Learning to classify in large committee machines

Dominic O'Kane[1,*] and Ole Winther[2,†]
[1]Theoretical Physics, 1 Keble Road, Oxford OX1 3NP, England
[2]Computational Neural Network Center, The Niels Bohr Institute, 2100 Copenhagen Ø, Denmark
(Received 28 March 1994)

The ability of a two-layer neural network to learn a specific non-linearly-separable classification task, the proximity problem, is investigated using a statistical mechanics approach. Both the tree and fully connected architectures are investigated in the limit where the number K of hidden units is large, but still much smaller than the number N of inputs. Both have continuous weights. Within the replica symmetric ansatz, we find that for zero temperature training, the tree architecture exhibits a strong overtraining effect. For nonzero temperature the asymptotic error is lowered, but it is still higher than the corresponding value for the simple perceptron. The fully connected architecture is considered for two regimes. First, for a finite number of examples we find a symmetry among the hidden units as each performs equally well. The asymptotic generalization error is finite, and minimal for $T \to \infty$ where it goes to the same value as for the simple perceptron. For a large number of examples we find a continuous transition to a phase with broken hidden-unit symmetry, which has an asymptotic generalization error equal to zero.

PACS number(s): 87.10.+e, 02.50.−r, 64.60.Cn

I. INTRODUCTION

Perhaps the most important property of feedforward neural networks [1] is their ability to learn a rule from examples. By this we mean that, given a set of input-output example pairs of data assumed to have been produced by some unknown rule, an appropriate choice of network architecture and algorithm will then allow us to train the network not only to be able to reproduce the training examples, but also to generalize, i.e., to correctly predict the output to new inputs on which it has not been trained. The statistical mechanical tools developed by Gardner and Derrida [2] can be used to determine the generalization ability of neural networks. This approach is powerful since, in many cases, it gives exact results for the typical generalization properties of specific network architectures learning specific rules [3]. Furthermore, it can help us to develop an insight into how a neural network actually generalizes.

To date, most work has focused on determining the generalization ability of the simple perceptron attempting to learn various rules. Unfortunately, the simple perceptron can only implement rules which are linearly separable; other rules are nonrealizable. However, it has been shown [4] that for any Boolean rule this severe restriction can be overcome by introducing a layer of "hidden" units, each connected to every input unit, to produce a two-layer neural network.

In this article, we study the ability of a specific two-layer neural network, the *committee machine,* to learn a rule which is not linearly separable. The rule takes the form of a classification task in which the network must learn to correctly associate a random input with one of M specified classes (prototypes). Previous study of this problem has been restricted to the single-layer perceptron [5,6], where the task is unlearnable when M scales with the number of connections. Note that this rule differs in nature from those previously studied for committee machines since it cannot be expressed as one committee machine being taught to reproduce the architecture of another [7]. Our main aims are twofold. First we wish to compare the performance of the "tree" version of the committee machine, for which, due to the restricted connectivity between the input and hidden layer, the result of [4] mentioned above does not hold, to that of the simple perceptron. This is of interest since it is not clear *a priori* whether the greater complexity of the tree architecture helps or hinders its generalization ability. Secondly, and more importantly, we wish to investigate the mechanisms which might allow the "fully connected" committee machine to learn this classification task perfectly. In both, we shall employ Gibbs learning such that we may consider the role of a stochastic noise in the training dynamics.

In what follows, we introduce the architecture of both the tree and fully connected versions of the committee machine. We then define the exact nature of the rule, and introduce the statistical mechanical formalism which will be used. In Sec. III we describe the solution for the tree committee machine, and in Sec. IV for the fully connected committee machine. We conclude with a discussion of our results.

II. THE MODEL

The committee machine is a simplified version of a general two-layer network. Two variations of its architec-

*Present address: Mathematics Department, Imperial College, 180 Queen's Gate, London, England.
†Author to whom all correspondence should be addressed.

ture, the tree and the fully connected, are shown in Fig. 1. In both, each hidden unit can output one of two values $\sigma_k = \pm 1$ $(k = 1, \ldots, K)$. A committee machine derives its name from the fact that the final output σ corresponds to the state of the majority of the K hidden units. We can therefore write

$$\sigma = \mathrm{sgn}\left[\frac{1}{\sqrt{K}} \sum_{k=1}^{K} \sigma_k\right], \quad \sigma_k = \mathrm{sgn}(h_k) , \tag{1}$$

where h_k is the local field arriving at hidden unit k and the prefactor of $1/\sqrt{K}$ ensures that the argument of the sgn function is $O(1)$. In the tree, each hidden unit receives inputs from only N/K of the input units, none of which are common to any other hidden unit, and so we have

$$h_k = \left[\frac{K}{N}\right]^{1/2} \sum_{i=N(k-1)/K+1}^{Nk/K} J_{ki} S_i , \tag{2}$$

where J_{ki} is the strength of the connection weight from input unit i to hidden unit k. However, in the fully connected version, each hidden unit is connected to all of the N inputs so that

$$h_k = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_{ki} S_i . \tag{3}$$

The physical significance of this difference is that in the tree committee machine no two hidden units are connected to the same part of the input space $\mathbf{S}$. As a consequence, the $\mathbf{J}$ inputting to each hidden unit are uncorrelated. Contrast this with the fully connected architecture in which all hidden units are connected to the same input space. Correlations will now arise between the $\mathbf{J}$ inputting to different hidden units which will cause the emergence of additional order parameters.

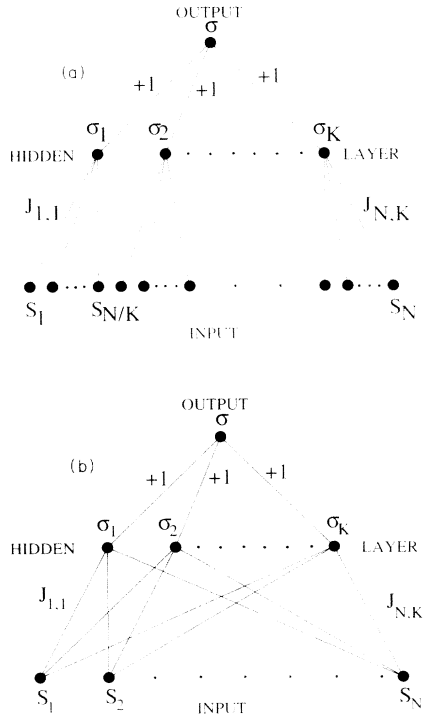

FIG. 1. Architecture of a committee machine. (a) The tree committee machine; each of the K hidden units experiences only a subset N/K of the inputs; (b) the fully connected committee machine in which each of the K hidden units experiences all N input units.

To define the classification task, we specify M categories or *prototypes* $\mathbf{S}^\mu = \{S_i^\mu\}$ $(\mu = 1, \ldots, M, i = 1, \ldots, N)$, with each S_i^μ chosen independently and randomly to be ± 1 with equal probability. To each we associate a particular output $\tau^\mu = \pm 1$. When a randomly chosen configuration is input, our aim is to make the network output the τ^μ corresponding to the category to which this input is closest in Hamming distance. We teach the system to do this by training it with P examples $\mathbf{S}^{\mu l}$ $(l = 1, \ldots, P)$ for each prototype with each example constrained to have a finite overlap m with its respective prototype pattern, i.e.,

$$\sum_{i=1}^{N} S_i^{\mu l} S_i^\mu = \mathbf{S}^{\mu l} \cdot \mathbf{S}^\mu = Nm . \tag{4}$$

Geometrically, we may see each prototype as a vector in N-dimensional space at the center of a cone with internal angle $\arccos m$. Its P example vectors lie on the surface of this cone.

The problem of learning is cast in the language of statistical physics through the introduction of a cost function which is minimized by the network $\mathcal{J}$ which best reproduces a set of training examples $\{\mathbf{S}^{\mu l}, \tau^\mu\}$ [8]. We define this "training energy" as

$$E_t(\mathcal{J}, \{\mathbf{S}^{\mu l}, \tau^\mu\}) = \sum_{\mu=1}^{M} \sum_{l=1}^{P} \Theta(-\tau^\mu \sigma(\mathcal{J}, \mathbf{S}^{\mu l})) , \tag{5}$$

where $\Theta(x)$ is the Heaviside function

$$\Theta(x) \equiv \begin{cases} 1 & \text{if } x > 0 , \\ 0 & \text{otherwise} . \end{cases} \tag{6}$$

We can associate this energy with a Monte Carlo dynamical process operating at finite temperature $T = \beta^{-1}$. T is therefore a measure of the amount of "noise" during training. As the final equilibrium states will then obey a Gibb's distribution, each network configuration $\mathcal{J}$ will, for a specific set of training examples, occur with a probability

$$\mathcal{P}(\mathcal{J}) = \frac{1}{Z} \exp[-\beta E_t(\mathcal{J}, \{\mathbf{S}^{\mu l}, \tau^\mu\})] , \tag{7}$$

where

$$Z = \int d\rho(\mathcal{J}) \exp[-\beta E_t(\mathcal{J}, \{\mathbf{S}^{\mu l}, \tau^\mu\})] \tag{8}$$

is the partition function and $\rho(\mathcal{J})$ contains all *a priori* constraints on possible connection-weight configurations $\mathcal{J}$. We denote the thermal average over $\mathcal{J}$ space performed with Eq. (7) by $\langle \cdots \rangle_\beta$. While the training energy tells us how well the network performs on the set of training examples, it is the ability of the network to successfully classify new inputs which is of interest to us.

We therefore define an error function which measures the generalization ability of the network for a specific set of M prototypes averaged over all possible inputs $\mathbf{S}$ constrained to obey Eq. (4):

$$\epsilon(\mathcal{J},\{\mathbf{S}^\mu,\tau^\mu\})=\frac{1}{M}\sum_{\mu=1}^{M}\langle\Theta(-\tau^\mu\sigma(\mathcal{J},\mathbf{S}))\rangle_{\mathbf{S},\mathbf{S}\cdot\mathbf{S}^\mu=Nm}\ . \tag{9}$$

As it is the typical properties of the system which are of interest, we must average over the quenched randomness in the choice of both prototypes and examples. The typical generalization error is then given by

$$\epsilon_g=\langle\langle\epsilon(\mathcal{J},\{\mathbf{S}^\mu,\tau^\mu\})\rangle_\beta\rangle_{\mathbf{S}^{\mu l},\mathbf{S}^\mu} \tag{10}$$

and the average training error

$$\epsilon_t=\frac{1}{PM}\langle\langle E_t(\mathcal{J},\{\mathbf{S}^{\mu l},\tau^\mu\})\rangle_\beta\rangle_{\mathbf{S}^{\mu l},\mathbf{S}^\mu}\ . \tag{11}$$

These quantities can be obtained from the disorder-averaged free energy per connection

$$-\beta f=\frac{1}{N}\langle\ln Z\rangle_{\mathbf{S}^{\mu l}\mathbf{S}^\mu}\ , \tag{12}$$

where Z has been defined in Eq. (8). For example,

$$\epsilon_t=\frac{1}{PM}\frac{\partial(\beta f)}{\partial\beta}\ . \tag{13}$$

Though it is not possible to perform this average directly, one can circumvent this problem by making use of the identity

$$\langle\ln Z\rangle=\lim_{n\to 0}\frac{\ln\langle Z^n\rangle}{n} \tag{14}$$

and the associated replica method [9]. In the work that follows we will be considering the scaling regime in which $N\to\infty$ with M/N finite. We will also be considering $P,K\gg 1$, but with both P/N and K/N going to zero.

III. THE TREE COMMITTEE MACHINE

We begin by considering the tree committee machine with N weights subject to a spherical constraint on the weights to each hidden unit

$$\rho(\mathcal{J})=\prod_{k=1}^{K}\delta\left[\sum_{i=N(k-1)/K+1}^{Nk/K}J_i^2-\frac{N}{K}\right]\ . \tag{15}$$

The free energy of the large tree committee machine is obtained following a procedure similar to that described in [7]. First we must use the integral representation of the δ function to introduce parameters which extract the quantities to be averaged from inside the Θ function in the training energy. One then performs the averages over the examples $\mathbf{S}^{\mu l}$ and prototypes $\mathbf{S}^\mu$ and introduces the appropriate order parameters. There are two sets of these. Dropping the replica index, we have first

$$\lambda_k^\mu=\gamma\tau^\mu\left[\frac{K}{N}\right]^{1/2}\sum_{i=N(k-1)/K+1}^{Nk/K}J_iS_i^\mu\ , \tag{16}$$

which represents the local stability field at hidden unit k produced by all P examples for prototype μ. It becomes a self-averaging order parameter in the limit $P\gg 1$ (we have defined $\gamma=m/\sqrt{1-m^2}$) and is directly related to the probability that a hidden unit give the correct output [see Eq. (21)]. The second set of order parameters

$$q_k^{ab}=\frac{K}{N}\sum_{i=N(k-1)/K+1}^{Nk/K}J_i^aJ_i^b \tag{17}$$

defines the overlap between two sets of weights inputting to hidden unit k in different replicas a and b. It is a measure of the number of different configurations of $\mathbf{J}_k$ which produce the correct output. Invoking replica symmetry

$$q_k^{ab}=q_k+(1-q_k)\delta_{ab} \tag{18}$$

we can then integrate over $\mathcal{J}$ and eliminate conjugate parameters. As the free energy factorizes in prototypes, we can drop the μ index. We assume "translational invariance" [10], i.e., that $q_k=q$ and $\lambda_k=\lambda$ for all $k=1,\ldots,K$. In the limit $N\to\infty$, the free energy is then given by

$$-\beta f=\operatorname*{extr}_{q,\lambda}\tfrac{1}{2}\ln(1-q)+\frac{q}{2(1-q)}-\frac{\alpha_0K\lambda^2}{2(1-q)}+\alpha_0\tilde{P}\int\prod_{k=1}^{K}Dt_k\ln\left\{\operatorname*{Tr}_{\{\tau_k=\pm1\}}\exp\left[-\beta\Theta\left[\frac{-1}{\sqrt{K}}\sum_{k=1}^{K}\tau_k\right]\right]\prod_{k=1}^{K}H\left[\tau_k\frac{t_k\sqrt{q}-\lambda}{\sqrt{1-q}}\right]\right\}, \tag{19}$$

where

$$Dt=\frac{dt}{\sqrt{2\pi}}e^{-t^2/2}\ \text{ and }\ H(x)=\int_x^\infty Dt \tag{20}$$

and we have introduced $\tilde{P}=P/\gamma^2$ and $\alpha_0=\gamma^2M/N$. We can write the generalization function defined in Eq. (9) in terms of the order parameter λ:

$$\epsilon_g(\lambda_k)=\operatorname*{Tr}_{\{\tau_k=\pm1\}}\Theta\left[-\frac{1}{\sqrt{K}}\sum_{k}^{K}\tau_k\right]\prod_{k=1}^{K}H(-\tau_k\lambda)\ . \tag{21}$$

This expression is quite straightforward. Since $H(\lambda)$ is the probability that a single hidden unit classifies wrongly, we see that we are summing over all possible internal representations which produce the wrong final output,

each weighted by the corresponding probability.

Both Eq. (19) and Eq. (21) are valid for any integer value of K. However, we wish to consider the limit $K \gg 1$. In this regime [or more precisely the regime where $K(1-q) \gg 1$], a self-consistent scaling requires that $\hat{\lambda} = \lambda \sqrt{K} \sim O(1)$. The last term in Eq. (19) then simplifies to become

$$\alpha_0 \tilde{P} \int_{-\infty}^{\infty} Dt \ln \left| e^{-\beta} + (1-e^{-\beta}) \times H\left[\frac{t\sqrt{q_{\text{eff}}} - \hat{\lambda}\sqrt{2/\pi}}{\sqrt{1-q_{\text{eff}}}} \right] \right| , \tag{22}$$

where

$$q_{\text{eff}} = \frac{2}{\pi} \arcsin q \ . \tag{23}$$

In this regime Eq. (21) also simplifies to become

$$\epsilon_g(\hat{\lambda}) = H\left[\hat{\lambda} \left[\frac{2}{\pi} \right]^{1/2} \right] \tag{24}$$

where $\hat{\lambda}$ corresponds to the value at the saddle point of the free energy for a given value of α_0, $\tilde{P}$, and β. We can interpret the $1/\sqrt{K}$ scaling of λ_k as a committee effect: the more committee members there are, the less well does each have to classify in order to reproduce all the training examples correctly.

We have investigated two temperature regimes. At $T=0$ we investigate the ability of the network to generalize when required to match the training examples exactly. Figure 2 shows the zero-temperature generalization error $\epsilon_g(T=0)$ as a function of the number of examples per prototype when $\alpha_0 = 1.6$. An evaluation of the saddle point equation in the limit $q \to 1$ [with $\tilde{P}(1-q)$ finite] gives the asymptotic value of the generalization error also shown on this figure. We observe that the capacity of the network is never reached in the studied regime: α_0 and $\tilde{P}$ are both $O(1)$, i.e., we can always reproduce the training examples without error. This is very different from the simple perceptron which has a finite capacity, and is due to the fact that the tree can perform many more dichotomies of the input space than can the simple perceptron. This has already been seen for the storage capacity [11]. We see that the generalization error initially falls as the number of examples increases. However, it is somewhat surprising to see that, beyond a certain value of $\tilde{P}$, it begins to rise towards an asymptotic value. We understand this as follows. As the problem is unlearnable for the tree committee machine, the $\mathcal{J}$ which generalizes the best has a finite training error. This means that the $\mathcal{J}$ which best reproduces the training examples has a certain limit to its ability, beyond which presenting new training examples no longer has the beneficial effect of providing information, but instead has the negative effect of imposing constraints which make its overall performance worse. This effect, called overtraining, was already seen in the case of the simple perceptron [6].

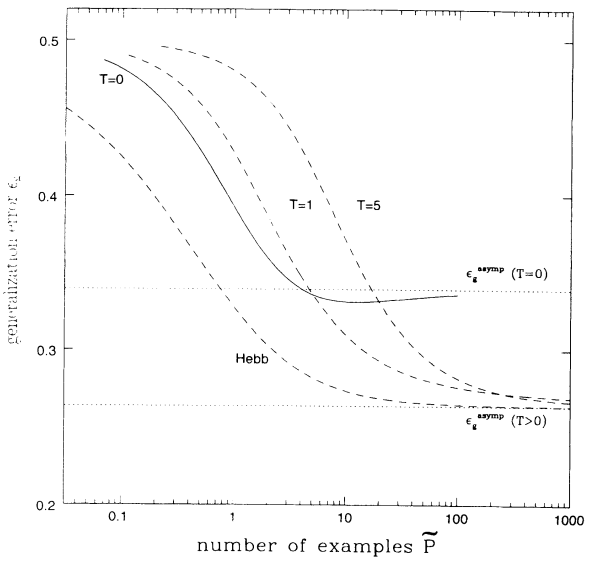

FIG. 2. Generalization error ϵ_g of the tree committee machine (CM) as a function of $\tilde{P}$ for $\alpha_0 = 1.6$. The $T=0$ curve reaches an asymptotic error of 0.34. As soon as $T>0$ the generalization curves tend to a common asymptotic generalization error of 0.26 (different from the $T=0$ value). We also show the generalization error curve for Hebbian weights.

The regime $T>0$ corresponds to training the network without requiring that the examples be learnt perfectly. This has the advantage of preventing overtraining by allowing the learning dynamics to find networks $\mathcal{J}$ which may not reproduce the example set perfectly, but which generalize better. This can be clearly seen from the curves in Fig. 2 which show the effect of training at two different temperatures $T=1$ and $T=5$. In the limit of large $\tilde{P}$, we can find the asymptotic generalization error to be $H(\sqrt{2/\pi\alpha_0})$. This limit is identical to that found for Hebbian weights, which is the optimal learning rule for a single-layer perceptron learning the proximity problem [12], and can be shown using similar arguments to be the optimal learning rule for the tree committee machine. From this it also follows that the optimal solution is translationally invariant.

Despite the fact that the tree committee machine is more "powerful" than the simple perceptron in terms of the number of random input-output mappings it can perform, its generalization ability is actually worse for all values of $\tilde{P}$. We therefore conclude that the tree architecture is a worse choice for learning this task and speculate that this is because the output is the result of K hidden units, each learning the task independently, i.e., we have K independent sets of N/K connections compared to one set of N connections in the simple perceptron. It should be pointed out that it is not possible to find a $\mathcal{J}$ which would reduce a tree committee machine to a perceptron for an arbitrary input. There is therefore no *a priori* reason to believe that the tree committee machine should perform at least as well as the perceptron.

IV. THE FULLY CONNECTED COMMITTEE MACHINE

In the fully connected network, each hidden unit receives inputs from all N units in the input layer. The

spherical normalization on the coupling strengths therefore takes the form

$$\rho(\mathcal{J})=\prod_{k=1}^{K}\delta\left[\sum_{i=1}^{N}J_{ki}^{2}-N\right]. \tag{25}$$

The derivation of the free energy follows closely that already described for the tree committee machine. A feature specific to the fully connected architecture is a coupling between hidden units due to the fact that they both receive inputs from all of the input space. Hence, in addition to the λ and q introduced for the tree [Eqs. (16) and (17) where the sum is now over all N input units], we require two new order parameters:

$$C_{jk}^{a}=\frac{1}{N}\sum_{i=1}^{N}J_{ji}^{a}J_{ki}^{a}\quad(j<k)\,, \tag{26}$$

$$D_{jk}^{ab}=\frac{1}{N}\sum_{i=1}^{N}J_{ji}^{a}J_{ki}^{b}\quad(j<k,a\neq b)\,. \tag{27}$$

As before, we impose replica symmetry,

$$D_{jk}^{ab}=D_{jk}\,. \tag{28}$$

Order parameter C_{jk} is interpreted as the typical overlap between weights coming from the same input unit, connected to different hidden units j and k in the same network $\mathcal{J}$ which minimizes the training energy. D_{jk} is the typical overlap between weights coming from the same input unit, connected to different units j and k in different network configurations which both minimize the training energy.

As before, the free energy factorizes in the prototypes. However, as different $\mathbf{J}_k$ are no longer independent, we do not immediately impose translational invariance for λ_k. In doing this, we are allowing for the possibility that different hidden units perform differently. However, we do impose translational invariance upon the remaining order parameters, i.e., $D_{jk}=D$ and $C_{jk}=C$. One can demonstrate that this simplification is not inconsistent with breaking the committee symmetry for λ. Its disadvantage is that it constrains us to search in only a subset of the space of solution networks, i.e., those networks which minimize the training energy and have different λ for different hidden units, but which must also have translational invariance in C and D. In the $N\to\infty$ limit, the free energy per connection $f=F/NK$ is written as the sum of an entropy term G_0 which measures the volume of $\mathcal{J}$ space minimizing the training energy, and an energy term G_1

$$-\beta f=\operatorname*{extr}_{q,C,D,\{\lambda_K\}}\left[G_0+\frac{\alpha_0}{K}G_1\right]. \tag{29}$$

We find

$$G_0=\frac{1}{2K}\ln[1-q+(K-1)(C-D)]+\frac{K-1}{2K}\ln(1-q-C+D)+\tfrac{1}{2}\ln(2\pi e)$$
$$-\frac{(C-D)[q+(K-1)D]}{2[1-q+(K-1)(C-D)](1-q-C+D)}+\frac{q-(\alpha_0/K)\sum_k^K\lambda_k^2}{2(1-q-C+D)}+\frac{\alpha_0(C-D)\left[\sum_k^K\lambda_k\right]^2}{2K[1-q+(K-1)(C-D)]} \tag{30}$$

and

$$G_1=\tilde{P}\int Dy\int\prod_{k=1}^{K}Dt_k\ln\left\{\int Dz\operatorname*{Tr}_{\{\tau_k=\pm1\}}\exp\left[-\beta\Theta\left[\frac{-1}{\sqrt{K}}\sum_k\tau_k\right]\right]\prod_k H\left[\tau_k\frac{t_k\sqrt{q-D}+z\sqrt{C-D}+y\sqrt{D}-\lambda_k}{\sqrt{1-q-C+D}}\right]\right\}. \tag{31}$$

To investigate the large K behavior of G_1 we have to make self-consistent scaling assumptions: $c=KC$ and $d=KD$. Such scalings have already been observed [11,7] in other studies of fully connected committee machines. We also make a simple scaling assumption for λ_k which makes G_1 scale independently of K in the large K limit and permits a limited breaking of the translational invariance:

$$\lambda_k=\begin{cases}\Delta & \text{in } \eta\sqrt{K} \text{ hidden units}\,,\\ \dfrac{\hat{\lambda}}{\sqrt{K}} & \text{for the remaining } K-\eta\sqrt{K}\,.\end{cases} \tag{32}$$

Parameters Δ, $\hat{\lambda}$, and η are all $O(1)$. Note that setting $\eta=0$ or $\Delta=0$ restores translational invariance for λ. As in the tree committee machine, we expand G_1 to first order in $K^{-1/2}$ to obtain (see the Appendix)

$$G_1=\tilde{P}\int Dt\ln\left[e^{-\beta}+(1-e^{-\beta})\times H\left[\frac{t\sqrt{Q_{\text{eff}}}-\lambda_{\text{eff}}}{\sqrt{1-Q_{\text{eff}}+(2/\pi)c}}\right]\right] \tag{33}$$

with

$$Q_{\text{eff}}=\frac{2}{\pi}(d+\arcsin q) \tag{34}$$

and

$$\lambda_{\text{eff}}=\left[\frac{2}{\pi}\right]^{1/2}\hat{\lambda}-\eta(2H(\Delta)-1)\,.$$

The G_0 term is also expressed as a function of the rescaled order parameters:

$$G_0=\frac{\ln(2\pi e)}{2}-\frac{\alpha_0\eta\Delta^2}{2\sqrt{K}\,(1-q_0)}+\frac{q_0-(\alpha_0/K)\hat{\lambda}^2}{2(1-q_0)}+\frac{\ln(1-q_0)}{2}+\frac{\alpha_0(c-d)(\eta\Delta+\hat{\lambda})^2}{2K(1-q_0+c-d)}$$
$$+\frac{1}{2K}\left[\ln(1-q_0+c-d)-\ln(1-q_0)-\frac{(c-d)(1+c)}{(1-q_0)(1-q_0+c-d)}\right], \tag{35}$$

where we define $q_0=q-D$ which is zero when two solutions which only differ by a permutation of hidden-unit index are part of a single connected space of solutions [13]. As a function of the rescaled parameters, the generalization function becomes

$$\epsilon_g(\hat{\lambda},\eta,\Delta,c)=H\left[\frac{\lambda_{\text{eff}}}{\sqrt{1+(2/\pi)c}}\right]. \tag{36}$$

The solution to these equations depends upon how the number of examples scales in relation to the number of hidden units. We therefore consider two regimes.

Regime I: $\tilde{P}\sim O(1)$

This scaling is equivalent to that investigated for the tree committee machine and means that $G_1\sim O(1)$. In order to have self-consistent saddle point equations, we therefore require that (the nonconstant part of) G_0 be $O(1/K)$. For this, we must have $\Delta=0$ (committee-prototype symmetry) and $q_0=0$ (permutation symmetry). The translational invariance of λ implies that all the hidden units will do equally well, i.e., each has the same probability for making a correct classification. Permutation symmetry tells us that the network is far below its capacity.

For $\alpha_0=1.6$ and $T=0$, we have solved the saddle point equations as a function of $\tilde{P}$ and inserted the corresponding values of c, d, and $\hat{\lambda}$ into Eq. (36) to give the generalization error ϵ_g shown in Fig. 3(a). We see that it decreases monotonically from 0.5 at $\tilde{P}=0$ to a finite asymptotic value as $P\gg 1$. The asymptotic generalization error as a function of temperature is shown in Fig. 4. It decreases with increasing temperature and it is interesting to note that, for $T\to\infty$, it goes to the same asymptotic value as for the simple perceptron. In the committee-prototype symmetric phase the fully connected committee machine always performs worse than the simple perceptron for the same number of examples. This is due to the fact that, when committee-prototype symmetry holds, the problem is unlearnable for the network, which is actually disadvantaged by its more complex architecture in the same way as the tree committee machine was in Sec. III. As before, we see that the best training is the noisiest, i.e., when $T\to\infty$.

Regime II: $\tilde{P}\sim O(\sqrt{K})$

When the number of examples per prototype is $O(\sqrt{K})$, we rescale $\tilde{P}_1=\tilde{P}/\sqrt{K}$. As G_1 is now $O(\sqrt{K})$, self-consistent saddle point equations require that deviations of the free energy about a constant term be $O(1/\sqrt{K})$. This demands that $q_0=0$, i.e., the network is still far below its capacity. In order to have solutions to the saddle point equations $\partial(\beta f)/\partial c=\partial(\beta f)/\partial d=0$ we require that to $O(1/\sqrt{K})$

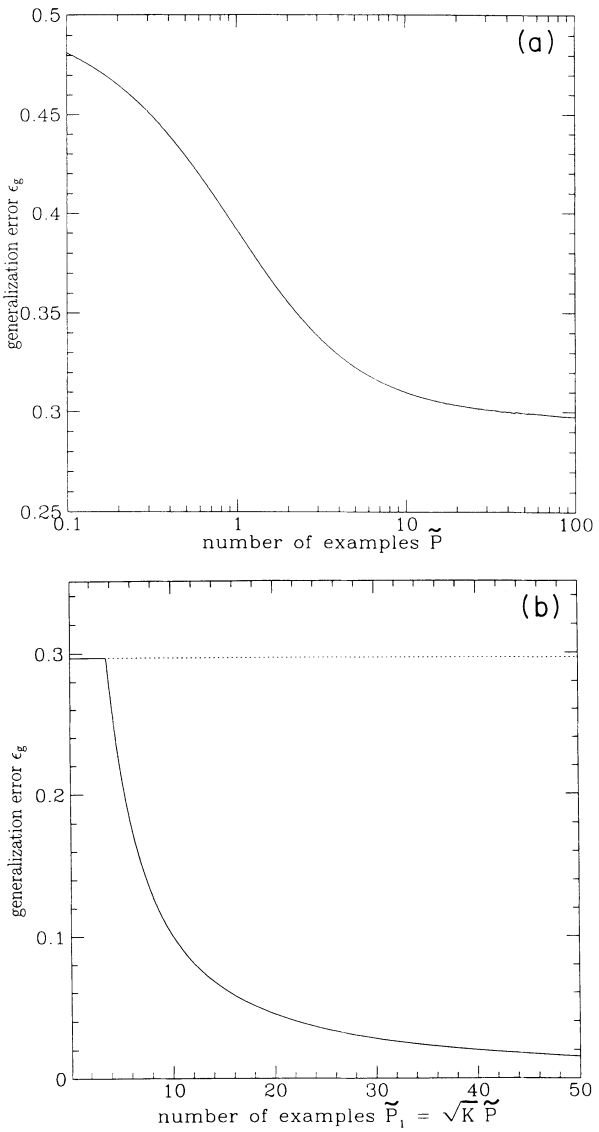

FIG. 3. Learning curve for fully connected committee machine. (a) The $\tilde{P}$ regime, and (b) the $\tilde{P}_1$ regime, both with $\alpha_0=1.6$ and $T=0$. For $\tilde{P}_1<3.64$ there exists one physical solution, the symmetric $\Delta=0$ solution with constant $\epsilon_g=0.27$. At $\tilde{P}_1=3.64$ the solution with broken symmetry $\Delta=-2.28$ becomes physical ($\eta>0$) and the system makes a continuous transition to this solution. The symmetric solution remains metastable for all $\tilde{P}_1<3.64$.

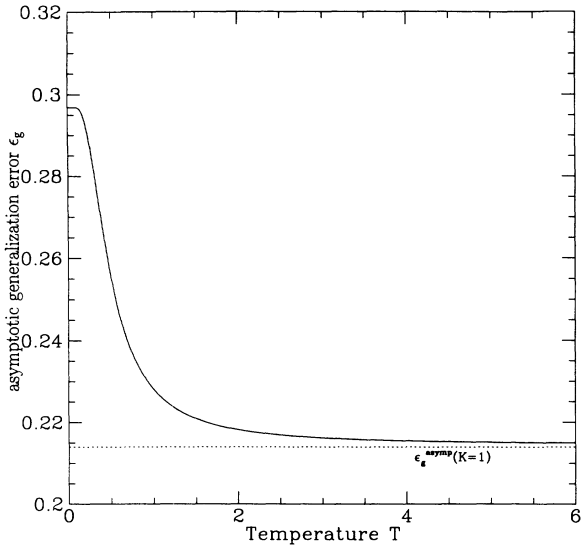

FIG. 4. Asymptotic generalization error $\epsilon_g(\tilde{P}\to\infty)$ of the fully connected CM in the symmetric phase as a function of training temperature T for $\alpha_0=1.6$.

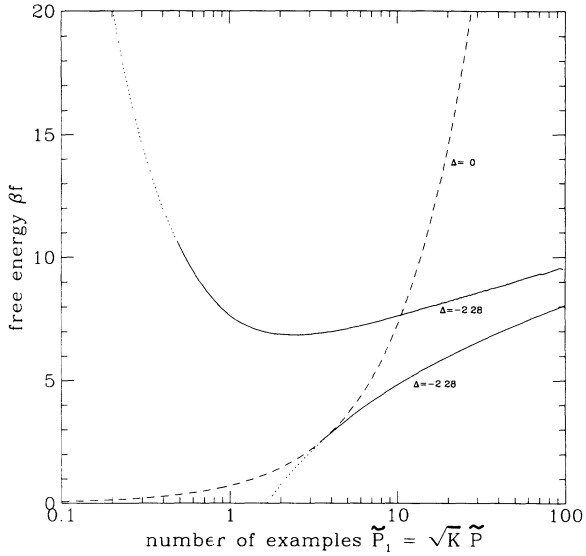

FIG. 5. Free energy of three solutions for the fully connected committee machine for $\alpha_0=1.6$ and $T=0$ in the $\tilde{P}_1$ regime. The dashed line represents the symmetric solution, the dotted lines the asymmetric solution with $\eta<0$, and the full lines the asymmetric solutions with $\eta>0$. The asymmetric solution with highest free energy corresponds to a maximum of the free energy.

$$d=\alpha_0(\eta\Delta+\hat{\lambda})^2 \quad \text{and} \quad c=d-1 \ . \tag{37}$$

Using the above relations the two remaining saddle point equations can be brought into a simplified form:

$$\Delta+(1-e^{-\Delta^2/2})\frac{\partial G_1}{\partial\hat{\lambda}}\frac{1}{\sqrt{K}}=0 \ , \tag{38}$$

$$\frac{\partial G_1}{\partial\hat{\lambda}}+2\alpha_0(\hat{\lambda}+\eta\Delta)\left[\frac{\partial G_1}{\partial c}+\frac{\partial G_1}{\partial d}\right]=0 \ . \tag{39}$$

The free energy must also be minimized with respect to the parameter η which describes the degree of breaking of the hidden-unit symmetry and can be considered as a continuous variable in the large K limit. The saddle point equation for η yields

$$\frac{\Delta^2}{2}+\left[\Delta+\left[\frac{\pi}{2}\right]^{1/2}[2H(\Delta)-1]\frac{\partial G_1}{\partial\hat{\lambda}}\frac{1}{\sqrt{K}}\right]=0 \ . \tag{40}$$

In this regime we have only considered zero-temperature learning. For $\alpha_0=1.6$, we have found three solutions to the saddle point equations. The free energy of these is shown in Fig. 5. The committee-prototype symmetric solution with $\Delta=0$ still exists and has a constant generalization error equal to the asymptotic value of the $\tilde{P}\sim O(1)$ regime.

Two additional solutions exist with Δ fixed at a value of -2.28 [this can be seen from combining Eqs. (38) and (40)]. This negative value signifies that $\eta\sqrt{K}$ hidden units are actually more likely to give the wrong answer. Whilst we find some solutions with $\eta<0$, we discount these on the grounds that they are unphysical. In this region, the physical solution has $\eta=0$. The solution with the lowest free energy has $\eta>0$ for $\tilde{P}_1>3.64$. On increasing $\tilde{P}_1$, the optimal η increases monotonically from zero and so we have a continuous transition to a phase with broken committee symmetry. The corresponding generalization error is shown in Fig. 3(b). In this broken symmetry phase, we have seen that a vanishing fraction of the hidden units generalize particularly badly. However, we find that this is more than compensated for by the rest of the hidden units, which classify correctly with a higher probability than in the committee-symmetric phase. Though still very small ($\sim 0.5+\hat{\lambda}/\sqrt{2\pi K}$) the overall effect is very significant since it results in a decrease of the asymptotic generalization error to zero—given enough examples, the network can learn this task perfectly. To first order, the asymptotic approach to zero has been found to scale as follows:

$$\epsilon_g(\tilde{P}_1\to\infty)\sim\frac{1}{\tilde{P}_1\sqrt{\ln\tilde{P}_1}} \ . \tag{41}$$

V. SUMMARY AND CONCLUSIONS

In this article we have studied the ability of two different architectures to learn a specific non-linearly-separable classification task.

For the simpler tree committee machine, the nonrealizable nature of the task and $T=0$ Gibbs learning combined to produce the phenomenon of overtraining, i.e., an increase in the number of training examples results in a reduction of the network generalization ability. Finite-temperature training improved the situation by reducing the value of the generalization error. However, this was still greater than the corresponding value for the simple

perceptron. We conclude that the tree architecture is very bad at learning this task and have speculated that this is due to the independent nature of the hidden units.

Due to the correlations arising from having overlapping receptive fields, the fully connected committee machine is a more interesting but also much more complicated machine to study. We have considered the case of finite-temperature Gibbs learning dynamics and introduced an ansatz to allow for a possible breaking of translational invariance for λ. When trained with a finite number of examples, it was found that the network prefers to remain in a symmetric phase, i.e., all hidden units have the same probability of classifying a novel example correctly. The asymptotic generalization error decreases with increasing T to become equal to that of the simple perceptron in the limit $T \to \infty$. A poorly generalizing committee-symmetry phase has also been observed in [7].

In the regime where the number of examples is large, the network undergoes a continuous transition to a phase in which there is a breaking of the translational symmetry. A fraction $O(1/\sqrt{K})$ of the K hidden units classifies a random input pattern wrongly with high probability while the rest classify even better than in the symmetric phase. This implies that on increasing the number of examples, an increasing number of hidden units "antilearn" the prototypes and in doing so allow the remaining hidden units to classify even better. As we are far below network capacity the training energy is zero for both the symmetric and symmetry-broken solutions. Therefore the breaking of this symmetry must be due to the number of symmetry-broken solutions exceeding that of symmetric solutions. This breaking of symmetry is important since it results in an asymptotic generalization equal to zero. In fact, it is completely to be expected that one must break the symmetry of the hidden units in order that a nonlinearly-separable problem be solved. Whilst perhaps the simplest, we cannot be sure that the symmetry-breaking ansatz adopted in Eq. (32) is unique, and suggest that it might be interesting to investigate alternatives.

All of these results have been obtained within the replica symmetric ansatz. Whilst we would expect replica symmetry to become invalid for the tree committee machine as $\tilde{P}$ becomes large, we would not expect this to affect our conclusion that the tree performs worse than the simple perceptron since this is already evident at very low $\tilde{P}$. In the case of the fully connected network, we are always considering a regime in which the network is well below its capacity such that the typical overlap between solution weight vectors, $q \sim O(1/K)$, is small. We therefore expect that the replica symmetric solution is valid and the results presented here are accurate if not exact. An analytical test of the stability of the solution using the AT [9] method is difficult.

It would be interesting to study the learning of this rule by a fully connected committee machine with only a finite number of hidden units. Though technically more difficult, this could give further insight into how symmetry breaking occurs and how the asymptotic generalization error depends on K.

ACKNOWLEDGMENTS

Both authors would like to thank H. Schwarze for sharing his insight with us, and T. Watkin for a conversation from which this work sprang. We also thank S. Solla for a critical reading of our manuscript. D.O'K. thanks the Niels Bohr Institute for their kind hospitality during a brief visit and O.W. thanks the group at the University of Oxford for an instructive and pleasant stay.

APPENDIX: THE FULLY CONNECTED COMMITTEE MACHINE FOR LARGE K

Beginning with the energy term of Eq. (31), we introduce the integral representation of the Θ function to factorize the trace over hidden units:

$$\operatorname*{Tr}_{\{\tau_k\}} \prod_{k=1}^{K} [e^{-ixK^{-1/2}\tau_k} H(A_k \tau_k)] = \prod_k \left[\cos\left(\frac{x}{\sqrt{K}} \right) + iF_k \sin\left(\frac{x}{\sqrt{K}} \right) \right], \quad \text{(A1)}$$

where $F_k = F(A_k) = 1 - 2H(A_k)$ and

$$A_k = \frac{t_k\sqrt{q-D} + z\sqrt{C-D} + y\sqrt{D} - \lambda_k}{\sqrt{1-q-C+D}} . \quad \text{(A2)}$$

Next we use the identity

$$F(a+\epsilon) = F(a) + \sqrt{2/\pi}\epsilon e^{-a^2/2} + O(\epsilon^2)$$

to expand the result to order $1/\sqrt{K}$ using the scaling assumptions for the order parameters.

$$F_{1 \le k \le \eta\sqrt{K}} = F\left(\frac{t_k\sqrt{q_0} - \Delta}{\sqrt{1-q_0}} \right) + \left(\frac{2}{\pi K} \right)^{1/2} \frac{z\sqrt{c-d} + y\sqrt{d}}{\sqrt{1-q_0}} \times \exp\left(-\frac{1}{2} \frac{(t_k\sqrt{q_0} - \Delta)^2}{1-q_0} \right) + O(1/K) ,$$

$$F_{\eta\sqrt{K} < k \le K} = F\left(\frac{t_k\sqrt{q_0}}{\sqrt{1-q_0}} \right) + \left(\frac{2}{\pi K} \right)^{1/2} \frac{z\sqrt{c-d} + y\sqrt{d} - \hat{\lambda}}{\sqrt{1-q_0}} \times \exp\left(-\frac{1}{2} \frac{t_k^2 q_0}{1-q_0} \right) + O(1/K) .$$

Expanding (A1) to second order in $O(1/\sqrt{K})$ yields

$$G_1 = -\tilde{P}\int Dy \int \prod_k Dt_k \ln \int Dz \left| e^{-\beta} + (1-e^{-\beta}) \times H\left[\frac{\frac{1}{\sqrt{K}} \sum_{k=1}^{K} F_k}{\left\{ 1-(1/K) \sum_{k=1}^{\eta\sqrt{K}} [F((\sqrt{q_0}t_k - \Delta)/(\sqrt{1-q_0}))]^2 - (1/K) \sum_{k=\eta\sqrt{K}}^{K} [F(\sqrt{q_0}t_k/(\sqrt{1-q_0}))]^2 \right\}^{1/2}} \right] \right| . \tag{A3}$$

As $K \to \infty$, we use the central limit theorem on each of the different sums in the argument of the H function to obtain Eq. (33).

[1] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, New York, 1991).

[2] E. Gardner and B. Derrida, J. Phys. A **22**, 1983 (1989).

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

[4] J. S. Denker, D. Schwartz, B. Wittner, S. Solla, R. Howard, L. Jackal, and J. Hopfield, Complex Syst. **1**, 877 (1987).

[5] P. Del Guidice, S. Franz, and M. Virasoro, J. Phys. (Paris) **50**, 121 (1989).

[6] D. Hansel and H. Sompolinsky, Europhys. Lett. **11**, 687 (1990).

[7] H. Schwarze, J. Phys. A **26**, 5781 (1993).

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

[9] M. Mézard, G. Parisi, and M. A. Virsoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).

[10] H. Schwarze and J. Hertz, Europhys. Lett. **20**, 375 (1992).

[11] A. Engel, H. M. Köhler, F. Tschepke, H. Vollmayer, and A. Zippelius, Phys. Rev. A **45**, 7590 (1992).

[12] T. Watkin, K. Y. M. Wong, and A. Rau, in *Proceedings of ICANN '93,* edited by S. Gielen and B. Kappen (Springer-Verlag, London, 1993), pp. 691–697.

[13] E. Barkai, D. Hansel, and H. Sompolinsky, Phys. Rev. A **45**, 4146 (1992).

[14] H. Schwarze, Ph.D. thesis, University of Copenhagen, 1993.