Variational studies and replica symmetry breaking in the generalization problem of the binary perceptron

Evaldo Botelho,[1] Cristiano R. Mattos,[1,2] and Nestor Caticha[1]
[1]Instituto de Física da Universidade de São Paulo, Caixa Postal 66318, São Paulo, SP 05315-970, Brazil
[2]Faculdade de Engenharia, Universidade Estadual Paulista, Caixa Postal 205, Guaratinguetá, SP 12500-000, Brazil

(Received 29 May 2000)

We analyze the average performance of a general class of learning algorithms for the nondeterministic polynomial time complete problem of rule extraction by a binary perceptron. The examples are generated by a rule implemented by a teacher network of similar architecture. A variational approach is used in trying to identify the potential energy that leads to the largest generalization in the thermodynamic limit. We restrict our search to algorithms that always satisfy the binary constraints. A replica symmetric ansatz leads to a learning algorithm which presents a phase transition in violation of an information theoretical bound. Stability analysis shows that this is due to a failure of the replica symmetric ansatz and the first step of replica symmetry breaking (RSB) is studied. The variational method does not determine a unique potential but it allows construction of a class with a unique minimum within each first order valley. Members of this class improve on the performance of Gibbs algorithm but fail to reach the Bayesian limit in the low generalization phase. They even fail to reach the performance of the best binary, an optimal clipping of the barycenter of version space. We find a trade-off between a good low performance and early onset of perfect generalization. Although the RSB may be locally stable we discuss the possibility that it fails to be the correct saddle point globally.

PACS number(s): 87.10.+e, 84.35.+i, 89.70.+c, 05.50.+q

I. INTRODUCTION

One topic of interest in the study of neural networks with the tools of statistical mechanics (SM) is that the process of information extraction from data can be modeled by a dynamical process of minimization of an energy function in the presence of noise. The use of techniques borrowed from the study of equilibrium disordered systems [1], such as, for example, replica methods, Thouless, Anderson, and Palmer equations, cavity analysis, and Monte Carlo simulations, as well as dynamical analysis techniques, has permitted a considerable understanding of typical properties of the thermodynamic limit (TL) of such systems as attractor and feedforward neural networks. In this paper we are interested in the equilibrium properties of the student-teacher problem [2–4] of rule extraction by a *binary Boolean perceptron*. This means that the weights as well as the output are restricted to be ± 1. We look at the generalization ability for the case of a realizable rule represented by a teacher of the same architecture as the student and in particular we concentrate on the determination of thermodynamic limit bounds and how to get them by the construction of an appropriate potential.

The binary perceptron has been attacked from many fronts. Studies of the computational complexity by Pitt and Valiant have shown that it belongs to the NP (nondeterministic polynomial time) complete class [5]. From a SM point of view it has been studied before by Gardner and Derrida [6] and Györgyi [7]. Their aim was to study the learning curve (generalization error) as a function of the number of examples where the training energy was chosen in the simplest way, i.e., error counting. For independent examples drawn from a uniform distribution, the main characteristic of this system in the TL is the first order transition to perfect generalization at $\alpha_c \simeq 1.25$, where as usual $\alpha = P/N$ measures the number of examples P in units of the number of inputs N. This shows the power of statistical mechanics methods, since the transition cannot be detected by, e.g., Vapnik-Chervonenkis analysis, which aims at a general type of result, such as those independent of the distribution of examples, and can therefore miss important features that appear only for particular but important cases. For a specific comment on this, see [8]

A relevant question, complementary in spirit to examination of the thermal equilibrium properties, delves into the dynamical aspects of actually determining the set of weights that minimize the training error. The studies of Horner [9] have shown that times exponential in N are required for learning. This is in agreement with the expectation that, since this is a computationally hard problem, it should have spin-glass properties. Polynomial time algorithms, such as simulated annealing of the error-counting energy, with a linear decrease of the temperature schedule, fail to reach global solutions.

Different approximations have been devised in order to overcome the problems associated with glassy dynamics. Based on the fact that learning the equivalent real-weights problem is not as slow, other studies have dealt with continuous approximations to the binary problem. Several groups looked into clipping strategies [10–13] and other transformations of the continuous perceptron, in particular one that optimally incorporates the information that the teacher weight vector points in the direction of a vertex of the unit N-dimensional hypercube [14]. Penney and Sherrington [15] have looked into how to reduce the effective dimension of the problem by clipping a partial set of real couplings and posterior learning of the remaining binary weights. Copelli *et al.* [16–18] have looked at a related problem in the unsupervised learning scenario. They employed their methods for the teacher-student case also. By extending an argument of

Watkin for the analogous problem with continuous weights, they looked into the generalization properties of the center of mass of the version space. Its performance should saturate the Bayes limit, but it is not itself a vertex on the hypercube. They also considered the Bayesian best binary (BB) vector by clipping the real component Bayes vector for both supervised and unsupervised learning.

On-line learning—a possibly efficient strategy to overcome slow dynamics since it makes no attempt to thermalize—has been shown to be ineffective if working in the binary coupling space directly [19]. Nevertheless, progress has been made in the direction of on-line learning. Solla and Winther [20] have shown how to incorporate, in the on-line Bayesian spirit suggested by Opper [21], information about the binary nature of the couplings. Their method leads to fast learning times but is not able to condense on the exact solution in finite times.

In this work we extend the work of Kinouchi and Caticha [22] for the off-line variational determination of a training energy. We look into the extension of the variational approach to the determination of training potentials, optimal—in the generalization sense—for the binary perceptron. A difference from the approach of Copelli *et al.* [16–18] is that our construction methods never leave the binary space. Although this in the end might turn out not to be a good strategy, it is theoretically interesting to see how far one can go on such a discrete problem without leaving the allowed space. In the case of random examples, [13] presents a variational calculation for potentials that, using the maximum stability perceptron as a teacher, aims at determining the continuous vector that on clipping determines the largest number of weights of the maximum stability binary perceptron.

The extension of the variational program to the binary case is not at all straightforward due to the failure of the replica symmetry (RS) ansatz. We studied first a replica symmetric variational calculation. The transition to perfect learning occurs at a much smaller value of α_c ($\simeq 0.5$). This would have been great news if not for the fact that a simple information theoretical bound for perfect generalization is $\alpha_c \geqslant 1$ [23]. Each example cannot carry more than just one bit of information and at least N examples are needed to determine the N independent bits encoded in the weight vector. The stability analysis showed the surprising failure of the RS ansatz from the very start at $\alpha=0$. The next step we report here is to perform a variational calculation at the level of a one-step replica symmetry breaking (RSB-1). The variational method does not determine a unique potential, but only fixes expected values of the modulation function. We study a restricted set of solutions, which form a one-parameter family. Our choice is based on physical similarities of this family to potentials already studied. This is interesting in itself as it indicates the possibility that potentials other than the type we study may be relevant. The phase diagrams are studied and discussed. The learning curves associated with this family show a trade-off between better generalization and earlier transitions to the perfect generalization phase.

This paper is organized as follows. In Sec. II A we present the general replica calculation results. In Secs. II B and II C the variational procedures under RS and RSB-1, respectively, are presented. Section III discusses the analysis of the RSB-1 free energy. This requires the determination of the effective potential, which is explicitly calculated in the Appendix. Section IV discusses results and presents conclusions.

II. OFF-LINE VARIATIONAL METHOD

A. General results

The statistical mechanics approach to determining the generalization ability of a network learning from examples a rule that itself is implemented by another network has been studied in several cases; for reviews, see [2–4]. Call the N-dimensional binary weight vectors of the teacher and student networks $\mathbf{B}$ and $\mathbf{J}$, respectively.

Off-line learning describes the infinite time limit of a learning algorithm that proceeds by minimization, in the presence of thermal noise, of an energy function $\sum_{\mu=1}^{P} V(\mathbf{J}\cdot\mathbf{S}_\mu,\sigma_\mu)$ that depends on quenched data represented by the P input-output example pairs $(\mathbf{S}_\mu,\sigma_\mu)$. The free energy for the replicated system presented with P examples, independently chosen from a uniform distribution, is given by

$$-\beta f = \mathrm{extr}_{q_{ab},\hat{q}_{ab},R_a,\hat{R}_a}[G_o\{q_{ab},\hat{q}_{ab},R_a,\hat{R}_a\} - \alpha G_r\{q_{ab},R_a\}], \tag{1}$$

where

$$G_o = \lim_{n\to 0}\frac{1}{n}\Bigg[-\sum_{a<b} q_{ab}\hat{q}_{ab} - \sum_a R_a\hat{R}_a + \frac{1}{N}\sum_{j=1}^{N}\ln\int\prod_a d\chi(J_j^a) \times\exp\Bigg(\sum_a \hat{R}_a B_j J_j^a + \frac{1}{2}\sum_{a\neq b}\hat{q}_{ab}J_j^a J_j^b\Bigg)\Bigg], \tag{2}$$

$$G_r = \lim_{n\to 0}\frac{1}{n}\ln\int Dt_2\int\left(\prod_a\frac{d\lambda_a d\hat{\lambda}_a}{2\pi}\right) \times\exp\left(-\beta\sum_a V(\lambda_a,t_2)\right) \times\exp\left(i\sum_a\hat{\lambda}_a(\lambda_a - R_a t_2) - \frac{1}{2}\sum_{a,b}(q_{ab}-R_aR_b)\hat{\lambda}_a\hat{\lambda}_b\right). \tag{3}$$

The interpretation of R and q at the extreme justifies the choice of method to study this problem. This is so since $R = \langle\mathbf{J}\cdot\mathbf{B}/N\rangle$, where the average is both thermal and over the disorder, is directly related to the generalization error e_G in the case of equivalent replicas. For the uniform distribution, $e_G = \arccos(R)/\pi$. The usual order parameters $q_{ab} = \langle\mathbf{J}_a\cdot\mathbf{J}_b/N\rangle$ describe the typical overlap between two students learning from the same data. The weight measure $d\chi(J_j^a)$ defines the particular problem. In this case $d\chi(J_j^a)$ is the counting measure over $J_j^a = \pm 1$.

Two related questions are discussed in this paper: What is the largest value that the overlap R can reach for fixed α? That is, what are the upper bounds for the generalization ability? Note that this is different in spirit from a Bayesian approach since here we ask for best performances within the constraints of a given architecture. This immediately leads to the second question: What energy function V should be chosen in order to achieve those bounds? The same questions have been answered satisfactorily for this machine with real weights; the first by Opper and Haussler [24] and the second by Kinouchi and Caticha [22]. In trying to answer them we have to go beyond the permutation replica symmetry that holds in that case, which we now analyze.

B. Replica symmetric analysis

It is reasonable to begin by assuming that the solution to the extremization problem is replica symmetry. The zero temperature limit can be taken by noticing that an optimal potential will most likely have a unique minimum and therefore it is natural to take the $q \rightarrow 1$ limit [25]. A sensible limit can be obtained by requiring that $x \equiv \beta(1-q)$, $y \equiv \hat{q}/\beta^2$, and $\omega \equiv \hat{r}/\beta$ remain finite as $\beta \rightarrow \infty$. The free energy is then the extreme of

$$f_{RS} = \frac{1}{2}xy + \{R - [1-2H(\eta)]\}\omega - \sqrt{\frac{2}{\pi}y}e^{-\eta^2/2}$$
$$+2\alpha \int DtH\left(\frac{-tR}{\sqrt{1-R^2}}\right)\left(V(\lambda_o) + \frac{(\lambda_o - t)^2}{2x}\right),$$

where $\eta = \omega/\sqrt{y}, Dt = \exp(-t^2/2)dt/\sqrt{2\pi}$, $H(x) = \int_x^\infty Dt$, and $\lambda_o(t)$ is defined as the minimum of $V(\lambda) + (\lambda - t)^2/2x$:

$$-x\left(\frac{\partial V(\lambda)}{\partial \lambda}\right)_{\lambda_o} = \lambda_o - t. \tag{4}$$

The saddle point equations lead to

$$R = 1 - 2H(\eta), x = \sqrt{\frac{2}{\pi y}}e^{-\eta^2/2},$$

$$e^{-\eta^2} = \pi\alpha \int DtH\left(\frac{-tR}{\sqrt{1-R^2}}\right)(\lambda_o - t)^2,$$

$$\eta e^{-\eta^2/2} = \alpha \int \frac{dt}{\sqrt{2\pi(1-R^2)}}e^{-t^2/2(1-R^2)}\lambda_o. \tag{5}$$

Following [22], we obtain the result that, in order to maximize R, the learning potential should be chosen such that the RS modulation function $F_{RS}(t) \equiv \lambda_o - t$ is given by

$$F_{RS}(t) = C\frac{e^{-(Rt)^2/2(1-R^2)}}{H(-Rt/\sqrt{1-R^2})}, \tag{6}$$

which can be seen to be very similar to the optimal modulation function for on-line learning in the real-weights perceptron, a most interesting result from a cavity perspective [26,27]. It is easy to see that

$$C = \frac{e^{-\eta^2/2}}{\pi\eta\sqrt{1-R^2}} \tag{7}$$

is a function of R but not of t. From Eq. (6) we can determine the potential V, since $F_{RS} = -x\partial V/\partial\lambda_o$. Thus

$$V(\lambda) = \mathcal{E}(t) - \frac{1}{2x}F_{RS}^2(t), \tag{8}$$

where $\mathcal{E}(t) = -[\sqrt{2\pi(1-R^2)}C/Rx]\ln H(-Rt/\sqrt{1-R^2})$ is proportional to the on-line optimal energy function. It resembles the equivalent off-line potential for the real-weights perceptron. However, the post-training stability may not be all positive, i.e., λ_o can be negative. This means that this algorithm would be called inconsistent since after training there still may be memorization errors. The learning curves (see Fig. 2 below) are obtained by solving the saddle point equations numerically. By comparing the free energy of the phase with incomplete learning $R<1$ with that for $R=1$, a first order phase transition at $\alpha_c = 0.53$ is found. As mentioned earlier, this is unacceptable and the cause of this unphysical result is the replica symmetric ansatz. A stability analysis [28] confirms this suspicion and we now turn to a one-step replica symmetry breaking ansatz

C. RSB-1 variational method

We proceed in the by now standard way [1]. Let q_{ab} be an $l \times l$ block matrix and call the block indices $\mu, \nu = 1,2, \ldots, l = n/m$. For replicas within the same block, the overlaps are taken to have the same overlap value q_1, while replicas belonging to different blocks have overlap q_o. The replicas are assumed to be equivalent and thus $R_a = \rho$, for all replicas. We do not use the same letter R to make explicit that it will have a different value from the previous section. The conjugated parameters (with carets) are assumed to have the same structure.

The entropic and energetic contributions to the free energy are

$$G_o = \frac{1}{2}[mq_o\hat{q}_o + (1-m)q_1\hat{q}_1 - \hat{q}_1] - \rho\hat{\rho} + \frac{1}{m}\int Dz_o \ln$$
$$\times \int Dz_1[2\cosh(\sqrt{\hat{q}_o}z_o + \sqrt{\hat{q}_1 - \hat{q}_o}z_1 + \hat{\rho})]^m, \tag{9}$$

$$G_r = -\frac{2}{m}\int_{-\infty}^{\infty} Dt_1 \int_0^\infty Dt_2 \ln \int_{-\infty}^{\infty} Dt_o\left\{\int \frac{d\lambda}{\sqrt{2\pi(1-q_1)}}\right.$$
$$\left.\times \exp\left[-\beta\left(V(\lambda) + \frac{(\lambda - t)^2}{2x_1}\right)\right]\right\}^m, \tag{10}$$

where

$$t \equiv \sqrt{q_o - \rho^2}t_1 + \rho t_2 + \sqrt{q_1 - q_o}t_o.$$

We are interested in the zero temperature case. A sensible limit can be found by requiring that within a valley the optimal potential has a unique minimum and that different valleys' minima have the same overlap. To achieve this we make the ansatz $\beta\rightarrow\infty, m\rightarrow 0, q_1\rightarrow 1$, and $\hat{q}_o$ and $\hat{R}\rightarrow\infty$, such that $x_o\equiv\beta m$, $x_1\equiv\beta(1-q_1), \hat{Q}_o\equiv m^2\hat{q}_o$, and $\hat{R}\equiv m\hat{\rho}$ remain finite.

To take the limit [29,30] we start with the entropic part

$$G_o=\frac{1}{m}\left[-\frac{(1-q_o)\hat{Q}_o}{2}-\rho\hat{R}+\int Dz_o \ln 2\cosh(\sqrt{\hat{Q}_o}z_o+\hat{R})\right] \tag{11}$$

and the energy part

$$G_r=-\frac{2}{m}\int_{-\infty}^{\infty}Dt_1\int_0^{\infty}Dt_2 \ln\int_{-\infty}^{\infty}Dt_o\, e^{-\beta m[V(\lambda_o)+(\lambda_o-t)^2/2x_1]} \tag{12}$$

where the λ integral is done by a saddle method and λ_o is such that

$$\partial V/\partial\lambda+\frac{(\lambda-t)}{x_1}=0 \quad \text{for } \lambda=\lambda_o. \tag{13}$$

It follows that

$$\begin{aligned}-x_o f=&-\frac{(1-q_o)\hat{Q}_o}{2}-\rho\hat{R}+\int Dz_o \ln 2\cosh(\sqrt{\hat{Q}_o}z_o+\hat{R})\\ &+2\alpha\int_{-\infty}^{\infty}Dt_1\int_0^{\infty}Dt_2 \ln\int_{-\infty}^{\infty}Dt_o\\ &\times e^{-x_o[V(\lambda_o)+(\lambda_o-t)^2/2x_1]}.\end{aligned} \tag{14}$$

The order parameters of the incomplete generalization phase are obtained by solving the saddle point equations:

$$\frac{\partial f}{\partial\hat{R}}=0\Rightarrow\rho=\int Dz_o \tanh(\sqrt{\hat{Q}_0}z_o+\hat{R})\equiv\phi_1(\hat{Q}_o,\hat{R}), \tag{15}$$

$$\frac{\partial f}{\partial\hat{Q}_0}=0\Rightarrow q_o=\int Dz_o \tanh^2(\sqrt{\hat{Q}_0}z_o+\hat{R})\equiv\phi_2(\hat{Q}_o,\hat{R}), \tag{16}$$

$$\frac{\partial f}{\partial\rho}=0\Rightarrow\hat{R}=2\alpha\frac{x_o}{x_1}\int Du\,\frac{e^{-u^2/2\Gamma}}{\sqrt{2\pi(1-\gamma^2)}}\langle(\lambda_o-t)\rangle_{t_o}, \tag{17}$$

$$\frac{\partial f}{\partial q_o}=0\Rightarrow\hat{Q}_o=2\alpha\frac{x_{o_2}}{x_1}\int Du\, H\left(-\frac{u}{\sqrt{\Gamma}}\right)\langle\lambda_o-t\rangle_{t_o}^2, \tag{18}$$

where $\gamma\equiv\rho/\sqrt{q_o}$, $\Gamma\equiv(1-\gamma^2)/\gamma^2$, and

$$\langle(\cdots)\rangle_{t_o}\equiv\frac{\int Dt_o e^{-x_o[V(\lambda_o)+(\lambda_o-t)^2/2x_1]}(\cdots)}{\int Dt_o e^{-x_o[V(\lambda_o)+(\lambda_o-t)^2/2x_1]}}. \tag{19}$$

The orthogonal transformation $(t_1,t_2)\rightarrow(u,v)$ defined by $u=(\sqrt{q_o-\rho^2}\,t_1+\rho t_2)/\sqrt{q_o}$, $v=(-\rho t_1+\sqrt{q_o-\rho^2}\,t_2)/\sqrt{q_o}$, and $t=\sqrt{q_o}\,u+\sqrt{1-q_o}\,t_o$ was used to simplify Eq. (18).

Introducing an effective modulation function

$$F(u)=\langle\lambda_o-t\rangle_{t_o}, \tag{20}$$

the saddle point equations can be written as

$$\hat{R}=2\alpha\frac{x_o}{x_1}\int Du\,\frac{e^{-u^2/2\Gamma}}{\sqrt{2\pi(1-\gamma^2)}}F(u),$$

$$\hat{Q}_o=2\alpha\left(\frac{x_o}{x_1}\right)^2\int Du\, H\left(-\frac{u}{\sqrt{\Gamma}}\right)F^2(u).$$

We can now determine the $F(u)$ that maximizes, for a fixed α, the overlap ρ. This is somewhat harder than in the real-weights case, where the solution to the variational problem is obtained very easily, just by inspection plus knowledge of the equivalent solution for the on-line problem. Nevertheless, the solution is very similar. So we look at the variations of Eq. (15) with respect to $F(u)$ subject to the constraint (16). Let ξ be a Lagrange multiplier (see also [31]). $F(u)$ can be determined by maximizing

$$\begin{aligned}\Phi=&\,\phi_1(\hat{Q}_o[q_o,\rho,F],\hat{R}[q_o,\rho,F])\\ &-\xi\,\phi_2(\hat{Q}_o[q_o,\rho,F],\hat{R}[q_o,\rho,F]),\end{aligned} \tag{21}$$

which yields

$$\frac{\delta\hat{Q}_o}{\delta F}=-\frac{(\partial\phi_1/\partial\hat{R}-\xi\partial\phi_2/\partial\hat{R})}{(\partial\phi_1/\partial\hat{Q}_o-\xi\partial\phi_2/\partial\hat{Q}_o)}\frac{\delta\hat{R}}{\delta F}\equiv k\frac{\delta\hat{R}}{\delta F}, \tag{22}$$

and by using the explicit form of $\hat{R}$ and $\hat{Q}_o$ and the definitions (15) and (16):

$$\int Du\, H\left(-\frac{u}{\sqrt{\Gamma}}\right)F(u)=\frac{x_1}{x_o}\frac{k}{2}\int Du\,\frac{e^{-u^2/2\Gamma}}{\sqrt{2\pi(1-\gamma^2)}}. \tag{23}$$

This does not determine $F(u)$ uniquely. A possible solution is

$$F(u)\equiv\langle\lambda_o-t\rangle_{t_o}=\frac{1}{2}\frac{x_1}{x_o}\frac{k}{\sqrt{2\pi(1-\gamma^2)}}\frac{e^{-u^2/2\Gamma}}{H(-u/\sqrt{\Gamma})}. \tag{24}$$

Any odd function of u could be added to the Gaussian in the numerator, but we choose to look at this form since as it stands it is proportional to the optimal modulation function that appears in off-line and on-line optimization of the real-weights perceptron and we see no physical motivation for other terms. This choice (24) leads to the saddle point equations

$$\hat{R}=\sqrt{\frac{2}{\pi}}\alpha\kappa I_o(\gamma), \tag{25}$$

$$\hat{Q}_o=2\alpha\kappa^2\sqrt{1-\gamma^2}I_o(\gamma), \tag{26}$$

where $\kappa\equiv\frac{1}{2}k/\sqrt{2\pi(1-\gamma^2)}$, $I_o(\gamma)\equiv\int Dv\, g(-\gamma v), \gamma\equiv\rho/\sqrt{q_o}$, and $g(x)\equiv e^{-x^2/2}/H(x)$. We can still choose κ, and

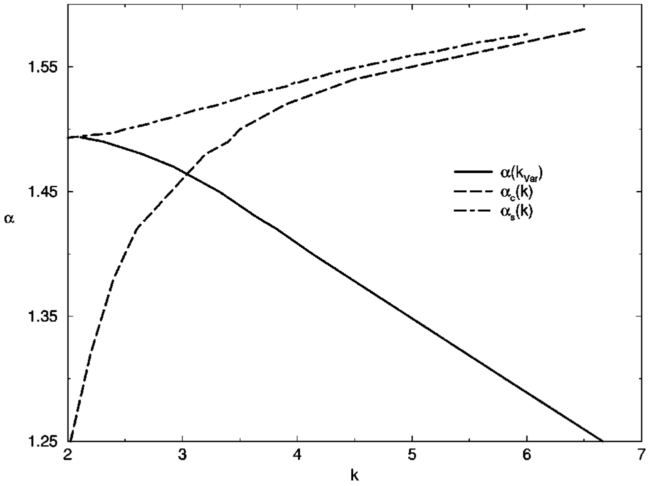

FIG. 1. The value of α as a function of hyperparameter k at which the low generalization phase performance is optimal (thick line). For each fixed k, we show also the value of α_s, the spinodal point (short-long dash), where the low generalization phase ceases to exist, and of α_c, the first order transition point (long dash), above which the perfect generalization phase has the lowest free energy.

we do so in order to pick the largest possible ρ for fixed α. (Fig. 1 shows α as a function of k_{var}, the value of k that leads to the smallest generalization error.) Note that this liberty is due to the fact that the value of q_o is not constrained; only its form is constrained by Eq. (16). The learning curves of the low generalization phase can be determined numerically [see Figs. 2(a) and 2(b)]. In Fig. 3 the overlap q_o is shown as a function of α for the best choice of the potential, i.e., using k_{var}. The determination of the thermodynamic phase, that is, the location of the first order transition, needs analysis of the free energy in both high and low generalization phases. While the free energy of the low generalization phase can be determined without explicit knowledge of the potential, that of the high generalization phase will require detailed knowledge of the potential. In the next section we show how this can be done.

III. THE FREE ENERGY AND THE POTENTIAL

A. The low generalization phase

To complete the analysis of the learning curve we must look into the behavior of the free energy. This is not straightforward since the form of the potential has not yet been determined, but only the expected value $\partial V/\partial \lambda$. It is quite interesting, as we now show, that the precise form of the potential is not needed to determine the free energy or the learning curve in this phase; just knowledge of the effective modulation function suffices, and even this is the same for any solution of Eq. (23).

The energy contribution to the free energy can be written as

$$-x_o f_1 = 2\alpha \int DuH\left(\frac{-u}{\sqrt{\Gamma}}\right) \ln \int Dt_o \times \exp\left[-x_o\left(V(\lambda_o) + \frac{(\lambda_o - t)^2}{2x_1}\right)\right],$$

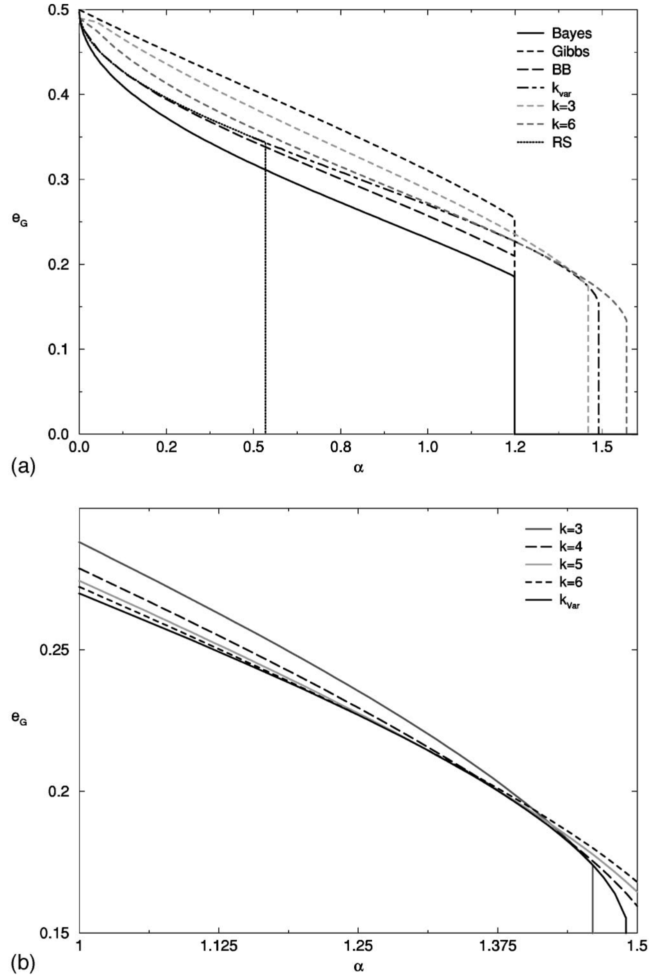

FIG. 2. (a) Learning curves. Generalization errors as functions of α for different algorithms. The Bayes (continuous line) and Gibbs (top short dashed curve) algorithm learning curves are included for comparison. Note that the Bayes algorithm is obtained by a student outside the binary space. BB is the best binary vector [18] obtained as the result of clipping the center of mass of the version space (Bayes). The learning curve (dots) obtained under the hypothesis of replica symmetry has a transition to perfect learning at the impossible value of $\alpha = 0.53$. The algorithm obtained with $k = 2$ gives the Gibbs result. For $k = 3$ and $k = 6$ the curves show a trade-off between earlier performance and later transition. For variable $k_{var}(\alpha)$ (Fig. 1) the best performance for potential based learning is obtained. See (b) for details. For small α the unphysical RS, the RSB-1 with k_{var}, and the BB curves are numerically indistinguishable. (b) Details of the learning curves. The black continuous line is obtained for $k_{var}(\alpha)$ (Fig. 1); it is the envelope, within the low generalization phase, of the family of algorithms obtained for $k \geqslant 2$.

provided the order parameters are understood as the solution of the saddle point equations. Integrating by parts, using the definitions of λ_o and $F(u)$,

$$-x_o f_1 = 2\alpha \frac{x_o}{x_1} \sqrt{q_o} \int_{-\infty}^{\infty} du F(u) \int_{-\infty}^{u} Dy H\left(\frac{-y}{\sqrt{\Gamma}}\right),$$

substituting Eq. (24) for $F(u)$, and integrating by parts again,

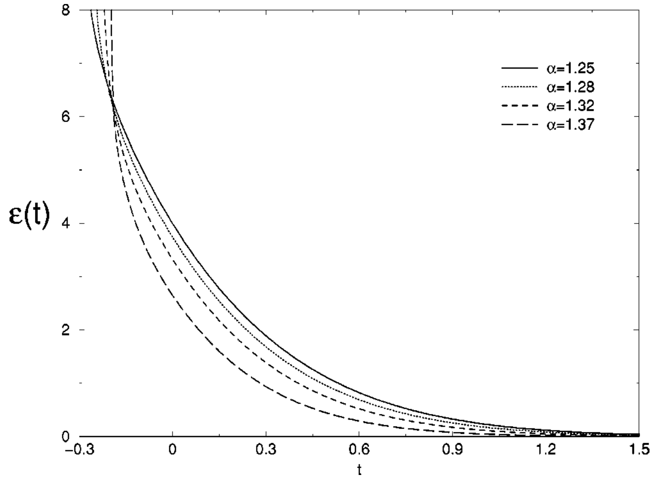

FIG. 3. q_o and ρ as a function of α. Note that q_o is different from 1 as soon as $\alpha\neq 0$ but is so close that the learning curves under the RS and RSB-1 methods [see Fig. 2(a)] are not very different for small α.

$$-x_o f_1 = 2\alpha\kappa\sqrt{2\pi q_o\Gamma}\int_{-\infty}^{\infty} Du H\left(\frac{-u}{\sqrt{\Gamma}}\right)\ln H\left(\frac{-u}{\sqrt{\Gamma}}\right) \tag{27}$$

is obtained.

B. The high generalization phase

We now look at the value of the free energy in the extreme of the allowed interval for the order parameters. The determination of the potential cannot be postponed since the free energy depends on it explicitly.

Equation (23), which results from the variational prescription, can be transformed into an integral equation for the effective potential $\mathcal{E}(t)=V(\lambda_o)+(\lambda_o-t)^2/2x_1$:

$$\int Dt_o e^{-x_o\mathcal{E}(t)} = \left[H\left(-\frac{u}{\sqrt{\Gamma}}\right)\right]^b, \tag{28}$$

where $t=\sqrt{q_o}u+\sqrt{1-q_o}t_o$ and

$$b=\sqrt{\frac{k^2 q_o}{4\gamma^2}}. \tag{29}$$

An expression for the potential, obtained in the Appendix, is

$$\mathcal{E}(t) = -\frac{1}{x_o}\ln\int Dx\left[H\left(-\frac{t}{\sqrt{\Gamma q_o}}-ix\sqrt{\frac{(1-q_o)}{\Gamma q_o}}\right)\right]^b, \tag{30}$$

where the effect of the one-step RSB is seen to be the introduction of a noiselike term in the stability *t*, which depends on the existence of the other valleys. Its influence goes to zero as $q_o\rightarrow 1$.

Despite its appearance, this expression is real, as can be seen by defining

$$\mathcal{C}(x,t) = \int DK\Theta\left(K+\frac{t}{\sqrt{\Gamma q_o}}\right)\cos\left(xK\sqrt{\frac{(1-q_o)}{\Gamma q_o}}\right), \tag{31}$$

$$\mathcal{S}(x,t) = \int DK\Theta\left(K+\frac{t}{\sqrt{\Gamma q_o}}\right)\sin\left(xK\sqrt{\frac{(1-q_o)}{\Gamma q_o}}\right). \tag{32}$$

Finally we obtain

$$\mathcal{E}(t) = -\frac{1}{x_o}\ln\int\frac{dx}{\sqrt{2\pi}}\exp-\left[\frac{1}{2}\left(1-b\frac{1-q_o}{\Gamma q_o}\right)x^2\right] \times[\mathcal{C}^2(x,t)+\mathcal{S}^2(x,t)]^{b/2}\cos\left(b\tan^{-1}\frac{\mathcal{S}(x,t)}{\mathcal{C}(x,t)}\right), \tag{33}$$

which can be used for numerical evaluations. It is easy to verify that the solution is real. At this point notice the structure of the potential that emerges from the calculation. If $q_o=1$, then we are back to the replica symmetric calculation; the *x* integral decouples and we are left with $-x_o\mathcal{E}(t)=b\ln H(-t/\sqrt{\Gamma})$, which is very similar to the optimal potentials that have been found for the real-weights perceptron, both on line and off line, and for the binary perceptron without RSB. However RSB introduces a new noiselike element.

One of the most striking features of this solution is that, like other variationally determined potentials, it depends on the values of the order parameters. This has been discussed elsewhere [32] and can be interpreted as the inclusion of the correct annealing along the learning process, generalizing the annealing of the learning rate that has been studied in on-line learning algorithms (e.g., [33,34,26]). These order parameters have the role of hyperparameters and will have the value determined by the solution of the saddle point equations. The fact that these values may differ from those that are determined by the thermodynamics is of fundamental importance. The learning process thus proceeds in the following way. Determine self-consistently the potential (functional form and hyperparameters) that will lead to maximum performance in the low generalization phase in such a manner that the respective values of the order parameters are equal to the hyperparameters. Then minimize the potential by some dynamical process, letting the temperature go to zero. We do not worry here about thermalization times, as these might diverge for a NP problem. The final result may land on the perfect generalization phase and therefore the order parameters will not have the same values as the hyperparameters. We denote the hyperparameters by starred quantities. Then we look at the limit $q_o,q_1\rightarrow 1,\beta\rightarrow\infty,m\rightarrow 0$ of Eq. (14) with $\mathcal{E}^*(t)$ calculated at the starred (saddle point equation) values, which gives

$$f=2\alpha\int_0^{\infty} Dt\mathcal{E}^*(t). \tag{34}$$

C. Learning curves

Equations (15), (16), (25), and (26) are used to build the low generalization learning curves. From the comparison of

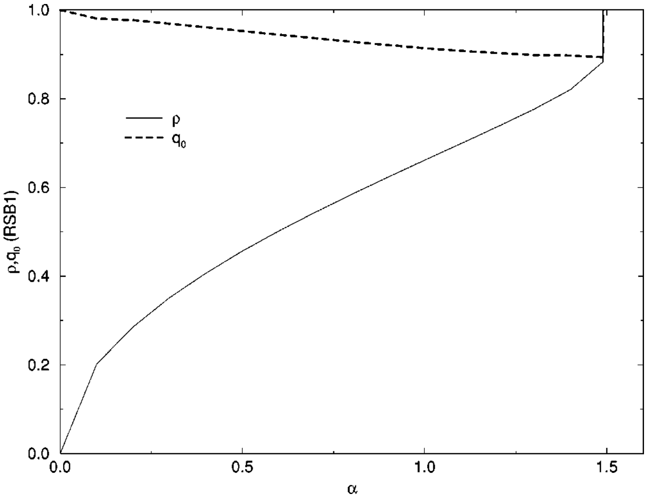

FIG. 4. The effective potential $\mathcal{E}(t)$ as a function of the stability t showing the annealing for different values of α.

the free energies, Eqs. (27) and (34), the phase transition is located. These equations have to be complemented with a choice of k. We can just look at the numerical value of k that leads to the smallest generalization error for fixed α, which we called $k_{var}(\alpha)$ (Fig. 1). We look also at the results obtained for fixed k. A Bayesian statistician will not be surprised that the Bayes bounds are not beaten. Neither is the (low phase) generalization error improved, nor the onset of the high generalization phase anticipated. These equations just lead in general to smaller errors than the Gibbs algorithm. However for small α the RS, the RSB-1 $[k_{var}(\alpha)]$, and the Bayesian best binary of Copelli *et al.* [18] are very close. At this point this agreement is only numerical, but it is possible that these algorithms have the exact same optimal (Bayesian) performance in the limit $\alpha \rightarrow 0$, which is similar to the results of [18] for unsupervised learning.

Figure 4 shows the potential [Eq. 33] for different values of b. At $b=1$ the potential turns into the error-counting potential, which gives the Gibbs performance. Then replica symmetry is restored. Below the value $b=1$, the potential cannot be determined.

IV. DISCUSSION AND CONCLUSIONS

The learning curves shown in Fig. 2 show that the potentials obtained variationally fail to reach the Bayes bound. This is in contrast to the continuous weight perceptron, where the Bayes limit is obtained by a network with the same architecture as the teacher. As shown by Copelli *et al.* [18] the Bayes algorithm is equivalent to a network with a weight vector given by the center of mass of the version space, which is not itself a binary vector. It follows that no method constrained to the hypercube will reach the Bayes limit. A similar failure to reach the Bayes limit was also reported by Winther *et al.* [35] for multilayer networks, where again the Bayes algorithm cannot be matched within the space of students with the same architecture as the teacher.

The variational method probes potentials from within a restricted class and it is therefore natural not to expect to find Bayesian performances if the Bayes algorithm does not belong to it. The performance of Bayesian inference restricted to binary vectors is saturated only for small α, but then even the simple Hebb algorithm has optimal generalization in this regime.

Within the low generalization phase the variational method is able to identify a class of potentials that lead to better performance than the Gibbs algorithm. There is a trade-off within the class, as can be seen in Figs. 2(a) and 2(b), between earlier transition to perfect generalization and better performance. A potential that leads to better generalization will have a delayed transition.

The potentials of this class do not work by imposing a zero memorization error, not even by minimization of the total error count. They tend to minimize the average overlap with the teacher, and since this is typically near the border (just by the geometry of the N-dimensional space) the approach to the teacher weight vector can be made through vectors outside the version space. The version space collapses to only one element at typically $\alpha \sim 1.25$, but the variational class, not considering the version space, will not detect this until later.

The replica symmetric variational study incorrectly predicts a transition at $\alpha < 1$. A stability analysis indicates that the replica symmetric ansatz is not adequate for all $\alpha > 0$. This is a little surprising, and the origin of the instability can be traced back to the requirement that $\beta \rightarrow \infty$ implies $q \rightarrow 1$ for $\alpha \neq 0$. This is in contrast with the case of real weights where, even in the presence of multiplicative noise, learning with the variational potential is replica symmetric. The effect of noise can break phase space into disconnected regions, which are essentially ignored by the robust learning of the optimal potential, which disregards outliers. A method that insists on minimizing the memorization error will certainly lead in such conditions to a replica symmetry breaking situation. The one-step broken replica symmetry leads to apparently consistent physical behavior. The stability analysis will be presented elsewhere [28]. A preliminary picture that emerges from such analysis is that, while one-step RSB is enough to give locally stable results and suggests a reasonable physical picture it may fail to be globally correct. We think that the RSB-1 calculation describes qualitatively the main features of this difficult problem, but a full continuous RSB scheme [28] will be necessary to understand the thermodynamic equilibrium bounds obtainable from the minimization of a potential. Even this will not tell the complete story, however, since issues dealing with effective learning times will still remain. This work, together with the results of [16–18,13], suggests that in the optimization of computationally hard discrete problems it might be a better strategy to first leave the space of configurations, in this case the vertices of the hypercube, then optimize in the hypersphere, and finally go back to the original space, instead of striving to respect the discreteness constraints at every step.

ACKNOWLEDGMENTS

The authors thank M. Copelli and O. Kinouchi for important discussions and suggestions and also the participants of the Workshop on Statistical Mechanics, Max Planck Institute, Dresden (1999) for discussions on this subject. E.B. and N.C. were partially sponsored by the Conselho Nacional de Desenvolvimento Cientı́fico e Tecnologico (CNPq) and

C.R.M. by the Fundação para o Desenvolvimento da Unesp (FUNDUNESP).

APPENDIX

Equation (23) leads to an integral equation for the potential:

$$\int Dt_o \exp[-x_o\mathcal{E}(t)] = \left[H\left(-\frac{u}{\sqrt{\Gamma}}\right)\right]^b \equiv A(u), \quad \text{(A1)}$$

where $t=\sqrt{q_o}u+\sqrt{1-q_o}t_o$, $b=\hat{k}\sqrt{q_o/4\gamma^2}$, and $Dt_o=(dt_o/\sqrt{2\pi})e^{-t_o^2/2}$. To obtain the effective potential, note that the integral on the left side is a convolution, so it seems natural to perform a Fourier transformation. The fact that $A(u)$ is not square integrable is bypassed by defining a new problem:

$$\int Dt_o \exp[-x_o\mathcal{E}_\xi(t)] = \left[H_\xi\left(-\frac{u}{\sqrt{\Gamma}}\right)\right]^b \equiv A_\xi(u) \quad \text{(A2)}$$

Here

$$H_\xi\left(-\frac{u}{\sqrt{\Gamma}}\right) = \int_{-\infty}^{\infty} Dy_i \Theta_\xi\left(y_i+\frac{u}{\sqrt{\Gamma}}\right)$$

where

$$\Theta_\xi\left(y+\frac{u}{\sqrt{\Gamma}}\right) = \exp\left[-\xi\left(y+\frac{u}{\sqrt{\Gamma}}\right)^2\right]$$

if $y>-u/\sqrt{\Gamma}$ and 0 otherwise. Once $\mathcal{E}_\xi(t)$ is found, we will take the regularizing parameter ξ to zero. After Fourier transforming, dividing by the Gaussian on the left, and Fourier transforming back, we get

$$E_\xi \equiv \exp[-x_o\mathcal{E}_\xi(t)] = \sqrt{q_o}\int \frac{dk}{2\pi}\left(e^{(1-q_o)k^2/2}\int e^{i\sqrt{q_o}ku}A_\xi(u)du\right)e^{-ikt}. \quad \text{(A3)}$$

To be able to perform the above integrals we use a simple replica trick, which consists in considering b as an integer,

$$\left[H_\xi\left(-\frac{u}{\sqrt{\Gamma}}\right)\right]^b = \prod_{i=1}^{b}\int_{-u/\sqrt{\Gamma}}^{\infty} Dy_i \exp\left[-\xi\left(y+\frac{u}{\sqrt{\Gamma}}\right)^2\right] = \prod_{i=1}^{b}\int_{-\infty}^{\infty} Dy_i \Theta_\xi\left(y_i+\frac{u}{\sqrt{\Gamma}}\right). \quad \text{(A4)}$$

Introduce the Fourier transforms

$$\Theta_\xi(y_i-x) = \int f_\xi(r_i)e^{ir_i(y_i-x)}dr_i, \quad e^{-y_i^2/2} = \int g(v_i)e^{-iv_iy_i}dv_i.$$

The Fourier transform of $A_\xi(u)$ is

$$\int e^{i\sqrt{q_o}ku}\left[H_\xi\left(-\frac{u}{\sqrt{\Gamma}}\right)\right]^b du = \int e^{i\sqrt{q_o}ku} \times\left\{\int\prod_{i=1}^{b}\left(\frac{dy_i}{\sqrt{2\pi}}dr_i dv_i f_\xi(r_i)g(v_i)\right) \times\exp\left[i\sum r_i\left(y_i+\frac{u}{\sqrt{\Gamma}}\right)-i\sum v_iy_i\right]\right\}du.$$

The u and y_i integrals are now automatic. Going back to the equation for the potential

$$E_\xi = \int\frac{\sqrt{q_o}dk}{(2\pi)^{-b/2}}e^{(1-q_o)k^2-ikt/2} \times\int_{-\infty}^{\infty}\left(\prod_{i=1}^{b}dr_if_\xi(r_i)g(r_i)\right)\delta\left(\sqrt{q_o}k+\sum\frac{r_i}{\sqrt{\Gamma}}\right),$$

we now integrate over k:

$$E_\xi = \int_{-\infty}^{\infty}\left(\prod_{i=1}^{b}\sqrt{2\pi}dr_if_\xi(r_i)g(r_i)\right) \times\exp\left[\frac{1}{2}\frac{(1-q_o)}{\Gamma q_o}\left(\sum r_i\right)^2\right]\exp-i\frac{\sum r_i}{\sqrt{\Gamma q_o}}t.$$

The linearization of the Σr_i is done by a standard trick. Use the fact that $g(r)$ is a Gaussian and that $f_\xi(r)=(1/2\pi)\int\Theta_\xi(y)e^{-iyr}dy$; then

$$E_\xi = \int\frac{dx}{\sqrt{2\pi}}e^{-x^2/2} \times\left(\int\frac{dy}{\sqrt{2\pi}}\Theta_\xi(y)\int_{-\infty}^{\infty}dre^{-r^2/2}e^{-x[\sqrt{(1-q_o)/\Gamma q_o}]r}e^{-iKr}\right)^b,$$

where $K=t/\sqrt{\Gamma q_o}+y$. Integrating over r,

$$E_\xi = \int\frac{dx}{\sqrt{2\pi}}e^{-x^2/2}\left[\int\frac{dy}{\sqrt{2\pi}}\Theta_\xi(y)\exp\frac{1}{2}\left(\frac{(1-q_o)}{\Gamma q_o}x^2-K^2 +2ixK\sqrt{\frac{(1-q_o)}{\Gamma q_o}}\right)\right]^b$$

and b can be again taken to be real. Changing the integration variable to K and taking ξ to zero,

$$E_0 = \int\frac{dx}{\sqrt{2\pi}}\exp-\left[\frac{1}{2}\left(1-\frac{(1-q_o)}{\Gamma q_o}b\right)x^2\right] \times\left[\int DK\Theta\left(K+\frac{t}{\sqrt{\Gamma q_o}}\right)\exp+ixK\sqrt{\frac{(1-q_o)}{\Gamma q_o}}\right]^b.$$

Extending the definition of the H function to complex arguments, the potential can be written formally as Eq. (30).

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

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

[3] M. Opper and W. Kinzel, in *Models of Neural Networks III*, edited by E. Domany, J. L. van Hemmen, and K. Schulten (Springer-Verlag, Berlin, 1996).

[4] A. Engels and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, 2000).

[5] L. Pitt and L. Valiant, J. Assoc. Comput. Mach. **35**, 965 (1988).

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

[7] G. Györgyi, Phys. Rev. Lett. **64**, 2957 (1990).

[8] D. Haussler, M. Kearns, H.S. Seung, and N. Tishby, Mac. Learning **25**, 195 (1996).

[9] H. Horner, Z. Phys. B: Condens. Matter **86**, 291 (1992); **87**, 371 (1992).

[10] M. Golea and M. Marchand, J. Phys. A **26**, 5751 (1993).

[11] C. Van den Broeck and M. Bouten, Europhys. Lett. **22**, 223 (1993).

[12] D. Bollé and G.M. Shim, Network Comput. Neural Syst. **6**, 619 (1995).

[13] L. Reimers, M. Bouten, and B. Van Rompaey, J. Phys. A **29**, 6247 (1996).

[14] M. Bouten, L. Reimers, and B. Van Rompaey, Phys. Rev. E **58**, 2378 (1998).

[15] R.W. Penney and D. Sherrington, J. Phys. A **26**, 6173 (1993).

[16] M. Copelli and C. Van den Broeck, Phys. Rev. E **61**, 6971 (2000).

[17] M. Copelli, M. Bouten, C. Van den Broeck, and B. Van Rompaey, Europhys. Lett. **47**, 139 (1999).

[18] M. Copelli, C. Van den Broeck, and M. Opper, J. Phys. A **32**, L555 (1999).

[19] W. Kinzel and R. Urbanczik, J. Phys. A **31**, L27 (1998).

[20] S. Solla and O. Winther, in *On-Line Learning in Neural Networks*, edited by D. Saad (Cambridge University Press, Cambridge, 1998).

[21] M. Opper, Phys. Rev. Lett. **77**, 4671 (1996); in *On-Line Learning in Neural Networks* (Ref. [20]).

[22] O. Kinouchi and N. Caticha, Phys. Rev. E **54**, R54 (1996).

[23] G.J. Bex, R. Seneerls, and C. Van den Broeck, Phys. Rev. E **51**, 6309 (1995).

[24] M. Opper and D. Haussler, Phys. Rev. Lett. **66**, 2677 (1991); in *Proceedings of the IVth Annual Workshop on Computational Learning Theory (COLT91)* (Morgan Kauffman, San Mateo CA, 1991).

[25] M. Bouten, J. Schieste, and C. Van den Broeck, Phys. Rev. E **52**, 1958 (1995).

[26] O. Kinouchi and N. Caticha, J. Phys. A **25**, 6243 (1992).

[27] N. Caticha and E. Araújo de Oliveira, Universidade de São Paulo report (unpublished).

[28] E. Botelho, C. Mattos, and N. Caticha (unpublished).

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

[30] H.S. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. A **45**, 6056 (1992).

[31] A. Buhot, J.M. Torres Moreno, and M.B. Gordon, Phys. Rev. E **55**, 7434 (1997).

[32] N. Caticha and O. Kinouchi, Philos. Mag. B **77**, 1565 (1998).

[33] S. Amari, Neural Networks **6**, 161 (1993).

[34] N. Barkai, H.S. Seung, and H. Sompolinsky, Phys. Rev. Lett. **75**, 1415 (1995).

[35] O. Winther, B. Lautrup, and J-B. Zhang, Phys. Rev. E **55**, 836 (1997).