Correlations between hidden units in multilayer neural networks and replica symmetry breaking

D. Malzahn and A. Engel
Institut für Theoretische Physik, Otto-von-Guericke-Universität, Postfach 4120, D-39016 Magdeburg, Federal Republic of Germany
(Received 19 February 1999)

We consider feed-forward neural networks with one hidden layer, tree architecture, and a fixed hidden-to-output Boolean function. Focusing on the saturation limit of the storage problem the influence of replica symmetry breaking on the distribution of local fields at the hidden units is investigated. These field distributions determine the probability of finding a specific activation pattern of the hidden units as well as the corresponding correlation coefficients and therefore quantify the division of labor among the hidden units. We find that although modifying the storage capacity and the distribution of local fields markedly replica symmetry breaking has only a minor effect on the correlation coefficients. Detailed numerical results are provided for the PARITY, COMMITTEE, and AND machines with $K=3$ hidden units and nonoverlapping receptive fields.
[S1063-651X(99)10507-5]

PACS number(s): 87.18.Sn, 05.20.−y

I. INTRODUCTION

Multilayer neural networks (MLN) are more powerful devices for information processing than the single-layer perceptron because of the possibility of *different* activation patterns, so-called internal representations (IR), at the hidden units for the *same* input-output mapping. It is well known that the correlations between the activities at the hidden units are crucial for the understanding of the storage and generalization properties of a MLN [1–6]. A particular simple situation to study these correlations is the implementation of random input-output mappings by the network, the so-called storage problem, near the storage capacity. Using the replica trick and assuming replica symmetry the correlation coefficients building up in this case were calculated in [6] and shown to be characteristic for the prewired Boolean function between hidden layer and output. Conversely, *prescribing* these correlations the storage properties of the networks change [7].

The assumption of replica symmetry (RS) in this calculation is somewhat doubtful. In fact it is well known that the storage capacity of MLN is strongly modified by replica symmetry breaking (RSB) [8–10], which is due to the very possibility of different internal representations. Moreover, even the distribution of the output field of a simple perceptron is influenced by RSB effects [11,12].

In the present paper we elucidate the impact of RSB on the correlation coefficients between the activity of different hidden units in MLN with one hidden layer and nonoverlapping receptive fields. The central quantity of interest is the joint probability distribution for the local fields at the hidden units. In the general part of this paper we show how this distribution can be calculated both in RS and in one-step RSB. For a detailed analysis we then specialize to MLN with $K=3$ hidden units and discuss, in particular, the PARITY, COMMITTEE and AND machines. Together with the corrections from one-step RSB the RS results give insight into the division of labor between different subperceptrons in MLN and the role of RSB. Calculating finally the correlation coefficients we find that although modifying the local field distribution markedly RSB gives rise to minor corrections to the correlation coefficients only.

II. GENERAL RESULTS

We consider feed-forward neural networks with N inputs $\boldsymbol{\xi}_k^\nu$, one hidden layer of K units $\tau_1,\tau_2,\ldots,\tau_K$, and a single output σ. The hidden units have nonoverlapping receptive fields of dimension N/K (tree structure). They are determined by the inputs via spherical coupling vectors $\mathbf{J}_k \in \mathbf{IR}^{N/K}, \mathbf{J}_k^2=N/K$ according to $\tau_k=\mathrm{sgn}(h_k)$ with $h_k=\mathbf{J}_k\boldsymbol{\xi}_k\sqrt{K/N}$ denoting the local fields. We call an activation pattern $(\tau_1^\nu,\tau_2^\nu,\ldots,\tau_K^\nu)$ of the hidden units an internal representation (IR). The output σ of the MLN is a fixed Boolean function $\sigma=F(\tau_1,\ldots,\tau_K)$ of the IR. Examples of special interest include the PARITY machine, $F(\{\tau_k^\nu\})=\Pi_{k=1}^K\tau_k^\nu$, the COMMITTEE machine, $F(\{\tau_k^\nu\})=\mathrm{sgn}(\Sigma_{k=1}^K\tau_k^\nu)$, and the AND machine, $F=+1$ if all $\tau_k=+1$; else $F=-1$.

All IR consistent with a desired output are called legal internal representations (LIR). The number of and similarity between LIR to a given output specifies the division of labor taking place between the different perceptrons forming the MLN. It is quantitatively characterized by the correlation coefficients

$$c_n=\langle\langle \sigma\tau_{i_1}\tau_{i_2}\cdots\tau_{i_n}\rangle\rangle, \tag{1}$$

$n=1,\ldots,K$, where $\langle\langle\cdots\rangle\rangle$ denotes the average over the inputs and the output and $i_1,\ldots,i_n$ is a subset of n natural numbers between 1 and K. For permutation symmetric Boolean functions, the c_n only depend on n and not on the particular choice of this subset.

We focus on the so-called storage problem in which the inputs $\boldsymbol{\xi}_k^\nu$ and the outputs σ^ν are generated independently at random according to the probability distributions

$$p(\sigma^\nu)=\frac{\delta(\sigma^\nu-1)+\delta(\sigma^\nu+1)}{2} \tag{2}$$

and

$$p(\xi_{k,i}^\nu)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(\xi_{k,i}^\nu)^2\right), \tag{3}$$

where $k=1,\ldots,K$, $i=1,\ldots,N/K$, and $\nu=1,\ldots,\alpha N$.

The basic quantity which gives us access to the probability of the LIR and to the correlation coefficients is the distribution $p(h_j)$ of the local fields h_j at the jth hidden unit. It is given by

$$p(h_j)=\left\langle\left\langle \frac{1}{\mathcal{Z}}\int\prod_{k=1}^{K} d\mu(\mathbf{J}_k)d\mu(\lambda_k^\nu) \times\prod_{\nu=1}^{\alpha N}\Theta(\sigma^\nu F(\mathrm{sgn}(\lambda_1^\nu),\ldots,\mathrm{sgn}(\lambda_K^\nu))) \times\delta(h_j-\lambda_j^1)\right\rangle\right\rangle_{\{\boldsymbol{\xi}_k^\nu\},\sigma^\nu}. \tag{4}$$

$\langle\langle\ldots\rangle\rangle$ denotes the average over all stored input-output patterns. $\mathcal{Z}$ denotes the partition function

$$\mathcal{Z}=\int\prod_{k=1}^{K} d\mu(\mathbf{J}_k)d\mu(\lambda_k^\nu) \times\prod_{\nu=1}^{\alpha N}\Theta(\sigma^\nu F(\mathrm{sgn}(\lambda_1^\nu),\ldots,\mathrm{sgn}(\lambda_K^\nu))), \tag{5}$$

$d\mu(\mathbf{J}_k)$ the measure on the Gardner sphere [13]

$$d\mu(\mathbf{J}_k)=\delta\left(\mathbf{J}_k^2-\frac{N}{K}\right)\frac{d\mathbf{J}_k}{\sqrt{2\pi e^{N/K}}}, \tag{6}$$

and $d\mu(\lambda_k^\nu)$ the integration measure

$$d\mu(\lambda_k^\nu)=\delta\left(\lambda_k^\nu-\mathbf{J}_k\boldsymbol{\xi}_k^\nu\sqrt{\frac{K}{N}}\right)d\lambda_k^\nu. \tag{7}$$

We use the replica trick $1/\mathcal{Z}=\lim_{n\to 0}\mathcal{Z}^{n-1}$ in Eq. (4) to perform the average over the inputs $\{\boldsymbol{\xi}_k^\nu\}$ and introduce the overlaps $q_k^{ab}=\mathbf{J}_k^a\mathbf{J}_k^b/(N/K)$ between different replicas a,b of a coupling vector $\mathbf{J}_k$ of hidden unit k. We will consider only permutation symmetric Booleans F. Hence all hidden units have the same statistical properties implying $p(h_k)=p(h)$ and $q_k^{ab}=q^{ab}$ with $k=1,\ldots,K$. Equation (4) takes on the form

$$p(h_j)=\lim_{n\to 0}\int\prod_{a<b} dq^{ab}\langle\langle p(h_j|\sigma)\rangle\rangle_\sigma\exp[(N/2)\ln\det(Q)+(\alpha N-1)\langle\langle\ln G_1(Q|\sigma)\rangle\rangle_\sigma] \tag{8}$$

in terms of the $(n\times n)$-dimensional order parameter matrix Q where $Q^{aa}=1$ and $Q^{ab}=q^{ab}$. Here

$$p(h_j|\sigma)=\int\prod_{k,a}\frac{d\lambda_k^a dx_k^a}{2\pi}\exp\left(\sum_{k,a}\left[ix_k^a\lambda_k^a-\frac{1}{2}(x_k^a)^2\right]-\sum_{k,a<b}x_k^a x_k^b q^{ab}\right)\prod_a\Theta(\sigma F(\{\mathrm{sgn}(\lambda_k^a)\}))\,\delta(h_j-\lambda_1^1), \tag{9}$$

and the expression for $G_1(Q|\sigma)$ is specified in the Appendix, Eq. (A6) together with some more details of the calculation.

In the limit $N\to\infty$ the integral (8) is dominated by the saddle point values of the order parameters q^{ab} which extremize the partition function

$$\mathcal{Z}=\exp\left(N\ \mathrm{extr}_{q^{ab}}\left\{\lim_{n\to 0}\frac{\frac{1}{2}\ln\det(Q)+\alpha\langle\langle\ln G_1(Q|\sigma)\rangle\rangle_\sigma}{n}\right\}\right). \tag{10}$$

In the following, we simplify Eqs. (9) and (10) using the assumption that the order parameter matrix Q is either replica symmetric or describes one-step replica symmetry breaking. We will always consider the saturation limit $\alpha\to\alpha_c$ since the expressions then simplify and the correlations become most characteristic in this limit. The RS case is specified by [14]

$$q^{ab}=\begin{cases}1 & \text{if } a=b\\ q & \text{else.}\end{cases} \tag{11}$$

The saturation limit $\alpha\to\alpha_c$ is characterized by the existence of a unique solution $\mathbf{J}_k$, e.g., $q\to 1$. We then get

$$p(h|\sigma)=\int\prod_{k=1}^{K}Dy_k\lim_{q\to 1}\left(\frac{\exp\left(-\frac{1}{2}(h+y_1\sqrt{q})^2/(1-q)\right)}{\sqrt{2\pi(1-q)}}\frac{\Phi_{\mathrm{LIR}}(\sigma|\delta_{\eta_1,\mathrm{sgn}(h)})}{\Phi_{\mathrm{LIR}}(\sigma)}\right) \tag{12}$$

for the conditional probability to find a specific value h of the postsynaptical potential under the constraint of a given output σ. The terms abbreviated by

$$\Phi_{\mathrm{LIR}}(\sigma|\delta_{\eta_1,\mathrm{sgn}(h)}) := \sum_{\text{all sets}(\eta_1,\ldots,\eta_K)} \delta_{\eta_1,\mathrm{sgn}(h)}\delta_{\sigma,F(\eta_1,\ldots,\eta_K)}\prod_{k=2}^{K} H\left(\eta_k y_k\sqrt{\frac{q}{1-q}}\right), \tag{13}$$

$$\Phi_{\mathrm{LIR}}(\sigma) := \sum_{\text{all sets}(\eta_1,\ldots,\eta_K)} \delta_{\sigma,F(\eta_1,\ldots,\eta_K)}\prod_{k=1}^{K} H\left(\eta_k y_k\sqrt{\frac{q}{1-q}}\right) \tag{14}$$

ensure that only LIR for the respective value of σ contribute to the sum in Eq. (12). As usual we have used the error function $H(x)=\int_x^\infty Dt$ with $Dt=\exp(-t^2/2)dt/\sqrt{2\pi}$.

Let us now turn to main features of the solution within the ansatz of one-step RSB. Then the following form for the order parameter matrix is assumed [14]:

$$q^{ab}=\begin{cases}1 & \text{if } a=b\\ q_1 & \text{if } |a-b|<m\\ q_0 & \text{else.}\end{cases} \tag{15}$$

Accordingly, there are two overlap scales characterizing the similarity between coupling vectors belonging to the same and different regions of the solution space, respectively.

Using this ansatz we find after standard manipulations [14] for the probability distribution of the local field for a specific output σ,

$$p(h|\sigma)=\int\prod_{k=1}^{K} Dy_k \frac{\displaystyle\int\prod_{k=1}^{K} Dz_k \frac{1}{\sqrt{2\pi(1-q_1)}}\exp\left(-\frac{(h+y_1\sqrt{q_0}+z_1\sqrt{q_1-q_0})^2}{2(1-q_1)}\right)\frac{\Phi_{\mathrm{LIR}}(\sigma|\delta_{\eta_1,\mathrm{sgn}(h)})}{\left(\Phi_{\mathrm{LIR}}(\sigma)\right)^{1-m}}}{\displaystyle\int\prod_{k=1}^{K} Dz_k\left[\Phi_{\mathrm{LIR}}(\sigma)\right]^m}, \tag{16}$$

where now

$$\Phi_{\mathrm{LIR}}(\sigma|\delta_{\eta_1,\mathrm{sgn}(h)}) := \sum_{\text{all sets}(\eta_1,\ldots,\eta_K)} \delta_{\eta_1,\mathrm{sgn}(h)}\delta_{\sigma,F(\eta_1,\ldots,\eta_K)}\prod_{k=2}^{K} H\left(\eta_k\frac{y_k\sqrt{q_0}+z_k\sqrt{q_1-q_0}}{\sqrt{1-q_1}}\right), \tag{17}$$

$$\Phi_{\mathrm{LIR}}(\sigma) := \sum_{\text{all sets}(\eta_1,\ldots,\eta_K)} \delta_{\sigma,F(\eta_1,\ldots,\eta_K)}\prod_{k=1}^{K} H\left(\eta_k\frac{y_k\sqrt{q_0}+z_k\sqrt{q_1-q_0}}{\sqrt{1-q_1}}\right). \tag{18}$$

These expressions simplify in the saturation limit $\alpha\to\alpha_c$ in which one finds $q_1\to 1$ and $m=w(1-q_1)\to 0$. The remaining order parameters w,q_0 are given by the saddle point equations corresponding to the following expression for the storage capacity α_c:

$$\alpha_c=\min_{q_0,w}\left[\frac{\ln[1+w(1-q_0)]+q_0 w/[1+w(1-q_0)]}{-2\lim\limits_{q_1\to 1}\left\langle\left\langle\int\prod_k Dy_k\ln\left\{\int\prod_k Dz_k\left(\Phi_{\mathrm{LIR}}(\sigma)\right)^m\right\}\right\rangle\right\rangle_\sigma}\right]. \tag{19}$$

As in the RS case the analytical and numerical analysis of these expressions for concrete situations needs some care (see next section).

To finally obtain $p(h)$ we must average Eqs. (12) and (16) over the two possible outputs $\sigma=\pm 1$,

$$p(h)=\langle\langle p(h|\sigma)\rangle\rangle_\sigma. \tag{20}$$

From this probability distribution we find the distributions $p(\tau_1,\ldots,\tau_K)$ of the LIR according to

$$p(\tau_1,\ldots,\tau_K)=\int_{-\infty}^{\infty}\prod_{k=1}^{K} dh_k\Theta(\tau_k h_k)p(h_k). \tag{21}$$

The correlation coefficients c_n, $n=1,\ldots,K$ are then given by

$$c_n=\sum_{\text{all sets}(\eta_1,\ldots,\eta_K)}\sigma\eta_1\eta_2\cdots\eta_n\delta_{\sigma,F(\eta_1,\eta_2,\ldots,\eta_K)} \times p(\eta_1,\eta_2,\ldots,\eta_K). \tag{22}$$

The Kronecker δ in Eq. (22) restricts the sum to all LIR of the output σ. Equation (22) is valid as long as the pattern load of the MLN does not exceed its saturation threshold α_c.

III. SPECIFIC EXAMPLES WITH $K=3$ HIDDEN UNITS

In this section we apply the general formalism developed above to the analysis of simple versions of three popular examples of MLN, namely, the PARITY, COMMITTEE, and AND machines, each with $K=3$ hidden units. We start with the RS results.

A. Replica symmetry

In COMMITTEE and PARITY machines there is for every LIR of output $\sigma=+1$ an IR with all signs reversed that realizes output $\sigma=-1$. Therefore $p(h)=p(h|+1)=p(h|-1)$ and the final average over σ in Eq. (20) is trivial. Analyzing Eqs. (13) and (14) in the limit $q\to 1$ one realizes that they depend on both the sign and values of all integration variables y_k. Expression (13) as well as Eq. (14) are either equal to one or exponentially small in some or all integration variables. The quotient of both figuring in Eq. (12) can hence become one, zero, or singular with respect to y_1. Whenever it is one the integral in Eq. (12) gives rise to $\delta(h+y_1)$ for $q\to 1$. Whenever the quotient is singular a contribution $\delta(h)$ results.

Keeping track of the different contributions arising in this way we find for the $K=3$ COMMITTEE machine

$$p(h)=\Theta(-h)\frac{e^{-h^2/2}}{\sqrt{2\pi}}H^2(h)+\frac{5}{24}\delta_+(h)+\Theta(h)\frac{e^{-h^2/2}}{\sqrt{2\pi}} \tag{23}$$

and for the PARITY machine

$$p(h)=\frac{1}{2}\frac{e^{-h^2/2}}{\sqrt{2\pi}}+\frac{e^{-h^2/2}}{\sqrt{2\pi}}\,4\int_0^{|h|}DtH(t)+\frac{1}{12}\delta_-(h)+\frac{1}{12}\delta_+(h). \tag{24}$$

Note that $p(h)$ for the PARITY machine is an even function due to the additional symmetry of the Boolean function F for this case.

In the AND machine the output $\sigma=+1$ can be realized by one LIR only whereas the output $\sigma=-1$ results from all the remaining 2^K-1 IR. Hence $p(h|+1)$ and $p(h|-1)$ differ significantly. In fact we find for the $K=3$ AND machine

$$p(h|+1)=\Theta(h)\frac{e^{-h^2/2}}{\sqrt{2\pi}}+\frac{1}{2}\delta_+(h), \tag{25}$$

$$p(h|-1)=\Theta(-h)\frac{e^{-h^2/2}}{\sqrt{2\pi}}+\frac{1}{24}\delta_-(h)+\Theta(h)\frac{e^{-h^2/2}}{\sqrt{2\pi}}[1-H^2(h)], \tag{26}$$

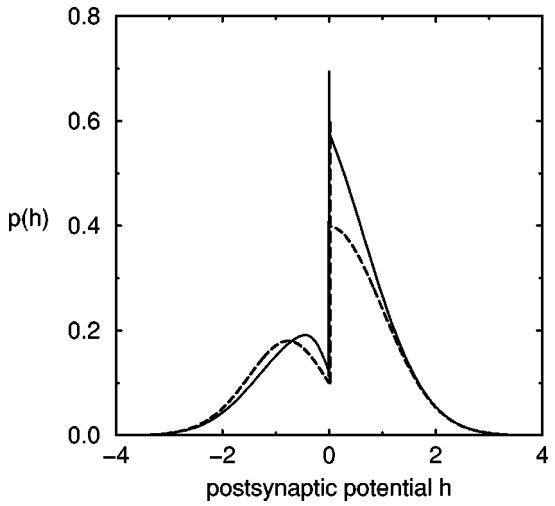

FIG. 1. Distribution of the local field h at the hidden units of a $K=3$ COMMITTEE tree in one-step RSB (bold) and RS (dashed). $\delta_+(h)$ is represented by adding its weight to the continuous part of the curve whereby, for a better presentation, the RS peak was shifted slightly to the right.

and $p(h)=[p(h|+1)+p(h|-1)]/2$. Note that we have introduced two different singular contributions $\delta_-(h)$ and $\delta_+(h)$ in Eqs. (23), (24) and Eqs. (25), (26). The reason for this is that the weight of $\delta_+(h)$ adds to the probability of positive local fields whereas the weight of $\delta_-(h)$ adds to that of negative local fields. This distinction will be important later when calculating the correlation coefficients from $p(h)$ [cf. Eq. (21)]. The results (23), (24), and (25), (26) are shown as the dashed lines in Figs. 1–3, respectively.

These RS results are in fact very intuitive and can be even quantitatively understood by assuming that the outcome of a Gardner calculation corresponds to the result of a learning process in which the initially wrong IR are eliminated with least adjustment [6]. Due to the permutation symmetry between the hidden units we may consider only the local field h_1 of the first unit of the hidden layer. Before learning the

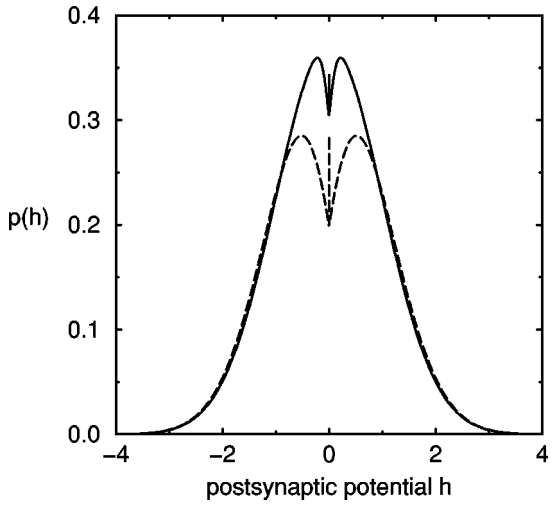

FIG. 2. Distribution of the local field h at the hidden units of a $K=3$ PARITY tree in one-step RSB (bold) and RS (dashed). $\delta_-(h)$ and $\delta_+(h)$ are represented by adding their weights to the continuous part of the curve.

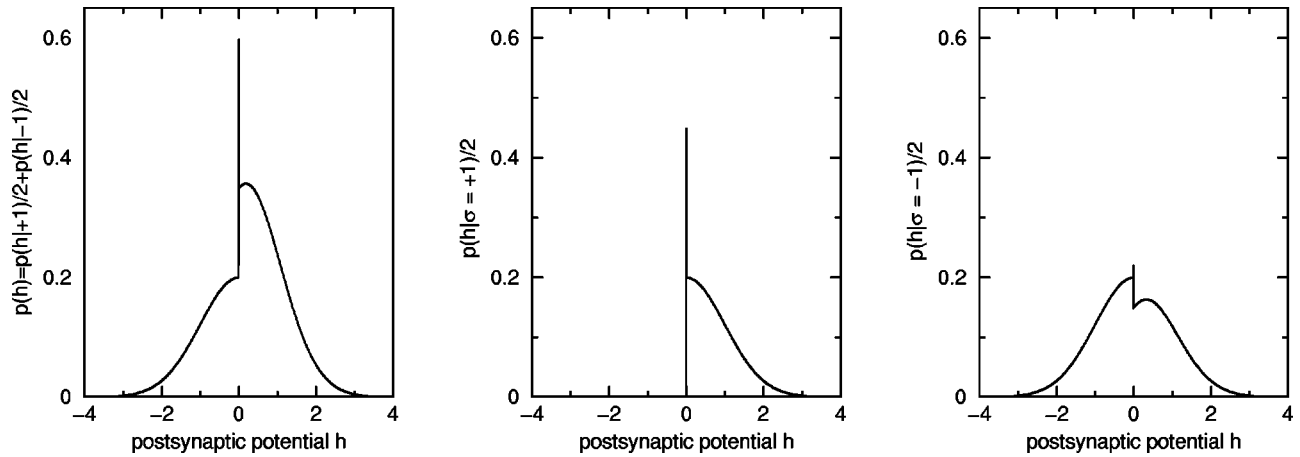

FIG. 3. Distribution $p(h)$ of the local field h at the hidden units of a $K=3$ AND tree in RS (left). The two panels to the right display its constituents $p(h|\sigma=+1)$ and $p(h|\sigma=-1)$ according to Eqs. (25) and (26). We found no RSB. $\delta_-(h)$ and $\delta_+(h)$ are represented by adding their weights to the continuous part of the curve.

couplings $\mathbf{J}_k$ are uncorrelated with the patterns and the local field h_1 is consequently Gaussian distributed with zero mean and unit variance.

Now consider, e.g., the PARITY machine. Due to the discussed symmetries it is sufficient to analyze the case $\sigma=+1$ and $h_1>0$. If h_2 and h_3 are equal in sign, which will occur with probability 1/2, there is no need to modify the couplings at all. This gives rise to the first term in Eq. (24) which is just the original Gaussian and describes the chance that a randomly found IR with $h_1>0$ is legal. If h_2 and h_3 differ in sign the IR is illegal and the couplings $\mathbf{J}_k$ have to be modified until one of the hidden units changes sign. In an optimal learning scenario the local field with the smallest magnitude would be selected and the corresponding coupling vector would be modified such that the field just barely changes sign. Hence h_1 remains still unmodified if either h_2 or h_3 is smaller in absolute value which gives rise to the second term in Eq. (24). Finally, if really h_1 is selected for the sign change, which will happen with probability 1/6 for symmetry reasons, it will after learning be either slightly smaller or slightly larger than zero, which is the origin of the last two terms in Eq. (24).

With a similar reasoning it is possible to rederive the RS result for the COMMITTEE machine. Again it is sufficient to consider the case $\sigma=+1$. If $h_1>0$ initially it will not be modified, which gives rise to the last term in Eq. (23). If, on the other hand, $h_1<0$, prior to learning it will not be modified only if both h_2 and h_3 are either positive from the start or easier to make positive than h_1. Hence a negative h_1 survives the learning process if the other two fields are both larger. This is described by the first term in Eq. (23). Finally, with probability 5/24 we find that $h_1<0$ and either h_2 or h_3 is even smaller than h_1 and therefore harder to correct. In this case the learning would shift h_1 to positive values as described by the second term in Eq. (23). The resulting distribution of local fields will hence have a dip for negative values of small absolute value clearly visible in Fig. 1.

The case of the AND machine is the simplest. The output $\sigma=+1$ requires all local fields to be positive. Hence positive fields are not modified, negative ones are shifted to 0^+ resulting immediately in Eq. (25) which is, of course, identical to the result for the single-layer perceptron [15,16]. In the case of a negative output $\sigma=-1$ only the IR $(+,+,+)$ is illegal and must be eliminated which is again done by changing the sign of the smallest field. This gives rise to Eq. (26).

It is finally interesting to compare the distribution of local fields found above with that for a single perceptron above saturation [17,11]. The individual perceptrons in a MLN certainly operate above their storage limit even when the storage capacity of the MLN is not yet reached. The most remarkable feature of the distribution of local fields for a perceptron above saturation minimizing the number of misclassified inputs is a *gap* separating positive from negative values. Being intimately related to the failure of any finite level of RSB for this problem this gap is believed to exist even in the solution with continuous RSB [12]. On the other hand, none of the distributions for MLN showed a gap.

As should be clear from the above qualitative discussion the reason for this is quite simple. The single perceptron above saturation has to reject some inputs as not correctly classifiable. In order to keep the number of these errors smallest it chooses those with negative fields of large absolute value. Inputs with initially only slightly negative local fields will be learned whereby their local fields shift to values just above zero. In this way the gap occurs. In MLN, on the other hand, there is no reason to shift all negative local fields of small absolute value because the correct output may be realized by the other hidden units. Therefore one will not find an interval of h values for which $p(h)$ is strictly zero. On the other hand, the tendency that predominantly fields of small absolute value will be modified in the learning process is clearly shown by the dips of the distribution functions around $h=0$ (cf. Figs. 1–3).

B. Replica symmetry breaking

Let us now discuss how the above results get modified by RSB. The analytical and subsequent numerical analysis of Eqs. (16)–(19) for the $K=3$ machines under consideration needs some care in order not to miss the various singular contributions. We have first to determine the values of the order parameters at the saddle point using Eq. (19). In the saturation limit $q_1\to 1$, $[\Phi_{\mathrm{LIR}}(\sigma)]^m$ is dominated by one specific LIR which is selected among all other LIR by the

TABLE I. Saturated $K=3$ machines: Integrated features of the probability distribution $p(h)$ of the local field. Corrections by one-step RSB are given in percent of the respective RS value. Dashes indicate that a respective singular contribution does not occur (COMMITTEE) or that we found no RSB (AND).

	Negative non-zero fields $\lim_{\epsilon\to 0}\int_{-\infty}^{-\lvert\epsilon\rvert} p(h)dh$		Singular contribution $\delta_-(h)$		Singular contribution $\delta_+(h)$		Positive non-zero fields $\lim_{\epsilon\to 0}\int_{\lvert\epsilon\rvert}^{\infty} p(h)dh$	
$K=3$ machine	RS	1-RSB	RS	1-RSB	RS	1-RSB	RS	1-RSB
COMMITTEE	7/24	−1.9%	–	–	5/24	−42.3%	1/2	+18.9%
PARITY	5/12	+10.3%	1/12	−51.8%	1/12	−51.8%	5/12	+10.3%
AND	1/4	–	1/48	–	1/4	–	23/48	–

sign and absolute value of the compound variables $v_k = y_k\sqrt{q_0} + z_k\sqrt{q_1 - q_0}$. $\Phi_{\mathrm{LIR}}(\sigma)]^m$ either tends to 1 or becomes exponentially small in one or more compound variables v_k. Transforming the integration from z_k space to v_k space allows us to reduce the K-fold z integral to a one-dimensional integral. This is performed numerically by Rhomberg integration whereas the outer y_k integrals are done using Gauss-Legendre quadrature [18].

The saddle point equation (19) is solved with a standard minimization routine (Powells method in two dimensions [18]). The values we get for the order parameters and for the storage capacity are consistent with those obtained earlier. For the $K=3$ PARITY machine we find $q_0=0$, $w\simeq 67.2$, and $\alpha_c^{\mathrm{RSB}}\simeq 5$ in agreement with [8]. In the case of the $K=3$ COMMITTEE machine we get $q_0\simeq 0.64$, $w\simeq 21.2$, and $\alpha_c^{\mathrm{RSB}} \simeq 3.14$, a result somewhat larger than reported previously [9,10]. The $K=3$ AND machine finally does not show RSB at all and we find accordingly $q_0\to 1$, $w\to\infty$ together with $\alpha_c^{\mathrm{AND}}=1.31$.

In a second step, we use these values of the order parameters w, q_0 to calculate the respective distribution of local fields (16). The distribution functions $p(h)$ obtained in this way are included as full lines in Figs. 1–3. Table I quantifies the main changes. The main modification of the distribution functions of local fields that occurs in one-step RSB is a redistribution of probability from the δ peaks at $h=\pm 0$ to the continuous part of the distribution around zero resulting in a reduction of the weight of the singular parts of roughly 50%. This gives rise to a less pronounced dip of the distribution functions around $h=0$ and is qualitatively similar to the RSB modifications for a single perceptron above saturation [11]. From the results for the PARITY machine it is conceivable that the central peak may get reduced further if higher orders of RSB are included and that it might eventually disappear completely in the full Parisi solution using continuous RSB. For all machines the probability of fields with large absolute values is hardly affected by RSB.

For the AND machine we did not find RSB at all. The numerical solution of the saddle point equations only gave the RSB result $q_0=1$, $w\to\infty$. We therefore suspect that replica symmetry is correct for the AND machine. This is also in accordance with the rule of thumb that RSB is necessary if the solution space is disconnected. In the AND machine the output $\sigma=+1$ can be realized only by one LIR which clearly corresponds to a connected (even convex) solution space. The output $\sigma=-1$ is realized by all remaining IR, which as the complement of the previous solution space must be connected too.

We have finally to clarify how much the modifications found for the distributions of local fields will change the probabilities of the internal representations and the correlation coefficients c_n depending only on the *sign* of the local fields.

This question is, in fact, nontrivial only in the case of the COMMITTEE machine. For the AND machine no RSB occurs at all and for the PARITY machine the correlation coefficients are completely determined by the symmetry of the Boolean function F between hidden units and output.

For the COMMITTEE machine we find that the probability of the LIR $(+,+,+)$ is shifted from its RS value 0.1250 to 0.1417, which is an increase by roughly 13%, whereas the probability of the three remaining LIR (consisting of two pluses and one minus each) is reduced by 1.9% from 0.2917 to 0.2861. Qualitatively this means that more inputs are stored with the LIR $(+,+,+)$ than the fraction 1/8 that had this LIR already by chance before learning. The learning process hence does not shift illegal IR just up to the decision boundary of the Boolean F but in some cases the correlations between inputs $\boldsymbol{\xi}$ and couplings $\mathbf{J}$ neglected in RS allow even the safer LIR $(+,+,+)$.

Using Eq. (22) we can now also calculate the correlation coefficients and find that c_1 increases by 2.7% from its RS value 5/12, c_2 decreases in absolute value by 13.3% from its RS value $-1/6$, and c_3 decreases in absolute value by 4.5% from its RS value $-3/4$. This confirms the prediction of [6] that although crucial for the storage capacity RSB will have only a minor influence on the correlation coefficients in MLN.

IV. SUMMARY

Generalizing the calculation of the distribution function of local fields for the single-layer perceptron we introduced a general formalism to determine the joint probability distribution $p(h_1,\ldots,h_K)$ of local fields at the K hidden units of a two-layer neural network of tree architecture with fixed Boolean function between hidden layer and output both in replica symmetry and in one-step replica symmetry breaking. Explicit results were obtained for the PARITY, COMMITTEE, and AND machine with $K=3$ hidden units in the saturation limit $\alpha\to\alpha_c$. Although the individual perceptrons are by far overloaded there is no gap in the distribution of local fields

as known from a single perceptron above saturation. There is no RSB for the AND machine which we attribute to the connected solution space for this architecture. For the PARITY and COMMITTEE machine we find as a result of RSB a slight redistribution of probability from the singular parts at $h=\pm 0$ to the continuous part around the origin. The correlation coefficients c_n characterizing the correlations between the legal internal representations are not modified by RSB for the PARITY machine since in this case they are fixed already by symmetries. For the COMMITTEE machine the changes of the correlation coefficients are rather small and the RS results derived in [6] may serve as useful approximations.

APPENDIX: REPLICA CALCULATION

In this appendix we give some more details on the calculation of the distribution function $p(h)$ of the local fields at the hidden units following Gardner's approach [15].

Introducing the replica trick $1/\mathcal{Z}=\lim_{n\to 0}\mathcal{Z}^{n-1}$ into Eq. (4) yields

$$p(h)=\lim_{n\to 0}\left\langle\left\langle\int\prod_{k,a} d\mu(\mathbf{J}_k^a)d\mu(\lambda_k^{\nu,a}) \times\prod_{\nu,a}\Theta(\sigma^\nu F(\{\mathrm{sgn}(\lambda_k^{\nu,a})\}))\,\delta(h-\lambda_1^{1,a})\right\rangle\right\rangle_{\{\boldsymbol{\xi}_k^\nu\},\sigma^\nu}, \tag{A1}$$

with replica index $a=1,\ldots,n$. In the integration measures (6), (7) we replace the δ functions by their integral form

$$\delta\left((\mathbf{J}_k^a)^2-\frac{N}{K}\right)=\int\frac{dE_k^a}{4\pi}\exp\left[-\frac{i}{2}E_k^a\left((\mathbf{J}_k^a)^2-\frac{N}{K}\right)\right], \tag{A2}$$

$$\delta\left(\lambda_k^{\nu,a}-\mathbf{J}_k^a\boldsymbol{\xi}_k^\nu\sqrt{\frac{K}{N}}\right)=\int\frac{dx_k^{\nu,a}}{2\pi}\exp \times\left[ix_k^{\nu,a}\left(\lambda_k^{\nu,a}-\mathbf{J}_k^a\boldsymbol{\xi}_k^\nu\sqrt{\frac{K}{N}}\right)\right]. \tag{A3}$$

We now perform the average over the Gaussian distributed patterns $\xi_{k,i}^\nu, i=1,\ldots,N/K$ and introduce the overlaps $q_k^{ab}=\mathbf{J}_k^a\mathbf{J}_k^b/(N/K)$ of different replicas of the same perceptron $\mathbf{J}_k$ as well as its conjugated variable F_k^{ab}. From the assumed permutation symmetry of the Boolean function F with respect to all hidden units we infer $q_k^{ab}=q^{ab}$, $F_k^{ab}=F^{ab}$, and $E_k^a=E^a$ for all $k=1,\ldots,K$. This gives rise to the form

$$p(h)=\lim_{n\to 0}\int\prod_a\frac{dE^a}{4\pi}\int\prod_{a<b}\frac{dq^{ab}dF^{ab}}{2\pi(K/N)} \times\langle\langle p(h|\sigma)\rangle\rangle_\sigma \times\exp\left[(N/2)\mathrm{tr}(QA)+(\alpha N-1)\langle\langle\ln G_1(Q|\sigma)\rangle\rangle_\sigma +N\ln G_2(A)\right], \tag{A4}$$

where Q and A denote the symmetric matrices $Q^{aa}=1$, $Q^{a\neq b}=q^{ab}$ and $A^{aa}=iE^a$, $A^{a\neq b}=-iF^{ab}$. Moreover,

$$p(h|\sigma)=\int\prod_{k,a}\frac{d\lambda_k^a dx_k^a}{2\pi}\exp\left(\sum_{k,a}\left[ix_k^a\lambda_k^a-\frac{1}{2}(x_k^a)^2\right] -\sum_{k,a<b}x_k^a x_k^b q^{ab}\right)\prod_a\Theta(\sigma F(\{\mathrm{sgn}(\lambda_k^a)\})) \times\delta(h-\lambda_1^1), \tag{A5}$$

$$G_1(Q|\sigma)=\int\prod_{k,a}\frac{d\lambda_k^a dx_k^a}{2\pi}\exp\left(\sum_{k,a}\left[ix_k^a\lambda_k^a-\frac{1}{2}(x_k^a)^2\right] -\sum_{k,a<b}x_k^a x_k^b q^{ab}\right)\prod_a\Theta(\sigma F(\{\mathrm{sgn}(\lambda_k^a)\})), \tag{A6}$$

$$G_2(A)=\int\prod_{k,a}\frac{dJ_k^a}{\sqrt{2\pi e}}\exp\left(-\frac{1}{2}\sum_{k;a,b}J_k^a A^{ab}J_k^b\right). \tag{A7}$$

In the limit $N\to\infty$ the integral in Eq. (A4) is dominated by the saddle point values of the order parameters E_a, F^{ab}, and q^{ab}. Solving the saddle point equation with respect to E^a and F^{ab} yields $A=Q^{-1}$. Hence Eq. (A4) takes the form

$$p(h)=\lim_{n\to 0}\int\prod_{a<b}dq^{ab} \times\langle\langle p(h|\sigma)\rangle\rangle_\sigma\exp[(N/2)\ln\det(Q)+(\alpha N -1)\langle\langle\ln G_1(Q|\sigma)\rangle\rangle_\sigma]. \tag{A8}$$

$p(h|\sigma)$ can be calculated by assuming either RS or one-step RSB for the matrix Q resulting in Eqs. (12) and (16), respectively.

The remaining saddle point condition for the matrix q^{ab} has in one-step RSB (15) the form

$$\mathrm{extr}_{q_0,q_1,m}\left[\frac{q_0}{2[1-q_1+m(q_1-q_0)]} +\frac{1}{2m}\ln\left(1+\frac{m(q_1-q_0)}{1-q_1}\right)+\frac{1}{2}\ln(1-q_1) +\frac{\alpha}{m}\left\langle\left\langle\int\prod_k Dy_k\ln\left\{\int\prod_k Dz_k\left(\Phi_{\mathrm{LIR}}(\sigma)\right)^m\right\}\right\rangle\right\rangle_\sigma\right]. \tag{A9}$$

It determines a set of order parameters $\{q_1,q_0,m\}$ for every pattern load below the storage capacity $\alpha\leq\alpha_c$. The abbreviation $\Phi_{\mathrm{LIR}}(\sigma)$ is defined by Eq. (18). The angular brackets $\langle\langle\cdots\rangle\rangle_\sigma$ indicate the average over the two possible outputs $\sigma=\pm 1$.

[1] M. Mezard and S. Patarnello, LPTENS report, 1989 (unpublished).
[2] M. Griniasty and T. Grossman, Phys. Rev. A **45**, 8924 (1992).
[3] A. Priel, M. Blatt, T. Grossman, E. Domany, and I. Kanter, Phys. Rev. E **50**, 577 (1994).
[4] B. Schottky, J. Phys. A **28**, 4515 (1995).
[5] R. Monasson and R. Zecchina, Phys. Rev. Lett. **75**, 2432 (1995).
[6] A. Engel, J. Phys. A **29**, L323 (1996).
[7] D. Malzahn, A. Engel, and I. Kanter, Phys. Rev. E **55**, 7369 (1997).
[8] E. Barkai, D. Hansel, and I. Kanter, Phys. Rev. Lett. **65**, 2312 (1990).
[9] E. Barkai, D. Hansel, and H. Sompolinsky, Phys. Rev. A **45**, 4146 (1992).
[10] A. Engel, H. M. Köhler, F. Tschepke, H. Vollmayr, and A. Zippelius, Phys. Rev. A **45**, 7590 (1992).
[11] P. Majer, A. Engel, and A. Zippelius, J. Phys. A **26**, 7405 (1993); W. Whyte and D. Sherrington, *ibid.* **29**, 3063 (1996).
[12] G. Györgyi and P. Reimann, Phys. Rev. Lett. **79**, 2746 (1997).
[13] E. Gardner, J. Phys. A **21**, 257 (1988); E. Gardner and B. Derrida, *ibid.* **21**, 271 (1988).
[14] M. Mézard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[15] E. Gardner, J. Phys. A **22**, 1969 (1989).
[16] M. Opper, Phys. Rev. A **38**, 3824 (1988).
[17] D. J. Amit, M. R. Evans, H. Horner, and K. Y. Wong, J. Phys. A **23**, 3361 (1990).
[18] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes* (Cambridge University Press, Cambridge, England, 1992).