"Cavity-approach" analysis of the neural-network learning problem

M. Griniasty*
The Racah Institute of Physics, The Hebrew University, Jerusalem 91904, Israel
(Received 31 July 1992)

We apply a "cavity-type" method for the analysis of the learning ability of single- and multilayer perceptrons. We show that the mean-field equations obtained in this way, which are identical to the equations derived previously by the replica method, describe not only the properties of the optimal network, but also a *learning process* which leads to this network. We discuss the applicability of our ideas to the construction of learning algorithms. Our interpretation of the mean-field theory also leads naturally to a new concept, "flexibility," which is a measure of the ability of the network to learn.

PACS number(s): 87.10.+e, 05.50.+q, 64.60.Cn

I. INTRODUCTION

It is now five years since Elizabeth Gardner introduced the statistical-mechanical approach to the learning problem of neural networks [1,2]. This method has been found to be a powerful theoretical tool and opened a new field of research. In this approach, the learning task is formulated as an optimization problem in the space of network connections: every configuration of weights $\{J\}$ is assigned with an energy $E(\{J\})$, which measures its success in achieving the learning task. The energy landscape is analyzed by a statistical-mechanical approach.

As an example let us consider the simple perceptron, which is a threshold-linear feed-forward network with N input units $S_i = \pm 1$, N *continuous* connections J_j $(j = 1,\ldots,N)$ and an output unit $\sigma = \pm 1$. The input output relation is given by

$$\sigma = \mathrm{sgn}\left(\sum_{j=1}^{N} J_j S_j\right).$$

The perceptron learning task is to find a set of connections J that map correctly a set of P input patterns $\{\xi_i^\mu\}$ $(i = 1,...,N\ ;\ \mu = 1,...,P)$ to their desired outputs σ^μ.

Gardner and Derrida [2] associate an energy with each configuration of connections

$$E(\{J\}) = \sum_{\mu=1}^{p} V(h^\mu)\ ,\quad h^\mu = \sigma^\mu \sum_{j=1}^{N} J_j \xi_j^\mu/\sqrt{N} \qquad (1)$$

and

$$V(h) = \theta(-h). \qquad (2)$$

where θ is the Heaviside step function and h^μ is denoted the stability of the pattern ξ^μ. The energy of a configuration of connections equals the number of patterns that are mapped incorrectly by this network. The optimal network is the one with the minimal energy, or the minimal number of errors.

In the statistical-mechanical approach one has to calculate the partition function

$$Z = \sum_{\{J\}} e^{-\beta E(\{J\})}\delta\left(\sum_j J_j^2 - N\right). \qquad (3)$$

β is the inverse temperature. The δ function confines the space of solutions to network configurations with a fixed norm, $J \cdot J = N$. The patterns are composed of random independent unbiased elements $\{\xi_i^\mu\}$. We do not know how to calculate the partition function for a particular set of patterns, but we can obtain results that describe the properties of a network that is trained with a typical set of random patterns. This is achieved by a quenched average over the patterns, an average of the logarithm of the partition function. The quenched average is achieved by the replica trick

$$\langle \ln Z\rangle = \lim_{n\to 0} \frac{\langle Z^n\rangle - 1}{n}. \qquad (4)$$

Angular brackets denote the average over the patterns.

As a result of this calculation one finds that the space of solutions (networks) for the simple perceptron depends on the fraction α between the number of patterns and the number of input units, $\alpha = P/N$. If α is less than some critical value α_c, there are many solutions in J space, each of zero energy. Above α_c the ground state is unique and the ground-state energy is positive. These results are obtained under the assumption of "replica symmetry." Other quantities which are calculated by this method are the distribution of fields [3], and the typical overlap among different solutions below α_c [1].

All these results are obtained as we solve the replica-derived mean-field equations. The problem with the replica approach is that although powerful, it obscures the the "physical" meaning of the theory. Replica theory has been successfully applied to the problem of spin glasses [4]. However, the same lack of physical clarity motivated Mezard, Parisi, and Virasoro [5] to develop

the alternative *cavity* approach. In this approach one equilibrates thermally a spin glass of N spins, then adds to the system a new spin and equilibrates the combined system. The mean-field equations are derived from self-consistency considerations on the newly added spin.

In this work we analyze the perceptron learning problem in the same spirit. The idea of the cavity method in the framework of neural networks is to optimize $E(J)$ for a system of P patterns, then add a new pattern and optimize again. This approach has been suggested first by Mezard [6]. There are some differences between the cavity approach of Mezard and the approach presented here. First, the Mezard calculation assumes finite temperature, and therefore many configurations in weight space should be taken into account. In this calculation we assume zero temperature and a unique ground state. This assumption leads to a theory which is equivalent to the replica theory above α_c. Second, Mezard adds a pattern and also a site. In the present calculation the number of sites is preserved. Third, and most important, in the present calculation the focus is on the *stabilities* of the patterns and their evolution under the addition of a new pattern, while Mezard focuses on the evolution of *weights*. As we shall see, our point of view makes the theory physically clear and explicitly equivalent to previous replica calculations.

This paper is set up as follows. In Sec. II we describe first the results of the replica analysis of a simple perceptron with continuous weights which are norm restricted. Thereafter, we derive the same equations by the cavity approach, and present a physical interpretation to their structure. In the second part of this section we analyze perceptrons which are not norm constrained. This serves as an introduction to Sec. III. In Sec. III we analyze multilayer perceptrons (MLP's). We consider two types of architectures: fully connected MLP's where each "hidden" unit receives input from all input units, and MLP's with local receptive fields, where each hidden unit receives input from a distinct group of input units. We explain the difference in the resulting mean-field theories. In Sec. IV we introduce a new concept, denoted "flexibility," which arises naturally from the mean-field theory, and which serves as a quantitative measure of the ability of a network to learn. In Sec. V we present learning algorithms which are motivated by the cavity interpretation of the mean-field theory.

II. SINGLE-LAYER FEED-FORWARD NETWORKS

A. Model A: simple perceptrons with a norm constraint

We present here a generalization of the approach of Gardner and Derrida, described previously, where $V(h)$ is a general cost function instead of the step function [7]. As in [2] we consider networks with a fixed norm: $J \cdot J = N$. The present group of models provides the simplest example for the application of the cavity method. We describe first the results of the replica method.

1. Results of the replica calculation

Since we are interested in the ground-state properties of the system we take the limit $\beta \to \infty$. In this limit we calculate the "free energy" G by the replica method,

$$G = \lim_{\beta\to\infty} \left\langle -\frac{1}{\beta N} \ln Z \right\rangle .$$

This calculation involves a marix of order parameters

$$q_{ab} = \frac{1}{N} \sum_{j=1}^{N} = J_j^a J_j^b .$$

The solution is obtained under the ansatz of replica symmetry which assumes $q_{ab} = q$ for $a \neq b$. Below α_c there are many network configurations that minimize the energy, each of which performs the learning task perfectly. q measures the typical overlap between two such solutions and since the networks are normalized its value is less than one. Above α_c the meaning of the replica-symmetric ansatz is that only one configuration of connections minimizes the energy. This optimal network does not perform the task perfectly and the energy is greater than the minimum of V (times P). Since the minimum is unique, as $\beta \to \infty$ we have $q \to 1$. In this region a new order parameter $x \equiv \beta(1-q)$ appears naturally. In terms of this parameter we may write G in the following form:

$$G = \alpha \int Dt\, F(h_0(x,t),x,t) - \frac{1}{2x}, \tag{5}$$

where

$$F(h,x,t) = V(h) + \frac{(h-t)^2}{2x}.$$

$h_0(x,t)$ is the value of h which minimizes F for fixed x and t.

The dependence of x on α is found by the mean-field equation

$$\frac{dG}{dx} = 0, \tag{6}$$

$$\alpha \int Dt (h_0(x,t) - t)^2 = 1. \tag{7}$$

The limit $x \to \infty$ corresponds to minimal ground-state energy and α that solves the equation in this is α_c. The distribution of stabilities at the ground state is given by

$$\rho(h) = \int Dt\, \delta(h - h_0(x,t)) \tag{8}$$

and the ground-state energy per pattern is given by

$$E/P = \int Dt\, V(h_0(x,t)). \tag{9}$$

Applying this machinery to the Gardner-Derrida step cost function one finds $\alpha_c = 2$ in agreement with a previously known result [8].

The replica equations look extremely simple. However,

it is not clear what is the "physical" meaning of the order parameter x and the Gaussian variable t. We derive now the statistical theory of learning by application of the "cavity" method above α_c.

2. *The cavity calculation*

Suppose we have to minimize the energy for a specific set of patterns ξ^μ $(\mu = 1, ..., P)$ (that is, to train the network with these patterns). We assume that $P/N > \alpha_c$ and, therefore, that only one network configuration, $J^\star$, minimizes the energy. This assumption makes our cavity calculation equivalent to the replica-symmetric solution. The ground-state energy is given by

$$E_{\mathrm{GS},P} = E_P(J^\star) = \sum_{\mu=1}^{p} V(h_\mu^\star). \tag{10}$$

Since the patterns are unbiased, we may assume without loss of generality, in the case of a simple perceptron, that the outputs are $+1$ for all patterns, and therefore

$$h_\mu^\star = J^\star \cdot \xi^\mu / \sqrt{N}.$$

The subscript GS stands for ground state. Let us add a new pattern ξ^0 and look for the ground state of the combined $P+1$ system.

For an arbitrary change ΔJ the energy of the combined system is

$$E_{P+1}(J^\star + \Delta J) = E_P(J^\star + \Delta J) + V\left[(J^\star + \Delta J)\cdot \xi^0/\sqrt{N}\right], \tag{11}$$

$$E_P(J^\star + \Delta J) \equiv \sum_{\mu=1}^{p} V(h_\mu^\star + \Delta J \cdot \xi^\mu/\sqrt{N}).$$

We assume that the new ground-state vector does not differ much from $J^\star$. The reason is that the typical change in the stability that is needed for the training of ξ^0 is of order one. A change of this order is achieved by $\Delta J = \xi^0/\sqrt{N}$. As a result of this change, the stability of another pattern ξ^μ $(\mu \neq 0)$ is changed by an amount $\xi^0 \cdot \xi^\mu/N$. This change is of the order $1/\sqrt{N}$ since the patterns are uncorrelated. Therefore, for $\mu \neq 0$ we may expand $V(h_\mu^\star + \Delta J \cdot \xi^\mu/\sqrt{N})$ around $h_\mu^\star$.

Since $J^\star$ minimizes E_P the first-order term vanishes and we find

$$E_P(J^\star + \Delta J) = E_{\mathrm{GS}}(P) + \Delta E_P(\Delta J), \tag{12}$$

$$\Delta E_P \equiv \tfrac{1}{2}\sum_{\mu=1}^{p} V''(h_\mu^\star)(\Delta J \cdot \xi^\mu/\sqrt{N})^2 \equiv \sum_{i,j} T_{ij}\Delta J_i \Delta J_j.$$

Terms of higher order do not contribute in the thermodynamic limit.

$J^\star$ minimizes E_P in the space of normalized J's and any small change ΔJ that is orthogonal to $J^\star$ (and therefore conserves the norm) should increase the energy of the P patterns. Therefore, ΔE_P is a positive-definite symmetric quadratic form in ΔJ in the $(N-1)$-dimensional subspace orthogonal to $J^\star$.

We can diagonalize T in this subspace and find its positive eigenvalues A_i and its eigenvectors U_i. In terms of these we write

$$\Delta E_P(\Delta J) = \sum_{i=1}^{N-1} A_i(\Delta J \cdot U_i)^2. \tag{13}$$

In terms of these we write

$$E_{P+1}(J^\star + \Delta J) = E_{\mathrm{GS},P} + \sum_{i=1}^{N} A_i(\Delta J \cdot U_i)^2 + V\left[(J^\star + \Delta J)\cdot \xi^0/\sqrt{N}\right]. \tag{14}$$

Generally $V(\Delta)$ is a nonlinear function. It is convenient to minimize E_{P+1} in two steps: First we minimize E_{P+1} with respect to ΔJ fixing the stability of the last pattern $h \equiv (J^\star + \Delta J)\cdot \xi^0/\sqrt{N}$ and at the second stage we minimize with respect to h,

$$E_{P+1}(\Delta J, h, \lambda_1, \lambda_2) = E_{\mathrm{GS},P} + \sum_{i=1}^{N} A_i(\Delta J \cdot U_i)^2 + \lambda_1[h - (J^\star + \Delta J)\cdot \xi^0] + V(h) + \lambda_2(J^\star \cdot \Delta J). \tag{15}$$

λ_1 defines h and λ_2 imposes the normalization constraint. Minimizing freely with respect to ΔJ and eliminating afterwards λ_2 and and λ_1 we find

$$E_{P+1}(h) = E_{\mathrm{GS},P} + \frac{(h-t)^2}{2x} + V(h), \tag{16}$$

where

$$2x \equiv \sum_{i=1}^{N-1} \frac{\left(U_i \cdot \xi^0/\sqrt{N}\right)^2}{A_i}$$

and

$$t \equiv J^\star \cdot \xi^0/\sqrt{N}.$$

Therefore,

$$E_{\mathrm{GS},P+1} - E_{\mathrm{GS},P} = \min_h \frac{(h-t)^2}{2x} + V(h). \tag{17}$$

The expression on the right-hand side (RHS) is nothing but the expression F that appears in the replica theory.

We are now in a position to interpret the meaning of variables that appear in the replica equations. t is the stability of the newly added pattern ξ^0 with respect to the old solution $J^\star$. In other words, this is the stability of

ξ^0 before training. Since $J^\star$ and ξ^0 are uncorrelated, and since the norm squared of $J^\star$ is N, t is a Gaussian random number with zero mean and unity average square, in agreement with the replica theory. As in the replica theory, we define h_0, which is the value of h that minimizes F. The meaning of h_0 is clear now: it is the stability of ξ^0 after training.

The value of h_0 is a result of a competition between the two energy terms in F: $V(h)$ is the energy associated with ξ^0 and $(h-t)^2/2x$ is the *minimal* energy increase of the P patterns if the stability of ξ^0 is changed from t to h. x serves as a "stiffness" parameter: the smaller x is the harder it is to change the network. We therefore expect that as P increases x should decrease.

In the replica calculation x depends only on α while in the cavity approach x depends on ξ^0 as well as on U_i and A_i, which are functions of the P patterns. Averaging x over ξ^0 we find

$$\langle 2x\rangle = \frac{1}{N}\sum_{i=1}^{N-1}\frac{1}{A_i} \equiv \int dA\,\rho(A)\frac{1}{A}. \tag{18}$$

The fluctuations in x with respect to ξ^0 are given by

$$\langle (2x)^2\rangle - \langle 2x\rangle^2 = \frac{2}{N^2}\sum_i \frac{1}{A_i}. \tag{19}$$

These fluctuations disappear in the large-N limit and we find that x depends only on the distribution of the eigenvalues A_i. The replica-symmetry ansatz actually bares two assumptions: the first is that the ground state is unique and the second is that the spectrum of T is self-averaging (or, at least, the trace of T^{-1}). We cannot prove that the distribution is sample independent; however, we shall present later an explicit example where this is indeed the case.

For most models, the distribution of the eigenvalues is unknown, and therefore cannot be used for the calculation of x. We shall present later on a self-consistency approach for the calculation of x, but before turning to this problem we show that if the value of x is known we can calculate by the cavity method everything that is calculable by the replica approach.

If the initial stability of ξ^0 is t then minimizing F we find $h_0(x,t)$. · The energy of the newly added pattern is therefore $V(h_0(x,t))$. The average energy of the last pattern is the average of $V(h_0(x,t))$ over t. However, after training, ξ^0 should not be different from the previous patterns, and therefore, the energy of the whole system is $P+1$ times the averaged energy of ξ^0,

$$E_{\mathrm{GS},P+1} = (P+1)\int Dt\,V(h_0(x,t)). \tag{20}$$

Since we know the function $h_0(x,t)$ we can also calculate the distribution of stabilities of the last pattern

$$\rho(h) = \int Dt\,\delta(h - h_0(x,t)). \tag{21}$$

Using, again, the equivalence of all patterns, this is also the distribution of stabilities of the whole $P+1$ system. The cavity expressions for the energy and the distribution of fields are identical to the results of the replica theory.

We are left with the determination of $x(P)$. The evolution of x with P is determined from a self-consistency consideration.

Let us calculate the average increase of the ground-state energy when a pattern is added to a system of P patterns. On one hand, the energy increase results from the change in x

$$\begin{aligned}\Delta E &\equiv E_{\mathrm{GS},P+1} - E_{\mathrm{GS},P}\\ &= (P+1)\int Dt\,V(h_0(x_p,t))\\ &\quad -P\int Dt\,V(h_0(x_{p-1},t)),\end{aligned} \tag{22}$$

where x_p is the value of x after training P patterns. On the other hand, from Eq. (17), the average energy increase can be written

$$\Delta E = \int Dt\,F(h_0(x_p,t),t). \tag{23}$$

Equating these two expressions we obtain a self-consistency equation for x,

$$\begin{aligned}\int Dt\left[\frac{(h_0(x,t)-t)^2}{2x}\right] &= P\int Dt\,V(h_0(x_p,t))\\ &\quad -P\int Dt\,V(h_0(x_{p-1},t))\\ &= P\frac{d}{dP}\int Dt\,V(h_0(x,t))\\ &= \alpha\frac{d}{d\alpha}\int Dt\,V(h_0(x,t)).\end{aligned} \tag{24}$$

This equation, which determines the evolution of x with α is the key for our whole theory.

On the RHS we have $dV(h_0)/dP = V'(h_0)\;dh_0/dP$. The dependence of h_0 on P is only via its dependence on x. Using the fact that h_0 minimizes F we have also

$$V'(h_0) = -\frac{d}{dh_0}\frac{(h_0-t)^2}{2x}$$

and the consistency equation is written

$$\int Dt\left[\frac{(h_0(x,t)-t)^2}{2x}\right] = -\alpha\int Dt\frac{dh_0}{d\alpha}\frac{d}{dh_0}\frac{(h_0-t)^2}{2x}. \tag{25}$$

Note that

$$\frac{dh_0}{d\alpha}\frac{d}{dh_0}\frac{(h_0-t)^2}{2x} = \frac{1}{2x}\frac{d}{d\alpha}(h_0-t)^2.$$

So the consistency equation may be written

$$\int Dt(h_0(x,t)-t)^2 = -\alpha\frac{d}{d\alpha}\int Dt(h_0-t)^2. \tag{26}$$

Integrating this equation we find

$$\alpha \int Dt(h_0(x,t)-t)^2 = C, \tag{27}$$

where C is an undetermined integration constant. This reproduces the replica mean-field equation that determines x [Eq. (7)] up to a constant C which should be 1.

To determine C we need a boundary condition. Such a boundary condition can be the knowledge of α_c. Using this information we can determine C since the value of x at α_c is known (it is infinity). However, cavity theory does not provide such information and C is left undetermined. There is, however, a simple argument that "explains" why $C=1$.

3. A heuristic calculation of the constant C

Suppose ξ^0 has initially a stability t and a stability $h_0(t)$ after training. Therefore,

$$\frac{1}{N}\Delta J \cdot \xi^0 = h_0(t) - t.$$

We assume that ΔJ is in the direction of ξ^0,

$$\Delta J = \frac{1}{\sqrt{N}} y^0 \xi^0 \quad , \quad y^0 = h_0(t) - t.$$

y^0 is denoted the "embedding strength" [9] of ξ^0. Its distribution is given by

$$\rho(y^0) = \int Dt\, \delta(y^0 - (h_0 - t)). \tag{28}$$

Since all patterns are equivalent we may write

$$J^\star = \frac{1}{\sqrt{N}} \sum_{\mu=0}^{P} y^\mu \xi^\mu. \tag{29}$$

y^μ are distributed like y^0. The equation that determines the value of x is the normalization constraint

$$|J|^2 = \left| \frac{1}{\sqrt{N}} \sum_{\mu=1}^{P} y^\mu \xi^\mu \right|^2 = N. \tag{30}$$

If we neglect the correlations among the patterns and keep only diagonal terms we find

$$\frac{1}{N} \sum_{\mu=1}^{P} (y^\mu)^2 = 1. \tag{31}$$

Replacing the sum with the distribution of embedding strengths we get

$$\alpha \int Dt (h_0 - t)^2 = 1, \tag{32}$$

which is the desired result.

In deriving this result we neglected the correlations among the patterns twice: first when we assumed that the change in J is in the direction of ξ^0 and second in calculation of the norm. Performing an even number of mistakes we succeeded to obtain the correct result.

B. Model B: Perceptrons without a normalization constraint

In this group of models the norm of the solution is not confined. This leads to the appearance of a new order parameter, the norm of the solution. This is a more complicated example of the application of the cavity method, and we present it here for two reasons: first, it is shown to be a generalization of model A, and second, it is an introduction for the application of the cavity method to the two-layer perceptron.

1. The replica calculation

The zero-temperature "free energy" of the system is derived by the replica method. The result is

$$G = \alpha \int Dt\, F(h_0(x, Rt), x, Rt) - \frac{R^2}{2x}, \tag{33}$$

where

$$F(h, x, Rt) = V(h) + \frac{(h - Rt)^2}{2x}$$

and $h_0(x, Rt)$ minimizes F for given x and Rt. We have now *two* order parameters x and R which depend on P. x plays the same role as in model A, and $R\sqrt{N}$ is the norm of the ground-state solution $J^\star$. Differentiating G with respect to x and R we obtain two mean-field equations:

$$\alpha \int Dt (h_0(x, Rt) - Rt)^2 = R^2 \tag{34}$$

and

$$\alpha \int Dt(-t)(h_0 - Rt) = R. \tag{35}$$

If we know R and x we can calculate the ground-state energy per pattern,

$$E/P = \int Dt\, V(h_0(x, Rt)) \tag{36}$$

and the distribution of stabilities

$$\rho(h) = \int Dt\, \delta(h - h_0(x, Rt)). \tag{37}$$

2. The cavity approach for model B

We start applying the cavity method on the same lines as in model A: $J^\star$ is the vector that minimizes the energy of P patterns. We assume that its norm is $R\sqrt{N}$. We add a new pattern ξ^0 and develop the energy of the combined

$P+1$ system around $J^\star$,

$$E_{P+1}(J^\star+\Delta J) = E_{\mathrm{GS},P} + \Delta E_P(\Delta J) + V\left[(J^\star+\Delta J)\cdot\xi^0/\sqrt{N}\right], \tag{38}$$

$$\Delta E_P = \tfrac{1}{2}\sum_{\mu=1}^{p} V''(h_\mu^\star)(\Delta J\cdot\xi^\mu/\sqrt{N})^2 \equiv \sum_{ij} T_{ij}\Delta J_i \Delta J_j.$$

Since there is no normalization constraint $J^\star$ is a true minimum of the energy. Diagonalizing T, which is a positive-definite matrix, we have

$$\Delta E_P(\Delta J) = \sum_{i=1}^{N} A_i(\Delta J\cdot U_i)^2.$$

As before we fix the stability of the last pattern to the value h and minimize in two steps,

$$E_{P+1}(\Delta J, h, \lambda) = E_{\mathrm{GS},P} + \sum_{i=1}^{N} A_i(\Delta J\cdot U_i)^2 + V(h) + \lambda\left[h - (J^\star+\Delta J)\cdot\xi^0/\sqrt{N}\right]. \tag{39}$$

Minimizing with respect to ΔJ and then with respect to λ we are left with

$$E_{\mathrm{GS},P+1} = E_{\mathrm{GS}}(P) + \frac{(h_0(x,Rt)-Rt)^2}{2x} + V(h_0), \tag{40}$$

where $h_0(x,Rt)$ denotes the minimum point of F and

$$2x = \sum_{i=1}^{N}\frac{(U_i\cdot\xi^0/\sqrt{N})^2}{A_i}.$$

The initial stability $J^\star\cdot\xi^0/\sqrt{N}$ is denoted Rt. Since the norm of $J^\star$ is $R\sqrt{N}$ the random number t is a normalized Gaussian variable.

Averaging x over ξ^0 we express x in terms of the spectrum of T

$$\langle 2x\rangle = \int dA\,\rho(A)\frac{1}{A} \tag{41}$$

and it can be shown, as in model A, that this quantity does not fluctuate with ξ^0 in the large-N limit.

As in model A we can rederive the replica expressions for the ground-state energy and the distribution of stabilities by application of the principle of equivalence of all patterns.

The task that we are left with is the calculation of the functions $x(\alpha)$ and $R(\alpha)$. We need two equations. The first one is, again, the self-consistency condition of the energy increase. In analogy with model A we find

$$\int Dt\left[\frac{(h_0(x,Rt)-Rt)^2}{2x}\right] = \alpha\frac{d}{d\alpha}\int Dt\,V(h_0(x,Rt)). \tag{42}$$

x and R are both functions of α. h_0 depends on α via its dependence on x and Rt. Using the fact that h_0 minimizes F this equation is written

$$\int Dt\left[\frac{(h_0(x,Rt)-Rt)^2}{2x}\right] = -\alpha\int Dt\frac{dh_0}{d\alpha}\frac{d}{dh_0}\frac{(h_0-Rt)^2}{2x}. \tag{43}$$

Note that

$$\frac{dh_0}{d\alpha}\frac{d}{dh_0}\frac{(h_0-Rt)^2}{2x} = \frac{1}{2x}\left\{\frac{d}{d\alpha}(h_0-Rt)^2 - \frac{dR}{d\alpha}\frac{\partial}{\partial R}(h_0-Rt)^2\right\}.$$

So we write

$$\int Dt(h_0(x,Rt)-Rt)^2 = -\alpha\int Dt\left\{\frac{d}{d\alpha}(h_0-Rt)^2 - \frac{dR}{d\alpha}(-2t)(h_0-Rt)\right\}. \tag{44}$$

This is the first cavity equation.

The second cavity equation is new and expresses the fact that $J^\star$ is a true minimum of the energy. We do not know $J^\star$ explicitly, but we do know the distribution of the stabilities of the patterns with respect to it

$$\rho(h) = \int Dt\,\delta(h - h_0(x,Rt))$$

and the average energy per pattern

$$E/P = \int dh\,\rho(h)V(h) = \int Dt\,V(h_0(x,Rt)).$$

If we *rescale* $J^\star$ by a factor s each stability $h^\mu = \sigma^\mu J^\star\xi^\mu$ will be multiplied by this factor and therefore the average energy per pattern will be

$$E = \int Dt\,V(sh_0(x,Rt)).$$

The second cavity equation expresses the fact that at the minimum point, under this scaling the energy should be stationary,

$$\frac{d}{ds}\int Dt\,V(\mathbf{s}h_0(x,Rt))|_{s=1} = 0 \tag{45}$$

or

$$\int Dt\,h_0V'(h_0) = 0. \tag{46}$$

We show now that the cavity equations are equivalent to

the replica mean-field equations. Since $V'(h_0) = -(h_0 - Rt)/x$ The second cavity equation can be written

$$\int Dt\, h_0(h_0 - Rt) = 0. \tag{47}$$

We plug this equation into the first cavity equation and find

$$-\alpha\frac{d}{d\alpha}\int Dt(h_0 - Rt)^2 + \alpha\frac{2}{R}\frac{dR}{d\alpha}\int Dt(h_0 - Rt)^2 = \int Dt(h_0 - Rt)^2 \tag{48}$$

or

$$-\alpha\frac{d}{d\alpha}\ln\int Dt(h_0 - Rt)^2 + \alpha\frac{d}{d\alpha}\ln(R^2) = 1. \tag{49}$$

Integrating this equation we find the first replica equation

$$\alpha\int Dt(h_0(x, Rt) - Rt)^2 = CR^2, \tag{50}$$

where C is again an undetermined constant.

Using this equation and the second cavity equation we get

$$\alpha\int Dt(-t)(h_0 - Rt) = CR, \tag{51}$$

which is the second replica equation. This completes the demonstration that the cavity and the replica equations are equivalent.

Note that the models with constrained norm are derived from the unconstrained norm models simply by setting $R = 1$ and keeping only the first mean-field equation.

As in model A, we have succeeded in deriving the cavity equations up to an integration constant. We can justify $C = 1$ for the first replica equation using the same argument as in model A.

3. A simple example

We use the quadratic model (no normalization)

$$V(h) = (h - 1)^2 \tag{52}$$

to demonstrate our ideas.

For $P < N$ there is always a set of connections for which the energy is zero since the number of linear equations that we have to solve, $h^\mu = 1$, is less than the number of variables J_j. It is clear therefore that $\alpha_c = 1$. We investigate the behavior of this model above saturation.

Suppose we have found a network configuration $J^\star$ which minimizes the energy of P patterns. We add a new pattern. Any change ΔJ increases the energy of the previous P patterns by

$$\Delta E_P = \tfrac{1}{2}\sum_{\mu=1}^{p} V''(h_\mu^\star)(\Delta J\cdot\xi^\mu/\sqrt{N})^2 \equiv \sum_{i,j} T_{ij}\Delta J_i\Delta J_j.$$

Note that since the model is quadratic, this is the *exact* value of ΔE_P.

In this model the matrix T is simply the correlation matrix

$$T_{ij} = \frac{1}{N}\sum_{\mu=1}^{P}\xi_i^\mu\xi_j^\mu.$$

The spectrum of this matrix is self-averaging and the distribution of eigenvalues is known [10]. For $\alpha > 1$ the eigenvalues are confined to the region $A_- < A < A_+$: $A_\pm = (1 \pm \sqrt{\alpha})^2$ and

$$\rho(A) = \sqrt{(A_+ - A)(A - A_-)}/(2\pi A). \tag{53}$$

Therefore we can calculate x explicitly by Eq. (41) and avoid the use of the first cavity equation. The result is

$$2x = \frac{1}{\alpha - 1}. \tag{54}$$

The energy of the $P + 1$ system is given by

$$E_{\mathrm{GS},P+1} = E_{\mathrm{GS},P} + \min_h\left[(\alpha - 1)(h - Rt)^2 + (h - 1)^2\right]. \tag{55}$$

From this we calculate the function $h_0(Rt)$ and find

$$h_0(Rt) = \frac{1 + (\alpha - 1)Rt}{\alpha}. \tag{56}$$

The value of $R(\alpha)$ is found by the second cavity equation [Eq. (47)]. The result is $R^2 = 1/(\alpha - 1)$. Knowing the dependence of R and x on α we calculate the energy per pattern as a function of α,

$$E/(P + 1) = \int Dt(h_0(Rt) - 1)^2 = \alpha - 1. \tag{57}$$

All these results are in complete agreement with the replica calculation.

III. MULTILAYER PERCEPTRONS

Cavity theory provides a new way of looking at multilayer perceptrons. The fully connected multilayer perceptron (MLP) consists of N binary input units S_i, a layer of k "hidden units" σ_i, and an output unit η.

A set of connections $J_{1j}, J_{2j}, ..., J_{kj}$ connects *all* input units to *every* hidden unit. The sign of a hidden unit is given by

$$\sigma_i = \mathrm{sgn}(h_i), \tag{58}$$

where the fields on the hidden units h_i are given by

$$h_i = \sum_j J_{ij}S_j/\sqrt{N}.$$

Each set of connections to one of the hidden units is denoted a *subnetwork* and may be regarded as a simple per-

ceptron. Between the hidden layer and the output there are other weights and possibly more hidden layers. All these connections should be modified in the learning process. The learning task is, again, to associate correctly a set of input patterns to their outputs.

In this work we assume that all the connections from the first hidden layer to the output are fixed and are effectively represented by a Boolean function B which maps each configuration of the hidden layer to the output

$$\eta = B(\bar{\sigma}) \ , \ \bar{\sigma} \equiv \{\sigma_i\}_{i=1,...,k}. \tag{59}$$

Only the first layer of connections is modified during the learning process.

A. Replica theory for the fully connected MLP

The ability of a MLP with a *fixed* Boolean function to learn can be analyzed by statistical mechanics. Suppose we have to train the network with P_+ patterns which have a (+) output and P_- patterns with a (−) output. As in the simple perceptron, we can associate an energy function with each network configuration $\{J_1, ..., J_k\}$

$$E(J_1, ..., J_k) = \sum_{\mu \in P_+} V_+(h_1^\mu, ..., h_k^\mu) + \sum_{\mu \in P_-} V_-(h_1^\mu, ..., h_k^\mu) \tag{60}$$

where $h_i^\mu = \sum_j J_{ij}\xi_j^\mu/\sqrt{N}$.

$V_+(h_1, ..., h_k) = 0$ if $B(\bar{\sigma}) = 1$, where the ith element of $\bar{\sigma}$ is given $\sigma_i = \mathrm{sgn}(h_i)$. If $B(\bar{\sigma}) = -1$ we have $V_+(h_1, ..., h_k) = 1$. V_- is given by $V_-(h_1, ..., h_k) = 1 - V_+(h_1, ..., h_k)$.

This is a generalization of the Gardner-Derrida step cost function. The energy of a network equals the number of patterns that are mapped incorrectly. However, as in the simple perceptron, the formalism that we describe here holds for arbitrary $V_\pm$ and not only to potentials which are a generalization of the step function.

The partition function is given by

$$Z = \sum_{J_1,...,J_k} e^{-\beta E(J_1,...,J_k)} \tag{61}$$

without a constraint on the norm. Norm-constrained models are derived from the present model in the same way that model A is derived from model B.

Applying the replica trick, assuming replica symmetry, we obtain the zero-temperature free energy G,

$$G = \lim_{\beta \to \infty} \left\langle -\frac{1}{\beta N} \ln Z \right\rangle.$$

The calculation of the free energy above α_c is almost identical with the calculation of Griniasty and Grossman [12] and the result in the limit of zero temperature is

$$G = \alpha_+ \int D_s\tau \left[V_+(h_1^+, ..., h_k^+) + \sum_{i,j=1}^{k} M_{ij}(h_i^+ - \tau_i)(h_j^+ - \tau_j) \right] + \alpha_- \int D_s\tau \left[V_-(h_1^-, ..., h_k^-) + \sum_{i,j=1}^{k} M_{ij}(h_i^- - \tau_i)(h_j^- - \tau_j) \right] - \sum_{i,j=1}^{k} M_{ij}S_{ij}, \tag{62}$$

where $\alpha_\pm = P_\pm/N$ and

$$D_s\tau \equiv \frac{d\tau_1, ..., d\tau_k}{(2\pi)^{k/2}\sqrt{\det S}} \exp -\tfrac{1}{2}\tau S^{-1}\tau.$$

S and M are symmetric matrices of dimension k. The saddle-point values of S_{ij} represent the overlaps between the subnetworks that minimize the energy

$$S_{ij} = \frac{1}{N}\sum_{m=1}^{N} J_{im}J_{jm}.$$

$\bar{h}^\pm$ and $\bar{\tau}$ are vectors of k elements. Each of the elements of $\bar{h}^+$ ($\bar{h}^-$) is a function of the vector $\bar{\tau}$. It is the value of $\bar{h}$ which minimizes F_+ (F_-)

$$F_\pm(M, \bar{\tau}, \bar{h}) = V_\pm(h_1, ..., h_k) + \sum_{i,j=1}^{k} M_{ij}(h - \tau)_i(h - \tau)_j. \tag{63}$$

Note the similarity to the corresponding expressions of model B [Eq. (33)]. We leave the derivation of the mean-field equations for the Appendix and turn to the cavity calculation.

B. The cavity calculation for fully connected MLP's

The cavity approach is a generalization of the cavity for model B. Suppose we train the network with $P = P_+ + P_-$ patterns and find the optimal set of subnetworks: $J_1^\star, ..., J_k^\star$. Let us add a new (+) pattern ξ^0. The initial fields are $\tau_i = J_i^\star\xi^0/\sqrt{N}$. These variables are random and *correlated*,

$$\langle \tau_i\tau_j \rangle_{\xi^0} = \frac{1}{N}\sum_{l=1}^{N} J_{il}J_{jl} \equiv S_{ij}. \tag{64}$$

This result is in agreement with the distribution $D_s\tau$ in the replica approach which generates the same correla-

tions among the τ variables,

$$\int D_s\tau\, \tau_i\tau_j = S_{ij}. \tag{65}$$

If each of the subnetworks is changed by ΔJ_i we have

$$E_{P+1}(\Delta J_1,...,\Delta J_k) = E_{\rm GS}(P) + \Delta E_P(\Delta J_1,....,\Delta J_k) + V_+(h_1,...,h_k), \tag{66}$$

where

$$h_i = (J_i^\star + \Delta J_i)\xi^0/\sqrt{N}$$

and ΔE_P is given by development of E_P around its minimum value to the second order

$$\Delta E_P(\Delta J_1,...,\Delta J_k) = \frac{1}{2N}\sum_{\mu=1}^{P}\sum_{i,j=1}^{k} V_\mu^{ij}(h^\star_{\mu,1},...,h^\star_{\mu,k}) \times(\Delta J_i\cdot\xi^\mu)(\Delta J_j\cdot\xi^\mu). \tag{67}$$

$V_\mu^{ij} \equiv (\partial^2 V_\mu/\partial h_i\partial h_j)$. $h^\star_{\mu,l} = J_l^\star\cdot\xi^\mu/\sqrt{N}$ and V_μ is either V_+ or V_-. This may also be written in a matrix form

$$\Delta E_P = \sum_{i,k,j,l} D_{ikjl}\Delta J_{ik}\Delta J_{jl}, \tag{68}$$

$$D_{ikjl} \equiv \frac{1}{2N}\sum_\mu V_\mu^{ij}\xi_k^\mu\xi_l^\mu.$$

We introduce k Lagrange multipliers λ_i to fix the value of h_i and write

$$E_{P+1}(\{\Delta J_i\},\{\lambda_i\},\{h_i\}) = E_{{\rm GS},P} + \Delta E_P(\{\Delta J_i\}) + V_+(h_1,...,h_k) + \sum_{i=1}^{k}\lambda_i[h_i - (J_i^\star + \Delta J_i)\xi^0/\sqrt{N}]. \tag{69}$$

First we minimize over ΔJ and find

$$E_{P+1}(\{\lambda_i\},\{h_i\}) = E_{{\rm GS},P} - \frac{1}{4N}\sum_{i,k,j,l} D^{-1}_{ikjl}\lambda_i\xi^0_k\lambda_j\xi^0_l + V_+(h_1,...,h_k) + \sum_{i=1}^{k}\lambda_i(h_i-\tau_i). \tag{70}$$

The second term on the RHS does not fluctuate with respect to ξ^0 in the thermodynamic limit and we replace it by its average

$$\left\langle \sum_{i,k,j,l} D^{-1}_{ikjl}\ \lambda_i\xi^0_k\ \lambda_j\xi^0_l \right\rangle = \sum_{i,k,j} D^{-1}_{ikjk}\lambda_i\lambda_j. \tag{71}$$

We define

$$\frac{1}{N}\sum_k D^{-1}_{ikjk} = M^{-1}_{ij}$$

and write

$$E_{P+1}(\{\lambda_i\},\{h_i\}) = E_{{\rm GS},P} - \tfrac{1}{4}\sum_{i,j} M^{-1}_{ij}\lambda_i\lambda_j + V_+(h_1,...,h_k) + \sum_{i=1}^{k}\lambda_i(h_i-\tau_i). \tag{72}$$

Eliminating the λ variables we are left with

$$E_{P+1}(\{h_i\}) = E_{{\rm GS},P} + V_+(h_1,...,h_k) + \sum_{i,j} M_{ij}(h_i-\tau_i)(h_j-\tau_j) \tag{73}$$

and

$$E_{{\rm GS},P+1} - E_{{\rm GS},P} = \min_{\bar h}\left[V_+(h_1,...,h_k) + \sum_{i,j} M_{ij}(h_i-\tau_i)(h_j-\tau_j)\right]. \tag{74}$$

As in previous cavity calculations, we find on the RHS the expression F_+ that appears in the corresponding replica calculation.

In analogy with the simple perceptron, the first term on the RHS is the energy associated with ξ^0 and the second term is the minimal increase in the energy of the P patterns if the fields induced by ξ^0 on each of the subneworks are changed from τ_i to h_i by a change of the weights. It is clear that the matrix M is positive definite, since the energy of the P patterns should increase any change of the fields.

If ξ^0 had a $(-)$ output we would arrive at F_- instead F_+. An important point is that the matrices M and S have the same value whether ξ^0 is a $(+)$ or a $(-)$ pattern, since they depend only on the original P patterns.

The minimum point of F_+ is denoted $\bar h^+(\bar\tau)$. These are the fields of ξ^0 after training. The *internal representation* of ξ^0 is given by

$$\sigma_i = {\rm sgn}(h_i^\pm). \tag{75}$$

We see that the initial fields of ξ^0 determine the internal representation that is associated with this pattern. This is a very appealing property from the point of view of real learning algorithms since the choice of the internal representations is a key problem in the training of multilayer perceptrons.

We leave the derivation of the mean-field equations, that define the values $M(\alpha)$ and $S(\alpha)$, to the Appendix and present the expressions for the distribution of fields and the ground-state energy assuming the knowledge of M and S. Using the equivalence of all the $(+)$ patterns we find, in analogy with the simple perceptron,

$$\rho_+(h_1,...,h_k) = \int D_s\tau\prod_{i=1}^{k}\delta\left[h_i - h_i^+(\bar\tau)\right]. \tag{76}$$

A similar expression defines ρ_- . The fields h_i are correlated and their distribution is a *joint* distribution.

The ground-state energy of the whole system is given by

$$E = P_+ \int D_s\tau\, V_+(h_1^+, ..., h_k^+) + P_- \int D_s\tau\, V_-(h_1^-, ..., h_k^-). \quad (77)$$

The mean-field theory of learning views the fully connected multilayer perceptron as *a system of coupled simple perceptrons.* The outputs of these perceptrons are coupled by the potentials $V_\pm$. The coupling is also expressed in the fact that the matrices M and S are nondiagonal. The nondiagonality of these matrices is connected with the fact that each of these perceptrons receives the same set of inputs. This point will be clarified in Sec. III C, where we present the analysis of MLP's with nonoverlapping receptive fields.

C. Nonoverlapping receptive fields

An alternative architecture that has been analyzed recently [13] is the MLP with nonoverlapping receptive fields (NRF). In this architecture, each of the k hidden units receives input from a *different* set of N input units. The input layer consists therefore of kN units. The output is, still, a fixed Boolean function of the hidden units. The replica symmetric free energy is given in this case by

$$G = \alpha_+ \int D_{s_1}\tau_1 \cdots D_{s_k}\tau_k \left[V_+(h_1^+, ..., h_k^+) + \sum_{i=1}^k M_{ii}(h_i^+ - \tau_i)^2 \right] + \alpha_- \int D_{s_1}\tau_1 \cdots D_{s_k}\tau_k \left[V_-(h_1^-, ..., h_k^-) + \sum_{i=1}^k M_{ii}(h_i^- - \tau_i)^2 \right] - \sum_{i=1}^k M_{ii}S_{ii}, \quad (78)$$

where $D_{s_i}\tau_i$ is a Gaussian weight of width square S_{ii} and $\alpha_\pm = P_\pm/N$.

For a fixed vector $\bar{\tau}$, $\bar{h}^\pm$ minimizes $F_\pm$

$$F_\pm(M, \tau) = V_\pm(h_1, ..., h_k) + \sum_{i=1}^k M_{ii}(h - \tau)_i^2, \quad (79)$$

where the potentials $V_\pm$ are derived from the Boolean function in the same way as in the fully connected case.

The free energy of the NRF architecture is identical to the free energy of the corresponding fully connected architecture, except that the order parameter matrices S and M are *diagonal.*

This is clearly explained by the cavity approach. Suppose we train the network with a set of P patterns and find a set of subnetworks $J_1^\star, ..., J_k^\star$. We introduce a new pattern ξ^0 which has initial random fields $\tau_i = J_1^\star \cdot \xi_i^0/\sqrt{N}$. Here ξ_i^0 is the group of N inputs that enters the ith hidden unit. Since ξ_i^0 are random and independent for different i, the initial fields τ_i are random independent Gaussian variables. This explains the diagonality of the matrix S. The diagonality of M results also from the fact that each hidden unit receives a different set of input patterns. If we repeat the steps starting with Eq. (66), we find, in analogy with Eq. (70)

$$E_{P+1}(\{\lambda_i\}, \{h_i\}) = E_{\mathrm{GS},P} - \frac{1}{4N} \sum_{i,k,j,l} D_{ikjl}^{-1} \lambda_i \xi_{ik}^0 \lambda_j \xi_{jl}^0 + V_+(h_1, ..., h_k) + \sum_{i=1}^k \lambda_i (h_i - \tau_i), \quad (80)$$

where

$$D_{ikjl} \equiv \frac{1}{2N} \sum_\mu V_\mu^{ij} \xi_{ik}^\mu \xi_{jl}^\mu$$

and ξ_{jl}^μ is the lth element of the group of inputs that enter hidden unit j. Performing the average over ξ^0

$$\left\langle \sum_{i,k,j,l} D_{ikjl}^{-1}\, \lambda_i \xi_{ik}^0\, \lambda_j \xi_{jl}^0 \right\rangle = \sum_{i,k} D_{ikjk}^{-1} \lambda_i \lambda_i \equiv \sum_i M_{ii}^{-1} \lambda_i^2, \quad (81)$$

we find that only the diagonal terms of M are nonvanishing.

We see that MLP's may be regarded as a set of interacting simple perceptrons. In both architectures (fully connected and NRF) the outputs of these perceptrons (the hidden units) are coupled by the Boolean function. In the fully connected architecture the perceptrons are coupled also in their inputs; this is expressed by the fact that the matrices M and S are nondiagonal while in the NRF network the inputs are decoupled.

D. Interpretation of the solution of the mean-field equations for the XOR and AND machines at the saturation limit

In order to demonstrate our ideas we describe here the solution of the mean-field equations for the fully connected XOR machine [14] and the AND machine [12] in the limit of critical capacity, in view of the cavity approach.

The XOR machine has two hidden units and an output which is their product. The corresponding cost functions are

$$V_\pm(h_1, h_2) = \theta(\mp h_1 h_2). \quad (82)$$

The AND machine has a positive output only if both hidden units are positive. The corresponding potentials are

$$V_+ = 1 - \theta(h_1 h_2) \ , \ V_- = \theta(h_1 h_2). \tag{83}$$

The solution is given in terms of the matrices M and S, whose dimension is 2 in our case, and the functions $\bar{h}^{\pm}(\bar{\tau})$.

In both the XOR and AND machines the Boolean functions are invariant under the permutation of its arguments. This implies that a solution of the form $M_{11} = M_{22}$ and $M_{12} = M_{21}$ exists.

We describe first the mean-field solution of the XOR machine. Approaching the saturation limit from above $\alpha \to \alpha_c^+$ the elements of M should vanish, since the addition of a new pattern should not increase the energy. Given the initial fields $\{\tau_1, \tau_2\}$ of a new (+) pattern, the fields after training $\{h_1^+; h_2^+\}$ are determined by the minimum of

$$F_+ = \theta(-h_1 h_2) + \sum_{i,j=1}^{2} M_{ij}(h-\tau)_i (h-\tau)_j. \tag{84}$$

Since $M_{ij} \to 0$ the minimum point lies in the region in $\bar{h}$ space where $V_+(\bar{h} = 0$, which we denote R_+. For the XOR machines R_+ consists of the first and third quadrants.

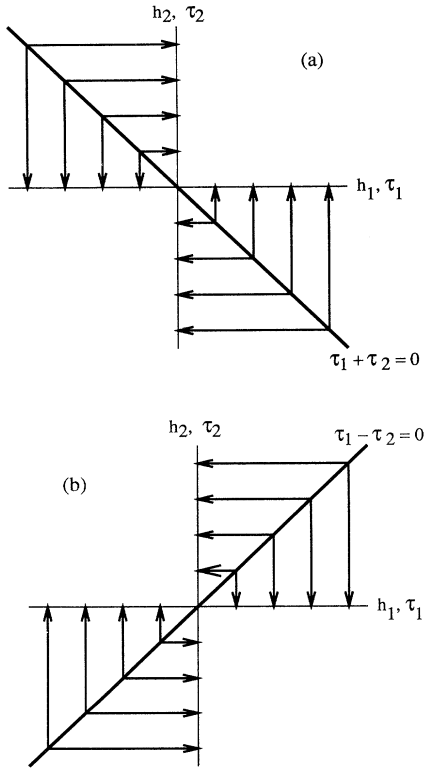

FIG. 1. (a) The optimal training flow of a (+) pattern in the XOR machine at saturation. ($\alpha_+ = \alpha_-$.) A pattern that falls initially in the "wrong" region (R_-) is carried in the learning process to R_+ (the first and third quadrants). The learning flow is to the closer quadrant and only one subnetwork is trained. If the pattern falls initially in R_+ it stays there (no flow, and no arrows). (b) The optimal training of a (−) pattern in the XOR machine.

The exact minimum point is determined by the residual values M_{ij}. The matrix M remains positive definite as we approach α_c, and therefore the quadratic form that appears in F is *concave* in $\bar{h}$. As a result, if $\tau \in R_-$, the minimum point lies on the boundary between R_+ and R_- and its value depends on the ratio M_{12}/M_{11} in the limit $\alpha \to \alpha_c$. Solving the mean-field equations for the XOR machine with an equal number of + and − patterns we find $M_{12}/M_{11} \to 0$. The dependence of $\bar{h}^+$ on $\bar{\tau}$ is shown in Fig. 1(a). The arrows denote the mapping from $\bar{\tau}$ to $\bar{h}^+$. In the R_+ region there are no arrows. The reason is that if $\tau \in R_+$ the concavity of the quadratic form in F implies that $\bar{h}^+ = \bar{\tau}$. The training flow describes the *cooperation* between the subnetworks J_1 and J_2 in the training of a + pattern. If the fields before training are in R_- then only one subnetwork is trained to correct the error. The chosen subnetwork is the one with the shorter Euclidean distance to R_+.

The training map of a (−) pattern is similar and is given in Fig. 1(b). Another result of the mean-field the-

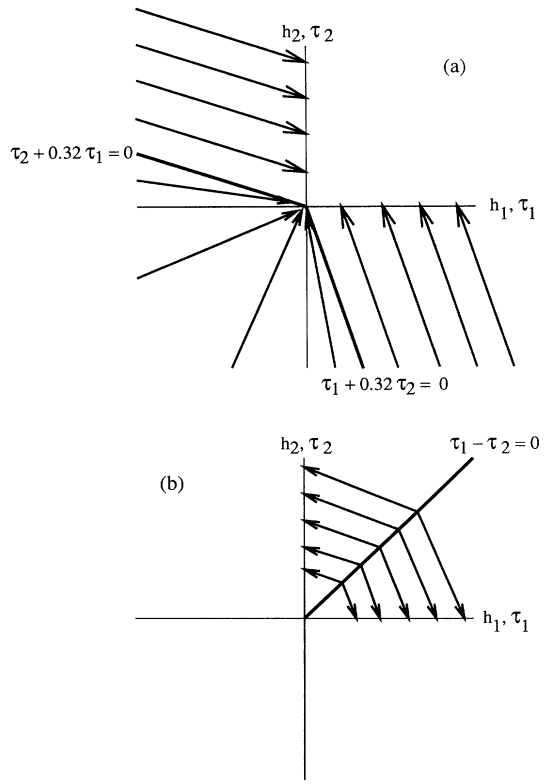

FIG. 2. (a) The optimal training flow of a (+) pattern in the AND machine at saturation. ($\alpha_+ = \alpha_-$.) The suggested learning strategy is nontrivial. A pattern that falls initially in the fourth quadrant is carried to the first quadrant by training the second subnetwork and *untraining* the first subnetwork. This is in contrast to the naive strategy, which is to train only the second subnetwork. (b) The optimal training flow of a (−) pattern in the AND machine. As in the case of a positive patter, the suggested learning strategy is nontrivial, and the learning flow does not take the shortest Euclidean path.

ory is that $S_{12} = 0$ which implies that the optimal subnetworks $J_1^\star, J_2^\star$ are orthogonal. The orthogonality is a result of the the learning strategy in which each of the subnetworks is trained with a *different* subset of the patterns.

A different scenario takes place in the AND machine. In this case R_+ is the first quadrant in $\bar{h}$ space. At the saturation limit, the elements of M vanish. As in the the XOR machine, the exact minimum point depends on the ratio M_{12}/M_{11} in this limit. Solving the mean-field equations (in the case of an equal number of + and − patterns) one finds $M_{12}/M_{11} \to 0.32$. The resulting map of the training of a + pattern $\bar{\tau} \to \bar{h}^+$ is given in Fig. 2(a). The optimal learning strategy that is offered by the mean-field solution is nontrivial. If τ falls in the fourth quadrant ($\tau_1 > 0 \;\; ; \;\; \tau_2 < 0$) then J_2 *learns* the pattern ($h_2^+ > \tau_2$) while J_1 *unlearns* the pattern partially ($h_1^+ < \tau_1$). Mean-field theory suggests a different learning strategy than the naive strategy, which is to train only the "wrong" subnetwork.

Another result of the mean-field theory is that the overlap S_{12} vanishes in the optimal network. This may be explained qualitatively by the learning maps for the (+) and (−) patterns [Figs. 2(a) and 2(b)]. Looking at the maps we see that learning (−) patterns that fall initially in the first quadrant *anti-correlates* the subnetworks. So does the training of + patterns that fall in the second and fourth quadrants. However, the training of a + pattern that falls initially in the third quadrant *correlates* the subnetworks. This leads finally to the result of zero overlap between the subnetworks.

IV. THE FLEXIBILITY OF NETWORKS

Let us consider again the simple perceptron with a norm constraint. Including the constraint, the cost function for a specific set of patterns is given by

$$E(J,\lambda) = \sum_{\mu=1}^{P} V(h^\mu) + \lambda(J\cdot J - N), \qquad h^\mu = \sigma^\mu J\xi^\mu/\sqrt{N}, \tag{85}$$

where λ is a Lagrange multiplier. On the other hand, according to mean-field theory, the pattern-averaged ground-state energy is given by the minimum over λ of $NG(\lambda)$ where

$$NG(\lambda) = P\int Dt \min_h \left[V(h) + \lambda(h-t)^2\right] - N\lambda \tag{86}$$

and we wrote λ instead of $1/2x$.

The similarity between expressions (85) and (86) suggests that in the mean-field theory patterns are represented by a Gaussian distribution of initial stabilities. We reverse now the averaging process. Let us consider a specific set of P patterns ξ^μ, and associate with each pattern a variable t^μ, which is the stability of ξ^μ before training, if it was *the last* pattern to be trained. The mean-field theory assumes that these variables are *independent* (for instance, in our cavity approach we infer from the averaged ground-state energy of a unique pattern to the ground state of the whole system). Therefore, we replace distribution of stabilities Dt in Eq. (86) by the variables t^μ, and write

$$NG(\lambda) = \sum_{\mu=1}^{P} V(h_0^\mu) + \lambda\left[\sum_{\mu=1}^{P}(h_0^\mu - t^\mu)^2 - N\right], \tag{87}$$

where h_0^μ minimizes $V(h^\mu) + \lambda(h^\mu - t^\mu)^2$.

We may write now the energy before the minimization over the h^μ variables

$$NG(\{h^\mu\},\lambda) = \sum_{\mu=1}^{P} V(h^\mu) + \lambda\left(\sum_{\mu=1}^{P}(h^\mu - t^\mu)^2 - N\right). \tag{88}$$

The minimization over the J variables in Eq. (85) is replaced by the minimization over the h^μ variables. Note that the specific values t^μ are unknown. The normalization constraint in Eq. (85) is replaced by the constraint

$$\sum_{\mu=1}^{P}(h^\mu - t^\mu)^2 = N, \tag{89}$$

which we denote the *flexibility* constraint of the network. Mean-field theory tells us that energy is minimized under a constraint on the sum of squares of the deviations of the stabilities after training from the stabilities before training.

We think that for networks with continuous weights, flexibility is an appropriate measure of the network's ability to learn. The question, "How much information can be stored in a network with continuous weights?" is as meaningless as the question, "How much information can we store on a sheet of paper of a given size?" The answer to the second question is that it depends on how small the letters are that we can read. Similarly, the amount of information that can be stored in a network depends on our ability to *interpret* or to *decode* the output. The "size of the letters" is analogous to the amount of flexibility that is needed to teach the network a pattern. A better "decoder" can detect information that has been coded with a smaller amount of learning effort, or, in other words, with a smaller expense of flexibility.

In order to demonstrate the power of the concept of flexibility, we use it to calculate the maximal capacity of the "parity decoder" multilayer perceptron with a nonoverlapping receptive fields architecture (Barkai, Hansel, and Kanter [13]). In this network the output is the *product* of the hidden units. As explained previously, this network may be regarded as a set of k simple perceptrons with coupled outputs. Suppose P patterns have been learned, and we have found k sets of connections $J_1^\star, ..., J_k^\star$ each normalized to N.

We present now a new pattern ξ^0 which should be mapped to +1. As explained previously, its initial fields,

$\tau_l = J_l^* \xi^0/\sqrt{N}$, are normalized independent Gaussian random numbers. There is a probability one-half that ξ^0 will be mapped correctly without training. However, if the product $\prod_i \tau_i$ is negative, an odd number of subnetworks should be trained so that $\prod_l h_l^+$ will be positive.

Each network, being a simple perceptron of norm square N, has total flexibility N. To minimize the flexibility expense in the training of ξ^0 we choose to train one subnetwork, and this is the subnetwork with the smallest $|\tau_l|$. At saturation each subnetwork J_l exhausts its flexibility and obeys the equation

$$\sum_{\mu=1}^{P} (h_l^{+\mu} - t_l^\mu)^2 = N. \tag{90}$$

Or, in the continuous version

$$\alpha \int D\tau_1 \cdots D\tau_l \cdots D\tau_k (h_l^+ - \tau_l)^2 = 1,$$

where $h_l^+ = 0$ (which means that it flipped its sign) if $|t_l|$ is the smallest, and $h_l^+ = \tau_l$ otherwise. So the equation for the lth subnetwork is written

$$\alpha \int_{\mathrm{R}} Dt_1 \cdots Dt_k (t_l)^2 = 1, \tag{91}$$

Where R is the region in the k-dimensional τ space where τ_l is the smallest. We have

$$\int_{\mathrm{R}} D\tau_1 \cdots D\tau_k (t_l)^2 = \int_{-\infty}^{\infty} D\tau\, \tau^2 [2H(|\tau|)]^{k-1} = 2\int_0^{\infty} D\tau\, \tau^2 [2H(\tau)]^{k-1}.$$

This must be multiplied by one-half since one-half of the patterns is mapped correctly without training. Finally we recover the replica-symmetric result for α_c

$$\alpha_c^{-1}(k) = \int_0^{\infty} D\tau\, \tau^2 [2H(\tau)]^{k-1}. \tag{92}$$

In this calculation we assumed that all the patterns have + output, but the same result is obtained if some are − patterns. This is in contrast to the fully connected MLP where the capacity does depend on the ratio between the number of (+) and (−) patterns.

For $k = 2$ we obtain $\alpha_c \simeq 11$, which is much more than the number of patterns that can be stored in two simple perceptrons ($11 > 2 \times 2$). The reason is that the parity decoder is efficient and the training of each pattern consumes less flexibility.

Flexibility is meaningful also for unnormalized models. Repeating the steps (86)–(88) for a perceptron without a norm constraint we find

$$E(\{J_j\}) = \sum_\mu V(h^\mu) \to NG(\{h^\mu\}, \lambda, R) = \sum_\mu V(h^\mu) + \lambda \left(\sum_\mu (h^\mu - \tau^\mu)^2 - NR^2 \right) \tag{93}$$

with initial stabilities τ^μ which have zero mean and variance R^2,

$$\frac{1}{P} \sum_\mu \tau^{\mu 2} = R^2.$$

This expresses the fact that the network can choose its norm, and therefore its flexibility, but on the other hand, the distribution of initial stabilities depends on the norm. The competition between these two factors determines the optimal norm of the network.

Analogously, for multilayer fully connected perceptrons we find the following correspondence:

$$E(J_{ij}) = \sum_{\mu \in P_+} V_+(\bar{h}^\mu) + \sum_{\mu \in P_-} V_-(\bar{h}^\mu)$$

$$\to NG(h_i^\mu, \lambda_{ij}, S_{ij}) = \sum_{\mu \in P_+} V_+(\bar{h}^\mu) + \sum_{\mu \in P_-} V_-(\bar{h}^\mu) + \sum_{ij} \lambda_{ij} \left(\sum_\mu (h_i^\mu - \tau_i^\mu)(h_j^\mu - \tau_j^\mu) - NS_{ij} \right) \tag{94}$$

with τ's that obey

$$\frac{1}{P} \sum_\mu \tau_i^\mu \tau_j^\mu = S_{ij}.$$

As in the previous case, there is a competition which finally determines the values S_{ij}.

V. "CAVITY" MOTIVATED ALGORITHMS

A. A learning algorithm for the simple perceptron

In this section we discuss the possibility of the application of the mean-field learning strategy for the construction of learning algorithms. There are some problems in the application of this idea:

(1) The theory describes the optimal learning process *around the solution* where an algorithm starts far away from the solution.

(2) The function $h_0(x,t)$ tells us the change in the weight vector in the direction of the pattern: $h_0 - t = \Delta J \xi^0 \sigma^0$ but the full change is not known.

(3) The theory describes the learning of a *new* pattern, how can we train a pattern that has been already trained?

We suggest an iterative algorithm which corrects the connections vector in the direction of the learned pattern:

after n steps the set of connections is given by

$$J(n) = \frac{1}{N} \sum_{\mu} y^{\mu}(n) \cdot \xi^{\mu} \sigma^{\mu}. \tag{95}$$

To correct for the pattern ξ^{ν} we define

$$\begin{aligned} t^{\nu}(n) &= \left[J - \frac{1}{n} y^{\nu}(n) \xi^{\nu} \sigma^{\nu} \right] \cdot \xi^{\nu} \sigma^{\nu} \\ &= h^{\nu}(n) - y^{\nu}(n). \end{aligned} \tag{96}$$

The subtraction of the contribution of ξ^{ν} to J, y^{ν}, allows one to treat this pattern as if it is a new one (this is not completely true since the embedding strengths of other patterns have been also affected by the presence of ξ^{ν}, but we did not find a better solution). The initial stability of ξ^{ν} is $t^{\nu}(n)$. We use now the function $h_0(t^{\nu}(n))$ to calculate the stability of ξ^{ν} after training. The function depends on a parameter $x(\alpha)$ which is determined by the mean-field equations. The new embedding strength is given by

$$y^{\nu}(n+1) = h^{\nu}(n) - t^{\nu}(n). \tag{97}$$

Note that the algorithm does not constrain the norm of the weight vector, and therefore corresponds to perceptron models of type B.

To demonstrate this algorithm, we choose

$$V(h) = (h-1)^2 \theta(h-1). \tag{98}$$

Analysis of this model [11] shows that for $\alpha \leq 2$, $x = \infty$ (which means no errors) and

$$h_0(t) = \begin{cases} 1, & t < 1 \\ t, & t > 1 \end{cases}$$

and the corresponding algorithm is

$$y^{\nu}(n+1) = \max\left(1 - [h^{\nu}(n) - y^{\nu}(n)], 0\right) \tag{99}$$

or equivalently

$$h^{\nu}(n+1) = \max\left([h^{\nu}(n) - y^{\nu}(n)], 1\right), \tag{100}$$

which is exactly the "Adatron" algorithm [11], which is a very efficient algorithm which finds the network with the highest minimal stability. It is also shown in [11] that this algorithm is guaranteed to converge. This demonstrates that a mean field theory leads to an efficient algorithm.

B. Learning algorithms for MLP's

We tried to construct a multilayer perceptron algorithm which is based on the mean-field theory. We consider a fully connected network with k hidden units. In analogy with the previous algorithm we write

$$J_i(n) = \frac{1}{N} \sum_{\mu} y_i^{\mu}(n) \xi^{\mu}, \tag{101}$$

where J_i is the ith subnetwork after n iterations.

We calculate now the fields of ξ^{ν} "before training"

$$\tau_i^{\nu}(n) = \left[J_i - \frac{1}{N} y_i^{\nu}(n) \xi^{\nu} \right] \xi^{\nu} \tag{102}$$

and use mean-field theory to calculate $\bar{h}^{\pm}(\bar{\tau}^{\nu})$ and the new embedding strengths

$$y_i^{\nu}(n+1) = h_i^{\pm}(\bar{\tau}^{\nu}(n)) - \tau_i^{\nu}(n). \tag{103}$$

We applied this algorithm to the AND machine, and compared is performance with the "minimal disturbance" algorithm, which is the algorithm which chooses to make the minimal correction needed.

It turns out that there is no essential difference in the maximal capacity achieved by these algorithms. For $N = 200$ we obtain $\alpha_c \simeq 3.0$, which is a bit less than the capacity achieved by the stochastic algorithm described in [12], $\alpha_c \simeq 3.3$. Although the mean-field algorithm does not show a clear advantage in this example, there is still room for further investigation.

VI. A SHORT SUMMARY AND DISCUSSION

In this work we suggest a unifying view of the mean-field theories of learning of simple and multilayer perceptrons. According to this view, which is obtained by the cavity approach, mean-field theory describes a learning process. This learning process is represented by a functional relation between the stability (or fields, in the case of MLP's) of a pattern before and after learning. The function is controlled by order parameters which are determined self-consistently by the mean-field equations. Since the order parameters are chosen optimally, the resulting "learning flow" is optimal with respect to the prescribed cost function.

Motivated by this interpretation, we try to construct new learning algorithms. In the case of a simple perceptron, we arrive at the already known and very efficient Adatron algorithm. We also suggest algorithms for MLP's. In particular, we investigate the training of a fully connected AND machine, which has a nontrivial optimal learning flow according to mean-field theory. Comparison with a "naive" algorithm shows no clear advantage to our new algorithm. However, since the correspondence between the mean-field theory and the algorithm is not unique, we think that there is still a place for further investigation in this direction.

Another consequence of the cavity approach is the concept of flexibility, which serves as a measure of the ability of a perceptron to change during the learning process. We demonstrate this concept by calculating the (replica symmetric) capacity of a multilayer network (nonoverlapping receptive fields XOR machine) using the principal of minimal flexibility expense of the simple perceptrons which construct the network.

We would like to stress that the concept of flexibility is a result of the assumption of a unique ground state, and that replica-symmetry-breaking effects alter the capacity of the multilayered network discussed above [13]. The inclusion of these effects in the framework of a cavity approach is a subject of future research. However, it would be also interesting to see whether our simplified

mean-field theory can be of use in constructing learning algorithms also for network architectures which are known to have many ground states.

ACKNOWLEDGMENTS

The author acknowledges discussions with S. Seung, M. Tsodyks, J-P Nadal, and especially H. Gutfreund. This work is supported by the Levi Eshkol Foundation of the Israeli Ministry of Science.

APPENDIX: THE MEAN-FIELD EQUATIONS OF THE FULLY CONNECTED MLP

We start again with the expression for the zero-temperature free energy G [Eq. (62)]. It is more convenient to change variables in the following way: Define a new symmetric matrix B of size k which is the square root of S. S is a correlation matrix, and therefore positive definite and has a square root. In fact, there are many roots to S but a unique *symmetric* root. Note that since B is symmetric, the number of order parameters is conserved. In terms of the matrices B and M the free energy is

$$G = \alpha_+ \left[\int Dt_1 \cdots Dt_k V_+(h_1^+, ..., h_k^+) + \sum_{i,j=1}^{k} M_{ij}(h^+ - Bt)_i (h^+ - Bt)_j \right]$$

$$+\alpha_- \left[\int Dt_1 \cdots Dt_k V_+(h_1^-, ..., h_k^-) + \sum_{i,j=1}^{k} M_{ij}(h^- - Bt)_i (h^- - Bt)_j \right] - \sum_{i,j=1}^{k} M_{ij} B_{ij}^2. \tag{A1}$$

Dt_i are normalized Gaussian weights and the variables τ_i have been replaced by $(Bt)_i$ which have the same correlations

$$\int Dt_1 \cdots Dt_k (Bt)_i (Bt)_j = B_{ij}^2 = S_{ij}.$$

The number of independent order parameters is $k(k+1)$. Differentiating with respect to M_{ij} we get for each ij an equation

$$\alpha_+ \int Dt_1 \cdots dt_k (h^+ - Bt)_i (h^+ - Bt)_j$$

$$+\alpha_- \int Dt_1 \cdots Dt_k (h^- - Bt)_i (h^- - Bt)_j = B_{ij}^2 = S_{ij}. \tag{A2}$$

These are equivalent to the first replica equation for model B. The second group of equations may be written

$$\alpha_+ \int Dt_1 \cdots dt_k \; t_l (h^+ - Bt)_j$$

$$+\alpha_- \int Dt_1 \cdots Dt_k \; t_l (h^+ - Bt)_j = -B_{lj}, \tag{A3}$$

which is equivalent to the second replica equation for model B.

We turn now to the derivation of the mean-field equations by the cavity approach. In terms of the variables M and B the ground-state energy of a system of $P = P_+ + P_-$ patterns is given by [see Eq. (77)]

$$E = P_+ \int Dt_1 \cdots Dt_k V_+(h_1^+, ..., h_k^+)$$

$$+P_- \int Dt_1 \cdots Dt_k V_-(h_1^-, ..., h_k^-), \tag{A4}$$

where $\bar{h}^+(M, Bt)$ [$\bar{h}^-(M, Bt)$] minimize F_+ [F_-] for a given vector of initial fields Bt

$$F_\pm(M, Bt, \bar{h}) = V_\pm(h_1, ..., h_k) + \sum_{i,j=1}^{k} M_{ij}(h - Bt)_i (h - Bt)_j. \tag{A5}$$

Each of the elements of M and B is a function of *both* P_+ and P_-, or, more precisely, of α_+ and α_-. We begin with the derivation of the second cavity equation. Since the set of subnetworks $J_1^\star, ..., J_k^\star$ should minimize the energy, we require that under rescaling of each of the subnetworks $\{J_i^\star \to sJ_i^\star\}$ the energy is stationarry. Therefore we derive k equations that are equivalent to the second cavity equation for model B.

$$P_+ \int Dt_1 \cdots Dt_k \; h_i^+ \; V_+^i(h_1^+, ..., h_k^+)$$

$$+P_- \int Dt_1 \cdots Dt_k \; h_i^- \; V_-^i(h_1^+, ..., h_k^+) = 0, \tag{A6}$$

$$V_\pm^i = \frac{\partial V_\pm}{\partial h_i}.$$

Using $F_{\pm}$ we write this equation in the form

$$P_+ \int Dt_1 \cdots Dt_k \; h_i^+ \sum_j M_{ij}(h^+ - Bt)_j$$

$$+P_- \int Dt_1 \cdots Dt_k \; h_i^- \sum_j M_{ij}(h^- - Bt)_j = 0. \quad \text{(A7)}$$

We regard the elements of the matrix M as independent variables, and we require the coefficient of each of these elements to vanish. Therefore, for every (ij) we have

$$P_+ \int Dt_1 \cdots Dt_k h_i^+ (h^+ - Bt)_j$$

$$+P_- \int Dt_1 \cdots Dt_k h_i^- (h^- - Bt)_j = 0. \quad \text{(A8)}$$

The first group of cavity equations is more complicated to derive. We add a $(+)$ pattern to a system of $P = P_+ + P_-$ patterns. It is straightforward to show that in analogy with previous calculations, the self-consistency equation is

$$P_+ \frac{d}{dP_+} \int Dt_1 \cdots Dt_k V_+(h_1^+, ..., h_k^+)$$

$$+P_- \frac{d}{dP_+} \int Dt_1 \cdots Dt_k V_-(h_1^+, ..., h_k^+)$$

$$= \int Dt_1 \cdots Dt_k \sum_{i,j=1}^{k} (h^+ - Bt)_i M_{ij} (h^+ - Bt)_j. \quad \text{(A9)}$$

Adding a $(-)$ pattern instead of a $(+)$ pattern we get

$$P_+ \frac{d}{dP_-} \int Dt_1 \cdots Dt_k V_+(h_1^+, ..., h_k^+)$$

$$+P_- \frac{d}{dP_-} \int Dt_1 \cdots Dt_k V_-(h_1^+, ..., h_k^+)$$

$$= \int Dt_1 \cdots Dt_k \sum_{ij=1}^{k} (h^- - Bt)_i M_{ij} (h^- - Bt)_j. \quad \text{(A10)}$$

The first cavity equation is obtained if we change the number of patterns but conserve the ratio P_+/P_-. We define

$$\frac{d}{dP} \equiv \frac{\alpha_+}{\alpha} \frac{d}{dP_+} + \frac{\alpha_-}{\alpha} \frac{d}{dP_-}.$$

This is a derivative in a direction of constant ratio. Multiplying Eq. (A9) by α_+/α and Eq. (A10) by α_-/α we have

$$P_+ \frac{d}{dP} \int Dt_1 \cdots Dt_k V_+ + P_- \frac{d}{dP} \int Dt_1 \cdots Dt_k V_-$$

$$= \frac{\alpha_+}{\alpha} \int Dt_1 \cdots Dt_k \sum_{i,j=1}^{k} (h^+ - Bt)_i M_{ij} (h^+ - Bt)_j$$

$$+ \frac{\alpha_-}{\alpha} \int Dt_1 \cdots Dt_k \sum_{i,j=1}^{k} (h^- - Bt)_i M_{ij} (h^- - Bt)_j. \quad \text{(A11)}$$

We use again $F_{\pm}$ to write the equation in the form

$$\sum_{i,j} M_{ij} \Big(-P_+ \left[\frac{d}{dP} - \frac{dB}{dP} \frac{d}{dB} \right] \int Dt_1 \cdots Dt_k (h^+ - Bt)_i (h^+ - Bt)_j$$

$$-P_- \left[\frac{d}{dP} - \frac{dB}{dP} \frac{d}{dB} \right] \int Dt_1 \cdots Dt_k (h^- - Bt)_i (h^- - Bt)_j \Big)$$

$$= \frac{\alpha_+}{\alpha} \int Dt_1 \cdots Dt_k \sum_{i,j=1}^{k} (h^+ - Bt)_i M_{ij} (h^+ - Bt)_j + \frac{\alpha_-}{\alpha} \int Dt_1 \cdots Dt_k \sum_{i,j=1}^{k} (h^- - Bt)_i M_{ij} (h^- - Bt)_j, \quad \text{(A12)}$$

where

$$\frac{dB}{dP} \frac{d}{dB} \equiv \sum_{i,j} \frac{dB_{ij}}{dP} \frac{d}{dB_{ij}}.$$

We require that the equality will hold for every M_{ij}. Therefore for every (ij) we have

$$-P_+ \left[\frac{d}{dP} - \frac{dB}{dP} \frac{d}{dB} \right] \int Dt_1 \cdots Dt_k (h^+ - Bt)_i (h^+ - Bt)_j - P_- \left[\frac{d}{dP} - \frac{dB}{dP} \frac{d}{dB} \right] \int Dt_1 \cdots Dt_k (h^- - Bt)_i (h^- - Bt)_j$$

$$= \frac{\alpha_+}{\alpha} \int Dt_1 \cdots Dt_k (h^+ - Bt)_i (h^+ - Bt)_j + \frac{\alpha_-}{\alpha} \int Dt_1 \cdots Dt_k (h^- - Bt)_i (h^- - Bt)_j. \quad \text{(A13)}$$

Since the second equation [Eq. (A8)] holds for every P, we have

$$\frac{d}{dP}\left[P_+\int Dt_1\cdots Dt_k h_i^+(h^+-Bt)_j + P_-\int Dt_1\cdots Dt_k h_i^-(h^--Bt)_j\right]=0. \tag{A14}$$

From this equation and, again, Eq. (A8) we derive the equation

$$P_+\frac{d}{dP}\int Dt_1\cdots Dt_k h_i^+(h^+-Bt)_j + P_-\frac{d}{dP}\int Dt_1\cdots Dt_k h_i^-(h^--Bt)_j=0. \tag{A15}$$

Using Eq. (A15) in the LHS of Eqs. (A13) and (A8) in the RHS we find

$$-P_+\frac{d}{dP}\int Dt_1\cdots Dt_k(-Bt)_i(h^+-Bt)_j + P_+\frac{dB}{dP}\frac{d}{dB}\int Dt_1\cdots Dt_k(h^+-Bt)_i(h^+-Bt)_j$$

$$-P_-\frac{d}{dP}\int Dt_1\cdots Dt_k(-Bt)_i(h^--Bt)_j + P_-\frac{dB}{dP}\frac{d}{dB}\int Dt_1\cdots Dt_k(h^--Bt)_i(h^--Bt)_j$$

$$=\frac{\alpha_+}{\alpha}\int Dt_1\cdots Dt_k(-Bt)_i(h^+-Bt)_j+\frac{\alpha_-}{\alpha}\int Dt_1\cdots Dt_k(-Bt)_i(h^--Bt)_j. \tag{A16}$$

Let us define

$$A_{ij}\equiv\frac{\alpha_+}{\alpha}\int Dt_1\cdots Dt_k t_i(h^+-Bt)_j+\frac{\alpha_-}{\alpha}\int Dt_1\cdots Dt_k t_i(h^--Bt)_j. \tag{A17}$$

Then Eq. (A16) is written in a matrix form:

$$\alpha BA'-\alpha AB'=-BA, \tag{A18}$$

where the prime denotes a derivative with respect to α.

Using Eq. (A8) we can show that

$$\frac{\alpha_+}{\alpha}\int Dt_1\cdots Dt_k(-Bt)_i(h^+-Bt)_j+\frac{\alpha_-}{\alpha}\int Dt_1\cdots Dt_k(-Bt)_i(h^--Bt)_j$$

$$=\frac{\alpha_+}{\alpha}\int Dt_1\cdots Dt_k(h^+-Bt)_i(-Bt)_j+\frac{\alpha_-}{\alpha}\int Dt_1\cdots Dt_k(h^--Bt)_i(-Bt)_j. \tag{A19}$$

This means that A and B commute,

$$AB=BA. \tag{A20}$$

Since A and B commute for every α, we assume

$$A_{ij}(\alpha)=B_{ij}(\alpha)f(\alpha), \tag{A21}$$

Plugging this ansatz into Eq. (A18) we find that f obeys

$$-\alpha f'=f$$

and the solution is $f(\alpha)=C/\alpha$ with some constant C. So the first cavity equation for every (ij) is

$$\alpha_+\int Dt_1\cdots Dt_k t_i(h^+-Bt)_j$$

$$+\alpha_-\int Dt_1\cdots Dt_k t_i(h^--Bt)_j=CB_{ij}. \tag{A22}$$

Combining the first and second cavity equations we find Eq. (A2) up to a constant C,

$$\alpha_+\int Dt_1\cdots dt_k(h^+-Bt)_i(h^+-Bt)_j$$

$$+\alpha_-\int Dt_1\cdots Dt_k(h^--Bt)_i(h^--Bt)_j=CS_{ij}. \tag{A23}$$

A straightforward generalization of the heuristic argument in model A, which neglects the overlaps among the patterns and regards Eq. (A23) as the condition

$$\frac{1}{N}J_i^\star\cdot J_j^\star=S_{ij},$$

explains "why" $C=1$.

* Present address: Ecole Supérieure de Physique et de Chimie Industrielles de la Ville de Paris, 10 rue Vauquelin, 75005 Paris, France.

[1] E. Gardner, J. Phys. A **21**, 257 (1988).
[2] E. Gardner and B. Derrida, J. Phys. A **21**, 271 (1988).
[3] E. Gardner, J. Phys. A **22**, 1969 (1989).
[4] See, for example, Ref. [5] and references therein.
[5] M. Mezard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[6] M. Mezard, J. Phys. A **22**, 2181 (1989).
[7] M. Griniasty and H. Gutfreund, J. Phys. A **24**, 715 (1991).
[8] T. Cover, IEEE Trans. Electron. Comput. **14**, 326 (1965).
[9] M. Opper, Europhys. Lett. **8**, 3824 (1989).
[10] See, for example, A. Crisanti and H. Sompolinsky, Phys. Rev. A **36**, 4922 (1987).
[11] J.K. Anlauf and M. Biehl, Europhys. Lett. **10**, 687 (1989).
[12] M. Griniasty and T. Grossman, Phys. Rev. A. **45**, 8924 (1992).
[13] E. Barkai, D. Hansel, and I. Kanter, Phys. Rev. Lett. **65**, 2312 (1990).
[14] M. Mezard and S. Patarnello (unpublished).