Neural network training without spurious minima

L. Diambra* and A. Plastino†
Physics Department, National University La Plata, Casilla de Correro 727, 1900 La Plata, Argentina
(Received 2 January 1996)

In training student perceptions, recourse to information theory concepts allows one to select the best working hypothesis and obtain an exact solution for the associated probability distribution. We apply this training scheme to perceptions with binary weights and show that no phase transition ensues. By recourse to our approach fast learning is guaranteed and trapping by spurious local minima is avoided. [S1063-651X(96)07505-8]

PACS number(s): 87.10.+e, 05.20.−y, 02.70.−c

I. INTRODUCTION

Neural networks have been proposed as models for many cognitive functions; associative memory, generalization, categorization, etc.. These functions appear as epiphenomena (emergent collective behavior) of the interconnected neural system. A well-documented situation is that of systems able to ''learn'' from examples. Great progress has been made by recourse to techniques of statistical mechanics in analyzing the performance of a student perceptron (SP) trained by a teacher perceptron (TP) [1–3].

Generalization is a characteristic ability of feedforward networks (the perceptron, in particular). They exhibit *inference* capacities, i.e., can produce outputs, corresponding to *new* inputs (not previously presented by the TP), on the basis of an adequately selected *working hypothesis* (WH). This hypothesis is, of course, represented by a set of synaptic weights W_i that, when appropriately implemented, yields good generalization performance. 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.

In the present effort, a recently introduced [4,5] maximum entropy method is applied to perceptrons with binary weights [6,7]. 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 (invertible) *transfer function* of the output neuron. We assume that the network space is restricted to vectors that satisfy the normalization $\Sigma_i W_i^2=N$. For each set of weights $\mathbf{W}$ the perceptron maps $\mathbf{S}$ on ζ. In order to select the WH for the SP, we infer the TP state from the training set $\{\mathbf{S}^\mu,\zeta_0^\mu\}$, with $\mu=1,\cdots,p$, provided by a TP with weights $\mathbf{W}_0$, and transfer function g_0 (our available information).

The usual training schemes are stochastic processes that can be viewed as a random walk on the training energy landscape. The training energy is defined by

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

where $\epsilon(\mathbf{W},\mathbf{S})$ is some measure of the deviation of the SP answer $g(\mathbf{S}\cdot\mathbf{W})$ from the TP one, represented by $g_0(\mathbf{S}\cdot\mathbf{W}_0)$. Levin, Tishby, and Sella [8], have shown that the stationary distribution of weights $P(\mathbf{W})$ is of a Gibbsian character: $Z^{-1}\exp[-E_t(W)/T]$. The training energy is, in most cases, a complicated function of $\mathbf{W}$, with multiple valleys and hills. In the (p,T) plane one encounters regions that contain an enormous number of metastables states (as the result of a strong frustration) [1]. The time required in order to surmount the free energy barrier is of the order of $t\approx e^{N\Delta f/T}$, where Δf is the height of the free energy. Consequently, regarded as a *relaxation phenomenon* the training process can be an *abnormally slow* one [9]. This, of course, constitutes a serious difficulty if one wishes to optimize the set of weights: the system can be trapped in a local minimum with a subsequent poor generalization performance.

Here we intend to show that these troubles can be avoided by recourse to information theory (IT) ideas [10,11], that have proved to be of utility in devising learning schemes [4]. In the present effort, the training process *will not* be regarded as a relaxation phenomenon but rather as an inference operation. One wishes to infer the $\mathbf{W}$ state of the SP from the information conveyed by the training set. Our specific suggestion is that of adopting as a WH the configuration of weights that maximizes the entropy associated with the concomitant probability distribution (PD). This PD, in turn, is to be obtained by recourse to IT ideas, within the framework of Jaynes' maximum entropy principle (MEP) [11]. More specifically: we wish to investigate the probability distribution that ensues in that situation in which *each* member of the training set is regarded as a constraint (for the entropy maximization procedure) [4].

The paper is organized as follows: in Sec. II we review the MEP method for the obtention of the associated probability distribution. The *a priori* probability distribution is introduced. In Sec. III we examine two different *a priori* probability distributions and an interesting limit case is analyzed. The generalization performance is discussed in Sec. IV, and some conclusions are drawn in Sec. V.

II. THE MEP APPROACH

In IT parlance, a given (fixed) set of observables, referred to as the ''relevant'' ones in order to build up the pertinent statistical operator, constitutes the so-called *observation level* [12]. In dealing with neural networks, one can use the infor-

*Electronic address: diambra@venus.fisica.unlp.edu.ar
†Electronic address: plastino@venus.fisica.unlp.edu.ar

mation contained in the set of examples in many different ways. Each of these leads to a different probability distribution which, of course, 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 [2]. The standard observation level is then given just by E_t. As our intention is that of concentrating efforts on the selection of the *best* working hypothesis, our idea here is that of constructing 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 p examples is regarded as a constraint, we can indeed consider an observation level consisting of p observables.

The MEP is now employed [10,11] in order to determine the probability distribution $P(\mathbf{W})$ on the basis of the information contained in the training set. We shall assume that *each* set of weights $\mathbf{W}$ is realized with probability $P(\mathbf{W})$. In other words, we introduce a normalized probability distribution over all possible sets $\mathbf{W}$. Of course,

$$\int P(\mathbf{W})d\mathbf{W}=1, \tag{2}$$

where $d\mathbf{W}=dW_1dW_2\cdots dW_N$, and the *relative* entropy is, in the usual way [10,11], associated to the probability distribution $P(\mathbf{W})$, i.e., the information measure (entropy) reads

$$H=-\int P(\mathbf{W})\ln\left[\frac{P(\mathbf{W})}{P_0(\mathbf{W})}\right]d\mathbf{W}, \tag{3}$$

where $P_0(\mathbf{W})$ is an appropriately *a priori* distribution [10,11,13]. The choice of P_0 depends on the assumptions made concerning the *a priori* $\mathbf{W}$ distributions and does not depend on the examples.

As stated, our main point is that of employing, in individual fashion, each of the p examples of the training set [4]. Thus p constraints are to be considered, given by

$$g^{-1}(\zeta_0^\mu)=\mathbf{S}^\mu\cdot\langle\mathbf{W}\rangle \tag{4}$$

and (3) is maximized subject to them [11], which is tantamount to search for the maximum of [11]

$$H'=-\int\left\{P(\mathbf{W})\ln\left[\frac{P(\mathbf{W})}{P_0(\mathbf{W})}\right]+\lambda_0P(\mathbf{W})+(\mathbf{S}^\mu)^t\boldsymbol{\lambda}\cdot\mathbf{W}P(\mathbf{W})\right\}d\mathbf{W}, \tag{5}$$

where λ_0 and $\boldsymbol{\lambda}$ are Lagrange multipliers associated, respectively, to the normalization condition (2) and to our p constraints (4). Variation of H' with respect to $P(\mathbf{W})$ immediately yields

$$P(\mathbf{W})=\exp[-(1+\lambda_0)-\boldsymbol{\Gamma}\cdot\mathbf{W}]P_0(\mathbf{W}), \tag{6}$$

with $\boldsymbol{\Gamma}=(\mathbf{S}^\mu)^t\boldsymbol{\lambda}$. This is the *a posteriori* probability distribution, obtained from the MEP method. The Lagrange multipliers are self-consistently determined from (4) once P_0 is properly selected.

III. PERCEPTRON WITH BINARY COUPLING

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 architecture. Two instances are to be considered.

(i) Assume first that nothing is known about PT weights. According to IT strictures we choose, following [4,5], P_0 is proportional to $\exp(-\mathbf{W}\cdot\mathbf{W}/2a)$. When we replace this choice in Eq. (6) we obtain a Gaussian form for the probability distribution, centered in $\langle\mathbf{W}\rangle=-a\boldsymbol{\Gamma}$, i.e.,

$$P(\mathbf{W})=\frac{1}{(2\pi a)^{N/2}}\exp\left[-\frac{1}{2a}(\mathbf{W}+a\boldsymbol{\Gamma})^2\right], \tag{7}$$

which is of the form $Z^{-1}\exp[-\beta E]$. The energy landscape exhibits a single minimum and the a parameter can be regarded as a temperature. Both the definition of $\boldsymbol{\Gamma}$ and of the constraints (4) allow for the elimination of the Lagrange multipliers $\boldsymbol{\lambda}$. One can thus express the $\langle\mathbf{W}\rangle$ solely in terms of the data set:

$$\langle\mathbf{W}\rangle=I_{ps}(\mathbf{S}^\mu)g^{-1}(\zeta_0^\mu), \tag{8}$$

where $I_{ps}(\mathbf{S}^\mu)=(\mathbf{S}^\mu)^t[\mathbf{S}^\mu(\mathbf{S}^\mu)^t]^{-1}$ is the Moore-Penrose pseudoinverse [14]. We choose the only minimum in the energy landscape as our working hypothesis. In this case, the minimum is the most probable configuration of weights compatible with the constraints (4) and corresponds to the mean value (8).

(ii) If it is *a priori* known that the TP possesses binary weights, it makes sense to examine the double-peaked probability distribution [15]

$$P_0=\prod_i^N\left\{\exp\left[-\frac{(W_i-1)^2}{2a}\right]+\exp\left[-\frac{(W_i+1)^2}{2a}\right]\right\}, \tag{9}$$

i.e., a *soft* form of an Ising constraint, which is *a posteriori* found to be a quite adequate selection. Using (6) and (9) we can express our probability distribution as the sum of two Gaussians, weighted by, respectively, $p_i^{\pm}=\exp(\pm\boldsymbol{\Gamma}_i)/2\cosh(\boldsymbol{\Gamma}_i)$, i.e.,

$$P(\mathbf{W})=\frac{1}{(2\pi a)^{N/2}}\prod_{i=1}^N\left[p_i^+\exp\left(-\frac{1}{2a}(W_i+a\boldsymbol{\Gamma}_i+1)^2\right)+p_i^-\exp\left(-\frac{1}{2a}(W_i+a\boldsymbol{\Gamma}_i-1)^2\right)\right], \tag{10}$$

The a parameter cannot be regarded as a temperature, but rather as an ''Ising constraint smoothness parameter.'' The multipliers λ_i are obtained after solving N uncoupled equations in the $\boldsymbol{\Gamma}_i$, given by

$$I_{ps}(\mathbf{S}^\mu)g^{-1}(\zeta_0^\mu)+a\boldsymbol{\Gamma}+\tanh(\boldsymbol{\Gamma})=0. \tag{11}$$

We will concentrate our attention upon the interesting limit $a\to0$, corresponding to that case in which the weights are restricted to adopting values equal to ±1. In this important limit the $\boldsymbol{\Gamma}_i$ can be expressed in *analytic fashion*, in

TABLE I. Probabilities inferred (p_i^+) for some weights from p examples in a case for which $g_0(x)=g(x)=\tanh(x)$ and $N=30$.

W_0	$p=3$	$p=6$	$p=12$	$p=21$	$p=27$	$p=30$
1	0.5910	0.9711	0.7795	0.6949	0.8710	1
−1	0.5189	0.6428	0.7130	0.2302	0.0117	0
−1	0.2945	0.2761	0.3053	0.1159	0.1837	0
1	0.7993	0.5799	0.5843	0.7991	0.9827	1
1	0.5098	0.8356	0.8528	0.8193	0.8337	1

terms of the training set information. We have $\mathbf{\Gamma}=-\tanh^{-1}[I_{ps}(\mathbf{S}^\mu)g^{-1}(\zeta_0^\mu)]$, and the probability distribution (10) acquires the appearance

$$P(\mathbf{W})=\prod_i^N \{p_i^+\delta(W_i+1)+p_i^-\delta(W_i-1)\}, \tag{12}$$

where the coefficients $p_i^\pm$ are the probabilities of having the ith weight adopting the ±1 values (δ stands for Dirac's distribution [16]). These probabilities can also be expressed in analytical fashion, for any invertible g, i.e.,

$$p_i^\pm=\frac{\exp\{\pm\tanh^{-1}[\{I_{ps}(\mathbf{S}^\mu)g^{-1}(\zeta_0^\mu)\}_i]\}}{2\cosh\{\tanh^{-1}[\{I_{ps}(\mathbf{S}^\mu)g^{-1}(\zeta_0^\mu)\}_i]\}}, \tag{13}$$

a result that does not depend upon the TP architecture (neither in what refers to the weights W_0 nor to the transfer function g_0). It depends *solely* on the training set and on the SP transfer function g. Table I lists the probabilities inferred for some weights with different values of α.

In facing the working hypothesis selection one has in mind the fact that, of course, one deals with binary weights and thus it does not make sense to use ''nonbinary'' quantities (e.g., mean values) as a guide in our choice. We must select a working hypothesis that maximizes (12) and thus choose $\mathbf{W}$ so that, if $p_i^+>p_i^-$ ($p_i^+<p_i^-$) then $W_i=1$ ($W_i=-1$). This recipe can be easily implemented. Just take

$$W_i=\mathrm{sgn}[p_i^+-p_i^-]=\mathrm{sgn}[\{I_{ps}(\mathbf{S}^\mu)g^{-1}(\zeta_0^\mu)\}_i]. \tag{14}$$

IV. RESULTS

If one follows the evolution of both (i) the generalization error and (ii) the training error (E_t) with the number of examples we obtain the *learning curves*. In order to evaluate the generalization performance the generalization error was defined in terms of the distances between the desired outputs ζ_0^μ and the actual outputs ζ^μ corresponding to the given inputs $\mathbf{S}$

$$\epsilon_g(\mathbf{W})=\frac{1}{2}\int d\mu(\mathbf{S})[g_0(\mathbf{W}_0,\mathbf{S})-g(\mathbf{W},\mathbf{S})]^2, \tag{15}$$

where $d\mu(\mathbf{S})$ denotes a measure in the input space. If the inputs are distributed independently with zero mean value and variance one, then $d\mu(\mathbf{S})=\Pi_i(2\pi)^{-1}e^{-S_i^2/2}dS_i$, and the generalization error can be expressed as an integral over two Gaussian variables x and y [1] given by

$$\epsilon_g(\mathbf{W})=\frac{1}{2}\int Dx\ Dy\{g_0(x)-g[(1-R^2)^{1/2}y+Rx]\}^2, \tag{16}$$

where $R=N^{-1}\mathbf{W}_0\cdot\mathbf{W}$ and Dx is a Gaussian measure. The behavior of the generalization error is completely determined by the order parameter $\mathbf{R}$.

We examine the behavior of ϵ_g in the different cases referred to above. In our simulation we deal with $N=80$ and average over 200 samples. Figure 1 depicts ϵ_g, for the transfer function $g_0(x)=g(x)=\tanh(x)$, in two situations:

(i) The TP is of a binary coupling type and we assume a Gaussian P_0 (dashed line). The weights are given by (8).

(ii) We employ the *a priori* probability distribution (9). The working hypothesis is again that of maximum likelihood and is given by (14). The solid line in Fig. 1 displays the associated ϵ_g values.

In both instances (albeit, in diverse fashion) ϵ_g vanishes when $\alpha=1$. If the SP and TP architectures are different, it is impossible for the former to *perfectly* learn rules. These rules can accordingly be called realizable or not realizable, depending upon whether perfect learning is (or is not) an attainable goal. We consider the case in which the SP-transfer function is not identical to that of the TP (see caption of Fig. 1). In this case, the ($R=1$)-value is reached for $\alpha=1$, but ϵ_g (cf. Fig. 1, dotted line) does not vanish. Indeed, it steadily diminishes and reaches a minimum value $\epsilon_{\min}$, which depends upon the concomitant transfer functions. On the other

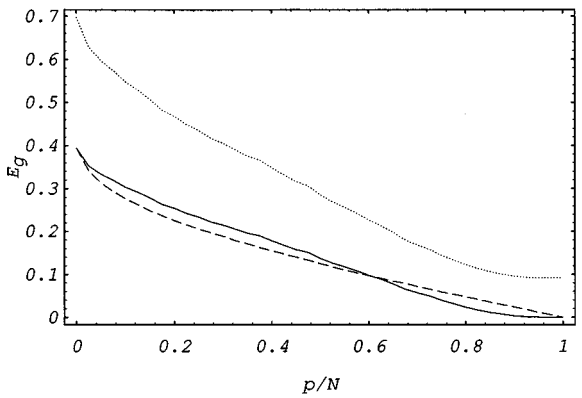

FIG. 1. Generalization error versus α for $N=80$ and average over 200 samples. An Ising *a priori* probability distribution is used in obtaining the results depicted by the solid line. Instead, those of the dashed line correspond to a Gaussian P_0. Results in a nonrealizable case with $g_0(x)=x$ and $g(x)=\tanh(x)$ are represented by the dotted line.

hand, no phase transition takes place here when $a \to 0$, in contrast with what one finds by studying the replica symmetric solution [1].

As another example, we consider a situation for which the unrealizability is due to a mismatch (in weight space) between TP and SP. Linear transfer functions are used, for the sake of simplicity. We assume that the weights $\mathbf{W}_0$ adopt unrestricted real values, whereas the SP weights are restricted to $W_i = \pm 1$. For a Gaussian $\mathbf{W}_0$ distribution, the maximal overlap $R_{\max}$ is obtained for $\mathbf{W} = \mathrm{sgn}[\mathbf{W}_0]$. In the thermodynamic limit, $R_{\max} = \sqrt{2/\pi}$. The symmetric replica solution [1] yields an asymptotic form for ϵ_g given by

$$\epsilon_g = \epsilon_{\min} + \frac{\epsilon_{\min} R_{\max}}{\alpha} + O(\alpha^{-2}), \tag{17}$$

with $\epsilon_{\min} = 1 - R_{\max} = 0.202$. Ours is a totally different scenario. We see that ϵ_g reaches a minimum $\epsilon_{\min}$ for $\alpha = 1$ (see Fig. 2).

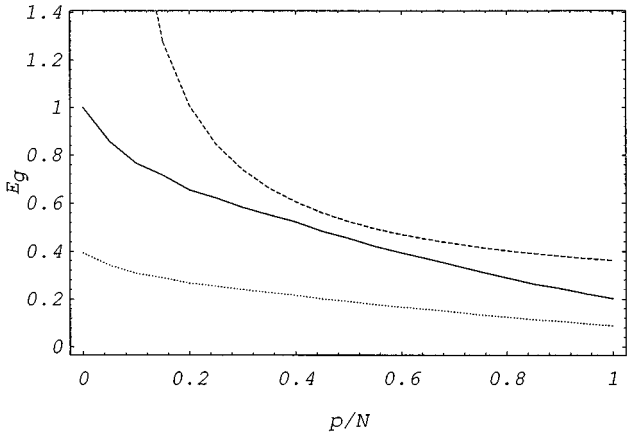

FIG. 2. Generalization error with mismatched weights: the solid line corresponds to our scheme and the dashed line to the symmetric replica solution, with $g_0(x) = g(x) = x$ in both cases. The dotted line corresponds to $g_0(x) = x$ and $g(x) = \tanh(x)$. Additional details are as in Fig. 1.

V. CONCLUSIONS

We conclude that that network's performance is very sensitive to the choice of our *a priori* probability distribution (APPB). Our approach takes advantage of this fact in the sense of allowing for the introduction of our previous knowledge concerning the nature of the TP weights in the APPB choice. In particular, if one employs a double-peaked a priori probability distribution, one can evaluate in analytical fashion the probabilities associated to each weight in terms of the available examples.

It is to be pointed out that our approach does not exhibit the phase transitions characteristic of the symmetric replica solution for the binary perceptron. Besides, at least in the perceptron case here investigated, frustration appears to be the result of poor ''administrative management'' of the available information. Our IT approach enables us to effectively employ *all* the available information, as *each* example is used as a constraint. The ensuing observation level becomes thus much richer than the standard one. Efficient management leads to better results, in neural network processes as in the ''real'' world.

ACKNOWLEDGMENTS

It is a pleasure to thank C. Mostaccio, J. M. Fernández, and M. Portesi for enlightening discussions. L. D. acknowledges the financial support of the Buenos Aires Scientific Commission (CICPBA) and A.P. that of the Argentine Research Council (CONICET).

[1] H. S. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. A **45**, 6056 (1992).
[2] T. Watkin, A. Rau, and M. Biehl, Rev. Mod. Phys. **65**, 499 (1993).
[3] S. Bös, W. Kinzel, and M. Opper, Phys. Rev. E **47**, 1384 (1993).
[4] L. Diambra, J. Fernandez, and A. Plastino, Phys. Rev. E **52**, 2887 (1995).
[5] L. Diambra and A. Plastino, Phys. Rev. E **53**, 1021 (1996).
[6] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing* (MIT, Cambridge, MA, 1986).
[7] F. Rosemblatt, *Principles of Neurodynamics* (Spartan, New York, 1962).
[8] E. Levin, N. Tishby, and S. Solla, Proc. IEEE **78**, 1574 (1990).
[9] In the spin glass phase, the metastable states are separated by energy barriers which diverge with N, and the termal fluctuation are frozen.
[10] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Chicago, Il, 1949).
[11] E. T. Jaynes, Phys. Rev. **106**, 620 (1957); **108**, 171 (1957).
[12] E. Fick and G. Sauermann, *Quantenstatistik dinamischer Prozesse, Band I* (Verlag Harri Deutsch, Francfurt/Main, 1983).
[13] R. D. Levine and M. Tribus, *The Maximum Entropy Principle* (MIT Press, Boston, MA, 1978).
[14] A. Albert, *Regression and Moore-Penrose Pseudoinverse* (Academic, New York, 1972).
[15] H. Sompolinsky, N. Tishby, and H. S. Seung, Phys. Rev. Lett. **65**, 1683 (1990).
[16] A. H. Zemanian, *Distribution Theory and Transform Analysis* (Dover, New York, 1987).