Learning from examples in fully connected committee machines

1993 J. Phys. A: Math. Gen. 26 4919

(http://iopscience.iop.org/0305-4470/26/19/024)

J. Phys. A: Math. Gen. 26 (1993) 4919–4936. Printed in the UK

Learning from examples in fully connected committee machines

H Schwarze and J Hertz

CONNECT, The Niels Bohr Institute and Nordita, Blegdamsvej 17, DK-2100 Copenhagen Ø, Denmark

Received 19 January 1993

Abstract. The problem of learning from examples in two-layer neural networks is studied within the framework of statistical mechanics. A fully connected committee machine is trained to implement a task which is not linearly separable. The generalization error as a function of the number of training examples per adjustable weight is calculated in the annealed approximation. For both binary and continuous weights we find a first-order transition with a discontinuous drop in the generalization error. The transitions occur due to a specialization of the hidden units from a symmetric state to one with each hidden unit in the student network specialized on a corresponding unit in the teacher network. The symmetric states of poor generalization remain metastable even for large training sets.

1. Introduction

The ability to extract an underlying rule from a given data set is a key feature of feedforward neural networks [1]. Such a network performs a mapping from any input configuration to the output parametrized by a set of weights. The goal of learning is to adapt the weights to model an unknown mapping which has produced a given set of input-output pairs.

There has been much interest in applying methods from statistical mechanics to study the learning properties of neural networks (for a recent review see, e.g., [2]). Two properties have been of particular interest. The capacity of a network reflects the maximum number of random input–output relations it can store, while the generalization error describes the quality of the modelling achieved by the network for a given rule. After the network has been trained on a set of examples generated by the target rule, the generalization error measures its average error on an arbitrary new input.

Following Gardner's approach [3] these properties have initially been explored for the simplest network, the single-layer perceptron [4] with one layer of weights connecting inputs and outputs [3, 5, 6, 7, 8]. As an extension of this work to more complex networks recent interest has focussed on networks with an additional layer of hidden units. While the computational power of single-layer perceptrons is limited to linearly separable tasks, networks with only one hidden layer can implement any Boolean or continuous function of the inputs [9, 10].

As an example of such an architecture the so-called committee machine [11] has been studied [12, 13, 14, 15, 16]. In this network the weights from the hidden to the output layer are fixed to +1 and the network realizes a majority decision of the hidden units. For binary weights this may already be regarded as the most general two-layer architecture,

0305-4470/93/194919+18$07.50

because any other combination of weights can be gauged to +1 by flipping the signs of the corresponding input-hidden weights.

Recent work has been concerned with the tree-version of this model, in which the hidden units receive their inputs from non-overlapping regions of the input layer. Both the capacity [12, 14] and the generalization problem [15, 16] have been studied for this restricted architecture. Due to the lack of correlations among the hidden units the tree committee machine exhibits generalization properties similar to a simple perceptron. In the limit of a large number of hidden units the difference to a simple perceptron can be expressed by an effective choice of the order parameters.

In the more general fully connected committee machine each hidden unit receives inputs from the entire input layer, yielding correlations between different hidden units. The capacity problem for this architecture has recently been studied in [12, 14]. For large committees correlations between hidden units tend to zero, revealing the same capacity per synapse as in the tree model [14]. In this paper we explore the learning of a rule in a fully connected committee machine, calculating the generalization error for a target rule which itself is defined by a committee machine and therefore not linearly separable. We do this within the annealed approximation and in the limit of a large network, in which both the number of inputs N and the number of hidden units K tend to infinity. We find that the generalization properties are different from those found in the tree version.

The paper is organized as follows. In section 2 the model is described in more detail, and the statistical mechanics approach is outlined. In sections 3 and 4 the calculation of the generalization error is presented for networks with both continuous and binary weights. Finally, section 5 gives a summary of the results and a brief discussion. The results of this work have been reported previously in preliminary form in [17].

2. The model

Throughout this paper we will study a two-layer network with N input units $S_i, i \in \{1, \ldots, N\}$, one layer of K *hidden units* $\sigma_l, l \in \{1, \ldots, K\}$, and a single output unit (see figure 1). The hidden units will be referred to as *students*. They are each simple perceptrons defined by N-dimensional weight vectors $\boldsymbol{W}_l$ with outputs

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

All students receive the same input and are connected to the overall output through weights which are all fixed to +1. Hence, the network output is given by

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

This architecture is called a *committee machine* [11], because the network output corresponds to the output of the majority of the hidden units. We study supervised learning of a given rule 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 uncorrelated, $N^{-1}\, \boldsymbol{V}_l \cdot \boldsymbol{V}_k = \delta_{lk}$. Note that in the thermodynamic limit the orthogonality condition does not have to be imposed explicitely, because K randomly drawn vectors (with $K \ll N$) will always be orthogonal. Furthermore, no corresponding restriction is imposed on the student weights, and the training dynamics can introduce correlations between different hidden units, because their receptive fields

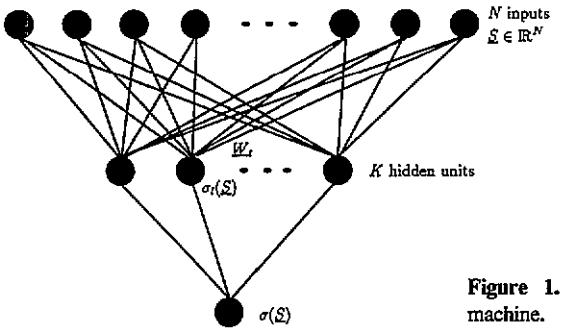

Figure 1. Fully connected committee machine.

are identical. The only information about the target rule available during training is a set of $P = \alpha K N$ input–output pairs $(\xi^\mu, \tau(\xi^\mu))$, $\mu \in \{1, \ldots, P\}$, of the teacher network. No explicit information about the states of the teacher hidden units is provided. The components of the training inputs ξ_i^μ are drawn independently from a Gaussian distribution with zero mean and unit variance. Note, however, that for large N our results do not depend on the specific form of the distribution.

The statistical mechanics approach reviewed in [2] formulates learning as a stochastic minimization process with a formal energy given by the *training error*, here defined as the number of misclassified training examples

$$E(\{\boldsymbol{W}_l\}) = \sum_{\mu=1}^{P} \Theta[-\sigma(\xi^\mu) \cdot \tau(\xi^\mu)] . \tag{3}$$

We consider a stochastic learning process that for long training times yields a Gibbs distribution in the space of couplings with measure

$$\mathrm{d}\rho_G(\{\boldsymbol{W}_l\}) = \frac{\mathrm{d}\rho_0(\{\boldsymbol{W}_l\})}{Z} \, \mathrm{e}^{-\beta E(\{\boldsymbol{W}_l\})} . \tag{4}$$

Constraints on the student weights are included in the normalized *a priori* measure $\mathrm{d}\rho_0(\{\boldsymbol{W}_l\})$. For continuous weights this is a product of delta functions constraining each $\boldsymbol{W}_l$ to length $\sqrt{N}$. For binary weights we just get a sum over all $\boldsymbol{W}_l \in \{\pm 1\}^N$. Z is the partition function

$$Z = \int \mathrm{d}\rho_0 \, (\{\boldsymbol{W}_l\}) \, \mathrm{e}^{-\beta E(\{\boldsymbol{W}_l\})} \tag{5}$$

and the formal temperature $T = 1/\beta$ describes the amount of noise during the training process.

Generally, the goal of learning is to achieve the best possible network performance for new inputs which were not used as training examples. The quantity one would like to minimize is the *generalization error*, the probability of misclassifying a new example randomly chosen from the entire distribution of inputs

$$\epsilon(\{\boldsymbol{W}_l\}) = \langle \Theta[-\sigma(\boldsymbol{S}) \cdot \tau(\boldsymbol{S})] \rangle_{\boldsymbol{S}} . \tag{6}$$

Here, we focus on the calculation of the average generalization error $\epsilon_g = \langle\!\langle \, \langle \epsilon(\{\boldsymbol{W}_l\}) \rangle_T \, \rangle\!\rangle$, where $\langle \cdots \rangle_T$ denotes a thermal average over the distribution (4) and $\langle\!\langle \cdots \rangle\!\rangle$ denotes a quenched average over all possible training sets.

Within the framework of statistical mechanics this quantity may be determined from the average free energy $F = -T \, \langle\!\langle \ln Z \rangle\!\rangle$. As a useful approximation we employ the annealed

approximation, which replaces the average of the logarithm of Z by the logarithm of the average of Z, yielding a lower bound for the free energy $F_{an} = -T \ln \langle\langle Z \rangle\rangle \leq F$. In general the annealed approximation is only valid at high temperatures. However, previous work indicates that for realizable rules, as in the present problem, this bound serves as a good approximation and predicts correctly the shape of learning curves at finite T [7, 18].

If the training examples are drawn independently from the same distribution, the average of the partition function over the training examples yields

$$\langle\langle Z \rangle\rangle = \int \mathrm{d}\rho_0 \left(\{\boldsymbol{W}_l\}\right) \mathrm{e}^{-P G(\{\boldsymbol{W}_l\})} \tag{7}$$

with an effective Hamiltonian $G(\{\boldsymbol{W}_l\})$, which can be written for Boolean outputs as [7]

$$G(\{\boldsymbol{W}_l\}) = -\ln[1 - (1 - \mathrm{e}^{-\beta})\,\epsilon(\{\boldsymbol{W}_l\})]\,. \tag{8}$$

As will be shown below, $\epsilon(\{\boldsymbol{W}_l\})$ can be expressed as a function of two sets of order parameters: the overlaps between the student and teacher weight vectors, $R_{lk} = N^{-1}\,\boldsymbol{W}_l \cdot \boldsymbol{V}_k$, and the mutual overlaps in the student network $C_{lk} = N^{-1}\,\boldsymbol{W}_l \cdot \boldsymbol{W}_k$. Therefore we can replace the integration over the weights $\boldsymbol{W}_l$ by integrations over R_{lk} and C_{lk}, introducing an additional 'entropy' factor

$$\exp\{N K G_0[R_{lk}, C_{lk}]\} = \int \mathrm{d}\rho_0(\{\boldsymbol{W}_l\}) \prod_{l,k} \delta\left(R_{lk} - \frac{1}{N}\boldsymbol{W}_l \cdot \boldsymbol{V}_k\right) \times \prod_{(l,k)} \delta\left(C_{lk} - \frac{1}{N}\boldsymbol{W}_l \cdot \boldsymbol{W}_k\right). \tag{9}$$

Now (7) becomes

$$\langle Z \rangle = \int [\mathrm{d}R_{lk}\mathrm{d}C_{lk}]\ \exp\{-N K(\alpha\, G[R_{lk}, C_{lk}] - G_0[R_{lk}, C_{lk}])\}$$

and in the thermodynamic limit ($N \to \infty$) the order parameters can be obtained by minimizing the free energy density f with respect to R_{lk} and C_{lk}, where

$$\beta f[R_{lk}, C_{lk}] = \alpha G[R_{lk}, C_{lk}] - G_0[R_{lk}, C_{lk}]\,. \tag{10}$$

Using the values of the order parameters at the minimum of f, we can calculate the average generalization error. In the following sections this will be described for committee machines with both continuous and binary weights in the limit of a large number of hidden units K.

3. Continuous weights

The calculation of the generalization error (6) and the entropy term (9) follows the methods introduced by Gardner [3] and Györgyi and Tishby [5] for simple perceptrons and subsequently extended to the case of two-layer architectures [12, 14, 15]. In order to proceed, we need to make symmetry assumptions for the order parameters R_{lk} and C_{lk} at the minimum of the free energy. Since the target rule in our problem has the same structure as the student network, the generalization error vanishes for the choice $R_{lk} = C_{lk} = \delta_{lk}$ of the order parameters. During training, the network only receives information about the overall target output rather than the outputs of the individual hidden units. It is therefore

reasonable to make a translationally-symmetric ansatz for the order parameters R_{lk} and C_{lk} but allowing for an additional specialization of the individual students which breaks this symmetry. We assume that each student has an overlap $R+\Delta$ with one of the teachers and an overlap R with the remaining ones, yielding

$$R_{lk} = R + \Delta\delta_{lk} \qquad C_{lk} = C + (1-C)\delta_{lk}\,. \tag{11}$$

With this ansatz the integral over weight space in the expression of the entropy term (9) can be performed after introducing integral representations of the δ-functions. A straightforward calculation using the saddle point method yields (see appendix A)

$$G_0(\Delta, R, C) = \frac{1}{2}(1+\ln 2\pi) + \frac{1}{2}\frac{K-1}{K}\ln(1-C-\Delta^2)$$
$$+ \frac{1}{2K}\ln[(1-C-\Delta^2) - K(KR^2 - C + 2\Delta R)]\,. \tag{12}$$

Note that the last term in (12) imposes geometrical constraints on the possible values for R and C. The argument of the logarithm has to be positive, yielding the conditions

$$KR^2 - C + 2\Delta R < \frac{1}{K} \qquad \text{and therefore} \quad R < \frac{1}{\sqrt{K}}\left[1 + \mathcal{O}\left(\frac{1}{K}\right)\right] \tag{13}$$

where we have assumed that $R, \Delta, C \geq 0$. The second condition can be interpreted as the maximal overlap of a single student vector with K orthogonal teacher vectors. For the following it will be convenient to introduce the abbreviation $D = K(KR^2 - C + 2\Delta R) < 1$.

The calculation of the generalization error for a given network (6) requires an average over the distribution of inputs. We introduce new variables for the internal fields $u_l = N^{-\frac{1}{2}}\boldsymbol{W}_l \cdot \boldsymbol{S}$ in the student network and $v_l = N^{-\frac{1}{2}}\boldsymbol{V}_l \cdot \boldsymbol{S}$ in the teacher network, respectively, by inserting corresponding δ-functions. Using integral representations of the δ-functions now yields

$$\epsilon(\{\boldsymbol{W}_l\}) = \int\prod_l \frac{\mathrm{d}u_l\,\mathrm{d}\hat{u}_l}{2\pi} \int\prod_l \frac{\mathrm{d}v_l\,\mathrm{d}\hat{v}_l}{2\pi} \exp\left(-\mathrm{i}\sum_l (u_l\hat{u}_l + v_l\hat{v}_l)\right)$$
$$\times\,\Theta\left(-K^{-1/2}\sum_l \mathrm{sign}(u_l)K^{-1/2}\sum_l \mathrm{sign}(v_l)\right)$$
$$\times\left\langle \exp\left(\frac{\mathrm{i}}{\sqrt{N}}\sum_l (\hat{u}_l\boldsymbol{W}_l + \hat{v}_l\boldsymbol{V}_l)\cdot\boldsymbol{S}\right)\right\rangle_S . \tag{14}$$

After averaging over the distribution of inputs and introducing the order parameters with the ansatz (11) the integrals over the $\hat{u}_l$'s and $\hat{v}_l$'s can be done. Furthermore we introduce internal representations $\{\sigma_l\}$ for the student network and $\{\tau_l\}$ for the teacher network to obtain (for details see appendix B)

$$\epsilon(\Delta, R, C) = 2\sum_{\{\sigma_l\}}\sum_{\{\tau_l\}} \Theta\left(-K^{-1/2}\sum_l \tau_l\right)\Theta\left(K^{-1/2}\sum_l \sigma_l\right)$$
$$\times \int_{-\infty}^{+\infty}\mathrm{D}t \int_{-\infty}^{+\infty}\prod_l \mathrm{D}v_l \prod_l \Theta(\tau_l v_l)\prod_l H\left(\sigma_l \frac{\mathrm{i}t\sqrt{D/K} - R\sum_k v_k - \Delta v_l}{\sqrt{(1-C)-\Delta^2}}\right) \tag{15}$$

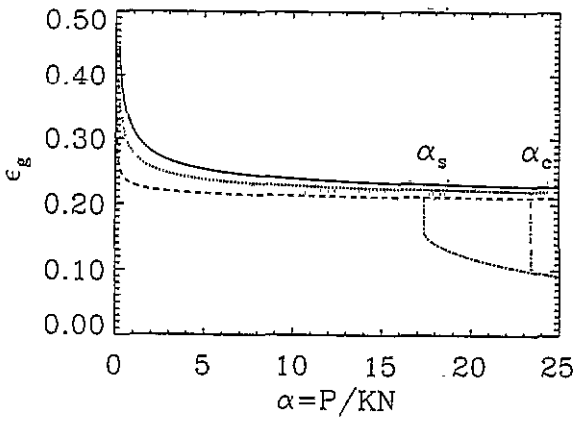

Figure 2. Learning curve for a fully connected committee machine with continuous weights at $T = 0.5$. The approach to the residual error is shown, including $1/\sqrt{K}$ corrections for $K = 5$ (solid line), $K = 11$ (dotted line), and $K = 100$ (dashed line). The broken line corresponds to the solution with $\Delta > 0$.

where $H(x) = \int_x^\infty \mathrm{Dz}$ and $\mathrm{Dz} = \mathrm{d}z\, \mathrm{e}^{-z^2/2}/\sqrt{2\pi}$.

Further simplifications of (15) for large K require assumptions about how the order parameters scale with the number of hidden units. The possible magnitude of the symmetric overlaps R and C is restricted by geometrical limitations, and the conditions (13) imposed by the entropy term provide upper bounds for the scaling of R and C. However, at the minimum of the free energy a different scaling is valid. To see this, we introduce the rescaled parameters

$$r = K^{3/4}R \qquad c = K^{1/2}C \tag{16}$$

and determine their values at the minimum of the free energy (10) self-consistently. With this scaling assumption we can proceed with the calculation of the generalization error for large K as described in appendix B. To order $1/\sqrt{K}$ we obtain

$$\epsilon(\Delta, r, c) = \frac{1}{\pi}\cos^{-1}\left(\sqrt{\frac{2}{\pi}}\frac{r + K^{-1/4}\sin^{-1}\Delta}{\sqrt{c + K^{-1/2}\pi/2}}\right). \tag{17}$$

The entropy (12), expressed as a function of r and c, is

$$\begin{aligned} G_0(\Delta, r, c) = {} & \frac{1}{2}(1 + \ln 2\pi) + \frac{1}{2}\frac{K-1}{K}\ln(1 - \Delta^2 - c/\sqrt{K}) \\ & + \frac{1}{2K}\ln[(1 - \Delta^2 - c/\sqrt{K}) - \sqrt{K}(r^2 - c + 2K^{-1/4}r\Delta)] \\ & + \mathcal{O}\left(\frac{1}{K^2}\right). \end{aligned} \tag{18}$$

With these expressions for the entropy and the generalization error, together with (8), we can minimize the free energy (10) to find the equilibrium values of the order parameters. Taking the derivatives of the free energy with respect to Δ, r and c yields after some algebra the conditions

$$\Delta = 0 \qquad r = \left[\frac{\alpha}{4\pi^2\gamma(\beta)}\left(\frac{\pi - 2}{2}\right)\right]^{1/4} + \mathcal{O}\left(\frac{1}{\sqrt{K}}\right) \qquad r^2 - c = \frac{1}{\sqrt{K}} - \mathcal{O}\left(\frac{1}{K}\right) \tag{19}$$

with $\gamma(\beta) = \sqrt{\pi/2 - 1}\,[(1 - \mathrm{e}^{-\beta})^{-1} - \epsilon_0]/(4\pi)$ and $\epsilon_0 = \pi^{-1}\cos^{-1}(\sqrt{2/\pi}) \approx 0.206$. At high temperatures $\gamma(\beta)$ is proportional to T. In terms of the original parameters equation

(19) implies $C - KR^2 < 0$. This corresponds to an effective anticorrelation of the hidden units, because student vectors which are uncorrelated except for their common overlap R to K orthogonal teacher vectors would yield $C - KR^2 = 0$. A similar anticorrelation was also found in the capacity calculation for this architecture [12, 14] and can be interpreted as a 'division of labour'[14] between hidden units. Inserting (19) in (17) yields the average generalization error as a function of α

$$\epsilon_g = \epsilon_0 + \sqrt{\frac{\gamma(\beta)}{\alpha K}} + \mathcal{O}\left(\frac{1}{K}\right) \qquad \epsilon_0 = \frac{1}{\pi}\cos^{-1}\left(\sqrt{\frac{2}{\pi}}\right) \approx 0.206\,. \tag{20}$$

For large values of α the leading correction in (20) vanishes and the generalization error approaches a non-vanishing constant value as shown in figure 2 for different values of K. An approach to the optimal value $\epsilon_g = 0$ cannot happen in this state, because it would require a close approach of each student to one of the teachers, implying $\Delta > 0$. On the other hand, a deviation from $\Delta = 0$ does not happen, because the solution with non-specialized students is favored by the entropy term (18), while Δ does not contribute to leading order to the generalization error (17).

So far we have assumed that the parameters r and c are of order one. However, if they are very small, Δ will dominate both terms in the free energy and a non-vanishing equilibrium value of Δ is possible. To investigate the free energy in this region of parameter space we introduce a new set of parameters, given by

$$\rho = KR = K^{1/4}r \qquad q = KC = \sqrt{K}c\,. \tag{21}$$

which are now assumed to be of order one. Inserting these into (17) and (18) yields

$$\begin{aligned}\epsilon(\Delta,\rho,q) &= \frac{1}{\pi}\cos^{-1}\left(\sqrt{\frac{2}{\pi}}\frac{\rho+\sin^{-1}\Delta}{\sqrt{q+\pi/2}}\right)\\ G_0(\Delta,\rho,q) &= \frac{1}{2}(1+\ln 2\pi) + \frac{1}{2}\ln(1-\Delta^2) - \frac{1}{2K}\frac{q}{1-\Delta^2}\\ &\quad + \frac{1}{2K}\ln[1-(\rho+\Delta)^2+q] - \frac{1}{2K}\ln(1-\Delta^2) + \mathcal{O}\left(\frac{1}{K^2}\right).\end{aligned} \tag{22}$$

The values of ρ and q at the minimum of the corresponding free energy can readily be obtained as

$$q = (\rho+\Delta)^2 - 1 + \mathcal{O}\left(\frac{1}{K}\right) \qquad \rho = \frac{\pi-2}{2}\frac{1}{\sin^{-1}\Delta - \Delta} - \Delta + \mathcal{O}\left(\frac{1}{K}\right). \tag{23}$$

Hence, to leading order in $1/K$ we are left with a free energy as a function of Δ

$$\begin{aligned}\beta f(\Delta) &= -\alpha\,\ln[1-(1-\mathrm{e}^{-\beta})\epsilon(\Delta)] - \frac{1}{2}(1+\ln 2\pi) - \frac{1}{2}\ln(1-\Delta^2)\\ \epsilon(\Delta) &= \frac{1}{\pi}\cos^{-1}\left[\sqrt{\frac{2}{\pi}}\sqrt{1+\frac{2}{\pi-2}(\Delta-\sin^{-1}\Delta)^2}\right].\end{aligned} \tag{24}$$

The term independent of α is simply the entropy of a simple perceptron with overlap Δ to the teacher. As shown in figure 3, for small values of α, $f(\Delta)$ has a single minimum at $\Delta = 0$ corresponding to the residual generalization error ϵ_0 (20). Note that for $\Delta \to 0$ the order parameters ρ and q diverge. This is due to the fact that for this solution the assumed scaling is not correct. As was shown above, the symmetric solution (19) is characterized by the scaling (16), which corresponds to infinite ρ and q. However, at a critical value α_s

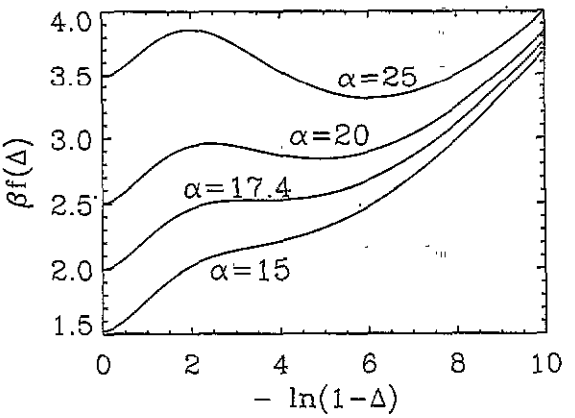

Figure 3. Free energy $f(\Delta)$ for continuous weights and different values of α. The free energy has always a minimum at $\Delta = 0$, but with increasing α a second minimum appears at Δ close to 1. For large α this becomes the global minimum of f. Note that the horizontal axis has a logarithmic scale in $(1 - \Delta)$ to show both minima in one plot.

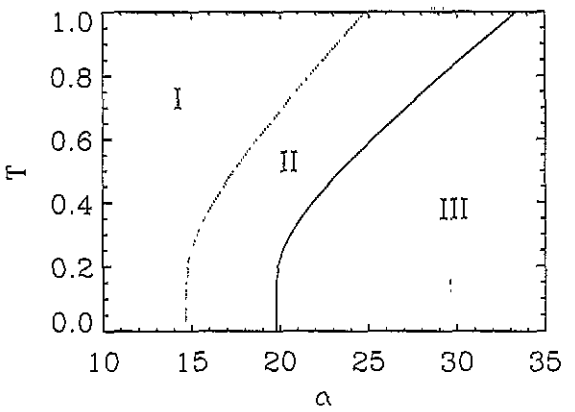

Figure 4. Annealed phase diagram for the large-K committee machine with continuous weights. The solid line marks the location of the phase transition, while the dotted line corresponds to the spinodal line. In region I $f(\Delta)$ has only a minimum at $\Delta = 0$, in region II the additional minimum with Δ close to 1 is metastable, and in region III this becomes the global minimum of $f(\Delta)$.

of the load parameter, a second local minimum appears at a value of Δ close to 1 and finite values for ρ and q. It becomes the global minimum at a higher $\alpha_c > \alpha_s$. Thus the system undergoes a discontinuous transition from a symmetric state to a specialized state in which each student has an overlap of order 1 with one of the teachers and an overlap of order $1/K$ with the remaining ones. The generalization error of the specialized state decays smoothly (see figure 2). In the limit $\alpha \to \infty$ we find $\Delta \to 1$ and $\epsilon_g \approx 2(1 - e^{-\beta})^{-1}/\alpha$. As can be seen from equation (23), the order parameters q and ρ vanish for $\Delta \to 1$, reflecting the structure of the target rule. Only when the network has learnt the rule perfectly does the student network adopt the orthogonality of the teacher vectors. The locations of the appearance of the second minimum in the free energy and the phase transition are shown in the phase diagram in figure 4.

The symmetric state remains metastable for all $\alpha > \alpha_c$, and a stochastic learning procedure starting with $\Delta = 0$ will first settle into this local minimum. For large N it will take an exponentially long time to cross the free energy barrier to the global minimum of f.

In the foregoing discussion we have considered a training set of size proportional to the number of adjustable weights in the network, and the load parameter $\alpha = P/(KN)$ was assumed to be finite. Although they reveal the interesting features of the generalization curve, our results are not valid for all values of α. The leading correction to the residual error in (20) is only small for $\alpha \gg 1/K$ and diverges for $\alpha \to 0$. However, for the region

where α itself is of order $1/K$ we can construct another self-consistent solution for the minimum of the free energy. In the following we assume that α is small, such that $\tilde{\alpha} = K\alpha$ is of order one. Using the same scaling assumption for the order parameters as in (21) we now have to minimize the free energy density

$$\beta f(\Delta, \rho, q) = -\frac{\tilde{\alpha}}{K} \ln[1 - (1 - \mathrm{e}^{-\beta})\epsilon(\Delta, \rho, q)] - G_0(\Delta, \rho, q) . \tag{25}$$

with $\epsilon(\Delta, \rho, q)$ and $G_0(\Delta, \rho, q)$ given by (22). This scaling of the order parameters ensures that in both contributions to f the terms depending on ρ and q are of the same order in $1/K$. However, the Δ-dependence is dominated by the entropy term, forcing the system to stay in the poorly-generalizing solution with $\Delta = 0$. At the minimum of the free energy we get the conditions

$$\begin{aligned} \Delta &= 0 \\ \rho^2 &= q \, \frac{1 + 2q/\pi}{(1 + 2q/\pi) - 2/\pi} \\ 0 &= \frac{2\tilde{\alpha}}{\pi^2} \frac{1}{(1 - \mathrm{e}^{-\beta})^{-1} - \epsilon(\Delta, \rho, q)} - \sqrt{\frac{\pi}{\pi - 2}} \sqrt{q(1 + 2q/\pi)^3} . \end{aligned} \tag{26}$$

The last equation can be solved numerically yielding a smoothly decreasing generalization error as a function of $\tilde{\alpha}$. The asymptotic solution for large values of $\tilde{\alpha}$ can easily be obtained and is determined by $q \propto \sqrt{\tilde{\alpha}}$, $\rho^2 = q(1 + \mathcal{O}(1/\sqrt{\tilde{\alpha}}))$. The generalization error behaves like $\epsilon_g = \epsilon_0 + \mathcal{O}(1/\sqrt{\tilde{\alpha}})$ approaching the residual error given by (20). There is no phase transition in this region of α. When the number of training examples is not proportional to the number of hidden units, the free energy (25) is dominated by the entropy term, while the term proportional to $\tilde{\alpha}$ only contributes to order $1/K$. Therefore, the average generalization error can not achieve its optimal value $\epsilon_g = 0$.

4. Binary weights

In this section we consider a committee machine in which all the weights are restricted to the values ± 1. First we study the scaling ansatz (21), which for continuous weights revealed the location of the phase transition. In the case of binary weights we only have to modify the entropy term (9), which now involves a sum over all configurations $\boldsymbol{W}_l \in \{\pm 1\}^N$. A calculation similar to the one preceding equation (24) leads again to the problem of minimizing a free energy as a function of Δ

$$\beta f(\Delta) = -\alpha \ln[1 - (1 - \mathrm{e}^{-\beta})\epsilon(\Delta)] + \frac{1 + \Delta}{2} \ln\left(\frac{1 + \Delta}{2}\right) + \frac{1 - \Delta}{2} \ln\left(\frac{1 - \Delta}{2}\right) \tag{27}$$

with the same expression for $\epsilon(\Delta)$ as in (24). The entropy term now corresponds to the entropy of a simple perceptron with binary weights. Due to this difference, f now has two local minima for all values of α, one at $\Delta = 0$ and the other at $\Delta = 1$ (see figure 5). For small α the minimum at $\Delta = 0$ is the global one. The corresponding state with non-specialized hidden units and with generalization error ϵ_0 is the equilibrium state. As for continuous weights, the correct description of this solution requires a different scaling of the order parameters, which will be described below. At a critical value of the load parameter, given by

$$\alpha_c = -\frac{\ln 2}{\ln[1 - (1 - \mathrm{e}^{-\beta})\epsilon_0]} \tag{28}$$

the free energy of this state vanishes. For larger values $\alpha > \alpha_c$ the minimum at $\Delta = 1$ has the lower free energy. We have a discontinuous transition from a symmetric state to a state of perfect generalization in which each student is perfectly aligned with one of the teachers. As in the continuous case, the minimum at $\Delta = 0$ does not vanish even for large α. Figures 6 and 7 show the large-K predictions for the generalization error of a binary-weight committee machine compared to Monte Carlo simulations with $K = 3$ and 5. The annealed phase diagram is shown in figure 8.

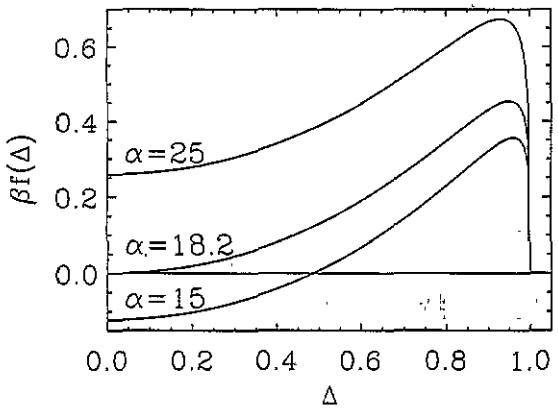

Figure 5. Free energy as function of Δ for binary weights and different values of α. The free energy has always two minima, one at $\Delta = 0$ and the other at $\Delta = 1$. At a critical value of α the relative heights of the two minima switches and a first-order transition occurs.

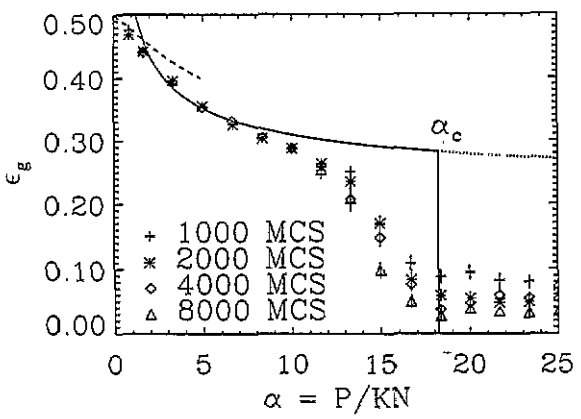

Figure 6. Generalization error for a binary-weight committee machine with $K = 3$ and at $T = 5$. The solid line shows the large-K prediction for the generalization error in the equilibrium state including $1/\sqrt{K}$ corrections. The generalization error of the metastable state is shown dotted. The dashed line corresponds to the solution valid for $\alpha \sim \mathcal{O}(1/K)$, where the solid line diverges. These results are compared to Monte Carlo simulations for $N = 45$ and $K = 3$, averaged over different runs and for different numbers of Monte Carlo steps (MCS).

Similar transitions due to a freezing of degrees of freedom have also been found in other binary models (e.g. [7, 15, 16, 18, 19]). In a committee machine with 3 hidden units learning a linearly separable task, this freezing was found to be asymmetric [13]: individual hidden units can align perfectly with the simple teacher perceptron, yielding a stepwise decrease in the generalization error. In order to investigate whether an asymmetric freezing also occurs in our model, we introduce a partially-symmetric ansatz for the order parameters. Following [13] we assume that $(1 - \eta)K, \eta \in [0, 1]$, students are perfectly aligned with one of the teachers respectively (each teacher corresponding to only one student). The remaining ηK students are assumed to have symmetric overlaps R with all the teachers and

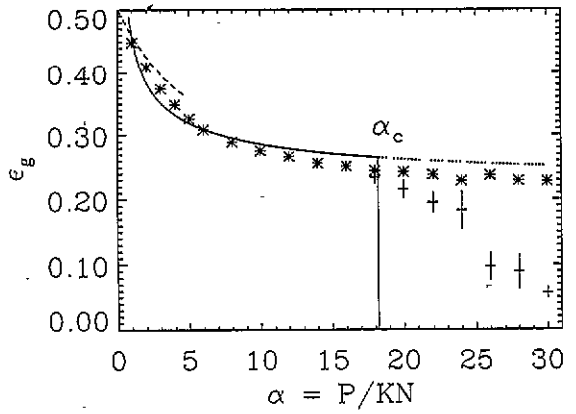

Figure 7. Generalization error for a binary-weight committee machine with $K = 5$ and at $T = 5$. The large-K predictions for the different regions are compared with Monte Carlo simulations ($N = 45$, $K = 5$, $T = 5$). The solid line corresponds to the equilibrium state, the dotted line to the metastable state and the dashed line to the small-α solution. The results of the simulations are averaged over all runs (+) and the runs in which no freezing occurred (*), respectively.

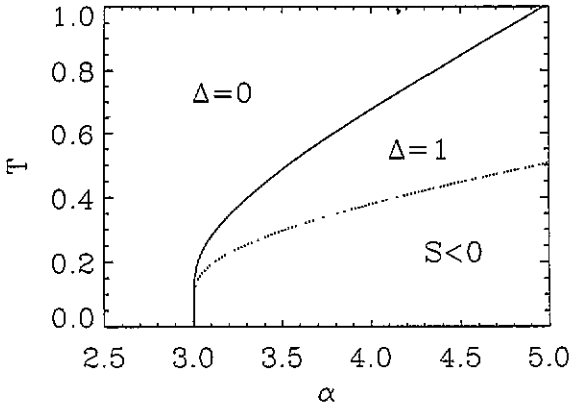

Figure 8. Annealed phase diagram for the large-K committee machine with binary weights. The solid line marks the location of the phase transition. In the region below the dotted line, the entropy of the metastable state is negative.

mutual overlaps C, yielding

$$R_{lk} = \begin{cases} R & \text{if } l \leq \eta K \\ \delta_{lk} & \text{otherwise} \end{cases} \qquad C_{lk} = \begin{cases} C + (1-C)\delta_{lk} & \text{if } l,k \leq \eta K \\ \delta_{lk} & \text{if } l,k > \eta K \\ R & \text{otherwise.} \end{cases} \tag{29}$$

Having made this ansatz we can proceed with the calculation of the entropy term G_0 and the generalization error ϵ as a function of the order parameters R and C. For the entropy term we get to order $1/K$

$$G_0(R,C) = \eta \ln 2 - \frac{\eta}{2}\frac{K-1}{K}C + \frac{1}{2K}\ln[(1-\eta C) - \eta K(KR^2 - C)]\,. \tag{30}$$

As for continuous weights, the logarithmic term imposes geometrical constraints on the parameters R and C

$$\begin{aligned} KR^2 - C &< \frac{1}{\eta K} \\ R &< \frac{1}{\sqrt{K}}\left[1 + \mathcal{O}\left(\frac{1}{\eta K}\right)\right]. \end{aligned} \tag{31}$$

For the generalization error we obtain similarly to the derivation of (15)

$$\epsilon(R,C)=2\sum_{\{\sigma_\lambda\}}\sum_{\{\tau_l\}}\Theta\left(-K^{-1/2}\sum_l\tau_l\right)\Theta\left[K^{-1/2}\left(\sum_\lambda\sigma_\lambda+\sum_\kappa\tau_\kappa\right)\right]$$

$$\times\int_{-\infty}^{+\infty}\mathrm{D}t\int_0^\infty\prod_l\mathrm{D}v_l\prod_\lambda H\left(\sigma_\lambda\frac{it\sqrt{KR^2-C}-R\sum_l\tau_l v_l}{\sqrt{1-C}}\right). \tag{32}$$

In the sums and products, l ranges over all hidden units, λ over the ηK unfrozen units and κ over the $(1-\eta)K$ frozen units. In order to obtain the dependence of ϵ on η for large K we make use of the constraints (32) and proceed similarly to the previous calculation. For large K we can expand the H-function and do the traces. The result is that we obtain to leading order in $1/K$ and for $\eta\gg K^{-1/4}$ simply $\epsilon(R,C)=\epsilon_0$, independent of η. Since the leading term of the entropy (30) is proportional to η, the free energy (10) increases if η deviates from 1. Therefore, any solution with a fraction $(1-\eta)$ of perfectly aligned hidden units, such that $K^{-1/4}\ll\eta<1$, has to be metastable, in contrast to the case of a finite-K committee machine trained on a linearly separable task [13].

The generalization error of the symmetric state approaches the residual value ϵ_0 as $\alpha\to\infty$. This approach can be examined more closely by calculating the generalization error for $\eta=1$. As a function of the rescaled parameters r and c as in (16), it is simply given by our previous expression (17) with $\Delta=0$

$$\epsilon(r,c)=\frac{1}{\pi}\cos^{-1}\left(\sqrt{\frac{2}{\pi}}\frac{r}{\sqrt{c+K^{-1/2}\pi/2}}\right). \tag{33}$$

The entropy (30) with $\eta=1$ as a function of r and c is

$$G_0(r,c)=\ln 2-\frac{1}{2}\frac{c}{\sqrt{K}}+\frac{1}{2K}\ln[1-\sqrt{K}(r^2-c)]+\mathcal{O}(K^{-3/2}). \tag{34}$$

We compare this expression to the corresponding entropy (18) in the continuous model with $\Delta=0$. To this order and up to a constant, they only differ in the dependence on c. However, the derivative of (35) with respect to c

$$\frac{\partial G_0}{\partial c}=-\frac{1}{2}\frac{1}{\sqrt{K}}+\frac{1}{2\sqrt{K}}\frac{1}{1-\sqrt{K}(r^2-c)} \tag{35}$$

only differs from the derivative of (18) to order $1/K$. This difference does not alter the solution (19) for the order parameters. In both cases the approach of the generalization error to the constant ϵ_0 is given by (20) as shown in figures 2, 6 and 7 for different values of K. In the large-K limit the discreteness of the weights only influences the behavior of the equilibrium state beyond the phase transition. While in the binary model a perfect alignment of the students with the teachers is possible, this cannot happen in the continuous-weight case.

We have performed Monte Carlo simulations to check our analytic results for binary weights using networks with $N=45$ input units and $K=3$ and 5 hidden units. Although we cannot expect good quantitative agreement with the large-K theory for these small committees, we found qualitative support for our findings. The simulations indicate a symmetric freezing of all hidden units at the critical value α_c. In some runs we found evidence for the occurrence of metastable states with only some students perfectly aligned with a teacher. However, an analytic description of these states would require a calculation

for finite K. Figure 6 shows results for $K = 3$ averaged over different simulations. The solid line corresponds to the predictions of (20) including $1/\sqrt{K}$ corrections. It shows the location of the phase transition and the divergence for small values of α. The location of the transition for $K = 3$ found in the simulation lies somewhat below the large-K prediction. A comparison of results for different numbers of Monte Carlo steps shows the influence of the metastable state even at high temperature ($T = 5$). Some simulations remain in the metastable state even for large α and long training times, yielding a rather smooth decay of the averaged generalization error to a small but non-vanishing value. This behavior is reflected in a double-peak structure of the distribution of the order parameters sampled over different simulations. The predictions for the metastable state can be checked by separately averaging over those simulations in which no freezing occurred as shown in figure 7 for $K = 5$.

The predictions of the annealed approximation can be checked by calculating the training error $\epsilon_t = \alpha^{-1}\partial(\beta f)/\partial\beta$ and the thermodynamic entropy $s = -\partial f/\partial T$. The annealed approximation predicts a simple relationship between the training error and the generalization error [7]

$$\epsilon_t = \frac{e^{-\beta}\,\epsilon_g}{1-(1-e^{-\beta})\,\epsilon_g}. \tag{36}$$

While we find a good agreement with our simulations at high temperatures, the annealed approximation fails to correctly predict the training error of the poorly-generalizing state at low temperatures. In particular, for zero temperature the annealed approximation predicts $\epsilon_t = 0$ for all α, clearly violating the correct limit $\epsilon_t, \epsilon_g \to \epsilon_0$ as $\alpha \to \infty$ [7]. Furthermore, the annealed approximation yields a region (below the dotted line in figure 8), where the metastable state has a negative entropy. This is unphysical in a binary model, and the annealed description of the metastable state cannot be adequate in this region. A full quenched theory with broken replica symmetry will be necessary to describe it. However, the annealed predictions for the generalization error of the poorly-generalizing state are in good agreement with our simulations even at low temperature and large values of α.

5. Summary

In summary, we have solved the generalization problem for a fully connected committee machine within the annealed approximation. The target rule was defined by another such network and thus not linearly separable. While the generalization properties of a committee machine with non-overlapping receptive fields [15, 16] are qualitatively similar to those of a simple perceptron, the fully connected architecture shows additional features.

The most important is the finding of a transition from an unspecialized or permutation-symmetric state, in which the overlaps between any one of its hidden unit weight vectors and those of all the hidden units of the teacher are equal, to a specialized one, in which each hidden unit weight vector becomes strongly correlated with that of a particular hidden unit of the teacher. This transition has no counterpart in single-layer machines, where such specialization has no meaning.

We have found that a committee machine trained on insufficient data ($\alpha < \alpha_c$) will always adopt the symmetric learning strategy with equal overlaps between each of its hidden unit weight vectors and those of all the teacher's hidden units. It is unable to learn the necessary specialization of its hidden units in which they come to match those of the teacher. Because the correct learning of the task requires this specialization, the symmetric solution always has a generalization error greater than $\epsilon_0 \approx 0.206$. Furthermore,

even when the training set is large enough to permit specialization ($\alpha > \alpha_c$), the system can be trapped in the metastable symmetric state. This metastability persists for arbitrarily large training sets.

The symmetry breaking occurs via a first-order transition for both continuous and binary weights. Thus the present problem differs from the corresponding one for single-layer machines, where a discontinuous transition occurs only in the binary case. However, even in the binary case the similarity is only superficial. In the single-layer machine the transition is simply the point where the free energies of the poorly-generalizing and correct solutions cross, while in the committee machine it involves breaking a permutation symmetry.

Although discontinuous transitions have previously been found in single-layer networks with continuous weights [6, 8], the reasons for them were rather different from those in the present model. Watkin and Rau [8] found a discontinuous transition in an unlearnable problem, and Sompolinsky *et al* [6] studied a quasi-discrete simple perceptron with a bimodal distribution of weights centered around ± 1. Their model interpolated between a spherical perceptron, in which no transition occurs, and a perceptron with binary weights. In contrast, the transition found here occurs in a learnable problem for spherical weights without further restrictions on the weight vectors and involves an intrinsically multilayer effect: the specialization of individual hidden units.

Recently Hansel *et al* [19] found a similar behavior in a parity machine with 2 hidden units and non-overlapping receptive fields. Due to the invariance of the network output under an inversion of the weight vector ($\boldsymbol{W} \to \boldsymbol{W}' = -\boldsymbol{W}$), their model always has a solution with vanishing student-teacher overlap. Therefore, the parity machine fails completely to generalize for small α. The output of the fully connected committee machine is invariant under a permutation of hidden units [12]. In contrast to the parity machine, this invariance allows for a translationally symmetric solution with small student-teacher overlaps and thus $\epsilon_g < 1/2$.

No state comparable to the symmetric, poorly-generalizing one was found in the non-overlapping architecture [15, 16], because there the specialization of the students was built into the model. In the fully connected committee a specialization of the individual students breaks the translational symmetry and only occurs if the training set is sufficiently large. Furthermore, the poorly-generalizing metastable state survives beyond the phase transition even for arbitrarily large α, another feature not found in the tree model.

The problem of learning a classification task which is not linearly separable is of practical importance, and two-layer networks are widely used in this context. In this paper we have studied the special case of a target rule whose structure is identical to the learning network, a situation which is the exception in practical applications. It would be desirable to extend this work to the learning of rules which do not match the structure of the learning network. This includes both unlearnable rules and functions which could be learned by smaller networks. Recently the learning of unlearnable rules by simple perceptrons has been studied in [7, 8]. An example of a small committee machine learning a function which could be realized by a simple perceptron has been studied in [13]. Furthermore, for a correct description of the metastable states at low temperatures the annealed approximation is not trustworthy and the quenched theory should be solved.

Acknowledgments

We thank S Solla for useful discussions. HS acknowledges support from the EC under the SCIENCE programme and from the Danish Natural Science Council and the Danish Technical Research Council through CONNECT.

Appendix A

In this appendix we give a more detailed description of the calculation of the entropy term (12) for continuous weights. Starting out from equation (9) we use integral representations of the δ-functions

$$\delta(x-a)=\int_{-i\infty}^{i\infty}\frac{d\hat{x}}{2\pi i}\,e^{\hat{x}(x-a)} \tag{A1}$$

introducing the parameters $\hat{R}_{lk}$, $\hat{C}_{lk}$ conjugate to the order parameters R_{lk}, C_{lk} and additional parameters E_l for the normalization condition $N=\boldsymbol{W}_l\cdot\boldsymbol{W}_l$. This leaves us with

$$\exp\{NKG_0[R_{lk},C_{lk}]\}=\int[d\hat{R}_{lk}\,d\hat{C}_{lk}\,dE_l]\,\exp\{NK\hat{G}_0[R_{lk},C_{lk},\hat{R}_{lk},\hat{C}_{lk},E_l]\}$$

where

$$\begin{aligned}\hat{G}_0&=\frac{1}{K}\sum_l E_l+\frac{1}{K}\sum_{l,k}R_{lk}\hat{R}_{lk}-\frac{1}{K}\sum_{(l,k)}C_{lk}\hat{C}_{lk}+\frac{1}{KN}\ln\mathcal{Z}\\ \mathcal{Z}&=\int\prod_l d\boldsymbol{W}_l\,\exp\Big\{-\sum_l E_l\,\boldsymbol{W}_l^2-\sum_{l,k}\hat{R}_{lk}\,\boldsymbol{W}_l\cdot\boldsymbol{V}_k+\sum_{(l,k)}\hat{C}_{lk}\,\boldsymbol{W}_l\cdot\boldsymbol{W}_k\Big\}.\end{aligned} \tag{A2}$$

In the thermodynamic limit ($N\to\infty$), the integrals over $\hat{R}_{lk}$, $\hat{C}_{lK}$ and E_l can be done using the saddle point methods. With the symmetry ansatz (11) for the order parameters and correspondingly for their conjugate parameters and E_l at the saddle point, (A2) reads

$$\begin{aligned}\hat{G}_0&=E+(\Delta+KR)\hat{R}+(\Delta+R)\hat{\Delta}-\frac{1}{2}(K-1)C\hat{C}+\frac{1}{KN}\ln\mathcal{Z}\\ \mathcal{Z}&=\int\prod_l d\boldsymbol{W}_l\,\exp\Big\{-\sum_l[E\,\boldsymbol{W}_l^2+(\sqrt{K}\hat{R}\,\nu+\hat{\Delta}\,\boldsymbol{V}_l)\cdot\boldsymbol{W}_l]\\ &\qquad+\hat{C}\Big(\sum_l\boldsymbol{W}_l\Big)^2-\frac{KN}{2}\hat{C}\Big\}\end{aligned} \tag{A3}$$

with the abbreviation $\nu=1/\sqrt{K}\,\sum_k\boldsymbol{V}_k$ (note that $\nu^2=1/K\,\sum_{l,k}\boldsymbol{V}_l\cdot\boldsymbol{V}_k=N$). After linearizing the square of the sum of weight vectors, using

$$\exp\Big\{\frac{1}{2}\hat{C}\Big(\sum_l\boldsymbol{W}_l\Big)^2\Big\}=\int\frac{dz}{(2\pi)^{N/2}}\,\exp\Big\{-\frac{1}{2}z^2+\sqrt{\hat{C}}\sum_l\boldsymbol{W}_l\cdot z\Big\} \tag{A4}$$

the integrals over weights and subsequently the integral over z can easily be done. We obtain at the saddle point

$$\begin{aligned}G_0(\Delta,R,C)=\operatorname*{extr}_{\hat{\Delta},\hat{R},\hat{C},E}\Big\{&E+(\Delta+KR)\hat{R}+(\Delta+R)\hat{\Delta}-\frac{1}{2}\hat{C}[1+(K-1)C]\\ &+\frac{1}{2}\,\frac{\hat{\Delta}^2}{2E}\,\frac{2E-(K-1)\hat{C}}{2E-K\hat{C}}+\frac{1}{2}\,\hat{R}\,\frac{K\hat{R}+2\hat{\Delta}}{2E-K\hat{C}}\\ &-\frac{1}{2}\,\frac{K-1}{K}\,\ln 2E-\frac{1}{2K}\ln(2E-K\hat{C})+\frac{1}{2}\ln 2\pi\Big\}\end{aligned} \tag{A5}$$

and the conditions

$$\hat{\Delta} = -\frac{\Delta}{(1-C)-\Delta^2} \qquad K\hat{R} = -\hat{\Delta} - \frac{\Delta + KR}{[1+(K-1)C]-[\Delta+KR]^2}$$
$$2E = \frac{1}{(1-C)-\Delta^2} \qquad K\hat{C} = 2E - \frac{1}{[1+(K-1)C]-[\Delta+KR]^2} \tag{A6}$$

leading to G_0 in the form (12).

Appendix B

In this appendix we present the derivation of the generalization error (17) starting from equation (14). The average over inputs in (14) yields, assuming the same Gaussian distribution of inputs as for the training examples

$$\left\langle \exp\left\{ \frac{\mathrm{i}}{\sqrt{N}} \sum_l (\hat{u}_l \boldsymbol{W}_l + \hat{v}_l \boldsymbol{V}_l) \cdot \boldsymbol{S} \right\} \right\rangle_S$$
$$= \int \frac{\mathrm{d}\boldsymbol{S}}{(2\pi)^{N/2}} \exp\left\{ -\frac{1}{2}\boldsymbol{S}^2 + \frac{\mathrm{i}}{\sqrt{N}} \sum_l (\hat{u}_l \boldsymbol{W}_l + \hat{v}_l \boldsymbol{V}_l) \cdot \boldsymbol{S} \right\}$$
$$= \exp\left\{ -\frac{1}{2N} \sum_{l,k} [\hat{u}_l \hat{u}_k \, \boldsymbol{W}_l \cdot \boldsymbol{W}_k + 2\hat{u}_l \hat{v}_k \, \boldsymbol{W}_l \cdot \boldsymbol{V}_k + \hat{v}_l \hat{v}_k \, \boldsymbol{V}_l \cdot \boldsymbol{V}_k] \right\}. \tag{B1}$$

Introducing the order parameters R_{lk} and C_{lk} with the symmetry ansatz (11) yields

$$\langle \cdots \rangle_S = \exp\left[-\frac{1}{2} C \left(\sum_l \hat{u}_l \right)^2 - \frac{1}{2}(1-C) \sum_l \hat{u}_l^2 \right.$$
$$\left. - R \left(\sum_l \hat{u}_l \right) \left(\sum_k \hat{v}_k \right) - \Delta \sum_l \hat{u}_l \hat{v}_l - \frac{1}{2} \sum_k \hat{v}_k^2 \right]. \tag{B2}$$

Now the integrals over the $\hat{u}_l$'s and $\hat{v}_l$'s can be done, if we use a relation analogous to (A4) to linearize squares of sums of $\hat{u}_l$'s in the exponent, yielding an additional Gaussian integral over t. So far we have

$$\epsilon(\Delta, R, C) = \int \prod_l \frac{\mathrm{d}u_l}{\sqrt{2\pi}} \int \prod_l \mathrm{D}v_l \int \mathrm{D}t \; \Theta\left[-\frac{1}{\sqrt{K}} \sum_l \mathrm{sign}(u_l) \frac{1}{\sqrt{K}} \sum_l \mathrm{sign}(v_l) \right]$$
$$\times \prod_l \left[[(1-C)-\Delta^2]^{-1/2} \exp\left\{ -\frac{1}{2} \frac{(u_l - \Delta v_l - R\sum_k v_k - \mathrm{i}t\sqrt{D/K})^2}{(1-C)-\Delta^2} \right\} \right]. \tag{B3}$$

The Θ-function can be factorized by introducing internal representations in the teacher $\{\tau_l = \pm 1\}$ and the student network $\{\sigma_l = \pm 1\}$, yielding

$$\Theta\left[-\frac{1}{\sqrt{K}} \sum_l \mathrm{sign}(u_l) \frac{1}{\sqrt{K}} \sum_l \mathrm{sign}(v_l) \right]$$
$$= \sum_{\{\tau_l\}} \sum_{\{\sigma_l\}} \Theta\left(-\frac{1}{\sqrt{K}} \sum_l \sigma_l \frac{1}{\sqrt{K}} \sum_l \tau_l \right) \prod_l \Theta(\sigma_l u_l) \Theta(\tau_l v_l). \tag{B4}$$

Now the u_l-integrals factorize and can be done. Making use of the identity $\Theta(ab) = \Theta(a)\Theta(b) + \Theta(-a)\Theta(-b)$ we obtain equation (15) for the generalization error of a committee machine with K hidden units.

We can factorize this expression in l if we control the sum $K^{-1/2}\sum_k v_k$ by a new variable s. Inserting the identity

$$1 = \int_{-\infty}^{+\infty} \mathrm{d}s\ \delta\left(s - K^{-1/2}\sum_k v_k\right) = \int_{-\infty}^{+\infty} \frac{\mathrm{d}s\ \mathrm{d}v}{2\pi}\ \exp\left\{-\mathrm{i}v\left(s - K^{-1/2}\sum_k v_k\right)\right\} \tag{B5}$$

we obtain

$$\epsilon(\Delta, R, C) = 2\sum_{\{\sigma_l\}}\sum_{\{\tau_l\}} \Theta\left(-K^{-1/2}\sum_l \tau_l\right)\Theta\left(K^{-1/2}\sum_l \sigma_l\right)$$
$$\times \int_{-\infty}^{+\infty} \mathrm{D}t \int_{-\infty}^{+\infty} \frac{\mathrm{d}s\ \mathrm{d}v}{2\pi}\ \mathrm{e}^{-\mathrm{i}vs} \prod_l \left[\int_{-\infty}^{+\infty} \mathrm{D}v_l\ \mathrm{e}^{\mathrm{i}vv_l/\sqrt{K}}\ \Theta(\tau_l v_l)\ H(-\sigma_l z_l)\right] \tag{B6}$$

where

$$z_l = \frac{\sqrt{K}Rs + \Delta v_l - \mathrm{i}t\sqrt{D/K}}{\sqrt{1 - C - \Delta^2}}. \tag{B7}$$

To do the traces, we introduce integral representations of the Θ-functions

$$\Theta(K^{-1/2}\sum_l \sigma_l) = \int_0^{\infty} \mathrm{d}\lambda \int_{-\infty}^{+\infty} \frac{\mathrm{d}x}{2\pi}\ \exp\left\{-\mathrm{i}x\left(\lambda - K^{-1/2}\sum_l \sigma_l\right)\right\} \tag{B8}$$

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

$$\begin{aligned}
\sum_{\{\sigma_l\}} \exp\left(\frac{\mathrm{i}x}{\sqrt{K}}\sum_l \sigma_l\right) \prod_l H(-\sigma_l z_l) &= \prod_l \left[\cos\frac{x}{\sqrt{K}} + \mathrm{i}\hat{H}(z_l)\ \sin\frac{x}{\sqrt{K}}\right] \\
\sum_{\{\tau_l\}} \exp\left(\frac{-\mathrm{i}y}{\sqrt{K}}\sum_l \tau_l\right) \prod_l \Theta(\tau_l v_l) &= \prod_l \left[\cos\frac{y}{\sqrt{K}} - \mathrm{i}\,\mathrm{sign}(v_l)\ \sin\frac{y}{\sqrt{K}}\right]
\end{aligned} \tag{B9}$$

with the notation $\hat{H}(z_l) = 1 - 2H(z_l)$. Now we are left with the integrals

$$I_l = \int_{-\infty}^{+\infty} \mathrm{D}v_l\ \mathrm{e}^{\mathrm{i}vv_l/\sqrt{K}} \left[\cos\frac{x}{\sqrt{K}} + \mathrm{i}\hat{H}(z_l)\ \sin\frac{x}{\sqrt{K}}\right]\left[\cos\frac{y}{\sqrt{K}} - \mathrm{i}\,\mathrm{sign}(v_l)\ \sin\frac{y}{\sqrt{K}}\right]. \tag{B10}$$

For large K and with the scaling (16) for the order parameters we can expand the integrand to order $1/K$, using the identity $\hat{H}(a+\varepsilon) = \hat{H}(a) + \sqrt{2/\pi}\,\varepsilon\,\mathrm{e}^{-a^2/2} - a\,\varepsilon^2\,\mathrm{e}^{-a^2/2}/\sqrt{2\pi} + \mathcal{O}(\varepsilon^3)$. To order $1/K$, we get for the integral

$$I_l = 1 - \frac{1}{K}\left[\frac{x^2}{2} + \frac{y^2}{2} + \frac{v^2}{2} - \frac{2}{\pi}\,xy\ \sin^{-1}\Delta\right.$$
$$\left. - \sqrt{\frac{2}{\pi}}\,(\mathrm{i}K^{1/4}rs + t\sqrt{D})\,x - \sqrt{\frac{2}{\pi}}\,(y - \Delta x)\,v\right] \tag{B11}$$

leading to

$$\epsilon(\Delta, r, c) = \int_{-\infty}^{+\infty} Dt\, Dx\, Dy\, Dv \int_{-\infty}^{+\infty} \frac{ds}{\sqrt{2\pi}} \exp\left\{-is\left(v - \sqrt{\frac{2}{\pi}} K^{1/4} r x\right)\right\} \int_0^{\infty} \frac{d\lambda\, d\mu}{2\pi} e^{-i\lambda y - i\mu x}$$

$$\times \exp\left[\sqrt{\frac{2}{\pi}}(y - \Delta x)\, v + \sqrt{\frac{2}{\pi} D}\, xt + \frac{2}{\pi} xy \sin^{-1}\Delta\right]. \tag{B12}$$

The s-integral just yields a δ-function, and all the remaining integrals can be done. We finally obtain equation (17) for the generalization error.

References

[1] Hertz J A, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Reading, MA: Addison-Wesley)
[2] Watkin T, Rau A and Biehl M 1993 *Rev. Mod. Phys.* **65** 499
[3] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
[4] Rosenblatt F 1962 *Principles of Neurodynamics* (New York: Spartan)
[5] Györgyi G and Tishby N 1990 *Neural Networks and Spin Glasses* ed K Thuemann and R Köberle (Singapore: World Scientific)
[6] Sompolinsky H, Tishby N and Seung H S 1990 *Phys. Rev. Lett.* **65** 1683
[7] Seung H S, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 6056
[8] Watkin T, Rau A 1992 *Phys. Rev.* A **45** 4102
[9] Denker J S *et al* 1987 *Complex Syst.* **1** 877
[10] Hecht-Nielsen R 1987 *IEEE First International Conference 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] Barkai E, Hansel D and Sompolinsky H 1992 *Phys. Rev.* A **45** 4146
[13] Schwarze H, Opper M and Kinzel W 1992 *Phys. Rev.* A **45** R6185
[14] Engel A, Köhler H M, Tschepke F, Vollmayr H, and Zippelius A 1992 *Phys. Rev.* A **45** 7590
[15] Schwarze H and Hertz J 1992 *Europhys. Lett.* **20** 375
[16] Mato G and Parga N 1992 *J. Phys. A: Math. Gen.* **25** 5047
[17] Schwarze H and Hertz J 1993 *Europhys. Lett.* **21** 785
[18] Sompolinsky H and Tishby N 1990 *Europhys. Lett.* **13** 567
[19] Hansel D, Mato G and Meunier C 1992 *Europhys. Lett.* **20** 471