Learning a rule in a multilayer neural network

Learning a rule in a multilayer neural network

H Schwarze
CONNECT, The Niels Bohr Institute, Blegdamsvej 17, DK-2100 Copenhagen Ø, Denmark

Received 4 June 1993

Abstract. The problem of learning from examples in multilayer networks is studied within the framework of statistical mechanics. Using the replica formalism we calculate the average generalization error of a fully connected committee machine in the limit of a large number of hidden units. If the number of training examples is proportional to the number of inputs in the network, the generalization error as a function of the training set size approaches a finite value. If the number of training examples is proportional to the number of weights in the network we find first-order phase transitions with a discontinuous drop in the generalization error for both binary and continuous weights.

1. Introduction

As nonlinear, parametric models for the solution of classification tasks and function approximation, feedforward neural networks have attracted considerable interest (e.g. [1]). Trained from examples of a given task, they are able to generalize, i.e. to compute the correct output for new, unknown inputs. Since the seminal work of Gardner [2] much effort has been put into studying the properties of feedforward networks within the framework of statistical mechanics (e.g. [3]). Most of this work has concentrated on the simplest such network, the simple perceptron [4] with only one layer of weights connecting the inputs with a single output unit. However, most applications of neural networks have to utilize architectures with hidden layers for which only a few general theoretical results are known (e.g. [5–7]). The computational power of networks with only one additional layer of hidden units is already dramatically increased compared with a simple perceptron. In principle, a network with one layer of sufficiently many hidden units can implement every Boolean [8] or continuous [9, 10] function of the inputs.

As an example of a two-layer network we will study the 'committee machine' [11]. This architecture has only one layer of adjustable weights, while the weights connecting the hidden units to the output are fixed to $+1$ so as to implement a majority decision of the hidden units. For binary weights this may already be regarded as the most general two-layer architecture, because any other combination of hidden-output weights can be gauged to $+1$ by flipping the signs of the corresponding input-hidden weights. Previous work has been concerned with some restricted versions of this model, such as learning geometrical tasks in machines with local connectivity in the input-hidden layer [12] and learning in committee machines with non-overlapping receptive fields [13, 14]. In this tree-like architecture there are no correlations between hidden units and its behaviour was found to be qualitatively similar to the simple perceptron. Furthermore, committee machines have been studied within the context of storing random input–output pairs [15, 16].

Recently, learning in fully connected committee machines has been studied within the annealed approximation [17–19], revealing properties which are qualitatively different

0305-4470/93/215781+14$07.50 © 1993 IOP Publishing Ltd

from the tree model. However, the annealed approximation (AA) is only valid at high temperatures, and a correct description of learning at low temperatures requires solution of the quenched theory. The purpose of this paper is to extend previous work towards a better understanding of the learning properties of multilayer networks. We present the calculation of the average generalization error of a fully connected committee machine in the thermodynamic limit ($N \to \infty$) employing the replica formalism, and we compare the results to the AA. In particular we will study a committee machine in the limit of a large number K of hidden units, but with $K \ll N$. The target rule is defined by another fully connected committee machine and is therefore realizable by the learning network.

In the following section we start with a definition of the model and briefly outline the statistical mechanics approach. In sections 3 and 4 we present the calculation of the average generalization error for both binary and continuous weights. The results are summarized in section 5.

2. The model

We will be concerned with a two-layer network with N inputs, K hidden units and a single output unit σ. Each hidden unit σ_l, $l \in \{1, \ldots, K\}$, is connected to the inputs $\boldsymbol{S} = (S_1, \ldots, S_N)$ through the weight vector $\boldsymbol{W}_l$ and performs the mapping

$$\sigma_l(\boldsymbol{W}_l, \boldsymbol{S}) = \operatorname{sign}\left(\frac{1}{\sqrt{N}} \boldsymbol{W}_l \cdot \boldsymbol{S}\right). \tag{1}$$

The hidden units may be regarded as outputs of simple perceptrons and will be referred to as 'students'. The overall network output is defined as the majority vote of the student committee, given by

$$\sigma(\{\boldsymbol{W}_l\}, \boldsymbol{S}) = \operatorname{sign}\left(\frac{1}{\sqrt{K}} \sum_{l=1}^{K} \sigma_l(\boldsymbol{W}_l, \boldsymbol{S})\right). \tag{2}$$

This network is trained from $P = \alpha K N$ input–output examples $(\boldsymbol{\xi}^\mu, \tau(\boldsymbol{\xi}^\mu))$, $\mu \in \{1, \ldots, P\}$, of the desired mapping τ. We study a realizable task defined by another committee machine with weight vectors $\boldsymbol{V}_l$, hidden units τ_l and an overall output $\tau(\boldsymbol{S})$ of the form (2). The teacher weight vectors are taken to be normalized to $\sqrt{N}$ and mutually orthogonal, $\boldsymbol{V}_l \cdot \boldsymbol{V}_k = N\delta_{lk}$. Note that orthogonality does not have to be imposed explicitly, because in the thermodynamic limit with $N \gg K$ and randomly drawn teacher vectors it will always be satisfied. The components of the training inputs ξ_i^μ are drawn independently from a Gaussian distribution with zero mean and unit variance. However, for large N our results are valid for a more general class of distributions, including binary inputs, with the same mean and variance.

The goal of learning is to find a network that performs well on unknown examples, which are not included in the training set. The network quality can be measured by the generalization error

$$\epsilon(\{\boldsymbol{W}_l\}) = \langle \Theta[-\sigma(\{\boldsymbol{W}_l\}, \boldsymbol{S})\tau(\boldsymbol{S})] \rangle_{\boldsymbol{S}} \tag{3}$$

the probability that a randomly chosen input is misclassified. However, a training algorithm has only access to the limited set of training examples from which one can construct the training error $E_t(\{\boldsymbol{W}_l\}) = \sum_\mu \Theta[-\sigma(\{\boldsymbol{W}_l\}, \boldsymbol{\xi}^\mu)\tau(\boldsymbol{\xi}^\mu)]$.

Following the statistical mechanics approach we will consider a stochastic learning algorithm which, for long training times, yields a Gibbs distribution of networks $\rho_{\mathrm{G}}(\{\boldsymbol{W}_l\}) = Z^{-1}\rho_0(\{\boldsymbol{W}_l\})\exp(-\beta E_{\mathrm{t}}(\{\boldsymbol{W}_l\}))$. Here, the formal temperature $T = 1/\beta$ determines the quantity of noise during training and the distribution $\rho_0(\{\boldsymbol{W}_l\})$ includes *a priori* constraints on the weights. The normalization constant Z is the partition function

$$Z = \int \mathrm{d}\rho_0\,(\{\boldsymbol{W}_l\})\mathrm{e}^{-\beta E_{\mathrm{t}}(\{\boldsymbol{W}_l\})}. \tag{4}$$

The average generalization and training errors at thermal equilibrium, averaged over all representations of the training examples, are given by

$$\epsilon_{\mathrm{g}} = \langle\!\langle\, \langle \epsilon(\{\boldsymbol{W}_l\})\rangle_T \,\rangle\!\rangle \tag{5}$$

and

$$\epsilon_{\mathrm{t}} = (1/P)\langle\!\langle\, \langle E_{\mathrm{t}}(\{\boldsymbol{W}_l\})\rangle_T \,\rangle\!\rangle \tag{6}$$

where $\langle\!\langle \ldots \rangle\!\rangle$ denotes a quenched average over the training examples and $\langle \ldots \rangle_T$ a thermal average. These quantities may be obtained from the average free energy $F = -T\langle\!\langle \ln Z \rangle\!\rangle$, which can be calculated within the standard replica formalism [2]. Following this approach, we calculate the average replicated partition function $\langle\!\langle Z^n \rangle\!\rangle$ in the thermodynamic limit ($N \to \infty$). From the analytical continuation of the result to $n \to 0$ we obtain $\langle\!\langle \ln Z \rangle\!\rangle = \lim_{n\to 0}(\langle\!\langle Z^n \rangle\!\rangle - 1)/n$. The average over training examples factorizes into single pattern averages and $\langle\!\langle Z^n \rangle\!\rangle$ can be written as

$$\langle\!\langle Z^n \rangle\!\rangle = \int \prod_{a=1}^{n}\prod_{l=1}^{K} \mathrm{d}\rho_0\,(\boldsymbol{W}_l^a)\exp[-\alpha K N G_{\mathrm{r}}^{(n)}(\{\boldsymbol{W}_l^a\})] \tag{7}$$

with $\alpha = P/KN$ and

$$G_{\mathrm{r}}^{(n)}(\{\boldsymbol{W}_l^a\}) = -\ln\left\langle \exp\left\{ -\beta\sum_a \Theta[-\sigma(\{\boldsymbol{W}_l^a\},\xi)\tau(\xi)]\right\}\right\rangle_{\xi}. \tag{8}$$

As will be shown in the appendix, the effective Hamiltonian $G_{\mathrm{r}}^{(n)}$ can be written as a function of the order parameters

$$R_{lk}^a = \frac{1}{N}\boldsymbol{W}_l^a \cdot \boldsymbol{V}_k \qquad D_{lk}^{ab} = \frac{1}{N}\boldsymbol{W}_l^a \cdot \boldsymbol{W}_k^b \qquad (a \neq b) \qquad C_{lk}^a = \frac{1}{N}\boldsymbol{W}_l^a \cdot \boldsymbol{W}_k^a. \tag{9}$$

Introducing these parameters through integrals over δ-functions allows us to rewrite $\langle\!\langle Z^n \rangle\!\rangle$ as

$$\begin{aligned}\langle\!\langle Z^n \rangle\!\rangle = \int \prod_{l,k,a} \frac{\mathrm{d}R_{lk}^a\,\mathrm{d}\hat{R}_{lk}^a}{2\pi\mathrm{i}} \int \prod_{l,k,(a,b)} \frac{\mathrm{d}D_{lk}^{ab}\,\mathrm{d}\hat{D}_{lk}^{ab}}{2\pi\mathrm{i}} \int \prod_{(l,k),a} \frac{\mathrm{d}C_{lk}^a\,\mathrm{d}\hat{C}_{lk}^a}{2\pi\mathrm{i}} \\ \times \exp\{-KN[\alpha G_{\mathrm{r}}^{(n)}(\{R_{lk}^a, D_{lk}^{ab}, C_{lk}^a\}) - G_0^{(n)}(\{R_{lk}^a, D_{lk}^{ab}, C_{lk}^a, \hat{R}_{lk}^a, \hat{D}_{lk}^{ab}, \hat{C}_{lk}^a\})]\}\end{aligned} \tag{10}$$

where the replica free energy $f^{(n)}$ is given by $\beta f^{(n)} = \alpha G_{\mathrm{r}}^{(n)} - G_0^{(n)}$ with the 'entropy' term

$$G_0^{(n)} = \frac{1}{K}\Bigg[\sum_{a,l,k} R^a_{lk}\hat{R}^a_{lk} + \sum_{l,k,(a,b)} D^{ab}_{lk}\hat{D}^{ab}_{lk} + \sum_{(l,k),a} C^a_{lk}\hat{C}^a_{lk} + \frac{1}{N}\ln\int\prod_{a,l}\mathrm{d}\rho_0(\boldsymbol{W}^a_l)\\ \times\exp\Bigg\{-\sum_{a,l,k}\hat{R}^a_{lk}\boldsymbol{W}^a_l\cdot\boldsymbol{V}_k - \sum_{l,k,(a,b)}\hat{D}^{ab}_{lk}\boldsymbol{W}^a_l\cdot\boldsymbol{W}^b_k - \sum_{(l,k),a}\hat{C}^a_{lk}\boldsymbol{W}^a_l\cdot\boldsymbol{W}^a_k\Bigg\}\Bigg]. \tag{11}$$

We can evaluate the integrals over order parameters using the saddle-point method and analytically continue the result to $n \to 0$ if we make symmetry assumptions for the order parameters. We make a replica-symmetric (RS) ansatz assuming that the order parameters do not depend on the replica indices. Furthermore, we assume partial committee symmetry allowing for a specialization of the hidden units on their respective teachers, writing

$$R^a_{lk} = R + \Delta\delta_{lk} \qquad D^{ab}_{lk} = D + q\delta_{lk} \qquad C^a_{lk} = C + (1-C)\delta_{lk} \tag{12}$$

and similarly for the conjugate parameters $\hat{R}^a_{lk}$, $\hat{D}^{ab}_{lk}$ and $\hat{C}^a_{lk}$. This ansatz is similar to that used in the capacity calculation for this architecture [15, 16]. It has the important property of describing both a solution with $\Delta = q = 0$, which is symmetric under permutation of hidden units, and a specialized solution with $\Delta, q \neq 0$ in which this symmetry is broken and each hidden unit is correlated with a particular hidden unit in the teacher network. Note that no solution corresponding to the permutation-symmetric one was found in the committee machine with non-overlapping receptive fields [13, 14], because there the specialization was built into the model through the assignment of different inputs to different hidden units. The values of Δ, q, R, D, C, $\hat{\Delta}$, $\hat{q}$, $\hat{R}$, $\hat{D}$ and $\hat{C}$ have to be determined at the saddle point of the replica free energy and allow the calculation of the average generalization and training errors (5) and (6). In the following sections this will be described for committee machines with continuous and binary weights.

3. Continuous weights

In a committee machine with continuous weights we introduce spherical constraints for the individual hidden unit weight vectors $\rho_0(\{\boldsymbol{W}^a_l\}) = \prod_{l,a}\delta(N - \boldsymbol{W}^a_l\cdot\boldsymbol{W}^a_l)$. These constraints are controlled in the standard way by additional parameters E^a_l. At the saddle point with the symmetry properties (12) a straightforward calculation similar to the one performed for the simple perceptron [7] and the AA of the present model [18] yields $G_0 = \lim_{n\to 0} G_0^{(n)}/n$. After eliminating the auxiliary variables $\hat{\Delta}$, $\hat{q}$, $\hat{R}$, $\hat{D}$, $\hat{C}$ and E we obtain

$$G_0(\Delta, q, R, D, C) = \frac{1}{2}\ln(2\pi) + \frac{1}{2}\frac{K-1}{K}\frac{1-\Delta^2-C}{1-q-C} + \frac{1}{2}\frac{K-1}{K}\ln(1-q-C)\\ + \frac{1}{2K}\left[\frac{1-(\Delta+KR)^2+(K-1)C}{1-q-C-K(D-C)} + \ln[1-q-C-K(D-C)]\right]. \tag{13}$$

The calculation of the energy term G_{r} for large K requires scaling assumptions for the order parameters. We introduce new parameters

$$\rho = KR \qquad c = KC \qquad d = KD. \tag{14}$$

As will be shown below, the values of these parameters at the saddle point can be obtained self-consistently, if they are assumed to be of order $\mathcal{O}(1)$. This scaling for R and C was found to reveal interesting properties of the learning curve within the AA [17–19], and it is in good agreement with our Monte Carlo simulations. With the ansatz (14), G_r reads to leading order in $1/K$ (see appendix A)

$$G_r(\Delta, q, \rho, d, c) = -2\int_{-\infty}^{+\infty} \mathrm{D}x\, H\left(\frac{R_{\text{eff}}x}{\sqrt{Q_{\text{eff}} - R_{\text{eff}}^2}}\right) \times \ln\left[\mathrm{e}^{-\beta} + (1+\mathrm{e}^{-\beta})H\left(\sqrt{\frac{Q_{\text{eff}}}{1+(2/\pi)c - Q_{\text{eff}}}}x\right)\right] \tag{15}$$

with

$$R_{\text{eff}} = (2/\pi)(\rho + \sin^{-1}\Delta) \qquad Q_{\text{eff}} = (2/\pi)(d + \sin^{-1} q) \tag{16}$$

and $\mathrm{D}x = \mathrm{d}x\, \mathrm{e}^{-x^2/2}/\sqrt{2\pi}$, $H(z) = \int_z^\infty \mathrm{D}x$. The entropy term as a function of the rescaled parameters is

$$G_0 = \frac{1}{2}[1 + \ln(2\pi)] + \frac{1}{2}\frac{q-\Delta^2}{1-q} + \frac{1}{2}\ln(1-q) - \frac{1}{2K}\left[\ln(1-q) - \ln(1-q-d+c) - \frac{c}{1-q}\frac{q-\Delta^2}{1-q} + \frac{1-\Delta^2+c}{1-q} - \frac{1-(\Delta+\rho)^2+c}{1-q-d+c}\right]. \tag{17}$$

To leading order in $1/K$, this is the same expression as that obtained for a simple perceptron [20].

The generalization error for a given network (3) can be obtained by a calculation similar to the one leading to (15), yielding [17]

$$\epsilon(\Delta, \rho, c) = \frac{1}{\pi}\cos^{-1}\left(\frac{R_{\text{eff}}}{\sqrt{1+(2/\pi)c}}\right) \tag{18}$$

with R_{eff} as in (16). Inserting the values of the order parameters at the saddle point of the free energy $\beta f = \alpha G_r - G_0$ finally leads to the average generalization error (5).

3.1. Small α

First we consider the limit of small training sets, in which the load parameter is of order $\mathcal{O}(1/K)$. In this limit, the number of training examples is proportional to the number of inputs, and we introduce $\tilde{\alpha} = \alpha K = P/N$. The saddle-point equations for the corresponding free energy $\beta f = (\tilde{\alpha}/K)G_r - G_0$ are, to order $\mathcal{O}(1/K)$, solved by $\Delta = q = 0$ and with ρ, d and c given by the numerical solution of the equations $0 = \partial f/\partial\rho = \partial f/\partial d = \partial f/\partial c$. Hence, *the system settles into a committee-symmetric solution without any specialization* of the hidden units. This solution cannot achieve perfect generalization and for $\tilde{\alpha} \to \infty$ the generalization error approaches a finite value (see figure 1).

At zero temperature the saddle-point equations are simplified by the relations $\rho = d$ and $c = 0$. These relations are similar to those found for the simple perceptron [20, 7] and reflect a symmetry between the teacher network and typical solutions for the student network [3]: at zero temperature, the partition function Z (4) measures the volume in weight space

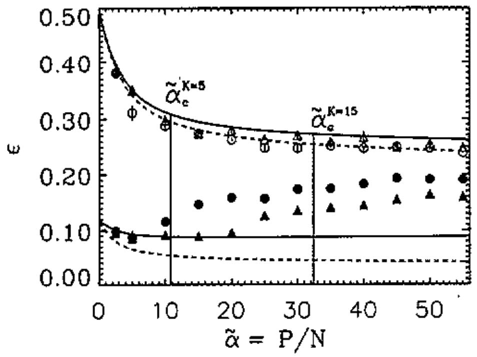

Figure 1. Generalization (upper curves) and training (lower curves) error as functions of $\tilde{\alpha} = P/N$. The full curves show the results of the RS calculation, while the broken curves correspond to the predictions of the AA. Note that these results are valid for both continuous and binary weights. The results of Monte Carlo simulations for the generalization (open symbols) and training (full symbols) errors are shown for $K = 5$ (circles) and $K = 15$ (triangles) with $T = 0.5$ and $N = 99$. The vertical lines indicate the predictions for $\tilde{\alpha}_c = K\alpha_c$ in the large-K theory of the binary model (31) for $K = 5$ and $K = 15$, respectively.

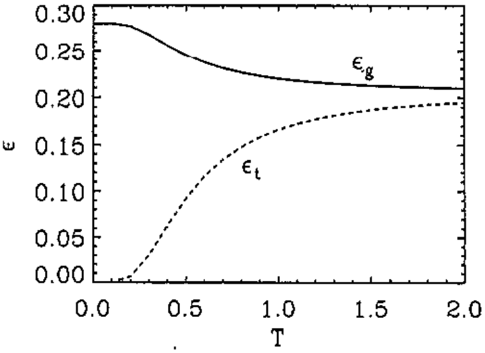

Figure 2. Asymptotic generalization and training errors for the committee-symmetric solution.

compatible with the training examples, and a stochastic training procedure corresponds to randomly placing weight vectors into this volume. Since the teacher weights are also drawn randomly from this volume, the typical overlaps between two solutions, as measured by the physical order parameter d, should be equal to the typical student–teacher overlap as measured by ρ. Furthermore, the internal overlaps between different hidden units within each solution vanish as in a randomly chosen teacher.

The asymptotic behaviour of this solution for $\tilde{\alpha} \to \infty$ (but with $\tilde{\alpha} \ll K$) is given by $1 - \rho \propto 1/\tilde{\alpha}$ and

$$\epsilon_g = \epsilon_0 + \frac{1}{\tilde{\alpha}}\left[\int \mathrm{D}x\, H^{-1}\left(\sqrt{2/(\pi+2)}x\right)\right]^{-1} + \mathcal{O}\left(\frac{1}{\tilde{\alpha}^2}\right) \tag{19}$$

where

$$\epsilon_0 = (1/\pi)\cos^{-1}(2/\pi) \approx 0.28. \tag{20}$$

For non-zero temperature the qualitative behaviour remains unchanged, but the asymptotic values of ρ and ϵ_g depend on T. The asymptotic student–teacher overlap increases with increasing temperature, while the asymptotic generalization error decreases and approaches $\epsilon'_0 = (1/\pi)\cos^{-1}(\sqrt{2/\pi}) \approx 0.20$ for $T \to \infty$ as shown in figure 2. This temperature dependence of the residual generalization ability can be compared with an improvement in the generalization ability at non-zero temperature in unlearnable problems (e.g. [7]). However, the behaviour for $\tilde{\alpha} \to \infty$ should be distinguished from the asymptotic approach to the optimal generalization error in the large-α limit [7]. The present problem is realizable and, accordingly, $\epsilon_{\rm opt} = 0$. However, in the small-α regime, this value cannot be achieved, and we do not find $\epsilon_g, \epsilon_t \to \epsilon_{\rm opt}$ within this committee-symmetric solution.

These results differ from the predictions of the AA [17, 19]. The AA does not give any temperature dependence in the residual generalization error. It predicts an approach to the value $\epsilon'_0 = (1/\pi)\cos^{-1}(\sqrt{2/\pi})$ for all temperatures, given by $\epsilon_g - \epsilon'_0 \propto 1/\sqrt{\tilde{\alpha}}$. Furthermore, in the AA the rescaled overlaps ρ and c diverge for $\tilde{\alpha} \to \infty$, while they remain finite in the RS solution at finite T.

3.2. *Finite* α

If the number of training examples is proportional to the number of adjustable weights in the network, we have to find saddle points for the free energy $\beta f = \alpha G_r - G_0$ with $\alpha \sim \mathcal{O}(1)$. In this regime the saddle-point equations $0 = \partial f/\partial\rho = \partial f/\partial d$ can only be solved for

$$d = (\Delta + \rho)^2 - q + \mathcal{O}(1/K) \qquad c = d + q - 1 + \mathcal{O}(1/K). \tag{21}$$

Using these relations, the remaining equations can, to leading order in $1/K$, be brought into the form

$$\frac{\Delta}{1-q} = \alpha\left(1 - \frac{1}{\sqrt{1-\Delta^2}}\right)\frac{\partial G_r}{\partial\rho} \tag{22}$$

$$-\frac{1}{2}\frac{q-\Delta^2}{(1-q)^2} = \alpha\left(1 - \frac{1}{\sqrt{1-q^2}}\right)\frac{\partial G_r}{\partial d} \tag{23}$$

$$0 = \frac{\partial G_r}{\partial\rho} + 2(\Delta+\rho)\left(\frac{\partial G_r}{\partial c} + \frac{\partial G_r}{\partial d}\right). \tag{24}$$

Note that equation (24) does not explicitly depend on α. It is easy to see that there is always a solution—the symmetric one—with $\Delta = q = 0$, $c = \rho^2 - 1$, $d = \rho^2$ and ρ given by (24). This solution corresponds to the residual generalization error shown in figure 2. Other solutions can be found numerically, and for $T = 0$ we find the following situation. At zero temperature the saddle-point equations (22)–(24) admit solutions which again show symmetry between a typical student and the teacher network, $\Delta = q$ and $\rho = d$. Together with equation (21), this leaves us with a free energy as a function of Δ. As shown in figure 3, $f(\Delta)$ has a local minimum at $\Delta = 0$ for all values of α. However, for $\alpha > \alpha_s = 7.17$ a second local minimum appears with $\Delta > 0$. In the region $\alpha_s < \alpha < \alpha_c = 7.65$ this minimum has a higher free energy than the symmetric solution at $\Delta = 0$, but it is the global minimum for $\alpha > \alpha_c$. Therefore, the system exhibits a first-order transition from a symmetric solution to one with specialized hidden units, accompanied by a discontinuous drop in the generalization error. For the specialized solution, the asymptotic behaviour of the generalization error can easily be obtained as

$$\epsilon_g = 2\sqrt{2}\left(\int \mathrm{D}t\, \frac{\mathrm{e}^{-t^2/2}}{H(t)}\right)^{-1}\frac{1}{\alpha} + \mathcal{O}\left(\frac{1}{\alpha^2}\right) = \frac{1.25}{\alpha} + \mathcal{O}\left(\frac{1}{\alpha^2}\right) \tag{25}$$

the same result found for the large-K tree committee machine [13].

For $\alpha > \alpha' = 9.27$ the saddle-point equations (22)–(24) have a pair of additional solutions which do not satisfy the symmetry $\Delta = q$. However, these solutions have free energies which are always higher than the free energy of the symmetric solution. Therefore, in the region $\alpha > \alpha_c$ we always identify the specialized solution with $\Delta = q$ and $\rho = d$ as the thermodynamic solution.

At $T > 0$ the situation remains qualitatively the same, and the complete RS phase diagram is shown in figure 4.

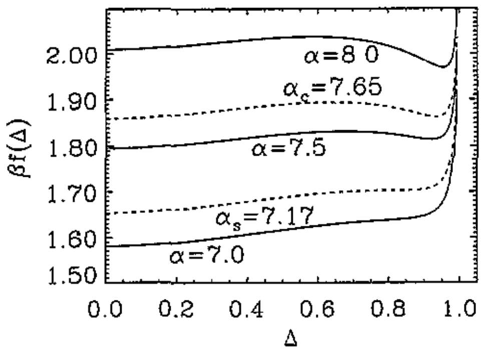

Figure 3. Free energy $f(\Delta)$ for continuous weights and different values of α. The free energy always has a minimum at $\Delta = 0$, but at $\alpha_s = 7.17$ a second minimum appears at Δ close to 1. At $\alpha_c = 7.65$ this becomes the global minimum of f.

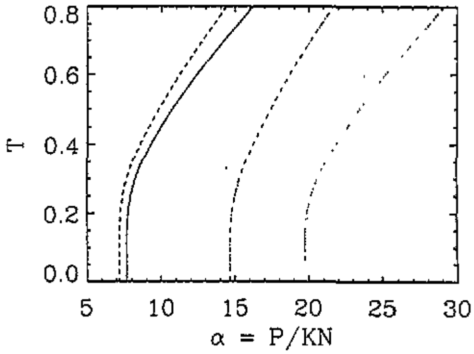

Figure 4. RS and annealed phase diagrams for the large-K committee machine with continuous weights. The two left lines show the RS results for the spinodal line' (– – –), where the second minimum appears, and the location of the phase transition (——). These results are compared to the predictions of the AA for the spinodal line (— · —) and the phase transition (· · · · · ·).

4. Binary weights

For binary weights, $W_l \in \{\pm 1\}^N$, the analysis proceeds in a similar way. Now the entropy term is given by

$$\begin{aligned} G_0 = \ln 2 - \Delta\hat{\Delta} - \frac{1}{2}(1-q)\hat{q} + \int \mathrm{D}t\, \ln\cosh\left(\hat{\Delta} + \sqrt{\hat{q}}t\right) \\ - \frac{1}{2K}\bigg[2\rho\hat{\Delta} - d\hat{q} + \ln(1-q) - \ln(1-q-d+c) + \frac{\rho^2}{1-q} \\ - \frac{c-d}{1-q}\left(\frac{q-\Delta^2}{1-q} - \frac{1-(\Delta+\rho)^2+c}{1-q-d+c}\right)\bigg] \end{aligned} \qquad (26)$$

where the auxiliary variables $\hat{\Delta}$ and $\hat{q}$ are, to order $1/K$, given by

$$\Delta = \int \mathrm{D}t\, \tanh\left(\hat{\Delta} + \sqrt{\hat{q}}t\right) \qquad q = \int \mathrm{D}t\, \tanh^2\left(\hat{\Delta} + \sqrt{\hat{q}}t\right). \qquad (27)$$

As for continuous weights the saddle-point equations in the small-α region with $\tilde{\alpha} = K\alpha \sim \mathcal{O}(1)$ only have a committee-symmetric solution with $\Delta = q = 0$. Furthermore, since the entropy terms in the continuous and binary model for $\Delta = q = 0$ only differ by a constant, both models show the same behaviour of the generalization error for $\alpha \sim \mathcal{O}(1/K)$.

Only if the number of training examples is proportional to the number of adjustable weights in the network does the discreteness of the weights influence the generalization properties. For binary weights, the system can reach its ground state with all student weight vectors perfectly aligned with their respective teacher vectors. Accordingly, the free energy in the binary model always has a local minimum with $\Delta = q = 1$ and $\rho = d = c = 0$. Additional solutions can be obtained from the saddle-point equations of the free energy. As for continuous weights, the equations $0 = \partial f/\partial \rho = \partial f/\partial d$ require the relations (21), and the remaining equations yield the conditions

$$\hat{\Delta} = \alpha \left(1 - \frac{1}{\sqrt{1-\Delta^2}}\right) \frac{\partial G_{\mathrm{r}}}{\partial \rho} \tag{28}$$

$$-\frac{1}{2}\hat{q} = \alpha \left(1 - \frac{1}{\sqrt{1-q^2}}\right) \frac{\partial G_{\mathrm{r}}}{\partial d} \tag{29}$$

$$0 = \frac{\partial G_{\mathrm{r}}}{\partial \rho} + 2(\Delta + \rho)\left(\frac{\partial G_{\mathrm{r}}}{\partial c} + \frac{\partial G_{\mathrm{r}}}{\partial d}\right). \tag{30}$$

Note that equation (30) is identical to the corresponding equation (24) in the continuous model. Once again, for all values of α these equations have a solution $\Delta = q = \hat{\Delta} = \hat{q} = 0$ and ρ given by (30). For small α this symmetric solution has a lower free energy than the perfectly generalizing one with $\Delta = q = 1$ and $f = 0$. However, the free energy of the symmetric solution becomes positive at a critical value of the load parameter, given by

$$\alpha_{\mathrm{c}} = -\frac{\ln 2}{2\int \mathrm{D}x\, H(\sqrt{2/(\pi-2)}x)\ln[\mathrm{e}^{-\beta} + (1-\mathrm{e}^{-\beta})H(\sqrt{2/(\pi-2)}\rho x)]} + \mathcal{O}\left(\frac{1}{K}\right) \tag{31}$$

where ρ is given by equation (30) for $\Delta = q = 0$. Hence, the system exhibits a first-order phase transition to perfect generalization similar to other learnable models with binary weights. Here, the poorly generalizing symmetric solution remains metastable even for large α, and a stochastic training algorithm can always get stuck in this solution. The phase diagram of the binary model is shown in figure 5. A careful numerical evaluation of the saddle-point equations (28)–(30) yields an additional solution at large values of α. However, as in the continuous model, this solution always has a higher free energy than the symmetric one and does not satisfy the symmetry $\Delta = q$ and $\rho = d$ at $T = 0$.

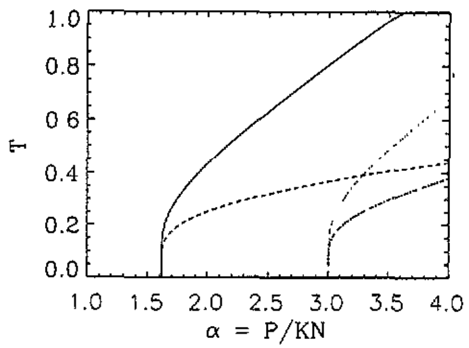

Figure 5. RS and annealed phase diagrams for the large-K committee machine with binary weights. The RS result for the location of the phase transition (——) and its zero-entropy line (– – –) are compared with the prediction of the AA for the phase transition (······) and its zero-entropy line (— · —).

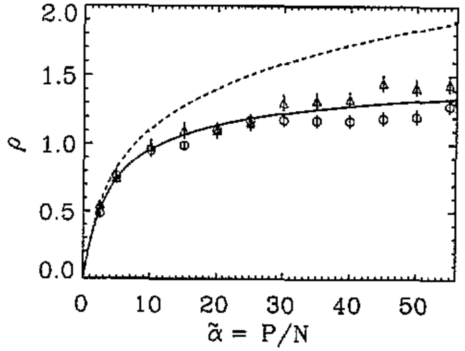

Figure 6. Rescaled student–teacher overlap $\rho = KR$ as a function of $\tilde{\alpha} = P/N$ for the RS solution (——) and in the AA (– – –) compared with Monte Carlo simulations with $K = 5$ (○) and $K = 15$ (△). The simulations were performed for $N = 99$ and $T = 0.5$.

In the binary model, the validity of the replica-symmetric ansatz can be checked by evaluating the thermodynamic entropy $s = \beta(\alpha\epsilon_t - f)$. For any given temperature the entropy of the symmetric state becomes negative at $\alpha_{s=0} = \ln(2)/(G_r - \beta\epsilon_t)$, where G_r is given by (15) and the training error of the symmetric state is

$$\epsilon_t = 2\mu_\beta \int \mathrm{D}x\, H\left(\sqrt{\frac{2}{\pi-2}}x\right) H\left(-\sqrt{\frac{2}{\pi-2}}\rho x\right) \Bigg/ \left[\mu_\beta + H\left(\sqrt{\frac{2}{\pi-2}}\rho x\right)\right] \qquad (32)$$

with $\mu_\beta = \mathrm{e}^{-\beta}/(1-\mathrm{e}^{-\beta})$. A negative entropy is unphysical in a discrete model and, correspondingly, the RS ansatz cannot be correct in the region $\alpha > \alpha_{s=0}$. In this region, we expect the existence of a one-step RS breaking solution with a structure similar to the one found for the simple perceptron [7].

We have performed Monte Carlo simulations to check our analytical findings for the binary model. Figure 1 shows results for $T = 0.5$ and networks with $K = 5$ and $K = 15$ hidden units, respectively. The generalization error is in good quantitative agreement with the theoretical results for the poorly generalizing solution both within the AA and the RS ansatz. However, at this temperature, the simulations stay in the metastable state beyond the predicted location of the thermodynamic transitions indicated by the vertical lines in figure 1. Furthermore, the RS solution only predicts the training error correctly for small values of $\tilde{\alpha}$. With increasing K the deviation from the theoretical line occurs at higher values of $\tilde{\alpha}$. This is another indication of the existence of RS breaking for finite α.

Figure 6 compares the predictions of the AA and the RS theory for the order parameter ρ. Clearly, the divergence of ρ for $\tilde{\alpha} \to \infty$ found in the AA [17] does not correctly describe the results, while there is a good quantitative agreement with the RS solution.

Even at high temperatures the Monte Carlo dynamics is strongly influenced by the presence of the metastable, poorly generalizing state. Only for sufficiently small systems do the simulations follow the thermodynamic transition, as shown in figure 7. For larger values of N a stochastic process would require increasingly long times to cross the free-energy barrier between two minima.

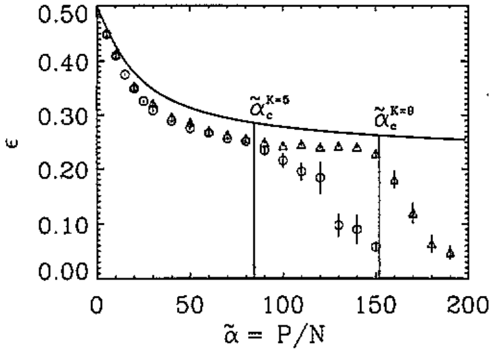

Figure 7. Generalization error for the binary committee machine as a function of $\tilde{\alpha} = P/N$ at $T = 5$. The vertical lines indicate the predictions of $\tilde{\alpha}_c = K\alpha_c$ in the large-K theory (31) for $K = 5$ and 9. The simulations were performed for $N = 75$, $K = 5$ (○) and $N = 25$, $K = 9$ (△).

5. Summary

In summary, we have studied the generalization properties of fully connected committee machines both analytically and numerically. Within an RS ansatz we have calculated the average generalization error as a function of the load parameter α. We have considered networks with both continuous and binary weights and with a large number of hidden units K.

In the limit of small training set sizes, $\alpha \sim \mathcal{O}(1/K)$, we found a committee-symmetric solution where each student weight vector has the same overlap with all the teacher vectors. For both binary and continuous weights the generalization error approaches a non-vanishing residual value. Only if the number of training examples is sufficiently large, $\alpha \sim \mathcal{O}(1)$, can the committee symmetry be broken in favour of a specialization of hidden units. We find first-order phase transitions in both the continuous and the binary model. While in the binary model the transition is accompanied by a perfect alignment of the hidden unit weight vectors with their respective teachers, this is not possible in a continuous model. Instead, we found each student vector would closely approach one of the teachers resulting in an algebraic decay of the generalization error. In both models the symmetric, poorly generalizing state remains metastable for arbitrarily large α.

A similar effect of a decrease in the generalization ability due to a symmetry in the network was also found in a tree parity machine [21]. While the symmetric state in the parity machine fails completely to generalize, the committee-symmetric solution in the present model allows for small student–teacher overlaps and a generalization error $\epsilon_g < 1/2$.

In the binary model a region of negative thermodynamic entropy suggests that RS has to be broken to describe the metastable, symmetric solution correctly at large α. In the continuous model, the RS saddle-point equations allow a committee symmetric solution for all values of α. However, at zero temperature this solution corresponds to a generalization *error $\epsilon_0 > 0$ (20), while the training error vanishes. In particular, we do not find $\epsilon_t \rightarrow \epsilon_g$* for $\alpha \rightarrow \infty$ within this state. This fact suggests the conjecture that this solution loses stability at some $\alpha^* \geqslant \alpha_c$ and a full broken RS calculation will be necessary for its correct description.

A comparison of the RS solution with the results obtained within the AA [17] shows that the AA gives a qualitatively correct description of the main features of the learning curve. However, it fails to predict the temperature dependence of the residual generalization error and gives an incorrect description of the approach to this value. While the AA predicts a divergence of the rescaled order parameters ρ and c, they approach finite values for every finite temperature. Furthermore, the quantitative predictions for the locations of the phase transitions differ considerably (see figures 4 and 5).

Finally, the analysis described in this paper was restricted to the learning of a realizable rule. It would be desirable to discard the symmetry between the target rule and the learning network and extend this work to unlearnable problems.

Acknowledgments

The author would like to thank M Biehl, J Hertz and S Solla for valuable discussions. He wishes to acknowledge the hospitality of the Laboratory of Neuropsychology, NIMH, Bethesda MD, USA, where part of this work was carried out. This work was supported by the EC under the SCIENCE programme and by the Danish Natural Science Council and the Danish Technical Research Council through CONNECT.

Appendix

In this appendix we describe the calculation of the energy term G_r (15). We start out from the general form (8) and introduce the internal fields $u_l^a = N^{-1/2}W_l^a \cdot \xi$ and $v_l = N^{-1/2}V_l \cdot \xi$

through integrals over δ-functions in the standard way, writing

$$\exp[-G_{\rm r}^{(n)}] = \int \prod_{l,a} \frac{\mathrm{d}u_l^a\,\mathrm{d}\hat{u}_l^a}{2\pi} \int \prod_{l} \frac{\mathrm{d}v_l\,\mathrm{d}\hat{v}_l}{2\pi} \left\langle \exp\left[\frac{\mathrm{i}}{\sqrt{N}}\left(\sum_{l,a} \hat{u}_l^a W_l^a + \sum_l \hat{v}_l V_l\right)\cdot \xi\right]\right\rangle_\xi$$

$$\times \exp\left\{ -\mathrm{i}\sum_{a,l} u_l^a \hat{u}_l^a - \mathrm{i}\sum_l v_l \hat{v}_l - \beta\sum_a \Theta\left[-\left(\frac{1}{\sqrt{K}}\sum_l \operatorname{sign} u_l^a\right)\left(\frac{1}{\sqrt{K}}\sum_l \operatorname{sign} v_l\right)\right]\right\}. \tag{33}$$

The components ξ_j^μ of the training examples are drawn from a Gaussian distribution with zero mean and unit variance. Hence, the average over inputs reduces to a simple Gaussian integral and yields

$$\left\langle \exp\left[\frac{\mathrm{i}}{\sqrt{N}}\left(\sum_{l,a} \hat{u}_l^a W_l^a + \sum_l \hat{v}_l V_l\right)\cdot \xi\right]\right\rangle_\xi = \exp\left(-\frac{1}{2}\sum_l (\hat{v}_l)^2 - \sum_{l,k,(a,b)} D_{lk}^{ab}\hat{u}_l^a\hat{u}_k^b - \sum_{a,(l,k)} C_{lk}^a \hat{u}_l^a \hat{u}_k^a - \sum_{a,l,k} R_{lk}^a \hat{u}_l^a \hat{v}_l\right) \tag{34}$$

where we have introduced the order parameters (9) and used the orthogonality and normalization of the teacher vectors. To rewrite the Boltzmann factor in (33) we use the identities $\mathrm{e}^{-\beta\Theta(-ab)} = \Theta(ab) + \mathrm{e}^{-\beta}\Theta(-ab)$ and $\Theta(-ab) = \Theta(-a)\Theta(b) + \Theta(a)\Theta(-b)$ and introduce internal representations, writing

$$\begin{aligned} \Theta\left(\frac{1}{\sqrt{K}}\sum_l \operatorname{sign} v_l\right) &= \sum_{\{\tau_l=\pm 1\}} \Theta\left(\frac{1}{\sqrt{K}}\sum_l \tau_l\right)\prod_l \Theta(\tau_l v_l) \\ \Theta\left(\frac{1}{\sqrt{K}}\sum_l \operatorname{sign} u_l^a\right) &= \sum_{\{\sigma_l^a=\pm 1\}} \Theta\left(\frac{1}{\sqrt{K}}\sum_l \sigma_l^a\right)\prod_l \Theta(\sigma_l^a u_l^a). \end{aligned} \tag{35}$$

At the RS saddle point (12), the integrals over the $\hat{v}_l$s can be done. Furthermore, the integrals over $\hat{u}_l^a$s and u_l^as can be simplified, and a straightforward calculation leads us to

$$\exp[-G_{\rm r}^{(n)}] = 2\sum_{\{\tau_l\}} \Theta\left(\frac{1}{\sqrt{K}}\sum_l \tau_l\right)\int \mathrm{D}y \int \prod_l \mathrm{D}t_l \int \prod_l \mathrm{D}v_l\,\Theta(\tau_l v_l)$$

$$\times \prod_a \left\{ \int \mathrm{D}s^a \sum_{\{\sigma_l^a\}}\left[\Theta\left(\frac{1}{\sqrt{K}}\sum_l \sigma_l^a\right) + \mathrm{e}^{-\beta}\Theta\left(-\frac{1}{\sqrt{K}}\sum_l \sigma_l^a\right)\right] \times \prod_l H(-\sigma_l^a z_l^a)\right\} \tag{36}$$

where

$$z_l^a = \frac{\Delta v_l + R\sum_k v_k + \sqrt{q-\Delta^2}\,t_l - \mathrm{i}\sqrt{KR^2+2R\Delta-D}\,y - \mathrm{i}\sqrt{D-C}\,s^a}{\sqrt{1-q-C}}. \tag{37}$$

Here we have used the abbreviations $\mathrm{D}x = \mathrm{e}^{-x^2/2}\,\mathrm{d}x/\sqrt{2\pi}$ and $H(x) = \int_x^\infty \mathrm{D}t$. To do the traces, we introduce integral representations of the Θ-functions

$$\Theta\Big(K^{-1/2}\sum_l \sigma_l^a\Big) = \int_0^\infty \mathrm{d}\lambda^a \int_{-\infty}^{+\infty} \frac{\mathrm{d}x^a}{2\pi} \exp\Big[-\mathrm{i}x^a\Big(\lambda^a - K^{-1/2}\sum_l \sigma_l\Big)\Big] \tag{38}$$

and similarly for $\Theta(K^{-1/2}\sum_l \tau_l)$ with the integration variables μ and y. Now the traces can be done, using

$$\begin{aligned}\sum_{\{\sigma_l^a\}} \exp\left[\frac{\mathrm{i}x^a}{\sqrt{K}}\sum_l \sigma_l^a\right]\prod_l H(-\sigma_l^a z_l^a) &= \prod_l \left[\cos\frac{x^a}{\sqrt{K}} + \mathrm{i}\hat{H}(z_l^a)\sin\frac{x^a}{\sqrt{K}}\right]\\ \sum_{\{\tau_l\}} \exp\left[\frac{\mathrm{i}y}{\sqrt{K}}\sum_l \tau_l\right]\prod_l \Theta(\tau_l v_l) &= \prod_l \left[\cos\frac{y}{\sqrt{K}} + \mathrm{i}\,\mathrm{sign}(v_l)\sin\frac{y}{\sqrt{K}}\right]\end{aligned} \tag{39}$$

with the definition $\hat{H}(x) = 1 - 2H(x)$. For large K we can expand the result, using the scaling ansatz (14) for the order parameters and

$$\begin{aligned}\hat{H}(z_l^a) = \hat{H}&\left(\frac{\Delta v_l + \sqrt{q-\Delta^2}\,t_l}{\sqrt{1-q}}\right)\\ &+ \frac{1}{\sqrt{K}}\sqrt{\frac{2}{\pi}}\,\frac{\rho K^{-1/2}\sum_k v_k - \mathrm{i}\sqrt{\rho^2+2\rho\Delta-d}\,y - \mathrm{i}\sqrt{d-c}\,s^a}{\sqrt{1-q}}\\ &\times \exp\left[-\frac{1}{2}\left(\frac{\Delta v_l + \sqrt{q-\Delta^2}\,t_l}{\sqrt{1-q}}\right)^2\right] + \mathcal{O}(1/K).\end{aligned} \tag{40}$$

The integrals over the v_ls can be factorized if we control the sum $K^{-1/2}\sum_k v_k$ by a new variable w, writing

$$1 = \int_{-\infty}^{+\infty} \mathrm{d}w\,\delta\Big(w - K^{-1/2}\sum_k v_k\Big) = \int_{-\infty}^{+\infty} \frac{\mathrm{d}w\,\mathrm{d}\nu}{2\pi}\exp\Big[-\mathrm{i}\nu\Big(w - K^{-1/2}\sum_k v_k\Big)\Big]. \tag{41}$$

Now, a straightforward calculation similar to the one in the AA [18] leads to

$$\begin{aligned}\exp[-G_{\mathrm{r}}^{(n)}] = 2\int \mathrm{D}x\, H&\left(\frac{R_{\mathrm{eff}}x}{\sqrt{Q_{\mathrm{eff}} - R_{\mathrm{eff}}^2}}\right)\\ &\times\left[\mathrm{e}^{-\beta} + (1+\mathrm{e}^{-\beta})H\left(\sqrt{\frac{Q_{\mathrm{eff}}}{1+(2/\pi)c - Q_{\mathrm{eff}}}}\,x\right)\right]^n\end{aligned} \tag{42}$$

with R_{eff} and Q_{eff} given by (16). Taking the limit $n \to 0$ finally yields equation (15) for $G_{\mathrm{r}} = \lim_{n\to 0} G_{\mathrm{r}}^{(n)}/n$.

References

[1] Hertz J A, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Redwood City, CA: Addison–Wesley)

[2] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257

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

[4] Rosenblatt F 1962 *Principles of Neurodynamics* (New York: Spartan)

[5] Levin E, Tishby N and Solla S A 1989 *Proc. 2nd Workshop on Computational Learning Theory* (San Mateo: Morgan Kaufmann)

[6] Krogh A and Hertz J 1992 *Advances in Neural Information Processing Systems* vol IV, ed J E Moody, S J Hanson and R P Lippmann (San Mateo: Morgan Kaufmann)

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

[8] Denker J S, Schwarz D, Wittner B, Solla S, Howard R, Jackel L and Hopfield J 1987 *Complex Systems* **1** 877

[9] de Figueiredo R J P 1980 *J. Math. Anal. Appl.* **38** 1227

[10] Hecht–Nielsen R 1987 *IEEE 1st Int. Conf. on Neural Networks* vol II, ed M Caudill and C Butler (New York: IEEE)

[11] Nilsson N J 1965 *Learning Machines* (New York: McGraw-Hill)

[12] Sompolinsky H and Tishby N 1990 *Europhys. Lett.* **13** 567

[13] Schwarze H and Hertz J *Europhys. Lett.* **20** 375

[14] Mato G and Parga N 1992 *J. Phys. A: Math. Gen.* **25** 5047

[15] Barkai E, Hansel D and Sompolinsky H 1992 *Phys. Rev.* A **45** 4146–61

[16] Engel A, Köhler H M, Tschepke F, Vollmayr H and Zippelius A *Phys. Rev.* A **45** 7590–607

[17] Schwarze H and Hertz J 1993 *Europhys. Lett.* **21** 785–90

[18] Schwarze H and Hertz J 1993 *J. Phys. A: Math. Gen.* at press

[19] Kang K, Oh J-H, Kwon C and Park Y 1993 *Preprint* Pohang Institute of Science and Technology, Korea

[20] Györgyi G and Tishby N 1990 *Neural Networks and Spin Glasses* ed K Thuemann and R Köberle (Singapore: World Scientific)

[21] Hansel D, Mato G and Meunier C 1992 *Europhys. Lett.* **20** 471