

where $V_\alpha(y)$ is the information potential of the joint output density.

As an example, consider a 2-source separation problem. The scatter plot of the samples after the nonlinearity with the information forces acting on them during training looks like Figure 4. The information forces acting on the information particles will guarantee that the joint output density tends to

the uniform distibution. This pair-wise interaction model for ICA and also for subspace projections was first proposed by Principe, Xu, and Fisher (Principe et al. 2000; Fisher 1997; Xu 1999).

7.3 *ICA via mutual information*

Mutual information is a quantity that measures the dependence between random variables and is minimized to achieve the value of zero when and only when all the random variables are independent. There are ICA/BSS algorithms that employ Shannon's definition of mutual information directly (Comon 1994; Yang and Amari 1997). Shannon's mutual information possesses a desirable additivity property with the joint and marginal entropies of the random variables under consideration.

$$I_S(y) = \sum_{o=1}^{n} H_S(y^o) - H_S(y) \tag{30}$$

Comon exploits this property along with the invariance of the joint entropy to rotations and proposes a two-stage algorithm, which consists of sphering (spatial whitening) followed by an axes-rotation stage (Comon 1994). Since joint entropy is constant under rotations, the cost function reduces to the sum of marginal entropies. The advantages associated with this approach include reduced number of adaptation parameters, elimination of nonlinearity tuning, and avoidance of high dimensional entropy estimation.

Although Renyi's mutual information and joint and marginal entropies do not satisfy the equality presented in (29), the sum of Renyi's marginal entropies minus the joint entropy can still be used as a measure of dependence (Hild II 2001). Thus, it is possible to perform ICA/BSS using the sphering-rotation scheme along with the sum of Renyi's marginal entropies for the criterion. The advantage gained by doing so is investigated in (Erdogmus et al. 2001), and it is shown that in fact Shannon's entropy does not yield the best separation for source distributions of different kurtosis values. In fact, for super-Gaussian sources the suggested entropy order for optimal performance is larger than two, while for sub-Gaussian sources the order is less than two, although any order can be used. For mixed kurtosis cases, an entropy order of two is suggested (Erdogmus et al. 2002). These recommendations have been supported by experimentation over the generalized Gaussian distributions, which have the following density function.

$$G_\upsilon(x) = C \cdot \exp(-|x|^\upsilon / (\upsilon E[|x|^\upsilon]) \tag{31}$$

Here, ν controls the kurtosis of the density and splits the family into super-Gaussian and sub-Gaussian distribution sets for the values ($\upsilon < 2$) and ($\upsilon > 2$), respectively.

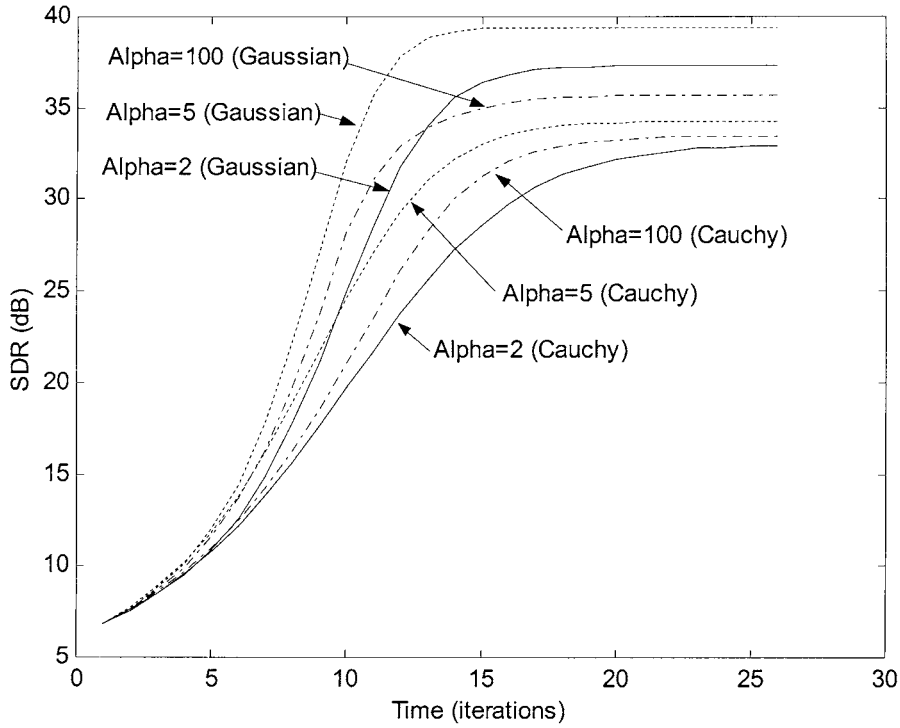

Figure 5. Separation performance vs iteration for various entropy orders and kernel function choices in a BBS problem. Adapted form (Erdogmus et al. 2001).

For example, consider a two-speech-source separation example using various values of entropy order and kernel function. The evolution of the separation performance is depicted in Figure 5 as a function of iterations. As the sources are both super-Gaussian, $\alpha = 5$ achieves the best and fastest learning of the separation matrix, whereas the other entropy orders also perform satisfactorily (over 20dB). This example also illustrates that the Gaussian kernel was a better match for this data set than the Cauchy kernel (Erdogmus et al. 2002).

Now consider a 10-source separation problem, where the sources consist of a musical piece, four female, and five male speakers. This example will demonstrate how the pair-wise interaction model exploits the information content in the data set efficiently compared to other methods. Clearly seen from Figure 6, the pair-wise interaction model requires much less data points to achieve the same level of separation performance of the other mehtods (Hild II et al. 2001b).

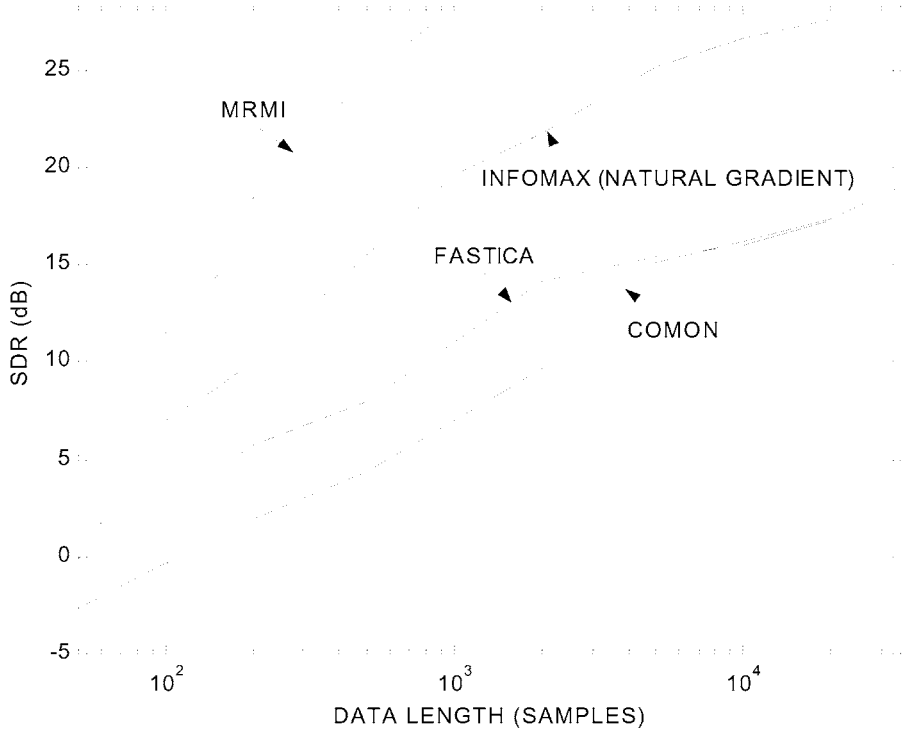

Figure 6. Comparison of data efficiency of minimization of Renyi's mutual information (MRMI) with Infomax, FastICA and Comon's MMI algorithms. Final separation performance vs number of training samples. Adapted from (Hild II et al. 2001).

The same argument can be stated for the stochastic gradient version of this ICA algorithm (Erdogmus et al. 2002). In fact, when compared with the Bell & Sejnowski (B&S) algorithm in a two-source separation example, there is a drastic performance gap between the adaptation performances of these two on-line ICA methods. As you can see in Figure 7, the SIG algorithm converges to a satisfactory separation solution in less than half a second after the speakers become active (the first 0.5 sec is silence, therefore there is no adaptation). On the other side, it takes the B&S algorithm 8 seconds to reach an acceptable separation level (Hild II et al. 2001).

8. Conclusions

Second order statistics have been satisfactory for many of the practical problems encountered in adaptive systems theory and therefore have long been employed as adaptation and optimality criteria. Recent interest in challen-

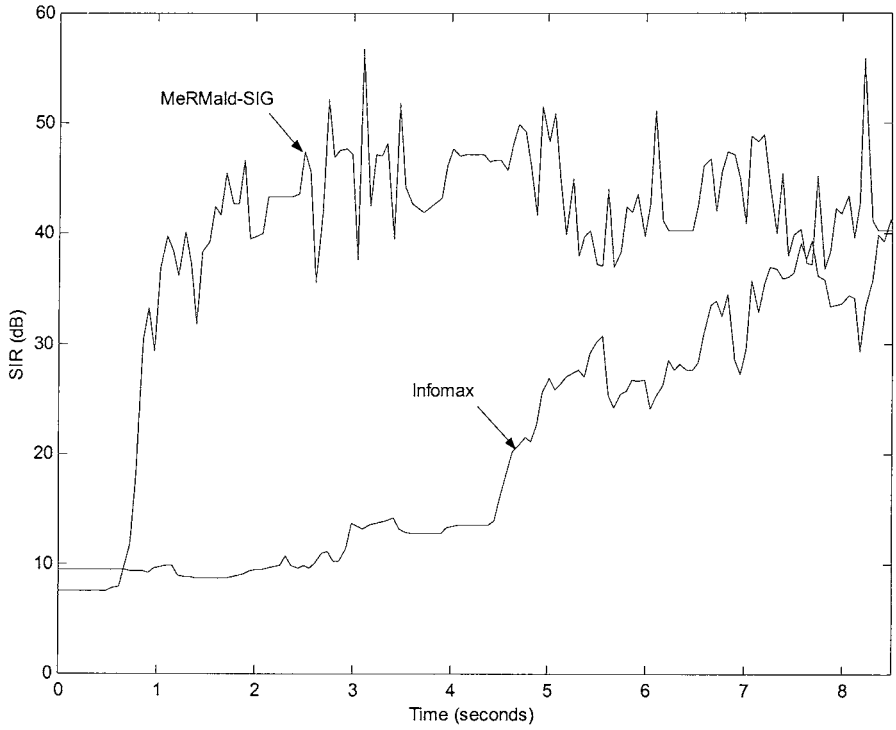

Figure 7. Convergence of MRMI-SIG compared with the Bell & Sejnowski algorithm in an online separation environment. Separation performance vs number of data samples (8 KHz).

ging signal processing and adaptive systems problems have shown that it is not always sufficient to consider second order statistics and in fact, higher orders become a necessity (e.g. ICA). Entropy intrinsically represents higher order statistics of a random variable; furthermore, it is appealing because it represents the uncertainty and the average information of the underlying random event. In this paper, we have reviewed some of the mathematical properties of a nonparametric Renyi's entropy estimator, which is specifically designed for adaptation purposes. As Shannon's entropy becomes a special case of Renyi's, the analysis also applies to Shannon's entropy. The utilization of Parzen windowing in the entropy estimator leads to an interesting analogy between the emerging adaptation algorithms and particle interactions in potential fields.

Since the nonparametric estimator uses a kernel pdf estimator, the important problem of kernel size selection is addressed. The equivalence between error entropy minimization in supervised learning and minimization of the distance between pdfs of desired and output signals is theoretically

shown and demonstrated in a time-series prediction example. A stochastic gradient algorithm (SIG) for entropy manipulation is presented and the performance of a hybrid training algorithm that combines the tracking capabilities of LMS with the regularization properties of SIG is demonstrated on a system identification problem. Finally, the efficient data usage of the entropy estimator and the fast convergence of the associated stochastic gradient algorithm are illustrated in a blind source separation example.

Acknowledgements

This work is partially supported by NSF grant ECS-9900394 and ONR N00014-01-1-0405.

References

Amari S (1985) Differential–Geometrical Methods in Statistics. Springer-Verlag, Berlin

Amari S (1998) Natural gradient works efficiently in learning. Neural Computation 10: 251–276

Bell A and Sejnowski T (1995) An information-maximization approach to blind separation and blind deconvolution. Neural Computation 7: 1129–1159

Bishop C (1995) Neural Networks for Pattern Recognition. Clarendon Press, Oxford

Comon P (1994) Independent component analysis, a new concept? Signal Proc. 36: 287–314

Cover T and Thomas J (1991) Elements of Information Theory. John Wiley, New York

Deco G and Obradovic D (1996) An Information-Theoretic Approach to Neural Computing. Springer, NY

Erdogmus D and Principe JC (2002) Generalized information potential criterion for adaptive system training. To appear in IEEE Transactions on Neural Networks

Erdogmus D and Principe JC (2001) An on-line adaptation algorithm for adaptive system training with minimum error entropy: Stochastic information gradient: 7–12, ICA

Erdogmus D, Hild II KE and Principe JC (2002) Blind Source Separation Using Renyi's α-Marginal Entropies. To appear in Neurocomputation

Erdogmus D, Hild II KE and Principe JC (2002) Do Hebbian synapses estimate entropy? Submitted, NNSP

Fisher JW (1997) Nonlinear extensions to the minimum average correlation energy filter. Ph.D. Dissertation, University of Florida

Fukunaga K (1972) An Introduction to Statistical Pattern Recognition. Academic Press, New York, NY

Haykin S (1984) Introduction to Adaptive Filters. MacMillan, NY

Hild II KE, Erdogmus D and Principe JC blind source separation using renyi's mutual information. IEEE Signal Processing Letters 8: 174–176

Hild II KE, Erdogmus D and Principe JC (2001) On-line minimum mutual information method for time-varying blind source separation: 126–131, ICA

Hyvarinen A (1999) Fast and robust fixed-point algorithms for independent component analysis. IEEE Transactions on Neural Networks 10: 626–634.

Hyvarinen A, Karhunen J and Oja E (2001) Independent Component Analysis. Wiley, New York

Lee TW, Bell AJ and Orglmeister R (1997) Blind source separation of real world signals. International Conference of Neural Networks 4: 2129–2134

Oja E (1983) Subspace Methods of Pattern Recognition. Wiley, New York

Parzen E (1967) On estimation of a probability density function and mode. In: Time Series Analysis Papers. Holden-Day, Inc., CA

Principe JC, Xu D and Fisher JW (2000) Information theoretic learning. In: Haykin S (ed) Unsupervised Adaptive Filtering, pp. 265–319. John Wiley & Son, New York

Renyi A (1970) Probability Theory. American Elsevier Publishing Company Inc., New York

Scharf LL (1990) Statistical Signal Processing: Detection, Estimation, and Time Series Analysis. Addison Wesley, New York

Shannon CE (1948) A mathematical theory of communications. Bell Sys. T. J. 27: 379–423, 623–656

Wessberg J, Stambaugh CR, Kralik JD, Beck PD, Laubach M, Chapin JK, Kim J, Biggs SJ, Srinivasan MA, Nicolelis MAL (2000) Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. Nature 408: 361–365

Widrow B and Stearns SD (1985) *Adaptive Signal Processing*. Prentice Hall, NJ

Wiener N (1949) Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications. MIT Press, Cambridge, MA

Xu D (1999) Energy, entropy, and information potential for neural computation. Ph.D. Dissertation, University of Florida

Yang H and Amari S (1997) Adaptive online learning algorithms for blind separation: maximum entropy and minimum mutual information. Neural Computation 9: 1457–1482