Perturbative treatments and learning techniques

L. Diambra* and A. Plastino†
Departamento de Física, Universidad Nacional de La Plata, C.C. 727, 1900 La Plata, Argentina

(Received 11 August 1995)

A second order approximation for studying the thermodynamics of the learning process in neural networks is proposed. Particular attention is paid to noise effects. We show that the inclusion of pair interactions between replicas considerably improves upon the well-known first-order approach.

PACS number(s): 87.10.+e, 02.50.−r, 05.20.−y

I. INTRODUCTION

The replica trick (RT) of statistical mechanics has become a very useful tool for the investigation of complex systems in general and, in particular, for studying learning and generalization processes in neural networks (NN) [1,2]. The trick overcomes the difficulty of performing an ensemble average of the logarithm of the partition function Z by that of averaging Z over n replicas of the original network, with n a very large integer. For feedforward, single layered NN [3,4] the RT ansatz has been proved to be a valuable tool for the learning of a rule on the basis of a suitable set of examples. In particular, the full quenched theory has been carefully studied within the framework of a replica symmetry approximation [5]. However, as the temperature drops, symmetry breakings may invalidate this approximation, so that an approach to the full quenched theory that incorporates disorder effects due to "improper" examples may be of some interest. In the present effort we wish to study, in the spirit of a second-order approximation, *noise effects* in the training set, unavoidable in any realistic setting. The noise will here be the result of letting just part of the examples to be produced by the perceptron teacher (PT). The rest are to be randomly selected ("bad" examples).

Two types of situations are to be confronted: learnable and unlearnable rules [6]. For the former, there is at least a vector in the concomitant weight space that can learn the rule in an exact fashion. The latter arises mostly in cases of architectural mismatch. In such a situation the training error can never vanish. The question to be answered is, can a perceptron trained under these circumstances correctly respond to queries posed by a PT? In other words, is the rule underlying the "good" examples a learnable one? We will show here that these questions can be adequately dealt with.

The paper is organized as follows. Section II is devoted to a brief recapitulation of basic concepts concerning the RT, while Sec. III deals with the thermodynamics of the situation that interests us here. A second order, high temperature approximation is derived. Boolean perceptrons with Ising weights on the learning curves are the subject of Sec. IV. Finally, some conclusions are drawn in Sec. V.

II. THE REPLICA METHOD

Consider learning by a single-layered perceptron [3,4] within a statistical mechanics environment [7–9]. Our NN has N input units S_i connected to a single output unit σ, whose state is given by

$$\sigma = g\left(\frac{1}{\sqrt{N}}\mathbf{W}\cdot\mathbf{S}\right), \tag{1}$$

where $g(x)$ is the transfer function. For each set $\mathbf{W}$ of weights, the NN maps $\mathbf{S}$ onto σ. Learning is said to take place whenever the W_i are so chosen that σ closely approaches the desired, correct map $\sigma_0(\mathbf{S}) = g[(1/\sqrt{N})\mathbf{W}_0\cdot\mathbf{S}]$. Within the supervised learning scheme [10] one reaches this goal by recourse to a cost function that is constructed on the basis of P examples

$$\{\mathbf{S}^l, t^l\}, \quad l=1,\dots,P. \tag{2}$$

Here we assume that a subset of P_1 examples are of a PT-generated character, so that $t^l=\sigma_0(\mathbf{S}^l)$, with $l=1,\dots,P_1$, while the remaining $P_2=P-P_1$ outputs t^l are of a random nature. Both the random inputs $\mathbf{S}^l$ and the random outputs t^l are themselves randomly selected according to probabilities $D(\mathbf{S})$ and $D(t)$, respectively, from the input and output spaces.

The training process can be associated to the minimization of the cost function (or training energy) cost E_t, defined by

$$E_t(\mathbf{W}) = \sum_{l=1}^{P} \epsilon(\mathbf{W},\mathbf{S}^l,t^l), \tag{3}$$

where $\epsilon(\mathbf{W},\mathbf{S},t)$ is the so-called error function, a measure of the deviation between actual and correct outputs. Here we focus our attention upon Boolean perceptrons (BP), for which $\epsilon(\mathbf{W},\mathbf{S},t)=\theta(-N^{-1/2}(\mathbf{W}\cdot\mathbf{S})t)$, θ standing for the Heaviside function.

The learning process can be regarded as a stochastic dynamic one, with the NN weights evolving according to a Langevin-like relaxation prescription that leads to a Gibbs' probability distribution for the weights [11–13]

$$P(\mathbf{W}) = Z^{-1}\exp[-\beta E_t(\mathbf{W})], \tag{4}$$

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

with $\beta = 1/T$ and T a "temperature" characterizing the noise level in the learning process. The normalization factor Z is, of course, the partition function

$$Z = \int d\mathbf{W} \exp[-\beta E(\mathbf{W})]. \tag{5}$$

Two different types of noise are to be dealt with here: that coming from the stochastic learning process itself and that arising from the randomness of the examples. The energy above depends upon the specific examples that have been selected. Thus, a double averaging is called for: a thermal one over the weight space with probability distribution $P(\mathbf{W})$, to be denoted by $\langle \cdots \rangle_T$ and a so-called "quenched average," over all possible *{input, output}* "pairs," to be represented by $\langle\langle \cdots \rangle\rangle \equiv \int \Pi_l d\mu(\mathbf{S}^l) d\mu(t^l)$, $d\mu(\mathbf{S}^l)$, where $(d\mu(t^l))$ is the special measure: $d\mu(\mathbf{S}^l) = D(\mathbf{S}^l) d(\mathbf{S}^l)$.

The NN free energy F and entropy S are given, respectively, by

$$F(T,P) = -T\langle\langle \ln Z \rangle\rangle \tag{6}$$

$$S(T,P) = -\left\langle\!\!\left\langle \int d\mathbf{W} P(\mathbf{W}) \ln P(\mathbf{W}) \right\rangle\!\!\right\rangle. \tag{7}$$

The network's performance over the space of the examples is characterized by the average generalization error ϵ_g. The performance with relation to the training set (2), on the other hand, is given by the average training error ϵ_t, i.e.,

$$\epsilon_t(T,P) = P^{-1} \langle\langle \langle E(\mathbf{W}) \rangle_T \rangle\rangle, \tag{8}$$

$$\epsilon_g(T,P) = \langle\langle \langle \epsilon(\mathbf{W}) \rangle_T \rangle\rangle, \tag{9}$$

where in (9) it is to be understood that the "examples averaging" is performed only over the inputs and not over the outputs, as our interest lies in quantifying the NN's performance with respect to the underlying rule. Graphs of either $\epsilon_t(T,P)$ and $\epsilon_g(T,P)$ versus P are called *learning curves.* The following important relation holds:

$$F = P\epsilon_t - TS. \tag{10}$$

The RT is the usual tool employed to evaluate the average over the examples [11–13]. It originated with reference to spin glasses [14,15], and is by now a common NN artifact [16–18]. The RT is to be recommended whenever it is feasible to evaluate averages for Z, but not for lnZ. The RT exploits the identity

$$\langle\langle \ln Z \rangle\rangle = \lim_{n \to 0} n^{-1} \ln \langle\langle Z^n \rangle\rangle, \tag{11}$$

where Z^n can be regarded as the partition function of n identical noninteracting systems (copies of the original one). We distinguish among them by recourse to a label $\gamma = 1, \ldots, n$. In performing the averaging process over the examples, coupling arises among the distinct copies. This is easily seen by recourse to an effective replica Hamiltonian. Interchanging the order of the multiple concomitant integrals one is in a position to write

$$F = -\beta^{-1} \lim_{n \to 0} n^{-1} \ln \int \prod_{\gamma=1}^{n} d\mathbf{W}_\gamma \exp\left[-N\alpha \left(\frac{1}{1+\rho} g[\mathbf{W}_\gamma] + \frac{\rho}{1+\rho} h[\mathbf{W}_\gamma] \right) \right], \tag{12}$$

where $\rho = P_2/P_1$. In (12), g stands for the contribution of the good (PT-related) examples

$$g[\mathbf{W}_\gamma] = -\ln \int D(\mathbf{S}) d\mathbf{S} \exp\left[-\beta \sum_{\gamma=1}^{n} \epsilon(\mathbf{W}_\gamma, \mathbf{S}) \right]. \tag{13}$$

Since the associated output is the PT one, we have (see above) $\epsilon(\mathbf{W}_\gamma, \mathbf{S}) = \theta(-N^{-1/2}(\mathbf{W}_\gamma \cdot \mathbf{S})(\mathbf{W}_0 \cdot \mathbf{S}))$. There is no need to average over t. Likewise, the "bad" contribution is

$$h[\mathbf{W}_\gamma] = -\ln \int \int D(t) dt D(\mathbf{S}) d\mathbf{S} \times \exp\left[-\beta \sum_{\gamma=1}^{n} \epsilon(\mathbf{W}_\gamma, \mathbf{S}, t) \right]. \tag{14}$$

Now, $H = (1/1+\rho) g[\mathbf{W}_\gamma] + (\rho/1+\rho) h[\mathbf{W}_\gamma]$ is an intensive quantity that does not depend upon the number of examples N of Eq. (12). We are in this way guaranteed that both the energy and the entropy are proportional to N, being thus in a position to describe all observables in terms of an effective replicated system. A full quenched treatment of this system, following the developments of Ref. [5], would involve considerable effort, so that it should be of utility to study a *simpler* approach. Borrowing ideas from other fields of physics (quantum mechanics, for example [19]), instead of dealing with H in its entirety we consider just its most significant *part* (at high temperatures), by recourse to a series expansion. Of course, we pay the customary price: the approximation is valid just for some appropriate range of the perturbative parameter (β in our case).

III. PERTURBATIVE TREATMENT

We shall expand H in powers β and then study the behavior of the different terms of the series. We have

$$H[\mathbf{W}_\gamma] = \beta H_1 + \frac{1}{2}\beta^2 H_2 + O(\beta^3), \tag{15}$$

with

$$H_1 = \frac{1}{1+\rho} \sum_{\gamma=1}^{n} e(\mathbf{W}_\gamma) + \frac{\rho}{1+\rho} \sum_{\gamma=1}^{n} \int D(t) dt\, D(\mathbf{S}) d\mathbf{S}\, \epsilon(\mathbf{W}_\gamma, \mathbf{S}, t), \tag{16}$$

where $e(\mathbf{W}) = \int D(\mathbf{S}) d\mathbf{S}\, \epsilon(\mathbf{W}, \mathbf{S})$ is the generalization function that depends only upon the overlap $R = N^{-1} \mathbf{W} \cdot \mathbf{W}_0$. For a Boolean perceptron, (i) the generalization function is given by [5,6] $\pi e(\mathbf{W}) = \arccos(R)$, and (ii) the second term in (16) equals unity. H_1 represents the "nonrandom" part of the training energy, while H_2 is responsible for the two-replica coupling arising out of the randomness of the training examples. When T diminishes this coupling becomes more and more important so that one needs to consider H_2 contributions. One has

$$H_2=\frac{1}{1+\rho}\sum_{\gamma,\delta=1}^{n}\left(e(\mathbf{W}_\gamma)e(\mathbf{W}_\delta)-\int D(\mathbf{S})d\mathbf{S}\;\epsilon(\mathbf{W}_\gamma,\mathbf{S})\epsilon(\mathbf{W}_\delta,\mathbf{S})\right)+\frac{\rho}{1+\rho}\sum_{\gamma,\delta=1}^{n}\left(1-\int D(t)dt\;D(\mathbf{S})d\mathbf{S}\times\epsilon(\mathbf{W}_\gamma,\mathbf{S},t)\epsilon(\mathbf{W}_\delta,\mathbf{S},t)\right). \qquad (17)$$

Of course, higher order terms in β are associated to three-replica coupling, four-replica ones, and so on. Replicas can be regarded as particles with N degrees of freedom. The first term in (15) describes the coupling of the *particles* with an external field, while the second one represents *two-particle* interactions via an effective potential depending upon the Hamming distance between the replicas. The temperature T is the associated coupling constant. It is reasonable to expect our expansion to yield an adequate treatment for $T>1$.

In the limit $\beta\rightarrow 0$ with $\alpha\beta$ constant, only H_1 survives and energy fluctuations arising out of the randomness in the examples, of the order $\sqrt{P}$, can be neglected [5]. Further, thermodynamic functions depend only upon the effective temperature T/α. This high temperature limit is interesting because it enables one to predict the existence of possible phase transitions to states of perfect generalization $R=1$ [5]. However, at lower temperatures, this high-T approach does not allow for predictions concerning transitions to other states (i.e., spin glass) that should exist according to studies made with the quenched complete theory [5]: one cannot neglect any longer the above mentioned energy fluctuations.

It is our goal here to introduce, within the present context, a perturbative treatment that enables one to incorporate the disorder effects produced by the randomness in the examples. To this effect, we shall consider the next order in β, which entails taking into account two-replica correlations. This leads to consideration of the integrals (details in the Appendix)

$$C_{1_{\gamma\delta}}=\int D(\mathbf{S})d\mathbf{S}\;\epsilon(\mathbf{W}_\gamma,\mathbf{S})\epsilon(\mathbf{W}_\delta,\mathbf{S}),$$

$$C_{2_{\gamma\delta}}=\int D(t)dt\;D(\mathbf{S})d\mathbf{S}\;\epsilon(\mathbf{W}_\gamma,\mathbf{S},t)\epsilon(\mathbf{W}_\delta,\mathbf{S},t). \qquad (18)$$

The relevant order parameter is here $Q_{\gamma\delta}=N^{-1}\mathbf{W}_\gamma\cdot\mathbf{W}_\delta$ (two-replica overlap), which does not appear, of course, at high temperatures, where replica-replica correlations can be neglected. At second order, the Hamiltonian [cf. (15) and (16)] reads

$$H=\frac{\beta}{\pi(1+\rho)}\sum_{\gamma}^{n}\arccos(R_\gamma)+\frac{n\beta\rho}{1+\rho}-\frac{\beta^2}{2(1+\rho)}\sum_{\gamma\delta}^{n}C_{1\gamma\delta}-\frac{\rho\beta^2}{2(1+\rho)}\sum_{\gamma\delta}^{n}C_{2\gamma\delta}, \qquad (19)$$

where second-order terms in the total number of replicas n have been eliminated. The dependence upon the weights is to be found just in the order parameters R_γ and $Q_{\gamma\delta}$. Second order calculations now involve companion variables $\hat{R}_\gamma$ and $\hat{Q}_{\gamma\delta}$. We have then

$$\langle\langle Z^n\rangle\rangle=\int\prod_{\gamma<\delta}dQ_{\gamma\delta}\int\prod_{\gamma}dR_\gamma\exp(-N\alpha H[R_\gamma,Q_{\gamma\delta}])\times\int\prod_{\gamma}d\mathbf{W}_\gamma\prod_{\gamma<\delta}\delta(NQ_{\gamma\delta}-\mathbf{W}_\gamma\cdot\mathbf{W}_\delta)\times\prod_{\gamma}\delta(NR_\gamma-\mathbf{W}_\gamma\cdot\mathbf{W}_0)$$
$$=\int\prod_{\gamma<\delta}\frac{dQ_{\gamma\delta}d\hat{Q}_{\gamma\delta}}{2\pi i}\int\prod_{\gamma}\frac{dR_\gamma d\hat{R}_\gamma}{2\pi i}\times\exp(-N(\alpha H-S_d)), \qquad (20)$$

where S_d is the logarithm on the density of nets whose overlaps are R_γ and $Q_{\gamma\delta}$

$$S_d=N^{-1}\ln\int\prod_{\gamma}d\mathbf{W}_\gamma\exp\left(\sum_{\gamma}\hat{R}_\gamma\mathbf{W}_\gamma\cdot\mathbf{W}_0+\sum_{\gamma<\delta}\hat{Q}_{\gamma\delta}W_\gamma\cdot\mathbf{W}_\delta\right)-\sum_{\gamma}R_\gamma\hat{R}-\sum_{\gamma<\delta}Q_{\gamma\delta}\hat{Q}_{\gamma\delta}. \qquad (21)$$

In the thermodynamic limit $N\rightarrow\infty$ the integral (20) receives an overwhelming contribution from the minima of the variables $R_\gamma,\hat{R}_\gamma$, $Q_{\gamma\delta}$, and $\hat{Q}_{\gamma\delta}$. Thus, we must evaluate

$$-\beta f=\lim_{n\rightarrow 0}\frac{1}{n}N^{-1}\ln\langle\langle Z^n\rangle\rangle=\min\{\alpha H[R_\gamma,Q_{\gamma\delta}]-S_d[R_\gamma,\hat{R}_\gamma,Q_{\gamma\delta},\hat{Q}_{\gamma\delta}]\}. \qquad (22)$$

Here some physical reasoning is needed in order to simplify things. It is reasonable to assume that all replicas have the same overlap with the teacher network and that, further, the overlap between two of them is the same for every pair. Thus, we should have

$$Q_{\gamma\delta}=\delta_{\gamma\delta}+(1-\delta_{\gamma\delta})q,$$
$$R_\gamma=R. \qquad (23)$$

With this approximation (replica symmetry) for R_γ, $Q_{\gamma\delta}$, $\hat{R}_\gamma$, and $\hat{Q}_{\gamma\delta}$, passing to the limit $n\rightarrow 0$ and using (10) we are in a position to write

$$f=\alpha\epsilon_t-Ts, \tag{24}$$

where

$$\epsilon_t=\frac{1}{\pi(1+\rho)}\arccos(R)+\frac{\rho}{1+\rho}-\frac{\beta(P_1+P_2)}{4P\pi}\left[\frac{\pi}{2}-\arctan\left(\frac{q}{\sqrt{1-q^2}}\right)\right], \tag{25}$$

$$s=\frac{1}{2}(q-1)\hat{q}-R\hat{R}+\int \mathbf{Dz}\ln\int d\mathbf{W}\exp[\mathbf{W}\cdot(\sqrt{\hat{q}}\mathbf{z}+\mathbf{W}_0\hat{R})]. \tag{26}$$

Of course, R, $\hat{R}$, q, and $\hat{q}$ must be self-consistently determined from the free energy *saddle points.*

IV. BOOLEAN PERCEPTRON WITH ISING WEIGHTS

A. Results

We consider here an "Ising-weights perceptron" (IWP). In this case we have $d\mathbf{W}=\Pi_i dW_i[\delta(W_i-1)+\delta(W_i+1)]$ and we evaluate (26)

$$s=-\frac{1}{2}(1-q)\hat{q}-R\hat{R}+\int Dz\ \ln 2\cosh[(\sqrt{\hat{q}}z+\hat{R})]. \tag{27}$$

Extremalization of the free energy (24) with respect to the parameters R, $\hat{R}$, q and $\hat{q}$ and eliminating $\hat{R}$ and $\hat{q}$, we obtain the pertinent saddle point equations

$$R=\int Dz\ \tanh\left(\sqrt{\frac{\beta^2\alpha}{2\pi}\frac{1}{\sqrt{1-q^2}}}z+\frac{\alpha\beta}{\pi(1+\rho)}\frac{1}{\sqrt{1-R^2}}\right),$$

$$q=\int Dz\ \tanh^2\left(\sqrt{\frac{\beta^2\alpha}{2\pi}}\frac{1}{\sqrt{1-q^2}}z+\frac{\alpha\beta}{\pi(1+\rho)}\frac{1}{\sqrt{1-R^2}}\right). \tag{28}$$

These equations describe a first-order, spinodal phase transition from a state with poor generalization to a state with $R=1$ when $\alpha=\alpha_{sp}$, the critical α value. This state corresponds to a perfect generalization phase and is reached even if ρ constitutes an appreciable proportion of the training examples. Figure 1 depicts the phase diagram for different noise levels $\rho=0$, 0.1, 0.3, and 0.5. It is seen that as ρ augments so does α_{sp}. The thermodynamic transition curve appreciably changes and α_{th} is considerably smaller, so that the metastable state of poor generalization is quite poor indeed. Anomalies in the phase diagram arise at low temperatures ($T=0.5$) when $\rho=0$. This is an approximation effect that increases as the noise augments.

The training error is given by (25). It does not vanish even if the system has undergone a phase transition. This happens because no $\mathbf{W}$ exists that solves (2). If we regard the generalization error as a restricted average over PT question-answer pairs, then the pertinent error is given by [5] $\epsilon_g=(1/\pi)\arccos(R)$. Figure 2 displays learning curves for different ρ values ($T=1$).

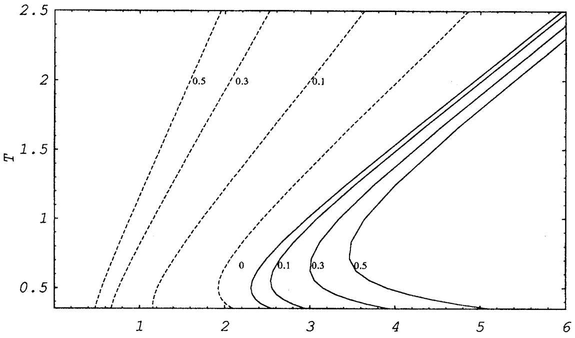

FIG. 1. Phase diagram obtained with a second order (in β) approximation for different values of the noise parameter ρ. The full line corresponds to the spinodal transition and the dashed line to the thermodynamic one.

Some features of our approach deserve particular mention. If we set $\rho=0$ in (25) and (28) we obtain a second-order approximation to an IWP trained *without noise* . In Fig. 2, the spinodal transition takes place at $\alpha_{sp}=2.95$, which agrees with the more elaborate complete quenched theory [5]. Our results, in addition, considerably improve upon the first-order ones, for which $\alpha_{sp}=2.08$. Our training errors ϵ_t and ϵ_g differ by an amount proportional to β [see Eq. (25)]. In the limit $\beta\rightarrow 0$, we recover the high temperature results, with the new and interesting relationship $q=R^2$ (which reads like a mean field recipe) as a bonus. This relationship cannot appear in the first-order treatment, which says nothing concerning q. The physical interpretation of the order parameter q is similar to that given to the Edwards-Anderson parameter in spin glasses [11,12]. It characterizes a typical overlap between two solutions to the constraints posed by (2). As α augments, more and more correlations are to be found among the different solutions, and q approaches unity. For $\alpha=\alpha_{sp}$, we have $q=1$ and the concomitant degeneration is broken. Figure 3 displays the behavior of both R and q as α varies.

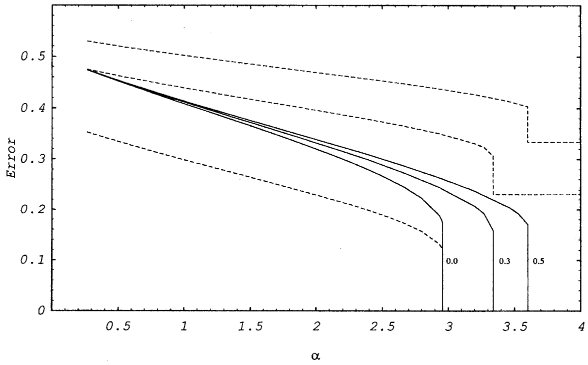

FIG. 2. For different noise levels the $(T=1)$ learning curves of our second order approach are drawn. Generalization errors (arbitrary units) are those of the full line, training ones (arbitrary units) correspond to the dashed line.

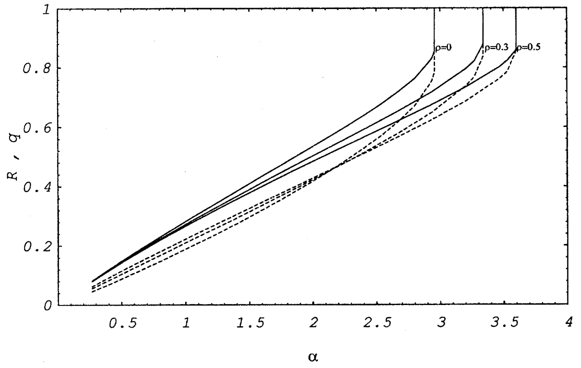

FIG. 3. Behavior of the order parameters as the α value changes (for $T=1$).

B. Spin-glass phase and symmetry breaking

The Hamiltonian (19) is invariant under permutations of the replica indexes. The replica symmetry approach (23) could then be expected to provide for a reasonable description. A spontaneous symmetry breaking could take place at low temperatures, however (the signature of a spin-glass phase), so that a complicated dependence of the replica indexes upon the parameter $Q_{\gamma\delta}$ would ensue. The spin-glass phase is characterized by a highly degenerate ground state which results as the consequence of a strong degree of frustration. The concomitant states are located in disconnected regions of configurations space, separated by barriers whose height diverges with N. As a consequence, the dynamical evolution of the $\mathbf{W}$ acquires a quite complex character and ergodicity is broken. An abnormally slow learning results.

The symmetry breaking can be dealt with Parisi's *one-step* approximation. The matrix $Q_{\gamma\delta}$ adopts an $m\times m$ block structure, obtained by a partition of the n replicas into n/m groups of m replicas each. The matrix elements $Q_{\gamma\delta}$ adopt the value q_1 if γ and δ belong to the same pure state, and the value q_0 otherwise. In the limit $n\rightarrow 0$, m is restricted to the $0\leqslant m\leqslant 1$ range. Since the spin-glass phase is to be found in the rather low temperature region ($T\leqslant 0.3$), going to second order is not enough for properly describing it. In the limit $q_1\rightarrow 1$, $\hat{q}_1\rightarrow\infty$ (β finite), however, the free energy can be cast into the form

$$f_{RSB}(q_0,\hat{q}_0,R,\hat{R},m,\beta)=\frac{1}{m}f_{RS}(q_0,m^2\hat{q}_0,R,m\hat{R},m\beta), \quad (29)$$

and the mathematical structure is seen to resemble that described in [5,16].

V. CONCLUSIONS

We conclude that even with a "bad" PT, perfect generalization can be attained. The neural network is still able to learn the rule underlying the "good" examples. In cases of architectural mismatch, on the other hand, rules cannot be correctly learned.

The bad news is that the metastable state of poor generalization is much poorer for our second-order approach than in the learning without noise. In any case, we hope to have convinced the reader that perturbative techniques, originally invented for dealing with problems of celestial mechanics and extensively used in many areas [19], are satisfactory *approximate* tools for investigating, at not too low temperatures, the thermodynamics of the learning process.

ACKNOWLEDGMENTS

The authors wish to thank Dr. R. Rossignoli for helpful conversations. The work of L.D. was supported by the Buenos Aires Scientific Commission (CICPBA) and that of A.P. by the Argentine Research Council (CONICET).

APPENDIX

We undertake here the calculation of the correlations of Eq. (18) (see [2]). We recast the first of them in the form

$$C_{1\gamma\delta}=\frac{1}{4}\int d\mathbf{S}\, D(\mathbf{S})[g(\mathbf{W}^{\sigma}\cdot\mathbf{S})-g(\mathbf{W}_o\cdot\mathbf{S})]^2 \times[g(\mathbf{W}^{\rho}\cdot\mathbf{S})-g(\mathbf{W}_o\cdot\mathbf{S})]^2, \quad (\mathrm{A1})$$

i.e.,

$$C_{1\gamma\delta}=\frac{1}{4}\int d\mathbf{S}\int d\mathbf{r}\;\delta(x-N^{-1/2}\mathbf{W}_{\gamma}\cdot\mathbf{S})$$
$$\times\,\delta(y-N^{-1/2}\mathbf{W}_{\sigma}\cdot\mathbf{S})\,\delta(z-N^{-1/2}\mathbf{W}_{o}\cdot\mathbf{S}) \quad (\mathrm{A2})$$
$$\times[g(x)-g(z)]^2[g(y)-g(z)]^2. \quad (\mathrm{A3})$$

By recourse to the representation

$$\delta(x)=(1/2\pi)\int dx'\exp(ixx')$$

of the δ function and remembering that $D\mathbf{S}=\Pi_i^N(dS_i/2\pi)\exp(-S_i^2/2)$, the integration process over $d\mathbf{S}$ leads to the (intermediate) result

$$\exp\left[-\left(\frac{1}{2}\mathbf{r}'\cdot\mathbf{r}'+x'z'R^{\sigma}+y'z'R^{\rho}+x'y'Q^{\sigma\rho}\right)\right], \quad (\mathrm{A4})$$

where $R^{\sigma}=(1/N)\mathbf{W}_{\gamma}\cdot\mathbf{W}_o y Q_{\gamma\sigma}=(1/N)\mathbf{W}_{\gamma}\cdot\mathbf{W}_{\sigma}$. For a Boolean perceptron we have $g(x)=\mathrm{sgn}(x)$, so that integration over the variables $\mathbf{r}$ and $\mathbf{r}'$ leads to

$$C_{1\gamma\delta}=\frac{1}{2\pi}\left\{\left[\frac{\pi}{2}+\arctan\left(\frac{Q_{\gamma\delta}}{\sqrt{1-Q_{\gamma\delta}^2}}\right)\right] -\left(\frac{\pi}{2}-\arcsin[1-(R^{\gamma})^2-(R^{\delta})^2]\right)\right\}. \quad (\mathrm{A5})$$

Taking for granted the replica symmetries $R_{\gamma}=R$ and

$$Q_{\gamma\sigma}=\begin{cases}1 & \sigma=\rho\\ q & \sigma\neq\rho\end{cases},$$

we find for the n diagonal terms, on the one hand,

$$\frac{n}{2}-\frac{n}{2\pi}\left(\frac{\pi}{2}-\arcsin(1-2R^2)\right),$$

and, for the n^2-n terms, on the other one (terms of second order in n neglected),

$$\frac{n}{2\pi}\left\{\left[\frac{\pi}{2}+\arctan\left(\frac{q}{\sqrt{1-q^2}}\right)\right]-\left(\frac{\pi}{2}-\arcsin(1-2R^2)\right)\right\},$$

so that

$$\sum_{\gamma\delta}^{n} C_{1\gamma\delta}=\frac{n}{2\pi}\left[\frac{\pi}{2}-\arctan\left(\frac{q}{\sqrt{1-q^2}}\right)\right]. \tag{A6}$$

$C_{2\gamma\delta}$ is evaluated in an entirely similar fashion and with the same result. Second order contributions in β are of identical form, both for "good" and for "bad" examples. They are weighted by the quantity of examples of each kind.

[1] E. Gardner, J. Phys A. **21**, 257 (1988).

[2] E. Gardner and B. Derrida, J. Phys A **21**, 271 (1988).

[3] D.E. Rumelhart, and J.L. McClelland, *Parallel Distributed Processing* (MIT, Cambridge, 1986).

[4] F. Rosemblatt, *Principles of Neurodynamics* (Spartan, New York, 1962).

[5] H.S. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. A **45**, 6056 (1992); H. Sompolinsky, N. Tishby, and H.S. Seung, Phys. Rev. Lett. **65**, 1683 (1990).

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

[7] N. Tishby, E. Levin, and S. Solla, in *Proceedings of the Internatinal Joint Conference on Neural Networks* (IEEE, New York, 1989), Vol. 2, pp. 403–409.

[8] E. Levin, N. Tishby, and S. Solla, Proc. IEEE **78** , 1568 (1990).

[9] J.A. Hertz, in *Statistical Mechanics of Neural Networks: Proceedings of the Eleventh Sitges Conference*, edited by L. Garrido (Springer, Berlin, 1990).

[10] J. Schrager, T. Hogg, and B.A. Hubermann, Science **242**, 414 (1988).

[11] S.F. Edwards and P.W. Anderson, J. Phys. F **5**, 965 (1980).

[12] M. Mezard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).

[13] G. Parisi, J. Phys. A **13**, 1101 (1980).

[14] D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett. **35**, 1792 (1975).

[15] S. Kirkpatrick and D. Sherrington, Phys. Rev. B **17**, 4384 (1978).

[16] W. Krauth and M. Mezard, J. Phys. (Paris) **50**, 3057 (1989).

[17] P. del Giudice, S. Franz, and M. A. Virasoro, J. Phys. (Paris) **50**, 121 (1989).

[18] G. Gyorgyi and N. Tishby, *Neural Networks and Spin Glasses*, edited by W. K. Theumann and R. Koberle (World Scientific, Singapore, 1990).

[19] C. H. Cohen-Tannoudji, B. Diu, and F. Laloe, *Quantum Mechanics* (Wiley, New York, 1977).