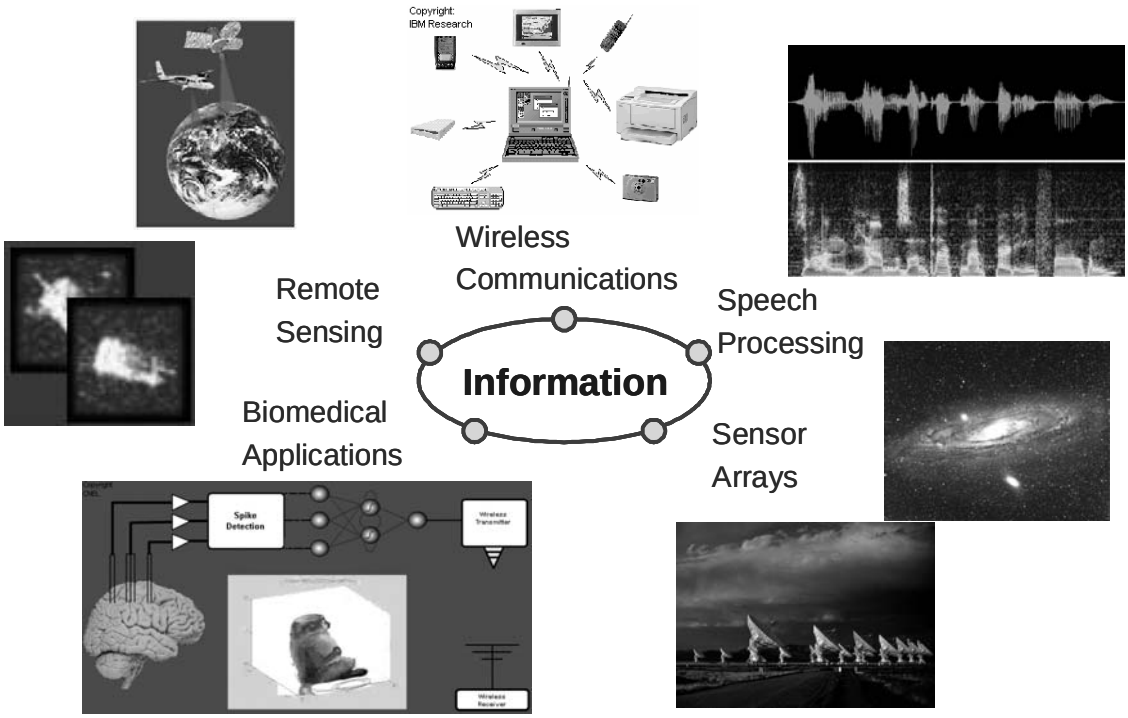
Information-Theoretic Learning

Tutorial

Jose C. Principe, Ph.D.

I- Introduction

One of the fundamental problems of our technology driven society is the huge amounts of data that are being generated by every segment of the society from factories, services, medicine and individuals alike (Fig 1). Unfortunately we humans seek information not data, and therefore a growing bottleneck is exactly how to extract information from data.



Information has many meanings in our colloquial language, however here, information means a precise mathematical quantity fully characterized in Information Theory (IT). We utilize this approach because it is very appropriate to deal with manipulation of information [35]. Shannon in a 1948 classical paper laid down the foundations of IT [36]. IT has had a tremendous impact in the design of *efficient and reliable* communication systems [8],[12] because it is able to answer two

key questions: what is the best possible (minimal) code for our data, and what is the maximal amount of information which can be transferred through a particular channel. In spite of its practical origins, IT is a deep mathematical theory concerned with the *very essence of the communication process [12]*. IT has also impacted statistics [22] and statistical mechanics by providing a clearer understanding of the nature of entropy as illustrated by Jaynes [19]. These advances are predicated however on the specification of the data distributions, which is not realistic for the design of learning machines. In the design of adaptive self-organizing systems, the primary objective is to develop algorithms that will learn an input-output relationship of interest on the basis of input patterns alone. We submit that a thrust to innovate IT is to develop methods to directly estimate entropy from a set of data. With entropic measures, we will be able to utilize the full probability density function for optimization and to lift the present restrictions of linearity and Gaussianity for the application of IT to real-world problems.

This document addresses the important issue of *extracting information directly from data*, which is at the core of the issue of learning from examples in both biological and artificial systems. The learning from examples scenario starts with a data set which globally conveys information about a real-world event, and the goal is to capture this information in the parameters of a learning machine. The information exists in a "distributed" mode in the data set, and appears "condensed" in the parameters of the learning machine after successful training. Learning in artificial neural networks and adaptive filters has used almost exclusively correlation (the L2 norm or mean-square error) as a criterion to compare the information carried by the signals and the response of the learning machine, but there is mounting evidence that correlation (a second order moment) is a poor measure to ascertain the equivalence of information between the desired response and the output of the mapper. The fundamental issue is to find the appropriate methodology to study this "change in state" and elucidate the issues of designing systems that are capable of producing the transfer of information as efficiently as possible.

Here we will develop information-theoretic criteria which can train directly from the samples linear or nonlinear mappers either for entropy or mutual information maximization or minimization. We will start by a brief review of Renyi's entropy and a description of information-theoretic learning (ITL). The Parzen window method of PDF estimation is fundamental in all our efforts to cre-

ate algorithms to manipulate entropy. The following section covers a more principled approach of designing practical information-theoretic criteria using Renyi's definition of entropy of order two (quadratic entropy). We show that quadratic entropy can be easily integrated with the Parzen window estimator. The pairwise data interactions for the computation of entropy are interpreted as an information potential field and are a powerful analogy between information theoretical learning and physics. We finally propose the integration of the Cauchy-Schwartz distance and an Euclidean difference with the Parzen window to provide estimators for mutual information. The mutual information criterion is very general and can be used either in a supervised or unsupervised learning framework.

Information Optimization Principles

The most common entropy optimization principles are Jayne's MaxEnt and Kullback's MinXEnt [20]. MaxEnt finds the distribution that maximizes Shannon's entropy subject to some explicit constraints. Hence, MaxEnt guarantees that we make no assumptions about possible missing information. MinXEnt finds a distribution, from all possible distributions satisfying the constraints, that minimizes the distance in probability space to the given distribution. The most widely used measure for MinXEnt is the Kullback-Leibler (K-L) cross-entropy. Effectively, K-L is a measure of directed divergence between the given and the unknown distribution (a directed divergence is a relaxed concept of distance since it does not need to be symmetric nor obey the triangular inequality). It turns out that MinXEnt (using the K-L divergence) with respect to the uniform target distribution is equivalent to the MaxEnt principle under the same constraints. However, they are intrinsically different since one maximizes uncertainty while the other minimizes directed divergence between PDFs. Moreover, MinXEnt is invariant to coordinate transformations which is an advantage for learning, while MaxEnt does not hold this characteristic in the continuous case. These principles have been applied using mostly Gaussian assumptions for the data distribution, which is not very realistic when adapting nonlinear systems.

Information-Theoretic Learning (ITL)

Consider the parametric mapping $g:\Re^K \to \Re^M$, of a random vector $\boldsymbol{X} \in \Re^K$ (normally M<K),

which is described by the following equation

$$\boldsymbol{Y} = g(\boldsymbol{X}, \boldsymbol{W}) \tag{1}$$

where $\mathbf{Y}$ is also a random vector $\boldsymbol{Y} \in \Re^M$, and $\boldsymbol{W}$ is a set of parameters. For each observation $\boldsymbol{x}_i$ of the random vector $\mathbf{X}$, the parametric mapper responds with $\boldsymbol{y}_i = g(\boldsymbol{x}_i, \boldsymbol{W})$. Our goal is to choose the parameters $\boldsymbol{W}$ of the mapping $g(.)$ such that a figure of merit based on information theory is optimized at the output space of the mapper (Fig. 2). This is what we call information-theoretic learning (ITL). Notice that we are *only* requiring the availability of observations $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ or random vectors without assuming any *a priori* model for their probability density functions (PDF). Notice also that the mapper can either be linear or non-linear, and that the criterion may or may not exploit an added external input normally called the desired response, i.e. information theoretic learning includes as special cases both the unsupervised and supervised frameworks. We also want the learning criterion to be external and independent of the mapper. Let us briefly review work done in this area.

By analogy to optimization in Euclidean space, we can adapt the parameters $\boldsymbol{W}$ of the mapper by

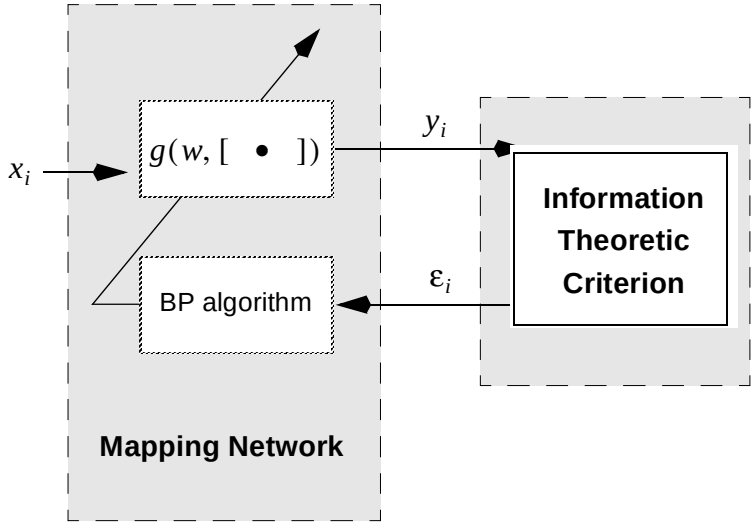


Figure 1: Training a mapper (linear or nonlinear) with ITL

manipulating the output distribution $p(\boldsymbol{Y})$: maximizing output entropy (MaxEnt) or minimizing the cross-entropy among the outputs or among the output and other signals (MinXEnt). The work of Bell & Sejnowski on blind source separation (BSS) [5] is an example of the application of the MaxEnt principle. In the neural network literature, the work of Barlow [3] and Atick [2] also uti-

lized entropy concepts for learning.

Optimization based on the MinXEnt principle is the one that is potentially more useful to solve engineering and in particular learning problems [9]. Comon [7], Deco and Obradovic [9], Cardoso [6] and Amari [37] among others utilized the MinXEnt principle to formulate and solve the blind source separation (BSS) problem. One solution to BSS is obtained by minimizing the mutual information (redundancy) among the outputs of a mapper $\boldsymbol{Y}$, which can be formulated as the K-L divergence between the joint PDF of $\boldsymbol{Y}$ and its factorized marginals as $I(y_1, \ldots y_n) = \sum_{i=1}^{n} H(y_i) - H(\boldsymbol{Y})$. The problem arises in estimating the joint output density $H(\boldsymbol{Y})$. These researchers utilize the well known [30] result of using a *linear network* to directly compute the output entropy from the input entropy as $H(\boldsymbol{Y}) = H(\boldsymbol{X}) + \log|det(\boldsymbol{W})|$ where $\boldsymbol{Y} = \boldsymbol{WX}$. Note that *a full rank k-to-k linear mapping* $\boldsymbol{W}$ is required in this approach which is a severe constraint for learning applications (for instance in sub-space mappings as required in classification). The next step is the estimation of the marginal entropy of each output $H(y_i)$ (a scalar problem). Comon [7] proposed the use of the Edgeworth expansion of the PDF and Amari [37] the Gram-Charlier expansion which are both well known and equivalent methods (in the limit) of estimating PDFs by the moment expansion method. In practice, the expansions must be truncated (a source of error) and higher order moments of the PDF estimated from the data, which becomes computationally expensive and requires large amounts of data for robust results. However, after the marginal PDFs are estimated, then a gradient based algorithm can be formulated to solve the BSS problem [37]. Although this method is very appealing from the point of view of a learning criterion, notice that it is not general because the criterion is not totally independent of the topology of the mapper. Recently, Amari and Cardoso proposed a semi-parametric model for BSS [1].

In the neural network literature there is still another information optimization principle, Linsker's principle of maximum information preservation (InfoMax), which is a special case of the information loss minimization principle of Plumbey [29]. Optimization with mutual information has not been extensively addressed in the optimization literature. Linsker was interested in finding a principle that self-organizes biological systems [24]. These systems are adaptable, so the issue is to find a criterion to adapt the parameters of the mapper $g(\boldsymbol{X},\boldsymbol{W})$. The goal is to determine the

parameters $\boldsymbol{W}$ such that the output variable $\boldsymbol{Y}$ conveys as much information as possible about $\boldsymbol{X}$: that is, a principle for self-organization should maximize the average mutual information between $\boldsymbol{X}$ and $\boldsymbol{Y}$ in the presence of noise. For a linear network and under Gaussianity assumptions, the mutual information is maximized by maximizing the output variance [24]. Recall that maximization of output variance is basically principal component analysis (PCA) for which there are known on-line and local algorithms [10], [27]. Hence, forseeably, a biological network could self-organize with such a principle. We can see that this method leads to interesting solutions but it depends on very restrictive assumptions about the PDFs and linearity of the mapping. *In fact Plumbey states [29] that the big challenge is to extend Linsker's work to arbitrary distributions and nonlinear networks*. This is exactly what we propose to accomplish in our work.

From a theoretical perspective, InfoMax is a different principle from MaxEnt and MinXEnt since it *maximizes* a divergence measure (mutual information). Linsker applied InfoMax between the input and the output of deterministic mappers, so the principle reduces to applying MaxEnt at the output of the mapper [24], [5]. But InfoMax can be applied to any pairs of random variables, such as the outputs of the mapper and any other external random variable. This new application is called here *information filtering*, since it designs a mapper to preserve information maximally about a source while attenuating other information available in the input data. Information filtering will be exploited later in the chapter for supervised learning applications.

One of the difficulties of these information-theoretic criteria is that analytic solutions are known only for very restricted cases, e.g. Gaussianity and linear volume preserving mappings (see also Deco [9]). Otherwise mathematical approximations and computationally complex algorithms result. A useful neurocomputing algorithm should be applied to any topology, utilizes the data directly (on a sample-by-sample basis or in batch) and a simple learning rule distilled from the mathematics, so we submit that none of the above algorithms to train a nonlinear mapper with MinXEnt criterion is "neural". In this respect Bell and Sejnowski's algorithm for maximization of output entropy is paradigmatic. It utilizes a nonlinear mapper (although restricted to be a perceptron), it adapts the weights with a simple batch rule that is not specific to the input data model (as the solution in [37]) and globally leads to a solution which maximizes an entropic measure. Recently, Linsker showed that there is a local rule for MaxEnt, which only requires extending the

perceptron with lateral connections [25]. This is the spirit of a neural computation [17] which we have been seeking all along.

In our opinion, the two fundamental issues in the application of information-theoretic criteria to neurocomputing or adaptive filtering are: the choice of the criterion for the quantitative measure of information, and the estimation of the probability density function (PDF) from data samples.

ITL as an Unifying Criterion for Learning

Figure 3 shows a block diagram of a unifying scheme for learning based on divergence and entropy. The only difference is the source of information which is shown as a switch with 3 positions.

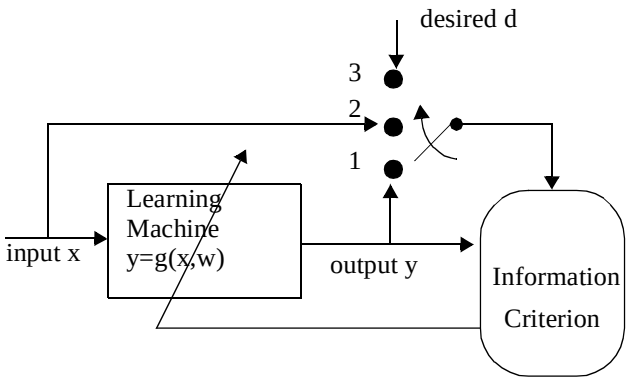


Figure 2: Unifying learning models with the mutual information criterion.

When the switch is in position 1 or 2 learning belongs to the *unsupervised type* (no formal desired response) and corresponds to manipulating the divergence or mutual information at the output of the learning system or between its input and output. A practical example with switch in position 1 is the on-going work on independent component analysis (ICA) where the goal is to minimize the mutual information among the multiple mapper outputs to yield independent components [15]. An example of the block diagram with switch in position 2 is Linsker's Infomax criterion [20] where the goal is to transfer as much information between the input and output of a mapper by maximizing the joint input-output mutual information. However, if the goal is to maximize the mutual information between the output of a mapper and an external desired response, then learning becomes supervised. This is achieved by setting the switch to position 3. Note that in this case

the desired response appears as one of the marginal pdfs in the mutual information criterion. The two outstanding cases belong both to function approximation: first, if the desired response is a set of indicator functions, the task is classification. However, the desired data is always quantified by means of its pdf, not by deriving a sample by sample error. Therefore we can think of this case as supervised learning without numeric targets, just class labels [37]. Second, if the desired response data is a continuous function then we named the application information filtering [26]. This name came from the realization that the learning machine is seeking a projection of the input space that best approximates (in an information sense) the desired response. In engineering this is the model used for Wiener filtering [16] but where the adaptive system is restricted to be a linear filter and the criterion is minimization of the error variance. Table I shows a more complete picture of ITL and can be used as a table of contents for browsing this website and companion documents.

Table 1: ITL as a Unifying Principle for Learning

Switch	One	Two	Three
mE	blind deconvolution		information filtering classification
ME	nonlinear PCA		
mMI	ICA blind source separation	novelty filtering	
MMI	clustering	Linsker infomax matched filters	feature extraction

II- Generalized Definitions of Entropy

Renyi's Entropy

Information theory is a mathematical formalization of our intuitive notion of information contained in messages. If a message is perfectly known *a priori*, its information content is zero. However, the less predictable a message is, the larger is its information content. Shannon, using an axiomatic approach [36] defined entropy of a probability distribution $P= (p_1, p_2,...p_N)$ as

$$H_S(P) = \sum_{k=1}^{N} p_k \log\left(\frac{1}{p_k}\right) \qquad \sum_{k=1}^{N} p_k = 1 \qquad p_k \geq 0 \tag{2}$$

that is, the average amount of information contained in a single observation of a random variable X which takes values $x_1, x_2,....x_N$ with probabilities $p_k=P(x=x_k)$, $k=1, 2,..., N$. Entropy measures the average amount of information conveyed by the event x, or alternatively, the amount of missing information on X when only its *a priori* distribution is given. Information theory has been widely applied to the design of communication systems [8], [12], [35]. But the definition of entropy can be derived even in a more abstract form. In the general theory of means [33], the mean of the real numbers $x_1, \ldots, x_N$ with positive weighting (not necessarily probabilities) $p_1, \ldots, p_N$ has the form:

$$\bar{x} = \varphi^{-1}\left(\sum_{k=1}^{N} p_k \varphi(x_k)\right) \tag{3}$$

where $\varphi(x)$ is a Kolmogorov-Nagumo function, which is an arbitrary continuous and strictly monotonic function defined on the real numbers. In general, an entropy measure H obeys the relation:

$$H = \varphi^{-1}\left(\sum_{k=1}^{N} p_k \varphi(I(p_k))\right) \tag{4}$$

where $I(p_k) = -\log(p_k)$ is Hartley's information measure [18]. In order to be an information

measure, $\varphi(.)$ can not be arbitrary since information is "additive". To meet the additivity condition, $\varphi(.)$ can be either $\varphi(x) = x$ or $\varphi(x) = 2^{(1-\alpha)x}$. If $\varphi(x) = x$ is selected, (3) will become Shannon's entropy. For $\varphi(x) = 2^{(1-\alpha)x}$ Renyi's entropy of order α is obtained [32], which we will denote by $H_{R\alpha}$

$$H_{R\alpha} = \frac{1}{1-\alpha}\log\left(\sum_{k=1}^{N} p_k^{\alpha}\right) \qquad \alpha > 0, \alpha \neq 1 \tag{5}$$

There is a well known relation between Shannon's and Renyi's entropy:

$$H_{R\alpha} \geq H_S \geq H_{R\beta}, \quad \text{if } 1 > \alpha > 0 \text{ and } \beta > 1$$
$$\lim_{\alpha \to 1} H_{R\alpha} = H_S$$

It is important to further relate Renyi's and Shannon's entropies. Let us consider the probability distribution $P= (p_1,p_2,...,p_N)$ as a point in a N-dimensional space. Due to the conditions on the probability measure ($p_k \geq 0$, $\sum_{k=1}^{N} p_k = 1$) P always lies in the first quadrant of an hyperplane in N dimensions intersecting each axis at the coordinate 1 (Fig. 1). The distance of P to the origin is the α root of

$$V_{\alpha} = \sum_{k=1}^{N} p_k^{\alpha} = \|P\|^{\alpha}$$

and the α root of V_{α} is called the α-norm of the probability distribution [16]. Renyi's entropy (4) can be written as a function of V_{α}

$$H_{R\alpha} = \frac{1}{1-\alpha}\log V_{\alpha} \tag{6}$$

When different values of α are selected in Renyi's family, the end result is to select different α-

norms. Shannon entropy can be considered as the limiting case $\alpha \to 1$ of the probability distribution norm. Notice that the limit provides an indeterminacy (zero over zero in (5)) but the result exists and is given by Shannon's entropy. With this view, Renyi's entropy is a monotonic function of the α-norm of the PDF and is essentially a monotonic function of the distance of the probability distribution to the origin. We have considerable freedom in choosing the α-norm [43]. When $\alpha = 2$, $H_{R2} = -\log \sum_{k=1}^{N} p_k^2$ is called quadratic entropy due to the quadratic form on the probability, and it corresponds to the 2-norm of the probability distribution.

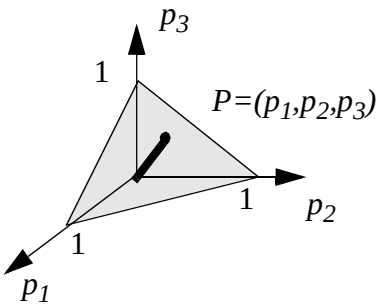


Figure 3: Geometric interpretation of entropy for N=3. The distance of P to the origin is related to the α-norm.

For the continuous random variable Y with PDF $f_Y(y)$, we can obtain the differential version of Renyi's entropy following a similar route to the Shannon differential entropy [32]:

$$\begin{cases} H_{R\alpha}(Y) = \dfrac{1}{1-\alpha}\log\left(\displaystyle\int_{-\infty}^{+\infty} f_Y(z)^{\alpha} dz\right) \\ H_{R2}(Y) = -\log\left(\displaystyle\int_{-\infty}^{+\infty} f_Y(z)^{2} dz\right) \end{cases} \tag{7}$$

Note that Renyi's quadratic entropy involves the use of the square of the PDF. An important observation is that this alternate definition of entropy is equivalent to Shannon's entropy for the goal of entropy maximization [21].

Renyi's entropy is just one example of a large class of alternate entropy definitions which have been called *generalized entropy measures* [20]. One may wonder why the interest in measures

more complex than Shannon's entropy or Kullback-Leibler direct divergence (also called cross-entropy). Here we will only provide a brief overview of this important question. The reader is referred to the entropy optimization literature for further study [20], [21], [32]. The reason to use generalized measures of entropy stems from practical aspects when modeling real world phenomena through entropy optimization algorithms. It has been found that when we apply the two basic optimization principles based on Shannon's entropy definition (which are Jayne's maximum entropy principle (MaxEnt) and Kullback's minimum cross-entropy principle (MinXEnt)) either just one solution from a spectrum of solutions is found, or not even "natural" solutions are found. To improve on this situation, researchers have proposed alternative definitions of entropy. An example of a generalized entropy measure in the digital signal processing arena is Burg's entropy estimator [4], which has been successfully applied in spectral analysis [26].

In our study of learning from examples, the interest in generalized entropy measures comes from a practical difficulty. We wish to directly estimate entropy from the data samples, without imposing assumptions about the PDF. Shannon's definition of entropy (the sum of terms which are weighted logarithms of probability) is not amenable to simple estimation algorithms, while Renyi's logarithm of the sum of the power of probability is much easier to estimate, and has been utilized in physics [15]. We will show in section 7.3 how a very effective algorithm can be derived. Renyi's entropy has been utilized successfully in nonlinear dynamics to estimate the correlation dimension of attractors. One important question stemming from the use of generalized entropy measures is the justification for the selected measure. We have not yet addressed this question in our research. At this point we can only state that the experimental results obtained with the use of Renyi's entropy estimator and its extension to mutual information have produced practical solutions to difficult problems in signal processing and pattern recognition. Since learning from examples is an inverse problem, we believe that the choice of an appropriate generalized entropy measure will play an important role in the quality of the final solution.

III- Information Theoretic Learning: Unsupervised learning with Renyi's Quadratic Entropy

ITL algorithms are based on a combination of a nonparametric PDF estimator and a procedure to compute entropy. In this section we will overcome the difficulty in approximating Shannon's entropy by utilizing Renyi's generalized entropy. Before we start the derivation of the algorithm let us state a property of Gaussian functions which will be very useful in the method.

Let $G(\boldsymbol{z}, \Sigma) = \frac{1}{(2\pi)^{M/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{z}^T\Sigma^{-1}\boldsymbol{z}\right)$ be the Gaussian kernel in M-dimensional space, where Σ is the covariance matrix, $\boldsymbol{z} \in R^M$. Let $\boldsymbol{y}_i \in R^M$ and $\boldsymbol{y}_j \in R^M$ be two data samples in the space, Σ_1 and Σ_2 be two covariance matrices for two Gaussian kernels in the space. Then it can be shown that the following relation holds:

$$\int_{-\infty}^{+\infty} G(\boldsymbol{z} - \boldsymbol{y}_i, \Sigma_1) G(\boldsymbol{z} - \boldsymbol{y}_j, \Sigma_2) d\boldsymbol{z} = G((\boldsymbol{y}_i - \boldsymbol{y}_j), (\Sigma_1 + \Sigma_2)) \quad (8)$$

Similarly, the integration of the product of three Gaussian kernels can also be obtained and so on. (20) can also be interpreted as a convolution between two Gaussian kernels centered at $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ and it is easy to see that the result should be a Gaussian function with a covariance equal to the sum of the individual covariances and centered at $\boldsymbol{d}_{ij} = (\boldsymbol{y}_i - \boldsymbol{y}_j)$.

Quadratic Entropy Cost function for Discrete Samples

Let $\boldsymbol{y}_i \in R^M$, $i = 1, \ldots, N$, be a set of samples from a random variable $Y \in R^M$ in M-dimensional space. An interesting question is what will be the entropy associated with this set of data samples, without pre-specifying the form of the PDF. Part of the answer lies in the methodology presented

in section 7.3 of estimating the data PDF by the Parzen window method using a Gaussian kernel:

$$\hat{f}_Y(\boldsymbol{z}, \{\boldsymbol{y}\}) = \frac{1}{N}\sum_{i=1}^{N} G(\boldsymbol{z}-\boldsymbol{y}_i, \sigma^2\boldsymbol{I}) \tag{9}$$

where $G(.,.)$ is the Gaussian kernel as above and $\sigma^2\boldsymbol{I}$ is the covariance matrix. When Shannon's entropy (1) is used along with this PDF estimation, an algorithm to estimate entropy becomes unrealistically complex as Viola [39] also realized. So, we conclude that Shannon's definition of information does not yield a practical measure for ITL. Fortunately, Renyi's quadratic entropy leads to a much simpler form. Using (21) in (6) we obtain an entropy estimator for a set of discrete data points $\{\boldsymbol{y}\}$ as

$$\begin{cases} H(\{\boldsymbol{y}\}) = H_{R2}(Y|\{\boldsymbol{y}\}) = -\log\left(\int_{-\infty}^{+\infty} f_Y(\boldsymbol{z})^2 d\boldsymbol{z}\right) = -\log V(\{\boldsymbol{y}\}) \\ V(\{\boldsymbol{y}\}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\int_{-\infty}^{+\infty} G(\boldsymbol{z}-\boldsymbol{y}_i, \sigma^2 I)G(\boldsymbol{z}-\boldsymbol{y}_j, \sigma^2 I)d\boldsymbol{z} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G(\boldsymbol{y}_i-\boldsymbol{y}_j, 2\sigma^2 I) \end{cases} \tag{10}$$

We will simplify the notation by representing $\boldsymbol{y}=\{\boldsymbol{y}\}$ whenever possible. The combination of Renyi's quadratic entropy with the Parzen window leads to an estimation of entropy by computing interactions among pairs of samples which is a practical cost function for ITL. There is no approximation in this evaluation (apart from the PDF estimation).

Quadratic Entropy and Information Potential

We wrote (22) in this way because there is a very interesting physical interpretation for this estimator of entropy. Let us assume that we place physical particles in the locations prescribed by the data samples $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$. For this reason we will call them information particles (IPCs). Since $G(\boldsymbol{y}_i-\boldsymbol{y}_j, 2\sigma^2\boldsymbol{I})$ is always positive and is inversely proportional to the distance between the IPCs, we can consider that a potential field was created in the space of interactions with a local field strength dictated by the Gaussian kernel (an exponential decay with the distance square) $V_{ij} = G(\boldsymbol{y}_i-\boldsymbol{y}_j, 2\sigma^2\boldsymbol{I}) = G(\boldsymbol{d}_{ij}, 2\sigma^2\boldsymbol{I})$. Physical particles interact with an inverse of distance

rule, but Renyi's quadratic entropy with the Gaussian kernel imposes a different interaction law. Control of the interaction law is possible by choosing different windows in the Parzen estimator.

The sum of interactions on the ith IPC is $V_i = \sum_j V_{ij} = \sum_j G(\boldsymbol{d}_{ij}, 2\sigma^2\boldsymbol{I})$. Now $V(\boldsymbol{y}) = \frac{1}{N^2}\sum_i\sum_j V_{ij}$, which is the sum of all pairs of interactions, can be regarded as an overall potential energy of the data set. *We will call this potential energy an information potential (IP)*. So maximizing entropy becomes equivalent to minimizing the IP. Our estimator for quadratic entropy is the negative logarithm of the IP. It was a pleasant surprise to verify that our quest for ITL algorithms ended up with a procedure that resembles the world of interacting physical particles which originated the concept of entropy.

We can also expect from (6) that this methodology can be applied to Renyi's entropy of higher order (α>2). In fact, Renyi's entropy of order α will compute interactions among α-tuples of samples, providing even more information about the complex structure of the data set. These interactions can be estimated with an extension of (20) when the Parzen window method implemented with the Gaussian kernel is utilized in the estimation. However, the complexity of the algorithm becomes increasingly prohibitive (O(N^{α})).

Information Forces

Just like in mechanics, the derivative of the potential energy is a force, in this case an information driven force that moves the data samples in the space of the interactions. Therefore,

$$\frac{\partial}{\partial \boldsymbol{y}_i} G(\boldsymbol{y}_i - \boldsymbol{y}_j, 2\sigma^2\boldsymbol{I}) = -G(\boldsymbol{y}_i - \boldsymbol{y}_j, 2\sigma^2\boldsymbol{I})(\boldsymbol{y}_i - \boldsymbol{y}_j)/(2\sigma^2) \tag{11}$$

can be regarded as the force $\boldsymbol{F}_{ij}$ that IPC $\boldsymbol{y}_j$ impinges upon $\boldsymbol{y}_i$, and *will be called an information force (IF)*. Figure 6 depicts the information forces created by a IPC. If all the data samples are free to move in a certain region of the space, then the information forces between each pair of IPCs will drive all the samples to a state with minimum IP. If we add all the contributions of the IF from the ensemble of samples on $\boldsymbol{y}_i$ we have the net effect of the information potential on sample $\boldsymbol{y}_i$,

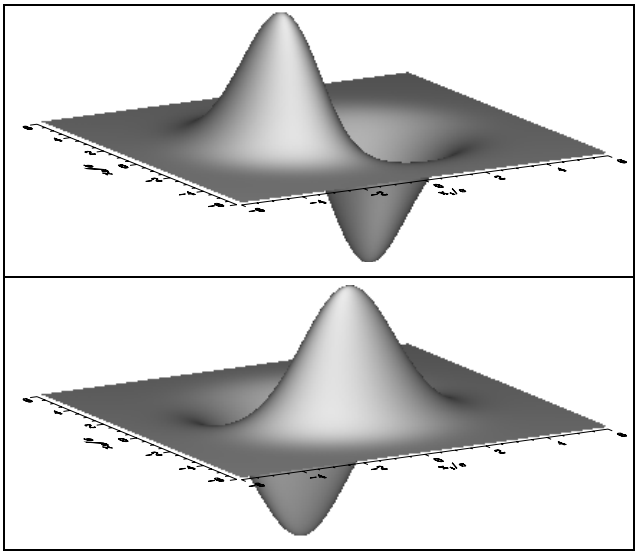


Figure 4: Two dimensional attractor functions. The x_1-component is shown at the top while the x_2-component is shown at the bottom. The function represents the local influence of each data point in the output space.

i.e.

$$\boldsymbol{F}_i = \frac{\partial}{\partial \boldsymbol{y}_i} V(\boldsymbol{y}) = -\frac{1}{N^2\sigma^2}\sum_{j=1}^{N} G(\boldsymbol{y}_i - \boldsymbol{y}_j, 2\sigma^2\boldsymbol{I})(\boldsymbol{y}_i - \boldsymbol{y}_j) = \frac{-1}{N^2\sigma^2}\sum_{j=1}^{N} V_{ij}\boldsymbol{d}_{ij} \tag{12}$$

"Force" Back-Propagation

The concept of IP creates a criterion for ITL, which is external to the mapper of Fig. 2. The only missing step is to integrate the criterion with the adaptation of a parametric mapper as the MLP. Suppose the IPCs $\boldsymbol{y}$ are the outputs of our parametric mapper of (7). If we want to adapt the MLP such that the mapping maximizes the entropy at the output $H(\boldsymbol{y})$, the problem is to find the MLP parameters w_{ij} so that the IP $V(\boldsymbol{y})$ is minimized. In this case, the IPCs are not free but are a function of the MLP parameters. So, the information forces applied to each IPC by the information potential can be back-propagated to the parameters using the chain rule [34], i.e.

$$\frac{\partial}{\partial w} V(\boldsymbol{y}) = \sum_{i=1}^{N}\left[\frac{\partial}{\partial \boldsymbol{y}_i} V(\boldsymbol{y})\right]^T \frac{\partial \boldsymbol{y}_i}{\partial w} = \sum_{i=1}^{N} F_i^T \frac{\partial}{\partial w} g(\boldsymbol{w}, \boldsymbol{x}_i) \tag{13}$$

where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iM})^T$ is the M-dimensional MLP output. Notice that from (25) the sensitiv-

ity of the output with respect to a MLP parameter $\frac{\partial \boldsymbol{y}_i}{\partial w}$ is the "*transmission mechanism" through which information forces are back-propagated to the parameter* (Fig. 13). From the analogy of (25) with the backpropagation formalism (see (11)) we conclude that $F_i = \varepsilon_i$, *that is, information forces take the place of the injected error in the backpropagation algorithm*. So, we obtain a general, nonparametric, and sample-based methodology to adapt arbitrary nonlinear (smooth and differentiable) mappings for entropy maximization (Fig. 13). Notice that we are adapting a MLP without a desired response, hence this is an unsupervised criterion. We have established an ITL criterion that *adapts the MLP with a global property of its output sample distribution*. It is very useful to analyze this expression in detail and compare it with the well known MSE. Note that MSE is computed with a single data sample/desired response combination. However, the entropy is estimated with *pairs of data samples*, that is, more information about the data set is being extracted here than with the MSE criterion (in a N sample data set there are $\binom{N}{2}$ different pairs).

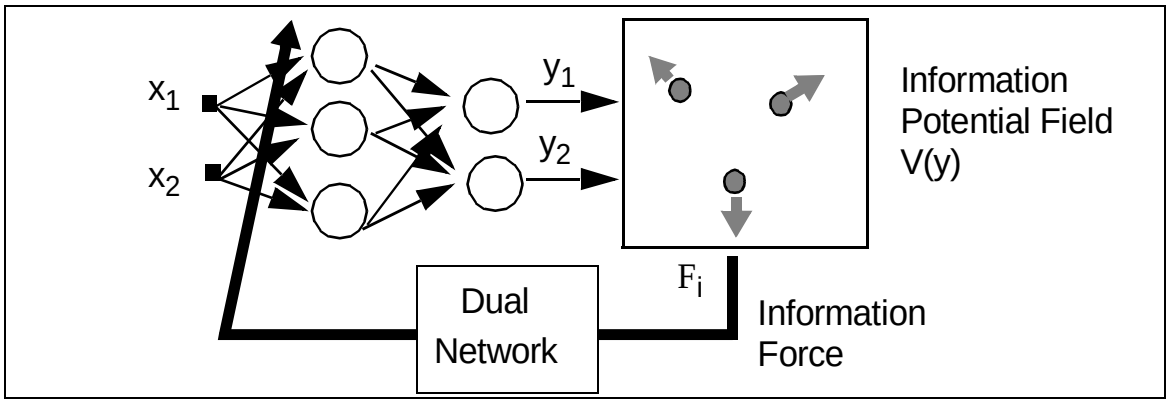


Figure 5: Training a MLP with the information potential

As a consequence, we can also expect that the algorithm will be computationally more expensive ($O(N^2)$).

This criterion can be utilized to directly implement Jayne's MaxEnt optimization principle, but instead of requiring analytic manipulations it solves the problem using the iterative approach so common in adaptive filtering and neurocomputing. The constraints in MaxEnt are here specified by the topology of the mapper. The weights of any MLP PE will be adapted with the backpropa-

gation algorithm [16] as

$$\Delta w_{ij} = \pm \eta \delta_j x_i$$

where η is the stepsize, x_i is the input to the PE, and δ_j is the local error at the PE (see [16]). If the goal is to maximize output entropy (as required by MaxEnt), the + sign is used, and if the purpose is to minimize output entropy, the - sign is required. Notice that this will change the interactions among IPCs in the output space from repulsion to attraction.

Ee conclude this section by stating that the methodology presented here lays down the framework to construct an "*entropy machine*", that is a learning machine that is capable of estimating entropy directly from samples in its output space, and can modify its weights through backpropagation to manipulate output entropy. An electronic implementation using the laws of physics to speed up the calculations is an intriguing possibility. The algorithm has complexity $O(N^2)$ since the criterion needs to examine the interactions among all pairs of output samples. Note that we are extending Bell and Sejnowski approach to ICA. Bell's approach is conceptually very elegant, but it cannot be easily extended to MLPs with arbitrary topologies nor to data distributions which are multimodal in nature. *On the other hand, Renyi's quadratic entropy becomes essentially a general-purpose criterion for entropy manipulation.*

IV Information-Theoretic Criteria: Unsupervised Learning with Quadratic Mutual Information

In the previous section we implemented a nonparametric method to solve MaxEnt. Here we will develop an ITL criterion to estimate the mutual information among random variables which enables the implementation of MinXEnt and InfoMax. Mutual Information is capable of quantifying the entropy between pairs of random variables so it is a more general measure than entropy and can be applied more flexibly to engineering problems. Mutual information at the output of a mapper can be computed as a difference of Shannon entropies $I(x, y) = H(y) - H(y|x)$. But we have to remember that Shannon entropy is not easily estimated from exemplars. Therefore this expression for *I(x,y)* can only be utilized in an approximate sense to estimate mutual information. An alternative to estimate mutual information is the Kullback-Leibler (KL) divergence [22]. The KL divergence between two PDFs $f(x)$ and $g(x)$ is:

$$K(f, g) = \int f(x) \log \frac{f(x)}{g(x)} dx \tag{14}$$

where implicitly Shannon's entropy is utilized. Likewise, based on Renyi's entropy, Renyi's divergence measure [32] with order α for two PDFs $f(x)$ and $g(x)$ is:

$$H_{R\alpha}(f, g) = \frac{1}{(\alpha - 1)} \log \int \frac{f(x)^{\alpha}}{g(x)^{\alpha - 1}} dx \tag{15}$$

The relation between the two divergence measures is:

$$\lim_{\alpha \to 1} H_{R\alpha}(f, g) = K(f, g)$$

that is, they are equivalent in the limit α=1. The K-L between two random variables Y_1 and Y_2 essentially estimates the divergence between the joint PDF and the factorized marginal PDFs, i.e.

$$I_S(Y_1, Y_2) = KL(f_{Y_1 Y_2}(z_1, z_2), f_{Y_1}(z_1) f_{Y_2}(z_2)) = \int\int f_{Y_1 Y_2}(z_1, z_2) \log \frac{f_{Y_1 Y_2}(z_1, z_2)}{f_{Y_1}(z_1) f_{Y_2}(z_2)} dz_1 dz_2 \tag{16}$$

where $f_{Y_1 Y_2}(z_1, z_2)$ is the joint PDF, $f_{Y_1}(z_1)$ and $f_{Y_2}(z_2)$ are marginal PDFs. From these formulas,

we can also observe that unfortunately none of them is quadratic in the PDF so they cannot be easily integrated with the information potential described in section III. Therefore, we propose below new distance measures between two PDFs which contain only quadratic terms to utilize the tools of IP and IF developed in section III. There are basically four different ways to write a distance measure using L_2 norms, but here we will concentrate on two:

1- Based on the Euclidean difference of vectors inequality we can write

$$\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - 2\boldsymbol{x}^T\boldsymbol{y} \geq 0 \tag{17}$$

2- Based on the Cauchy-Schwartz inequality (inner produce distance) we can write

$$\log \frac{\|\boldsymbol{x}\|^2 \|\boldsymbol{y}\|^2}{(\boldsymbol{x}^T\boldsymbol{y})^2} \geq 0 \tag{18}$$

Notice that both expressions utilize the same quadratic quantities, namely the length of each vector and their dot product. We will utilize these distance measures to approximate the K-L directed divergence between PDFs, with the added advantage that each term can be estimated with the IP formalism developed in the previous section.

For instance, based on the Cauchy-Schwartz inequality (30), we propose to measure the divergence of two PDFs $f(x)$ and $g(x)$ as

$$I_{CS}(f, g) = \log \frac{(\int f(x)^2 dx)(\int g(x)^2 dx)}{(\int f(x)g(x)dx)^2} \tag{19}$$

It is easy to show that $I_{CS}(f, g) \geq 0$ (non-negativity) and the equality holds true if and only if $f(x) = g(x)$ (identity) if *f(x)* and *g(x)* are PDFs. So (31) is also a divergence and estimates the distance between the joint quadratic entropy and the product of the quadratic entropy marginals. But it does not preserve all the properties of the K-L divergence. Likewise we can propose to estimate the divergence between two PDFs $f(x)$ and $g(x)$ based on the Euclidean distance as

$$I_{ED}(f, g) = \int f(x)^2 dx + \int g(x)^2 dx - 2\int f(x)g(x)dx \tag{20}$$

Estimators for Quadratic Mutual Information

For two random variables Y_1 and Y_2 (with marginal PDFs $f_{Y_1}(z_1)$, $f_{Y_2}(z_2)$ and joint PDF $f_{Y_1Y_2}(z_1, z_2)$), the "quadratic mutual information" based on the distance measure (31) becomes:

$$I_{CS}(Y_1, Y_2) = \log \frac{(\iint f_{Y_1Y_2}(z_1, z_2)^2 dz_1 dz_2)(\iint f_{Y_1}(z_1)^2 f_{Y_2}(z_2)^2 dz_1 dz_2)}{(\iint f_{Y_1Y_2}(z_1, z_2) f_{Y_1}(z_1) f_{Y_2}(z_2) dz_1 dz_2)^2} \tag{21}$$

It is obvious that $I_{CS}(Y_1, Y_2)$ is an appropriate measure for the independence of two variables (minimization of mutual information). We also have experimental evidence that $I_{CS}(Y_1, Y_2)$ is an appropriate measure for dependence of two variables (maximization of mutual information). Although we are unable to provide yet a strict justification that $I_{CS}(Y_1, Y_2)$ is appropriate to measure dependence, we will call (33) "Chauchy-Schwartz Quadratic Mutual Information" or CS-QMI for convenience. Now, suppose that we observe a set of data samples $\{\boldsymbol{y}_1\} = \{\boldsymbol{y}_{i1}, i = 1, \ldots, N\}$ for the variable $\boldsymbol{Y}_1$, $\{\boldsymbol{y}_2\} = \{\boldsymbol{y}_{i2}, i = 1, \ldots, N\}$ for the variable $\boldsymbol{Y}_2$. Let $\boldsymbol{y}'_i = (\boldsymbol{y}_{i1}, \boldsymbol{y}_{i2})^T$. Then $\boldsymbol{y} = \{\boldsymbol{y}'_i, i = 1, \ldots, N\}$ are data samples for the joint variable $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)^T$. Based on the Parzen window method (8), the joint PDF and marginal PDF can be estimated as:

$$\begin{cases} \hat{f}_{Y_1Y_2}(\boldsymbol{z}_1, \boldsymbol{z}_2) = \frac{1}{N}\sum_{i=1}^{N} G(\boldsymbol{z}_1 - \boldsymbol{y}_{i1}, \sigma^2 \boldsymbol{I}) G(\boldsymbol{z}_2 - \boldsymbol{y}_{i2}, \sigma^2 I) \\ \hat{f}_{Y_1}(\boldsymbol{z}_1) = \frac{1}{N}\sum_{i=1}^{N} G(\boldsymbol{z}_1 - \boldsymbol{y}_{i1}, \sigma^2 \boldsymbol{I}) \\ \hat{f}_{Y_2}(\boldsymbol{z}_2) = \frac{1}{N}\sum_{i=1}^{N} G(\boldsymbol{z}_2 - \boldsymbol{y}_{i2}, \sigma^2 \boldsymbol{I}) \end{cases} \tag{22}$$

Combining (33), (34) and using (22), we obtain the following expressions to estimate the Qua-

dratic Mutual Information $I_{CS}(Y_1, Y_2)$ based on a set of data samples:

$$I_{CS}((\boldsymbol{Y}_1, \boldsymbol{Y}_2)|\boldsymbol{y}) = \log \frac{V(\boldsymbol{y}) V^1(\{\boldsymbol{y}_1\}) V^2(\{\boldsymbol{y}_2\})}{V_{nc}(\boldsymbol{y})^2}$$

where

$$\begin{cases} V(\boldsymbol{y}) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\prod_{l=1}^{2} G(\boldsymbol{y}_{il} - \boldsymbol{y}_{jl}, 2\sigma^2 \boldsymbol{I}) \right) \\ V^l(\boldsymbol{y}_j, \{\boldsymbol{y}_l\}) = \frac{1}{N} \sum_{i=1}^{N} G(\boldsymbol{y}_{jl} - \boldsymbol{y}_{il}, 2\sigma^2 \boldsymbol{I}), \quad l = 1, 2 \\ V^l(\{\boldsymbol{y}_l\}) = \frac{1}{N} \sum_{j=1}^{N} V^l(\boldsymbol{y}_j, \{\boldsymbol{y}_l\}) \qquad l = 1, 2 \\ V_{nc}(\boldsymbol{y}) = \frac{1}{N} \sum_{j=1}^{N} \left(\prod_{l=1}^{2} V^l(\boldsymbol{y}_j, \{\boldsymbol{y}_l\}) \right) \end{cases} \tag{23}$$

In order to interpret these expressions in terms of information potentials we have to introduce some further definitions: We will use the term *marginal* when the IP is calculated in the subspace of each of the variables $\boldsymbol{y}_1$ or $\boldsymbol{y}_2$, and *partial* when only some of the IPCs are used. With this in mind, $V(\boldsymbol{y})$ is the joint information potential (JIP) in the joint space, $V_i^l = V^l(\boldsymbol{y}_i, \{\boldsymbol{y}_l\})$ is the partial marginal information potential (PMIP) because it is the potential of the sample $\boldsymbol{y}_j$ in its corresponding marginal information potential field (indexed by l). $V^l = V^l(\{\boldsymbol{y}_l\})$ is the l-th marginal information potential (MIP) because it averages all the partial marginal information potentials for one index l, and $V_{nc}(\boldsymbol{y})$ is the un-normalized *cross-information potential (UCIP) because it measures the interactions between the partial marginal information potentials* [40]. We will utilize the simplified notation herein. All these potentials can be computed from sums of pairs of interactions among the IPCs in each of the marginal fields, namely V^l_{ij}. PMIP (V^l_i), and MIP (V^l) have the same definitions as for Renyi's entropy but now qualified by the superscript l to describe which field we are referring to.

Actually, the argument of the logarithm in the first equation of (35) can be regarded as a normal-

ization for the UCIP, that is $V_{nc}(\boldsymbol{y})$ normalized by the joint information potential and the marginal information potentials. The *cross-information potential (CIP)* can then be defined as:

$$V_c(\boldsymbol{y}) = \frac{V_{nc}(\boldsymbol{y})}{V(\boldsymbol{y})} \frac{V_{nc}(\boldsymbol{y})}{V^1(\boldsymbol{y})V^2(\boldsymbol{y})} \tag{24}$$

So quadratic mutual information is measured by the CIP. With the CIP concept and the comparison between (22) and (36), we obtain consistent definitions from entropy to cross-entropy as shown by

$$\begin{cases} H(\boldsymbol{Y}|\boldsymbol{y}) = -\log V(\boldsymbol{y}) \\ I_{CS}((\boldsymbol{Y}_1, \boldsymbol{Y}_2)|\boldsymbol{y}) = -\log V_c(\boldsymbol{y}) \end{cases} \tag{25}$$

which relates the quadratic entropy with the IP and the quadratic mutual information with the CIP [40]. Therefore, maximizing the quadratic entropy is equivalent to minimizing IP, while maximizing the quadratic mutual information is equivalent to minimizing the CIP. Likewise, minimizing the quadratic mutual information is equivalent to maximizing the CIP. If we write the CIP as a function of the individual fields we obtain

$$V_c(\boldsymbol{y}) = \frac{\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} V_{ij}^1 V_{ij}^2\right)(V^1 V^2)}{\left(\frac{1}{N}\sum_{i=1}^{N} V_i^{\,1} V_i^2\right)^2}$$

and conclude that $V_c(\boldsymbol{y})$ is a *generalized measure of crosscorrelation* between the MIPs at different levels (at the individual IPC interactions, at the partial marginal information potential and marginal information potential levels).

The quadratic mutual information described above can easily be extended to the case with multiple variables $\boldsymbol{Y}_1, \ldots \boldsymbol{Y}_k$. as

$$
\begin{cases}
C((\boldsymbol{Y}_1, \boldsymbol{Y}_2)|\{\boldsymbol{y}_i\}) = -\log V_c(\{\boldsymbol{y}_i\}) \\
V(\boldsymbol{y}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G(\boldsymbol{y}_i - \boldsymbol{y}_j, 2\sigma^2\boldsymbol{I}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\prod_{l=1}^{k} G(\boldsymbol{y}_{il} - \boldsymbol{y}_{jl}, 2\sigma^2\boldsymbol{I})\right) \\
V^l(\boldsymbol{y}_j, \{\boldsymbol{y}_l\}) = \frac{1}{N}\sum_{i=1}^{N} G(\boldsymbol{y}_{jl} - \boldsymbol{y}_{il}, 2\sigma^2\boldsymbol{I}), \quad l = 1, \ldots, k \\
V^l(\{\boldsymbol{y}_l\}) = \frac{1}{N}\sum_{j=1}^{N} V^l(\boldsymbol{y}_j, \{\boldsymbol{y}_l\}), \quad l = 1, \ldots, k \\
V_{nc}(\boldsymbol{y}) = \frac{1}{N}\sum_{j=1}^{N}\left(\prod_{l=1}^{k} V^l(\boldsymbol{y}_j, \{\boldsymbol{y}_l\})\right) \\
V_c(\boldsymbol{y}) = \frac{V_{nc}(\boldsymbol{y})^2}{V(\boldsymbol{y})\prod_{l=1}^{k} V_l(\{\boldsymbol{y}_l\})}
\end{cases}
$$

“Forces” in the Cross-Information Potential

The cross-information potential is more complex than the information potential [40]. Three different information potentials contribute to the cross-information potential of (35), namely the JIP (V^l_{ij}), the PMIP (V^l_i), and the MIP (V^l). So, the force applied to each IPC $\boldsymbol{y}_i$ which is the derivative of the IP comes from three independent sources, which we call the *marginal information forces* (MIF). The overall marginal force from *k=1,2* that the IPC $\boldsymbol{y}_i$ receives is, according to (35),

$$
\frac{\partial}{\partial \boldsymbol{y}_{ik}} C((\boldsymbol{Y}_1, \boldsymbol{Y}_2)|\boldsymbol{y}) = \frac{1}{V(\boldsymbol{y})}\frac{\partial}{\partial \boldsymbol{y}_{ik}}V(\boldsymbol{y}) + \frac{1}{V^k(\{\boldsymbol{y}_k\})}\frac{\partial}{\partial \boldsymbol{y}_{ik}}V^k(\{\boldsymbol{y}_k\}) - 2\left(\frac{1}{V_{nc}(\boldsymbol{y})}\frac{\partial}{\partial \boldsymbol{y}_{ik}}V_{nc}(\boldsymbol{y})\right) \tag{26}
$$

Notice that the forces from each source are normalized by their corresponding information potentials to balance them out. This is a consequence of the logarithm in the definition of $I_{CS}(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$. Each marginal force *k* that operates on the data sample $\boldsymbol{y}_i$ can be calculated according to the fol-

lowing formulas obtained by differentiating the corresponding fields [40]

$$\begin{cases} \dfrac{\partial}{\partial \boldsymbol{y}_{ik}} V(\boldsymbol{y}) = -\dfrac{1}{N^2}\sum_{j=1}^{N}\left(\prod_{k=1}^{2} G(\boldsymbol{y}_{ik}-\boldsymbol{y}_{jk}, 2\sigma^2\boldsymbol{I})\right)\dfrac{\boldsymbol{y}_{ik}-\boldsymbol{y}_{jk}}{\sigma^2} \\ \dfrac{\partial}{\partial \boldsymbol{y}_{ik}} V^k(\{\boldsymbol{y}_k\}) = -\dfrac{1}{N^2}\sum_{j=1}^{N} G(\boldsymbol{y}_{ik}-\boldsymbol{y}_{jk}, 2\sigma^2\boldsymbol{I})\dfrac{\boldsymbol{y}_{ik}-\boldsymbol{y}_{jk}}{\sigma^2} \\ \dfrac{\partial}{\partial \boldsymbol{y}_{ik}} V_{nc}(\boldsymbol{y}) = -\dfrac{1}{N^2}\sum_{j=1}^{N}\left(\prod_{l\neq k} V^l(\boldsymbol{y}_j, \{\boldsymbol{y}_l\})\right) G(\boldsymbol{y}_{ik}-\boldsymbol{y}_{jk}, 2\sigma^2\boldsymbol{I})\dfrac{\boldsymbol{y}_{ik}-\boldsymbol{y}_{jk}}{\sigma^2} \end{cases} \tag{27}$$

Once the forces that each IPC receives are calculated by (39), they represent the injected error which can again be back-propagated to all the parameters of the mapper with backpropagation so that the adaptation with quadratic mutual information takes place. The marginal force for the two variable case is finally given by

$$F^k{}_i = -\frac{1}{N^2\sigma^2}\left[\frac{\sum_j V_{ij}^1 V_{ij}^2 d_{ij}}{\sum_i\sum_j V_{ij}^1 V_{ij}^2} + \frac{\sum_j V_{ij}^k d_{ij}}{\sum_i\sum_j V_{ij}^k} - \frac{2\sum_j V_j^1 V_{ij}^2 d_{ij}}{\sum_j V_j{}^1 V_j{}^2}\right] \tag{28}$$

Quadratic Mutual Information with the Euclidean Difference measure

We can also utilize (32) to express quadratic mutual information using the Euclidean difference (ED-QMI) of vectors inequality

$$I_{ED}(Y_1, Y_2) = \left(\iint f_{Y_1Y_2}(z_1, z_2)^2 dz_1 dz_2\right) + \left(\iint f_{Y_1}(z_1)^2 f_{Y_2}(z_2)^2 dz_1 dz_2\right) - 2\left(\iint f_{Y_1Y_2}(z_1, z_2) f_{Y_1}(z_1) f_{Y_2}(z_2) dz_1 dz_2\right) \tag{29}$$

Obviously, $I_{ED}(Y_1, Y_2) \geq 0$ and equality holds if and only if Y_1 and Y_2 are statistically independent, so it is also a divergence. Basically (41) measures the Euclidean distance between the joint pdf and the factorized marginals. With the previous definitions it is not difficult to obtain

$$\begin{aligned} I_{ED}((\boldsymbol{Y}_1, \boldsymbol{Y}_2)|\boldsymbol{y}) &= V_{ED}(\boldsymbol{y}) \\ V_{ED}(\boldsymbol{y}) &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} V_{ij}^1 V_{ij}^2 - \frac{2}{N}\sum_{i=1}^{N} V_i{}^1 V_i^2 + V^1 V^2 \end{aligned} \tag{30}$$

Although (35) and (42) differ in form, we can see that V_c is still an overall measure of cross-correlation between two marginal IPs. We have found experimentally that $I_{ED}(Y_1, Y_2)$ is better behaved than $I_{CS}(Y_1, Y_2)$ for maximization of the quadratic mutual information, while they both provide similar results for the minimization of quadratic mutual information.

It is also not difficult to obtain the formula for the calculation of the information force produced by the CIP field in the case of the Euclidean difference measure of (41)

$$\begin{aligned} c_{ij}^k &= V_{ij}^k - V_i^k - V_j^k + V^k, \quad k = 1, 2 \\ F_i^l &= \frac{\partial V_{ED}}{\partial y_i^l} = \frac{-1}{N^2\sigma^2}\sum_{j=1}^{N} c_{ij}^k V_{ij}^l d_{ij}^l \\ i &= 1, \ldots, N, \quad l \neq k, \quad l = 1, 2 \end{aligned} \tag{31}$$

where c_{ij}^k are cross matrices which serve as force modifiers.

Interpretation of the CIP

Another way to look at the CIP comes from the expression of the factorized marginal PDFs. From (34), we have:

$$f_{Y_1}(\boldsymbol{z}) f_{Y_2}(\boldsymbol{z}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G(\boldsymbol{z} - \boldsymbol{y}_{i1}, \sigma^2\boldsymbol{I}) G(\boldsymbol{z} - \boldsymbol{y}_{j2}, \sigma^2\boldsymbol{I}) \tag{32}$$

This suggests that in the joint space, there are N^2 "virtual IPCs" $\{(\boldsymbol{y}_{i1}, \boldsymbol{y}_{j2})^T, \ i, j = 1, \ldots, N\}$ whose coordinates are given by each of the coordinates of the IPCs, that is, for every real IPC location ($\boldsymbol{y}_{i1}$,$\boldsymbol{y}_{j2}$), N virtual IPCs are placed at points given by the coordinate y_{i1} of the real IPC and y_{j2}, j=1,...N, of all the other real IPCs. The PDF of (44) is exactly the factorized marginal PDFs of the IPCs. The relation between all types of IPCs is illustrated in Figure 17 for two

extreme cases.

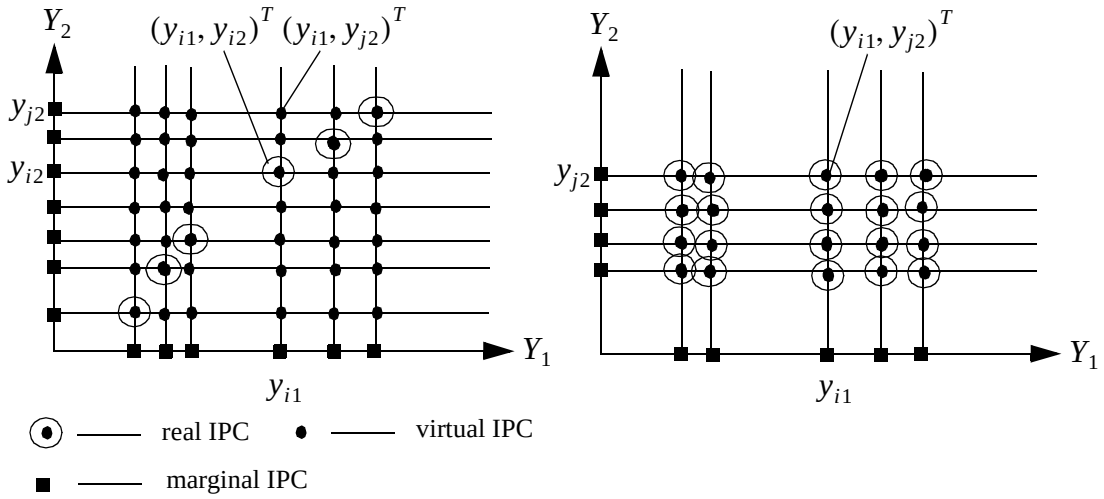


Figure 6: Illustration of "real IPC" & "virtual IPC"

In the left panel, the IPCs are distributed along a diagonal line. In this case the virtual IPCs are maximally scattered in the joint field, and the difference between the distribution of the real IPCs and virtual IPCs is maximized. In the right panel of Fig. 17 the IPCs are in a more compact distribution in the joint field. In this case the virtual IPCs occupy the same locations as the real IPCs. In this case the two fields are the same and the CIP is zero, which corresponds to the case of statistical independence of the two marginal variables $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$. All the other distributions of IPCs will provide intermediate conditions between these two extremes.

From the above description, we can re-interpret the CIP as the square of the Euclidean distance between the IP (formed by real IPCs) and the virtual IP fields (formed by virtual IPCs). CIP is a general measure for the statistical relation between two variables (based merely on the given data). It may also be noted that both $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ can be multidimensional variables, and their dimensions can be even different.

IV Conclusions

This chapter describes our efforts to develop an information-theoretic criterion which can be utilized in adaptive filtering and neurocomputing. The optimization criteria should be external to the mapper, and should work directly with the information contained in the samples, without any further assumptions. We found the answer in a combination of a nonparametric density estimator (Parzen windows) and easily computable definitions of entropy and mutual information. Although Shannon's entropy definition is the only one that obeys all the properties of information, alternate definitions have been shown of practical value. We have explored the parametric definition of entropy proposed by Renyi, and settled on the member of the family with order $\alpha = 2$, or quadratic entropy. Renyi's quadratic entropy can be readily integrated with the Parzen window estimator, yielding without any approximation (besides the PDF estimation step) an optimization criterion that is appropriate for our concept of "neural processing".

We explained how the idea of PDF estimation with Parzen windows leads to the integrated square error (ISE) method which was the first reported practical non-parametric method for ITL. ISE is a criterion external to the mapper so it can be used with any mapper, linear or nonlinear. An analysis of the computations showed that the PDF estimation can be bypassed and the criterion computed with local interactions among samples with an influence function. The algorithm has a computational complexity of $O(N^2)$, where N is the number of training patterns. We showed that the method is practical and works well, extending the work of Bell and Sejnowski for blind source separation. But the criterion seems to have other very interesting properties (such as neighborhood preservation) which have not been explored.

With Renyi's quadratic entropy we have a more principled approach to directly manipulate entropy. We provided an interpretation of the local interactions among pairs of samples as an information potential field. The injected error for the mapper can also be interpreted as an information force. This physical analogy raises hope of building an "entropy machine" based on this approach. We also showed the relationship between the information potential and the influence function obtained in ISE.

An important conclusion of this work is the form of the Renyi's entropy estimator. Although entropy is a function of the PDF, we do not need to estimate the PDF to estimate Renyi's entropy. This is due to the fact that Renyi's entropy is a function of the norm of the PDF which can be estimated directly by interactions among α-plets of data samples. A similar simplification happens in the design of classifiers, where the a posteriori probability is estimated without the need to directly estimate the PDF. This is a saving grace that will make ITL practical.

Mutual information was estimated using the Cauchy-Schwartz (CS-QMI) and the Euclidean Distance (ED-QMI) inequalities as measures of divergence between the joint density and the factorized marginals. This is a proxy (approximation) for the Kullback-Leibler divergence, but has the advantage of being easily integrated with the Parzen window method to implement sample estimators. We showed that the minimization of this cost function efficiently separates instantaneously mixed speech sources. The idea is to minimize the mutual information among the outputs of the de-mixing filter, as described in previous chapters. But with the information potential method we are using solely the information from samples, so we can potentially separate nonlinearly mixed sources.

But quadratic mutual information transcends the independent component analysis application. It can also be used for supervised learning by interpreting the variables as the desired response and the output of the mapper. We showed that the maximization of the quadratic mutual information works as an *information filtering criterion* to estimate pose from vehicle images in synthetic aperture radar. We also showed how to adapt an MLP layer-by-layer without error backpropagation. Each layer of the MLP is interpreted as an information filter with the explicit goal of maximizing mutual information between the desired response and the output of the layer. No backpropagation of errors is necessary to discover complex mappings.

Many challenging steps lie ahead in this area of research, but we hope to have shown that information-theoretic learning criteria are flexible, usable and provide more information about the data than the mean-square error (MSE) criterion which is still the workhorse of neurocomputing.

Acknowledgment: This work was partially supported by DARPA Grant F33615-97-1-1019 and NSF grant ECS- 9510715.

References

[1] Amari S., and Cardoso J., “Blind source separation - semiparametric statistical approach”, IEEE Trans. Signal PRoc., vol 45, #11, 2692-2700, 1997.

[2] Atick J.,”Could information theory provide an ecological theory of sensory processing?, Network 3, 213-251, 1992

[3] Barlow H., Kaushal T., Mitchison G.,”Finding minimum entropy codes”, Neural Computation vol 1, #3, 412-423, 1989.

[4] Burg J., “Maximum entropy spectral analysis”, Proc 37th Meeting Society of Exploration Geophysics, 1967.

[5] Bell, A.J. and Sejnowski, T.J. “An information-maximization approach to blind separation and blind deconvolution”, Neural Computation, Vol.7, no.6, pp1129-1159, 1995

[6] Cardoso, J.-F. “Infomax and Maximum Likelihood for Blind Source Separation”, IEEE Signal Processing Letters, Vol. 4. No. 4. April 1997, pp112-114

[7] Comon P., “Independent component analysis: a new concept?”, Signal Proc. vol 36, #3, pp 287-314, 1994.

[8] Cover T and Thomas J., “Elements of Information Theory”, Wiley, 1991.

[9] Deco, G., and Obradovic D., “An Information-Theoretic Approach to Neural Computing”, New York, Springer, 1996

[10] Diamantaras K. and Kung S., “Principal Component Neural Networks, Theory and Applications”, John Wiley & Sons, Inc, New York, 1996

[11] Duda, R.O., Hart P.E. “Pattern Classification and Scene Analysis”, John Wiley & Sons, 1973.

[12] Fano R., “Transmission of information”, MIT Press, 1961.

[13] Fisher J.W.III “Nonlinear Extensions to the Minimum Average Correlation Energy Filter” Ph.D dissertation, Dept. of ECE, University of Florida, 1997

[14] Fisher J., Principe J., “Blind source separation by interactions of output signals”, IEEE Workshop on Sig. Proc., DSP98, Utah.

[15] Grassberger P., Procaccia I., “Characterization of strange attractors”, Phys. rev. Letters, vol 50, #5, 346-349, 1983.

[16] Golub G. and Van Loan F., “Matrix computations”, Johns Hopkins Press, 1989.

[17] Haykin S., “Neural Networks, A Comprehensive Foundation”, Macmillan Publishing Company, 1994

[18] Hartley, R.V. “Transmission of information”. Bell System Technical Journal, 7, 1928.

[19] Jaynes E., “Information theory and statistical mechanics”, Phys. Rev., vol 106, pp 620-630, 1957.

[20] Kapur J., Kesavan H., “Entropy Optimization Principles and Applications”, Associated Press, 1992.

[21] Kapur, J.N. “Measures of Information and Their Applications”. John Wiley & Sons. 1994

[22] Kullback S. “Information Theory and Statistics”, Dover Publications, Inc., New York, 1968

[23] Kohonen T., “The self-organizing map”, Springer Verlag, 1995.

[24] Linsker R. “An application of the principle of maximum information preservation to linear systems”, in Advances in Neural Information Processing Systems 1, Touretzky D.S. (ed), Morgan-Kauffman

[25] Linsker R., “A local learning rule that enables information maximization for arbitrary input distributions”, Neural Computation, 9, 1661-1665, 1997.

[26] Marple S. “Modern spectral estimation”, Prentice-Hall, 1988.

[27] Oja, E. “A simplified neuron model as a principal component analyzer” Journal of Mathematical Biology 15. 267-273. 1982

[28] Parzen, E. “On the estimation of a probability density function and the mode”, Ann. Math. Stat. 33, 1962, p1065

[29] Plumbey M., Fallside F., “Sensory adaptation: an information theoretic viewpoint”, Int. Conf. on Neural Nets, vol 2, 598, 1989.

[30] Principe J., Fisher J., “Entropy manipulation of arbitrary nonlinear mappings”, Proc. IEEE Workshop Neural Nets for Signal Proc., 14-23, Amelia Island, 1997

[31] Papoulis A. ,”Probability, Random Variables and Stochastic Processes”, McGrawHill, 1965.

[32] Renyi, A. “Some Fundamental Questions of Information Theory”. Selected Papers of Alfred Renyi, Vol 2, pp 526-552, Akademia Kiado, Budapest, 1976.

[33] Renyi, A. “On Measures of Entropy and Information”. Selected Papers of Alfred Renyi, Vol. 2. pp 565-580, Akademia Kiado, Budapest, 1976.

[34] Rumelhart, D.E., Hinton, G.E. and Williams, J.R. “Learning representations by back-propagating errors”, Nature (London), 323, 1986, pp533-536.

[35] Shannon C. and Weaver W., “The mathematical theory of communication”, University of Illinois Press, 1949.

[36] Shannon, C.E. “A mathematical theory of communication”. Bell Sys. Tech. J. 27, 1948, pp379-423, 623-653

[37] Yang, H.H. and Amari S. “Adaptive on-line learning algorithms for blind separation -- maximum entropy and minimum mutual information”, Neural Computation, Vol 9, No. 7, 1997

[38] Vapnik V.N. “The Nature of Statistical Learning Theory”, Springer, 1995

[39] Viola P., Schraudolph N., Sejnowski T., “Empirical entropy manipulation for real-world problems”, Proc. Neural Info. Proc. Sys. (NIPS 8) Conf., 851-857, 1995.

[40] Xu, D., “From decorrelation to statistical independence”, Ph.D. proposal, U. of Florida, May 1997.

[41] Xu D, Fisher J., Principe J., “A mutual information approach to pose estimation”, Algorithms for Synthetic Aperture Radar Imagery V, vol 3370, 218229, SPIE 98, 1998.

[42] Xu D., Principe J, Fisher J., Wu H-C, “A Novel Measure for Independent Component Analysis (ICA)” in Proc. ICASSP’98 vol II 1161-1164, 1998.

[43] Xu D., “Energy, Entropy and Information Potential in NeuroComputing” Ph.D. Dissertation, U. of Florida, 1998.