Information theory approach to learning of the perceptron rule

L. Diambra[1,*] and J. Fernández[2]

[1]Departamento de Fisiologia e Biofísica, ICB Universidade de São Paulo, cep 05315-970, São Paulo, São Paulo, Brazil

[2]Departamento de Física, Universidad Nacional de La Plata, casilla de correo 67, 1900 La Plata, Argentina

(Received 31 January 2001; revised manuscript received 17 May 2001; published 20 September 2001)

By recourse to a method based on information theory, we have studied the generalization problem in perceptrons. We considered different *a priori* distributions about the weights of the teacher perceptron. Our approach allows us to define the information gain from the examples used in the training procedure. The information gain can be used to choose a convenient example set for training the perceptron and to select the transfer function of the student perceptron.

DOI: 10.1103/PhysRevE.64.046106 PACS number(s): 02.50.−r, 05.20.−y, 87.10.+e, 02.70.−c

I. INTRODUCTION

Neural networks exhibit remarkable properties for data processing, having found use in a wide variety of environments such as identification and classification of physical objects, time series processing, and image reconstruction. Given a representative set of examples, with an effective learning scheme, such systems can indeed capture the essential relationships and correlations that govern the pertinent class of input-output associations. This is evidenced both by accurate performance on training examples and by reliable generalizations or predictions for novel input patterns. Thus, trained networks are able to produce outputs corresponding to new inputs on the basis of an adequately selected *working hypothesis*. This working hypothesis is represented by a set of synaptic weights denoted by $\mathbf{W}^*$. Much effort has consequently been devoted to the task of developing suitable training algorithms that are able to adjust the synaptic weights so as to enable the network to infer the correct answer when presented with a new input (see [1,2] for a review).

Information theory (IT) [3] has proved to be of utility in devising learning techniques for perceptrons [4,5], and provides a powerful framework for discussing questions related to the learning process, such as (i) how to incorporate our *a priori* information about the teacher perceptron (TP); (ii) how to select the appropriate working hypothesis for the student perceptron (SP); and (iii) how to choose convenient examples for the training procedure.

Usually, training schemes are based on gradient descent algorithms on the training energy landscape E_t. The training energy is defined by a cost function

$$E_t(\mathbf{W}) = \sum_{\mu=1}^{p} \epsilon(\mathbf{W},\mathbf{S}^\mu) \tag{1}$$

where $\epsilon(\mathbf{W},\mathbf{S}^\mu)$ is some measure of the deviation and p is the number of examples. This scheme is liable to become trapped in local minima of the energy surface with subsequent poor generalization performance. In order to avoid this difficulty, a further generalization has been considered through incorporation of stochastic elements in the dynamics. In this refinement, the space of weights is explored by a stochastic learning process, i.e., a random walk on the training energy landscape [1]. Levin, Tishby, and Solla [6] showed that the stationary distribution of weights $P(\mathbf{W})$ is of Gibbsian character: $Z^{-1}\exp[-E_t(\mathbf{W})/T]$.

The training energy is, in most cases, a complicated function of $\mathbf{W}$, with multiple valleys and hills. In particular, for perceptrons with binary weights, one encounters regions in the (p,T) plane that contain an enormous number of metastable states as the result of strong frustration (while there is no indication of frustration for the continuous perceptron). Consequently, regarded as a relaxation phenomenon the training process can be an abnormally slow one [7]. This, of course, constitutes a serious difficulty if one wishes to optimize the set of weights because the system can be trapped in a local minimum. We show that these troubles can be avoided by regarding the training process as an *inference operation* rather than as a relaxation phenomenon. The inference process is to be accomplished according to Occam's razor, i.e., with the minimum number of assumptions compatible with the available data. Thus, the probability distribution is to be obtained by recourse to IT ideas, within the framework of Jaynes' maximum entropy principle (MEP) [8–10]. More specifically, we wish to investigate the probability distribution that ensues in a situation in which each member of the training set is regarded as a constraint for the entropy maximization procedure.

In the present work, the MEP is applied to the training of perceptrons supervised by a TP, with weight $\mathbf{W}_0$ and transfer function g_0, that provides a set of examples $D_p=\{\mathbf{S}^\mu,\zeta_0^\mu\}$, with $\mu=1,\ldots,p$. We consider here perceptrons with N input units S_i connected to an output unit ζ whose state is determined according to $\zeta=g(\mathbf{S}\cdot\mathbf{W})$, where $g(x)$ is the transfer function of the output neuron. For each set of weights $\mathbf{W}$ the perceptron maps $\mathbf{S}$ on ζ. In order to select the working hypothesis $\mathbf{W}^*$ for the SP, we infer the *a posteriori* distribution of weights $P(\mathbf{W}|D_p)$, and then we adopt as the working hypothesis $\mathbf{W}^*$ the configuration of weights that maximizes the *a posteriori* probability distribution $P(\mathbf{W}|D_p)$ (maximum likelihood criterion). The present approach offers an information measure as a bonus. This quantity, named the information gain, is defined from the *a posteriori* distribution $P(\mathbf{W}|D_p)$ which carries information about the example set

*Electronic address: diambra@fisio.icb.usp.br

used in the inferring procedure. We found that the information gain could be a useful tool to analyze and help in choosing convenient examples for the SP learning. Moreover, the information gain can tell us how suitable is a given transfer function g for reproducing the TP rule.

We organize our presentation as follows. In Sec. II we review the MEP method for obtaining the associated *a posteriori* probability distribution. The *a priori* probability distribution and the observation level concept are introduced. In Sec. III we examine how to incorporate our *a priori* information about the weights of the TP, using Gaussian and two-peaked *a priori* distributions. We also introduce an information gain measure. In Sec. IV we analyze the generalization performance of an ''average'' perceptron as well as perceptrons with weights from the maximum likelihood criterion. Further, we illustrate how the information gain can be a useful tool to help in choosing a convenient example set to be used in the learning procedure. Finally, some conclusions are drawn in Sec. V.

II. IMPLEMENTATION OF OCCAM'S RAZOR

We now describe the information theory implementation of Occam's razor in order to determine the probability distribution from the information contained in the D_p, assuming g as the transfer function of the SP. In IT parlance, a given set of observables, referred to as the relevant set for building up the pertinent statistical operator, constitutes the so-called observation level. In dealing with neural networks, one can use the information contained in the set of examples in many different ways. Each of these leads to a different probability distribution which exhibits diverse properties. The standard choice is to consider just one observable, the training energy E_t, obtained by recourse to an expression that involves the whole set of training examples. The standard observation level is then given just by E_t. As our intention is to concentrate our effort on the selection of the best working hypothesis, our idea here is to construct a more involved observation level that uses the information contained in the training set in a more efficient fashion than the standard one. If each one of the p examples is regarded as a constraint, we can indeed consider an observation level consisting of p observables.

Within the statistical physics framework, learning takes place through a modification of the probability distribution on the weight space due to incoming data. We shall assume that, given an example set D_p, each set $\mathbf{W}$ is realized with probability $P(\mathbf{W}|D_p,g)$. The entropy associated with the probability distribution is given by

$$H(D_p,g)=-\int P(\mathbf{W}|D_p,g)\ln[P(\mathbf{W}|D_p,g)]d\mathbf{W}, \quad (2)$$

while the *relative* entropy related to an *a priori* probability distribution P_0 is given by

$$H_r(D_p,g|P_0)=-\int P(\mathbf{W}|D_p,g)\ln\left[\frac{P(\mathbf{W}|D_p,g)}{P_0(\mathbf{W})}\right]d\mathbf{W}, \quad (3)$$

where $P_0(\mathbf{W})$ is an appropriate *a priori* distribution. The negative relative entropy H_r, known as the Kullback-Leibler distance [11], defines the information gained after p examples from D_p have been presented to the SP. The choice of P_0 does not depend on the examples. It depends on our knowledge of the weights of the TP, or some additional constraint imposed on the SP. As stated before, our main point is that we are employing, in an individual fashion, each of the p examples of the training set. Thus $p+1$ constraints are to be considered, given by

$$\int P(\mathbf{W}|D_p,g)d\mathbf{W}=1, \quad (4)$$

$$g^{-1}(\zeta_0^\mu)=\mathbf{S}^\mu\cdot\langle\mathbf{W}\rangle^t, \quad (5)$$

where $\langle\mathbf{W}\rangle^t$ is the transpose of $\langle\mathbf{W}\rangle$. If $p<N$, many weights $\mathbf{W}$ are compatible with our available information. Among all possible distributions $P(\mathbf{W}|D_p,g)$ that fulfill the requirements of Eqs. (4) and (5), we have to select the $P(\mathbf{W}|D_p,g)$ that contains no unjustified bias. Thus, following the central tenets of MEP, the relative entropy is maximized subjected to the constraints (4) and (5), which is tantamount to a search for the maximum of

$$-\int\left\{P(\mathbf{W}|D_p,g)\ln\left[\frac{P(\mathbf{W}|D_p,g)}{P_0(\mathbf{W})}\right]+\lambda_0P(\mathbf{W}|D_p,g)\right.$$
$$\left.+\mathbf{W}\cdot(\mathbf{S}^\mu)^t\boldsymbol{\lambda}P(\mathbf{W}|D_p,g)\right\}d\mathbf{W}, \quad (6)$$

where λ_0 and the $\boldsymbol{\lambda}$ are Lagrange multipliers associated, respectively, with the normalization condition (4) and with our p constraints (5). Variation of Eq. (6) with respect to $P(\mathbf{W}|D_p,g)$ immediately yields the *a posteriori* probability distribution

$$P(\mathbf{W}|D_p,g)=\exp[-(1+\lambda_0)]\exp(-\mathbf{W}\cdot\boldsymbol{\Gamma})P_0(\mathbf{W}), \quad (7)$$

where $\boldsymbol{\Gamma}=(\mathbf{S}^\mu)^t\boldsymbol{\lambda}$. Once P_0 is properly selected, the Lagrange multipliers $\boldsymbol{\lambda}$ are self-consistently determined from Eq. (5) and $e^{(1+\lambda_0)}$, which defines the partition function Z:

$$Z=\int\exp(-\mathbf{W}\cdot\boldsymbol{\Gamma})P_0(\mathbf{W})d\mathbf{W}. \quad (8)$$

Notice that maximizing the relative entropy is equivalent to minimizing the information gain. Expecting a good generalization performance when the information gain is minimized might seem counterintuitive. It is convenient to be aware of the inductive character of any process of gaining knowledge using IT tools. The main tenet of MEP consists in avoiding the introduction of any unnecessary hypothesis. In this sense, the minimization procedure of the information gain disregards all assumptions that are not supported by the training set.

III. SELECTING *A PRIORI* DISTRIBUTIONS

A judicious selection of the *a priori* probability distribution P_0 now becomes mandatory. In order to adequately select P_0, we must rely on our knowledge concerning the TP weights. Two instances are to be considered: (i) We assume that nothing is known about TP weights, except their finiteness, so we use a Gaussian *a priori* distribution; (ii) we know that the TP has binary weights and then we use a two-peaked *a priori* distribution.

In the first case, according to IT strictures we choose P_0 proportional to $\exp(-\mathbf{W}\cdot\mathbf{W}/2a)$. When we replace this choice in Eq. (7) we obtain a Gaussian form for the *a posteriori* distribution centered at $\langle\mathbf{W}\rangle = -a\mathbf{\Gamma}$, i.e.,

$$P(\mathbf{W}|D_p,g) = (2\pi a)^{-N/2}\exp\left[-\frac{(\mathbf{W}+a\mathbf{\Gamma})^2}{2a}\right], \qquad (9)$$

which is of the form $Z^{-1}\exp[-\beta E]$. The energy landscape exhibits a single minimum. Both the definition of $\mathbf{\Gamma}$ and the constraints (5) allow for the elimination of the Lagrange multipliers λ. Thus, one can express the $\langle\mathbf{W}\rangle$ solely in terms of data sets:

$$\langle\mathbf{W}\rangle = I_{ps}(\mathbf{S}^\mu)g^{-1}(\zeta_0^\mu), \qquad (10)$$

where $I_{ps}(\mathbf{S}^\mu)$ is the Moore-Penrose pseudoinverse of the rectangular input matrix $\mathbf{S}^\mu$ [12]. For $p \leqslant N$ the Moore-Penrose pseudoinverse is defined by $I_{ps}(\mathbf{S}^\mu) = (\mathbf{S}^\mu)^t[\mathbf{S}^\mu(\mathbf{S}^\mu)^t]^{-1}$. It should be remarked that when $p > N$ (overdetermined case) the pseudoinverse of the input matrix $\mathbf{S}^\mu$ is defined by $I_{ps}(\mathbf{S}^\mu) = [(\mathbf{S}^\mu)^t\mathbf{S}^\mu]^{-1}(\mathbf{S}^\mu)^t$, and it is equivalent to the least squares solution [13]. Now, we select the most probable set of weights compatible with the constraints as our working hypothesis, i.e., $\mathbf{W}^* = \langle\mathbf{W}\rangle$.

From Eqs. (2) and (3), we evaluate the entropy H and the information gain per weight, I_g, respectively,

$$H(D_p,g) = \frac{N}{2}[\ln(2\pi a)+1], \qquad (11)$$

$$I_g(D_p,g|P_0) = \frac{\langle\mathbf{W}\rangle\cdot\langle\mathbf{W}\rangle^t}{2aN}. \qquad (12)$$

The information gain per weight (12) depends on the parameter *a*, but only as a scaling factor. In this case, the average over examples of I_g rises linearly with increasing $\alpha = p/N$.

One more interesting case occurs when we know before hand that the TP has binary weights. It makes sense to examine the double-peaked probability distribution described by

$$P_0(\mathbf{W}) = (2\pi a)^{-N/2}\prod_i^N \left\{\frac{1}{2}\exp\left[-\frac{(W_i-1)^2}{2a}\right] + \frac{1}{2} \times\exp\left[-\frac{(W_i+1)^2}{2a}\right]\right\}, \qquad (13)$$

i.e., a *soft* form of an Ising constraint. We display in Fig. 1 the *a priori* distribution $P_0(\mathbf{W})$ for different values of the parameter *a*. When we replace this choice in Eq. (7) we obtain our *a posteriori* probability distribution as a sum of two Gaussians, weighted by $p_i^{\pm} = \exp(\pm\Gamma_i)/2\cosh(\Gamma_i)$, i.e.,

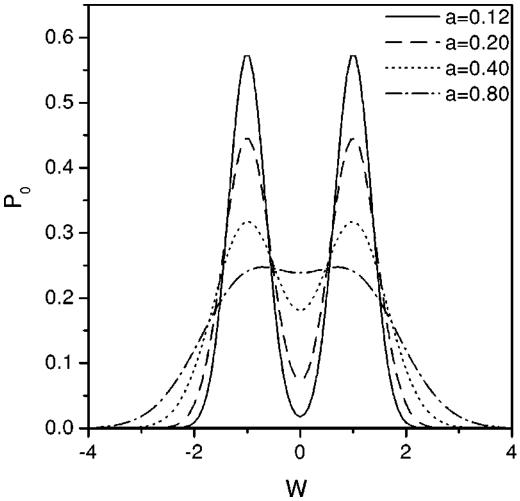

FIG. 1. Two-peaked *a priori* distribution (13) for different values of *a*: $a = 0.12$, solid line; $a = 0.20$, dashed line; $a = 0.40$, dotted line; and $a = 0.80$, dash-dotted line.

$$P(\mathbf{W}|D_p,g) = \frac{1}{(2\pi a)^{N/2}}\prod_i^N \left\{p_i^+\exp\left[-\frac{(W_i+a\Gamma_i-1)^2}{2a}\right] + p_i^-\exp\left[-\frac{(W_i+a\Gamma_i+1)^2}{2a}\right]\right\}. \qquad (14)$$

The parameter *a* can be regarded as an Ising constraint smoothness parameter. The multipliers λ_i are obtained after solving N uncoupled equations in the Γ_i, given by

$$I_{ps}(\mathbf{S}^\mu)g^{-1}(\zeta_0^\mu) + a\mathbf{\Gamma} + \tanh(\mathbf{\Gamma}) = 0. \qquad (15)$$

In the working hypothesis selection, we have in mind again the maximum likelihood criterion. Thus, our specific selection $\mathbf{W}^*$ is to be accomplished maximizing Eq. (14) where $\mathbf{\Gamma}$ is obtained by solving Eq. (15).

Thereby, the entropy H of the *a posteriori* distribution (14) is given by

$$H(D_p,g) = \frac{N}{2}\left[\ln(2\pi a) + \frac{2}{a} + 1\right] + \sum_i^N \ln[\cosh(\Gamma_i)] - \left\langle \ln\left[\cosh\left(\frac{\mathbf{W}}{a}\right)\right]\right\rangle, \qquad (16)$$

while the information gain per weight, I_g, with respect to the *a priori* distribution (13) is given by

$$I_g(D_p,g|P_0) = N^{-1}\sum_i^N \left(\frac{a}{2}\Gamma_i^2 + \Gamma_i\tanh(\Gamma_i) - \ln[\cosh(\Gamma_i)]\right). \qquad (17)$$

Therefore, $I_g(D_p,g|P_0)$ is expressed in terms of the example set D_p used in the learning procedure, and of the transfer

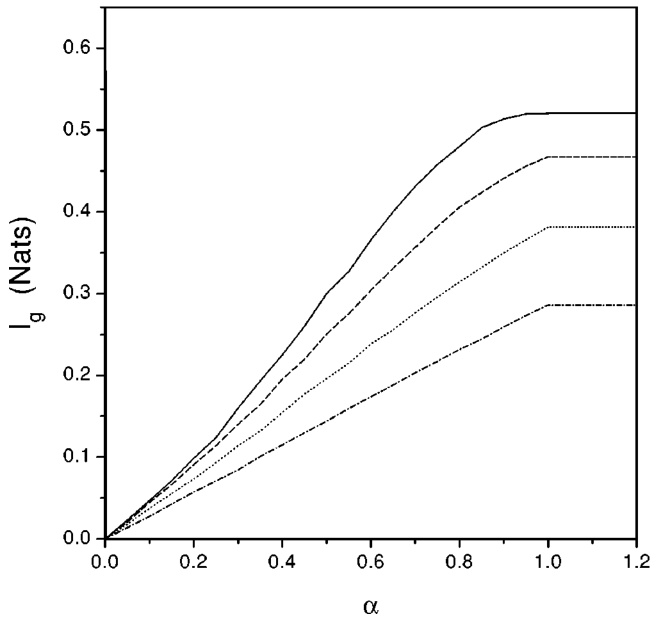

FIG. 2. Information gain per weight as a function of α for $N=100$, averaged over 500 training sets. TP has binary weights and $g_0(x)=g(x)=x$. We use the same *a priori* distributions displayed in Fig. 1: $a=0.12$, solid line; $a=0.20$, dashed line; $a=0.40$, dotted line; and $a=0.80$, dash-dotted line.

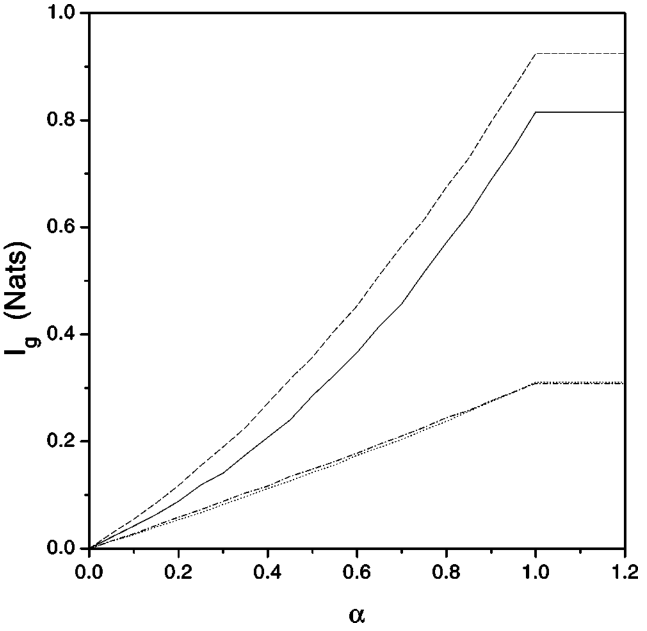

FIG. 3. Information gain per weight as a function of α for $N=100$, averaged over 500 training sets. TP has Gaussian weights $g_0(x)=g(x)=x$. We use the same *a priori* distributions displayed in Fig. 1: $a=0.12$, solid line; $a=0.20$, dashed line; $a=0.40$, dotted line; and $a=0.80$, dash-dotted line.

function of the SP. We compute I_g numerically as a function of α for a TP in two situations: (i) with binary weights, and (ii) with Gaussian weights. In both cases we perform the simulations with $N=100$, and then we average over 500 training sets. For the sake of simplicity, linear transfer functions are used for the SP and TP. Figure 2 depicts I_g for the SP from the examples prepared by a TP with binary weights, for the same values of parameter a displayed in Fig. 1. We can see that the information gain rate is constant for high values of a (typically greater than 0.4), in agreement with the Gaussian case. For $a=0.12$, we found saturation beyond $\alpha=0.90$, indicated by a very small slope in the I_g curve. For smaller values of a (not shown in the figure) the information gain rate became negative at a given value of α that depends on the value of the parameter a. For $\alpha=1$ we found a second-order transition to a phase where the SP cannot gain more information than that underlying the $p=N$ examples. Figure 3 depicts I_g for the SP from examples prepared by a TP with Gaussian weights as a function of α, for the same values of parameter a displayed in Fig. 1 and Fig. 2. We can see that, as in Fig. 2, the I_g rate is constant for higher a values, as expected for the Gaussian *a priori* distribution case. But, for smaller values of a, I_g grows exponentially, in contrast to the case with binary weights.

The limit $a\rightarrow 0$ corresponds to that case in which the weights are restricted to adopt values equal to ± 1; it was previously studied in [10]. We present below a few of those results for completeness. In this important limit the Γ_i can be expressed in analytical fashion, in terms of the training set information. We have $\boldsymbol{\Gamma}=-\tanh^{-1}[I_{ps}(\mathbf{S}^{\mu})g^{-1}(\zeta_0^{\mu})]$, and the *a posteriori* distribution (14) acquires the appearance

$$P(\mathbf{W}|D_p,g)=\prod_i^N \{p_i^+\delta(W_i-1)+p_i^-\delta(W_i+1)\}, \qquad (18)$$

where the coefficients $p_i^{\pm}$ are the probabilities of having the ith weight adopting the ± 1 values, and δ stands for Dirac's distribution. These probabilities can also be expressed in analytical fashion:

$$p_i^{\pm}=\frac{\exp\{\pm\tanh^{-1}[\{I_{ps}(\mathbf{S}^{\mu})g^{-1}(\zeta_0^{\mu})\}_i]\}}{2\cosh\{\tanh^{-1}[\{I_{ps}(\mathbf{S}^{\mu})g^{-1}(\zeta_0^{\mu})\}_i]\}}. \qquad (19)$$

Now we again use the maximum likelihood criterion in order to select the working hypothesis $\mathbf{W}^*$, i.e., we choose $W_i^*=1$ ($W_i^*=-1$), if $p_i^+>p_i^-$ ($p_i^+<p_i^-$). This recipe can be easily implemented. Just take $W_i^*=\mathrm{sgn}[p_i^+-p_i^-]$ or

$$W_i^*=\mathrm{sgn}[\{I_{ps}(\mathbf{S}^{\mu})g^{-1}(\zeta_0^{\mu})\}_i]. \qquad (20)$$

Notice that, in this case, the most probable set of weights is not equal to the mean value $\langle\mathbf{W}\rangle$, which falls outside the binary support. So it does not make sense to use "nonbinary" quantities as working hypothesis. The same problem can arise for the spherical perceptron (where, in general, $|\langle\mathbf{W}\rangle|\leqslant|\mathbf{W}_0|$) and any other perceptron whose probability distribution of weights has a support not extending to the full $\mathbf{W}$ space.

In the limit $a\rightarrow 0$, the entropy H and the information gain are not well defined, because the argument of $\tanh^{-1}$ in Eq. (19) can be greater than 1.

IV. GENERALIZATION ABILITY

The training energy measures the network's performance on a limited set of examples, whereas the ultimate goal is to find a network that performs well on all inputs, not just those in the training set. The performance of a given network **W** on the whole input space is measured by the generalization function. It is defined as the average error of the network over the whole input space, i.e.,

$$\epsilon_g(\mathbf{W}) = \frac{1}{2}\int d\mu(\mathbf{S})[g_0(\mathbf{W}_0\cdot\mathbf{S}) - g(\mathbf{W}\cdot\mathbf{S})]^2, \tag{21}$$

where $d\mu(\mathbf{S})$ denotes a measure in the input space. If the inputs are distributed independently with zero mean value and variance 1, then $d\mu(\mathbf{S}) = \Pi_i(2\pi)^{-1}e^{-S_i^2/2}dS_i$, and the generalization error can be expressed by

$$\epsilon_g(R) = \frac{1}{4\pi}\int \frac{dxdy}{\sqrt{1-R^2}}\exp\left[-\frac{x^2+y^2-2xyR}{2(1-R^2)}\right][g_0(x) - g(y)]^2, \tag{22}$$

where $R = \mathbf{W}_0\cdot\mathbf{W}/(|\mathbf{W}_0||\mathbf{W}|)$. The behavior of ϵ_g is completely determined by the parameter R. Let us consider here the generalization error of a SP with the same transfer function as the TP. In the linear case, where $g_0(x) = g(x) = x$, the generalization error is simply

$$\epsilon_g = 1 - R. \tag{23}$$

The second transfer function to be considered is the case $g_0(x) = g(x) = \sinh(x)/\sqrt{e}$. In this case, the generalization error is given by

$$\epsilon_g(R) = \sinh(1) - \sinh(R). \tag{24}$$

The powerful formalism of equilibrium statistical mechanics may now be applied to calculate averages of Eq. (22) with respect to the measure $P(\mathbf{W})$. Such averages yield information about the typical generalization performance of a network, governed by $P(\mathbf{W})$. Thus the expected value of the overlap of the SP with the TP is given by $R = \mathbf{W}_0\cdot\langle\mathbf{W}\rangle/(|\mathbf{W}_0||\langle\mathbf{W}\rangle|)$. For all cases studied here, $\langle\mathbf{W}\rangle = I_{ps}(\mathbf{S}^\mu)g^{-1}(\zeta_0^\mu)$, i.e., the mean overlap does not depend on the selected *a priori* distribution and is given by $R = \alpha$ for $\alpha \leqslant 1$. In Fig. 4, we display the generalization error of the average perceptron, with the linear transfer function as a solid line and the transfer function $\sinh(x)/\sqrt{e}$ as a dashed line.

In the working hypothesis selection we apply the maximum likelihood criterion. In the case of a Gaussian *a priori* distribution, the most probable configuration of weights is given by the mean value (10) and the generalization error curves coincide with those shown in Fig. 4. When both TP and SP have Gaussian weights the Gibbs learning scheme gives perfect generalization at $\alpha = 1$, equivalently to the approach presented here. But, when we use the doubled-peaked *a priori* distributions displayed in Fig. 1, we have a variety of scenarios. For these cases, there is no analytical solution, so we examine the different situations using simulation. In all our simulations we used $N = 100$ and we averaged over 500 training sets. Using the doubled-peaked *a priori* distributions, we consider here two situations: (i) learning a rule underlying the TP with binary weights and (ii) learning a rule underlying the TP with Gaussian weights. For the sake of simplicity, linear transfer functions are used for the SP and TP in both situations.

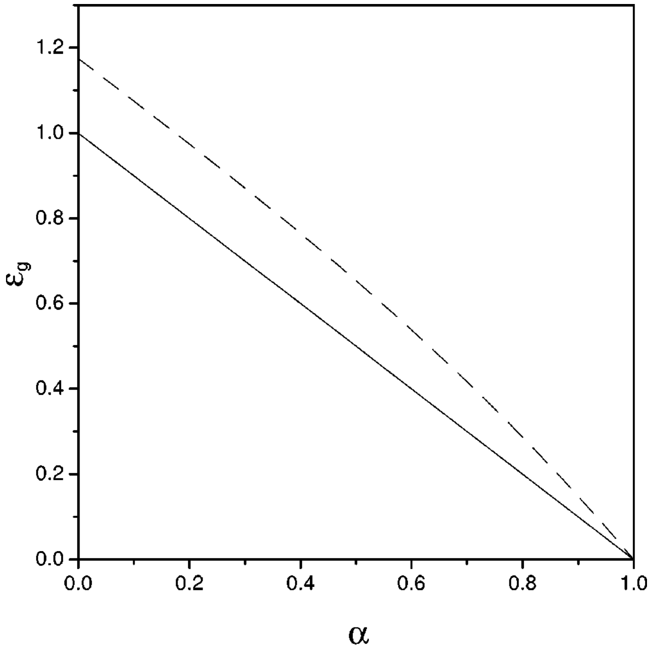

FIG. 4. Typical generalization error for a perceptron with linear transfer function (solid line), and for a perceptron with transfer function $\sinh(x)/\sqrt{e}$ (dashed line).

In the first case we have a realizable rule for the SP [14]. Figure 5 displays the generalization error for the same *a priori* distributions displayed in Fig. 1: $a = 0.12$ as the solid line, $a = 0.20$ dashed, $a = 0.40$ dotted, and $a = 0.80$ dash-dotted. We found a finite value of the error generalization at $\alpha = 1$ denoted by ϵ_{min}. The ϵ_{min} value depends on the *a priori* distribution through the parameter a as shown in Fig. 6 (solid line). The ϵ_{min} vanishes for very small (binary SP), and very large (Gaussian SP) values of the parameter a. This result suggests that inappropriate selection of the *a priori* distribution can lead to increased ϵ_{min} with consequent poor generalization performance. In this case, there is a loss of information when we interpret the example of the training set D_p as constraints (5). There is no loss of information when we choose the *a priori* distribution P_0 appropriately. In order to establish a comparison we briefly review some results of the Gibbs learning scheme at $T = 0$ for this case. For any $\alpha > 0$, the training energy possesses only one global minimum $R = 1$. However, the training energy may still possess low-R metastables states. For small α, the energy landscape far away from the optimal overlap $R = 1$ is rough, implying slow dynamic learning. Seung *et al.* [7] have shown by numerical simulations that for $\alpha = 2.39$ there are no local minima at all, and that for $\alpha > 1$ the system converges rapidly to $R = 1$ from almost all initial conditions. In this case, establishing a com-

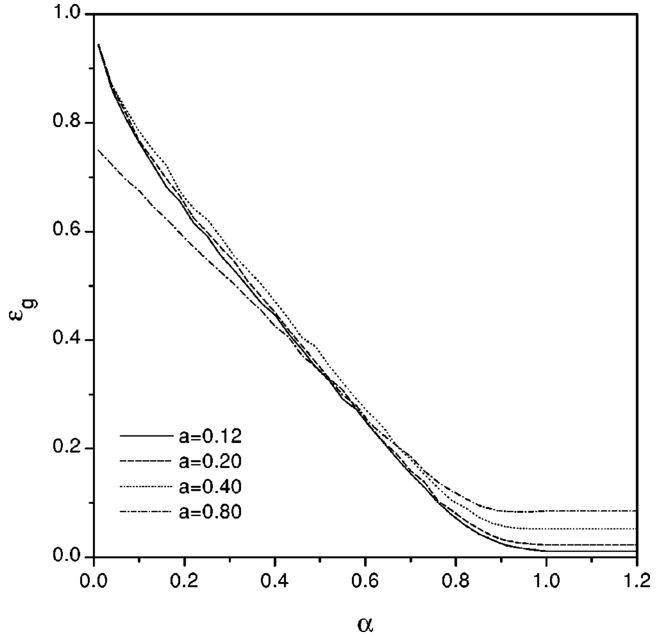

FIG. 5. Generalization error as a function of α for $N=100$, averaged over 500 training sets. The examples were generated by a TP with binary weights. We use the same *a priori* distributions displayed in Fig. 1: $a=0.12$, solid line; $a=0.20$, dashed line; $a=0.40$, dotted line; and $a=0.80$, dash-dotted line.

plete comparison is difficult due to the different scenarios presented by the two protocols, but we can say that, when we have the appropriate P_0, the protocol based on information theory and the Gibbs learning scheme at $T=0$ have a similar performance.

In the second case, the TP weights $\mathbf{W}_0$ are normally distributed, so the rule is realizable only for $a \to \infty$. Figure 7 shows ϵ_g versus α for the same *a priori* distributions displayed in Fig. 1: $a=0.12$ as a solid line, $a=0.20$ dashed, $a=0.40$ dotted, and $a=0.80$ dash-dotted. For these values of a, we have continuous learning with a finite ϵ_{min} value at $\alpha=1$. ϵ_{min} vanishes for higher values of a as shown in Fig. 6 (dashed line), but, in contrast with the earlier example, ϵ_{min} does not diminish for smaller values of a. In the limit $a \to 0$ the SP weights are restricted to binary values, so the rule is unrealizable due to a mismatch of weights between the TP and SP. In this case ϵ_{min} in the thermodynamic limit is 0.202 (dotted line in Fig. 6), the symmetric replica approach [7] yields an asymptotic learning $\epsilon_g \to \epsilon_{min}$ when $\alpha \to \infty$. These results suggest, at least in this instance, that the information theoretical approach is a more effective learning technique than the classical Gibbs learning scheme.

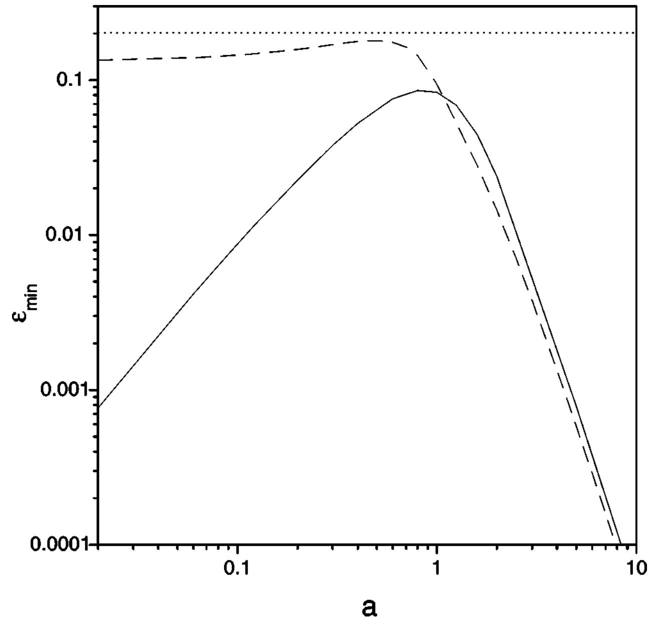

FIG. 6. ϵ_{min} versus a for $N=100$, averaged over 500 training sets. TP with binary weights as solid line and TP with Gaussian weights as dashed line. ϵ_{min} in the limit $a \to 0$ is displayed as the dotted line.

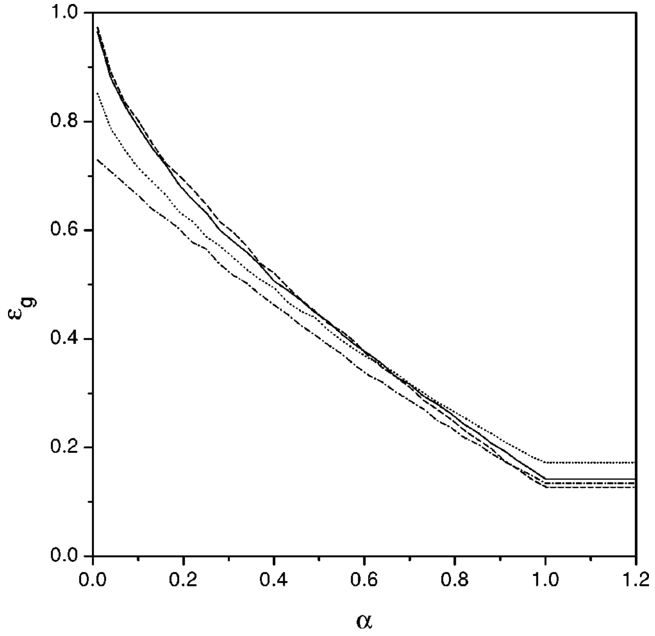

FIG. 7. Generalization error as a function of α for $N=100$, averaged over 500 training sets. The examples were generated by a TP whose weights are normally distributed. We use the same *a priori* distributions displayed in Fig. 1: $a=0.12$, solid line; $a=0.20$, dashed line; $a=0.40$ dotted line; and $a=0.80$, dash-dotted line.

Selection of examples

The expressions (12) and (17) establish the information gain I_g from the examples used in the training procedure. I_g depends also on the transfer function g used by the SP, and on the *a priori* distribution. Thus, I_g can capture some features associated with the training set. We think that the information gain could be used to evaluate the example set in different situations, and the suitability of the transfer function used by the SP. Below, we explore some of these possibilities.

We evaluated I_g for different example sets using a two-peaked *a priori* distribution with $a=0.05$ and linear transfer function for the SP and TP. In Table I, we show the I_g values obtained from three different example sets: (i) from 50 linear

TABLE I. I_g values obtained from 50 linear independent (LI) examples, and from 50 non-LI examples at the top. I_g values obtained from 50 LI examples to which we add Gaussian noise with three different standard deviations (SD) at the bottom.

Clean data	I_g	σ
LI	0.410	0.089
Non-LI	0.147	0.051
Noisy data (SD)	I_g	σ
0.10	0.426	0.090
0.05	0.426	0.090
0.01	0.426	0.090

independent (LI) examples, (ii) from a set with 25 LI examples plus 25 examples that are linear combinations of the first 25, and (iii) from 50 LI examples to which we added three different levels of Gaussian noise. Again we used $N=100$ and we averaged over 500 training sets. As one expects, the information gained using non-LI examples is less than when use LI examples. In the case of noisy examples the information gain is greater than for clean examples, because the information needed to describe this set is greater. These simple examples show that I_g works well to evaluate the information contained in the example set before training the SP. This kind of study could be relevant when we need to select the examples, and has the advantage of avoiding training procedures with the consecutive comparison of the performance of two perceptrons trained with different example sets.

The problem of learning from examples reduces to determining adequate weights, given a transfer function form g which we choose in one way or another. As discussed by Rissanen [15], there is no algorithmic way to determine this transfer function. However, the information gain I_g depends both upon the specific choice of the training set, and (in an indirect way) on the transfer function used by the SP. One may ask the following question: Is it possible that the information gain determines the suitability of a given transfer function of the SP, in order to reproduce the TP?

We can try to answer the above question by testing the hypothesis in the following manner. Consider a binary TP with transfer function $g_0(x)=\tanh(x)$ and $N=100$, which generates 75 independent examples. We compute, using a sharp two-peaked *a priori* distribution $(a=0.05)$, the information gained from these examples by two SP's, one with $g(x)=\tanh(x)$ and the other with a linear transfer function $g(x)=x$. In the case of a SP with the same transfer function as the TP, we have $I_g=0.669$ with a standard deviation of 0.061. In the case of a SP with $g(x)=x$, the information gained is $I_g=0.229$ with a standard deviation of 0.050. We can see that the information gained in this particular case is significantly greater when we choose the right transfer function for the SP.

V. DISCUSSION AND CONCLUSIONS

We have illustrated the IT approach to learning a rule from examples by perceptrons. We have regarded different *a priori* distributions in two scenarios: a TP with binary weights and a TP with Gaussian weights. We conclude that the network's performance is very sensitive to the choice of our *a priori* distribution. Our approach takes advantage of this fact in the sense of allowing the introduction of our previous knowledge concerning the nature of the TP weights. In particular, one can evaluate the probabilities associated with each weight in terms of the available examples. It should be pointed out that, in the limit $a\rightarrow 0$, our approach does not exhibit the phase transition from poor generalization to perfect generalization characteristic of the symmetric replica solution for the binary perceptron [7]. In addition, at least in the perceptron case investigated here, frustration appears to be the result of poor ''administrative management'' of the available examples. Our IT approach enables us to effectively employ all the available information, so that each example is used as a constraint. Thus, the ensuing observation level becomes much richer than the standard one. Efficient management leads to better results in neural processes as in the real world.

The IT approach also seems to offer promising perspective as a learning protocol [16,17]; the methodology presented here introduces an information measure as a bonus. We used some examples to show that this quantity can be a useful tool to explore both the example set to be used in the learning procedure, and the transfer function form used by the SP. We wish to remark that the I_g is easy and not expensive to compute. It is expressed only in terms of the training set, while measures like the mean square error are computed over many of test examples. This is certainly a notable facet of our approach that greatly helps in studying the learning process in a variety of situations, without paying an excessive computational cost.

The learning protocol presented here constitutes an extra learning technique for perceptrons, which should be of interest not only for basic research but also for applications to many interesting real world problems.

ACKNOWLEDGMENTS

The authors acknowledge fruitful discussions with N. Caticha. L.D. acknowledges the financial support of FAPESP Grant No. 99/07186-3 (Brazil) and J.F. acknowledges the financial support of UNLP (Argentina).

[1] T. Watkin, A. Rau, and M. Biehl, Rev. Mod. Phys. **65**, 499 (1993).

[2] A. Engel and C. P. L Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, 2001).

[3] C. E. Shannon and W. Weaver, in *The Mathematical Theory of Communication* (University of Illinois Press, Chicago, 1949).

[4] L. Diambra, J. Fernández, and A. Plastino, Phys. Rev. E **52**, 2887 (1995).

[5] L. Diambra and A. Plastino, Phys. Rev. E **52**, 4557 (1995).
[6] E. Levin, N. Tishby, and S. Solla, Proc. IEEE **78**, 1574 (1990).
[7] H. S. Seung, H. Sompolinsky, and H. Tishby, Phys. Rev. A **45**, 6056 (1992).
[8] E.T. Jaynes, Phys. Rev. **108**, 171 (1957).
[9] R. D. Levine and M. Tribus, *The Maximum Entropy Principle* (MIT Press, Cambridge, MA, 1978).
[10] L. Diambra and A. Plastino, Phys. Rev. E **53**, 5190 (1996).
[11] S. Kullback and R. A. Leibler, Ann. Math. Stat. **22**, 79 (1951).
[12] A. Albert, *Regression and Moore-Penrose Pseudoinverse* (Academic Press, New York, 1972).
[13] It is possible to define, in a more general way, the Moore-Penrose pseudoinverse using singular value decomposition.
[14] We say that a rule is realizable when the architectures of the SP and PT are matching.
[15] J. Rissanen, *Stochastic Complexity in Statistical Inquiry* (World Scientific, Singapore, 1989).
[16] L. Diambra and A. Plastino, Phys. Rev. E **53**, 1021 (1996).
[17] L. Diambra, A. Capurro, and A. Plastino, Phys. Lett. A **241**, 61 (1998).