# Generalization in Fully Connected Committee Machines

H. Schwarze*and J. Hertz†
CONNECT, The Niels Bohr Institute and Nordita
Blegdamsvej 17, DK-2100 Copenhagen Ø, Denmark

September 17, 1992

**Abstract**

We study supervised learning in a fully connected committee machine trained to implement a rule of the same structure. The generalization error as a function of the number of training examples per weight is calculated within the annealed approximation. For binary weights we find a discontinuous transition from poor to perfect generalization. Beyond this transition metastable states exist even for large training sets. The scaling of the order parameters with the number of hidden units depends on the size of the training set. For continuous weights we find a discontinuous transition from a committee–symmetric solution to one with specialized hidden units.

PACS. 87.10, 02.50, 64.60C

Feedforward neural networks are able to extract a rule from a set of training examples and generalize this knowledge to new inputs. Calculating the generalization ability of neural networks has therefore been a subject of considerable interest (eg.[1, 2]). The simplest network, a single layer perceptron with one layer of input units and a single output unit, is only able to implement linearly separable tasks. Most practical applications use networks with additional layers of hidden units, and recent work has concentrated on applying the methods introduced by Gardner [3] to networks with one hidden layer. This is not a severe restriction because a network with only one hidden layer can implement any Boolean or continuous function of the inputs [4, 5]. Previous work includes some general results on networks whose outputs are continuous functions of their inputs [2, 6] and calculations for restricted architectures [7]. A more general Boolean machine is the committee machine [8, 9, 10], in which all the hidden–output weights are fixed to +1 and the overall output implements a majority decision of the outputs of the hidden units. For

*Email: holm@nordita.dk
†Present address: Laboratory of Neuropsychology, NIMH, Bethesda, MD 20892, USA

networks with binary weights, the committee machine may already be regarded as the most general two–layer machine, since any combination of hidden–output weights can then be gauged to +1 by adjusting the signs of the input–hidden weights. In the 'tree' version of this model different committee members receive inputs from nonoverlapping regions of the input layer and are therefore uncorrelated. The qualitative behavior of this model has been found to be similar to a simple perceptron [9].

Here we extend this work to fully connected committee machines, allowing every committee member to receive inputs from the entire input layer. The network is trained to implement a rule which is given by another committee machine and is therefore not linearly separable. We study a large–net limit in which the number of inputs $N$ and the number of hidden units $K$ both go to infinity, but with $K \ll N$. We calculate the generalization error $\epsilon_g$ (the average error on an example not in the training set) as a function of the ratio $\alpha = P/(KN)$ of the training set size to the number of adjustable weights in the network. This will be done in the annealed approximation, which becomes exact for high temperatures (high noise during training). For learnable tasks, as in the present problem, the annealed approximation is known to give qualitative correct results even for low temperatures and predicts correctly the shape of learning curves [2]. A restricted version of this model with 3 hidden units has been studied in the high–temperature limit learning a linearly separable function [10]. Furthermore, the capacity problem for this model has recently been studied [11, 12].

We consider a committee machine with $N$ inputs $\underline{S} \in R^N$ and $K$ hidden units. Each hidden unit can be regarded as a simple perceptron (*student*) with weight vector $\underline{W}_l$ receiving the net input $u_l = N^{-1/2}\, \underline{W}_l \cdot \underline{S}$. The overall network output is given by

$$\sigma(\underline{S}) = \mathrm{sign}\left(\frac{1}{\sqrt{K}} \sum_{l=1}^{K} \mathrm{sign}\, u_l\right). \tag{1}$$

The network training is based on a set of $P = \alpha KN$ input–output pairs $(\underline{\xi}^\mu, \tau(\underline{\xi}^\mu)), \mu \in 1, \ldots, P$ of the target rule. In the present problem the target rule is defined by another committee machine with weight vectors (*teachers*) $\underline{V}_l$, internal fields $v_l = N^{-1/2}\, \underline{V}_l \cdot \underline{S}$, and an overall output $\tau(\underline{S})$ of the form (1). We assume that the teacher weight vectors are normalized to $\sqrt{N}$ and mutually uncorrelated, $N^{-1}\, \underline{V}_l \cdot \underline{V}_k = \delta_{lk}$.

Following the statistical mechanics approach as reviewed in [1], we consider a stochastic training process, which for long training times generates a Gibbs distribution in the space of couplings with the corresponding partition function

$$Z = \int d\rho_0(\{\underline{W}_l\})\, e^{-\beta E(\{\underline{W}_l\})}. \tag{2}$$

Here $E$ denotes the training error defined as the number of wrongly classified training examples $E = \sum_{\mu=1}^{P} \Theta[-\sigma(\underline{\xi}^\mu) \cdot \tau(\underline{\xi}^\mu)]$. The formal temperature $T = 1/\beta$ determines the amount of noise during the training process. Constraints on the student weights are included in the normalized *a priori* measure $d\rho_0(\{\underline{W}_l\})$. For continuous weights this is a product of delta–functions constraining each $\underline{W}_l$ to length $\sqrt{N}$. For binary weights we just get a sum over all $\underline{W}_l \in \{-1, +1\}^N$.

The average generalization error may be determined from the free energy averaged over all possible training sets $F = -T \langle\langle \ln Z \rangle\rangle$. In this paper 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 $F_{an} = -T \ln \langle\langle Z \rangle\rangle \leq F$. Performing the average of $Z$ and introducing the order parameters $C_{lk} = N^{-1} \underline{W}_l \cdot \underline{W}_k$ and $R_{lk} = N^{-1} \underline{W}_l \cdot \underline{V}_k$ leads in the thermodynamic limit ($N \to \infty$) to the problem of minimizing the free energy density [2]

$$-\beta f(R_{lk}, C_{lk}) = \alpha \ln\left[1 - \left(1 - e^{-\beta}\right) \epsilon(R_{lk}, C_{lk})\right] + G_0(R_{lk}, C_{lk}). \tag{3}$$

Here $\epsilon$ is the generalization error $\epsilon(R_{lk}, C_{lk}) = \langle \Theta[-\sigma(\underline{S}) \cdot \tau(\underline{S})] \rangle_{\underline{S}}$, and the entropy term $G_0$ measures the volume of weight space consistent with given $R_{lk}$ and $C_{lk}$.

**Binary weights:** During training all hidden units in the student network obtain identical information about the overall target output implying symmetric behavior of the hidden units. However, the ground state with every student $\underline{W}_l$ being perfectly aligned with one teacher $\underline{V}_l$ is not symmetric. We therefore make a partially symmetric ansatz for the order parameters which accounts for both situations and is in agreement with our Monte Carlo simulations. Following [10] 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 $\eta K$ students have symmetric overlaps $R$ with all the teachers and mutual overlaps $C$, yielding

$$R_{lk} = \begin{cases} R & \text{if } l \leq \eta K \\ \delta_{lk} & \text{otherwise} \end{cases}, \quad C_{lk} = \begin{cases} C + (1-C)\delta_{lk} & \text{if } l,k \leq \eta K \\ \delta_{lk} & \text{if } l,k > \eta K \\ R & \text{otherwise} \end{cases}. \tag{4}$$

Having made this ansatz we can proceed with the calculation of the entropy term $G_0$ and the generalization error $\epsilon$ as a function of the order parameters $R$ and $C$. In the large–$K$ limit, we obtain to order $1/K$

$$G_0(R,C) = \eta \ln 2 - \frac{\eta}{2}\frac{K-1}{K} C + \frac{1}{2K} \ln\left[(1-\eta C) - \eta K \left(K R^2 - C\right)\right]. \tag{5}$$

After introducing internal representations $\{\tau_l\}$ for the teacher network and $\{\sigma_\lambda\}$, $\lambda \in \{1, \ldots, \eta K\}$, for the nonfrozen students, a straightforward calculation yields the generalization error

$$\begin{aligned} \epsilon(R,C) &= 2 \sum_{\{\sigma_\lambda\}} \sum_{\{\tau_l\}} \Theta\left(-K^{-1/2} \sum_l \tau_l\right) \Theta\left[K^{-1/2}\left(\sum_\lambda \sigma_\lambda + \sum_\kappa \tau_\kappa\right)\right] \\ &\times \int_{-\infty}^{+\infty} Dt \int_0^\infty \prod_l Dv_l \prod_\lambda H\left(\sigma_\lambda \frac{it\sqrt{K R^2 - C} - R \sum_l \tau_l v_l}{\sqrt{1-C}}\right), \end{aligned} \tag{6}$$

where $H(x) = \int_x^\infty Dz$ and $Dz = dz\, e^{-z^2/2}/\sqrt{2\pi}$. In sums and products $l$ ranges over all hidden units, $\lambda$ over the $\eta K$ unfrozen units and $\kappa$ over the $(1-\eta)K$ frozen units.

Further simplifications of (6) for large–$K$ require assumptions about how the order parameters scale with $K$. We introduce new parameters $r = K^{3/4}R$ and $c = \sqrt{K}C$, which are assumed to be of order one, and we determine their values at the minimum of the free energy self–consistently. The entropy term $G_0$ puts an additional restriction on $r$ and $c$, since the argument of the logarithm in (5) can only be positive if $\eta\,(r^2 - c) < 1/\sqrt{K}$.

To do the traces over internal representations in (6) we introduce integral representations for the $\Theta$–functions and control the sum $\sum_l \tau_l v_l$ through a new variable by inserting an integral over a $\delta$–function. For large $K$ the remaining integrals can be done yielding to order $1/\sqrt{K}$

$$\epsilon(r,c) = \frac{1}{\pi}\arccos\left(\frac{\frac{2}{\pi}\,\eta\,r + \frac{1}{K^{1/4}}\,(1-\eta)}{\sqrt{\frac{2}{\pi}\,\eta^2\,r^2 + \frac{1}{K^{1/4}}\,\frac{4}{\pi}\,\eta(1-\eta)\,r + \frac{1}{\sqrt{K}}\left[1 - \frac{2}{\pi}\,\eta^2\,\sqrt{K}(r^2-c)\right]}}\right). \qquad (7)$$

The generalization error is dominated by the constant term

$$\epsilon_0 = \pi^{-1}\ \arccos\sqrt{2/\pi}\ \approx\ 0.2\,. \qquad (8)$$

Deviations from this value are of order $K^{-1/2}$ in the fully symmetric case ($\eta = 1$) and of order $K^{-1/4}$ if some of the hidden units are frozen ($\eta < 1$). To leading order the annealed free energy is simply

$$-\beta f = \alpha \ln\left[1 - \left(1 - e^{-\beta}\right)\epsilon_0\right] + \eta\ \ln 2, \qquad (9)$$

independent of $r$ and $c$. For given $\alpha$, $f$ reaches its minimum for the fully symmetric case $\eta = 1$; a partial freezing of hidden units ($\eta < 1$) has a higher free energy and therefore only describes metastable states. This is in contrast to the case of a committee machine learning a linearly separable rule, where an asymmetric freezing was found [10]. If all hidden units are perfectly aligned with the teacher ($\eta = 0$), both terms in the free energy vanish. Therefore this state corresponds to the equilibrium situation, if the free energy of the symmetric state is positive. At a critical value of the load parameter $\alpha_c = |\ \ln 2/\ln[1-(1-e^{-\beta})\,\epsilon_0]\ |$ the system exhibits a first order phase transition to a state of perfect generalization (figure 1).

To study the properties of the fully symmetric state ($\eta = 1$) in more detail we determine the equilibrium values of the order parameters $r$ and $c$. Minimizing the free energy using (5) and (7) yields

$$r = \left[\frac{\alpha}{4\pi^2\gamma(\beta)}\left(\frac{\pi-2}{2}\right)\right]^{1/4} + \mathcal{O}\left(\frac{1}{\sqrt{K}}\right), \quad r^2 - c = \frac{1}{\sqrt{K}} - \mathcal{O}\left(\frac{1}{K}\right), \qquad (10)$$

with $\gamma(\beta) = \sqrt{\pi/2-1}\,[(1-e^{-\beta})^{-1} - \epsilon_0]/(4\pi)$, in agreement with our previous ansatz. This solution exists even for very large values of $\alpha$, indicating a metastable state beyond the phase transition at $\alpha_c$. Inserting (10) in (7) yields the average generalization error as a

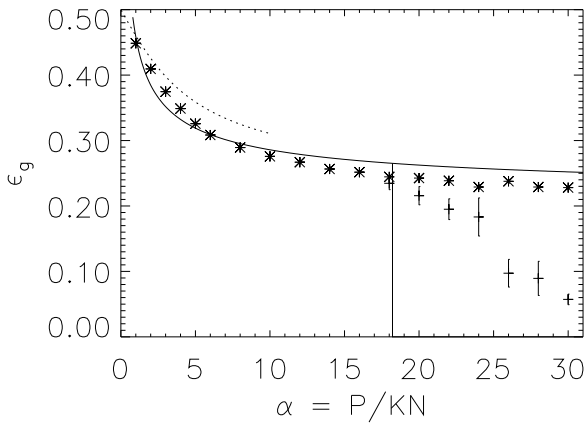


Figure1: Generalization error for binary weights at $T = 5$. The large–$K$ theory for different regions of $\alpha$ is compared with simulations for $K = 5$ averaged over all simulations (+) and the nonfrozen simulations ($*$) respectively. The solid line shows the finite–$\alpha$ results including $1/\sqrt{K}$–corrections. The dotted line shows the small–$\alpha$ solution.

function of $\alpha$, $\epsilon_g = \epsilon_0 + \sqrt{\gamma(\beta)}/\sqrt{\alpha K} + \mathcal{O}(1/K)$. For large values of $\alpha$ the leading correction vanishes and the generalization error approaches a nonvanishing constant value (8). For small $\alpha$, on the other hand, this term causes a breakdown of the approximation based on our scaling ansatz. The leading corrections to the generalization error are only small for $\alpha \gg 1/K$ and diverge for $\alpha \to 0$.

For the region where $\alpha$ itself is of order $1/K$ we can construct another self–consistent approximation based on a different scaling assumption for the order parameters. We expect the overlap $R$ between student and teacher hidden units to be smaller than in the finite–$\alpha$ case and assume that $R$ is now of order $1/K$. The logarithmic term in (5) then requires that $C$ is of the same order and therefore $\rho = KR$ and $q = KC$ are of order 1. Inserting these parameters in (6), a calculation similar to the previous one now yields for $K \to \infty$

$$\epsilon(\rho, q) = \frac{1}{\pi} \arccos \left( \frac{2\rho/\pi}{\sqrt{1 + 2q/\pi}} \right). \tag{11}$$

Minimizing the corresponding free energy yields a smoothly decreasing generalization error $\epsilon_g$ as a function of $\tilde{\alpha} = \alpha K$. Its asymptotic behavior for $\tilde{\alpha} \to \infty$ is given by $\epsilon_g \approx \epsilon_0 + \mathcal{O}(1/\sqrt{\tilde{\alpha}})$, in agreement with our solution for finite $\alpha$.

We have performed Monte Carlo simulations to check our analytic results 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 show a symmetric freezing of all hidden units and a double peak structure in the distribution of the order parameters beyond the transition at $\alpha_c$. The predictions for the metastable state can be checked by separately averaging over those simulations in which no freezing occurred. Figure 1 shows

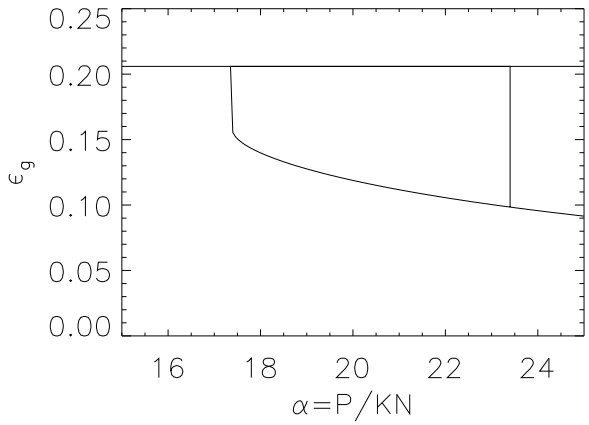


Figure2: Generalization error for continuous weights and $T = 0.5$. The horizontal line corresponds to the residual error $\epsilon_0$.

results for $K = 5$ compared to our analytic results. The solid line corresponds to the predictions of the first ansatz for $K = 5$, showing the location of the phase transition and the divergence for small values of $\alpha$.

**Continuous weights:** In the case of continuous weights we need only modify the expression for the entropy term $G_0$, which involves an integration over the weight space. In both regimes of $\alpha$ we find the same dependence on the order parameters as for binary weights to order $1/K$. The effect of the discreteness of the weights vanishes for large $K$ and the generalization error shows the same approach to $\epsilon_0$. In contrast to the binary case, a freezing into the teacher weights cannot happen here, because this solution would have a diverging entropy. However, we can construct a solution in which the symmetric behavior of the committee is broken in favour of a specialization of the individual hidden units. We assume that each hidden unit has an overlap $R$ to one of the teachers and a small overlap $\rho/K$ to all the remaining ones. Therefore we now have $R_{lk} = \rho/K + \Delta\delta_{lk}$ and $C_{lk} = q/K + (1 - q/K)\delta_{lk}$, with $\Delta = R - \rho/K$. We use these parameters to calculate the free energy similarly to the previous cases. At the minimum of $f$ we can eliminate the parameters $r$ and $q$ and are left with a free energy as a function of $\Delta$

$$\begin{aligned} -\beta\, f(\Delta) &= \alpha\, \ln[1 - (1 - e^{-\beta})\,\epsilon(\Delta)] + \frac{1}{2}(1 + \ln 2\pi) + \frac{1}{2}\ln(1 - \Delta^2) \\ \epsilon(\Delta) &= \frac{1}{\pi}\arccos\left[\sqrt{\frac{2}{\pi}}\sqrt{1 + \frac{2}{\pi - 2}(\Delta - \arcsin\Delta)^2}\right]. \end{aligned} \tag{12}$$

For small values of $\alpha$, $f$ only has a minimum at $\Delta = 0$ corresponding to the residual generalization error $\epsilon_0$ (8). At a critical value $\alpha_s$ of the load parameter, a second local minimum appears at a value of $\Delta$ close to 1. It becomes the global minimum at $\alpha_c > \alpha_s$.

The system undergoes a discontinuous transition from a symmetric state to a specialized state, in which each student has an overlap of order 1 to one of the teachers and an overlap of order $1/K$ to the remaining ones. The symmetric state remains metastable for all $\alpha > \alpha_c$. The generalization error of the specialized state decays smoothly with an asymptotic behavior given by $\epsilon_g \approx 2/\alpha$ (figure 2). Hence, for both binary and continuous weights we find a nontrivial behavior of the generalization error due to a specialization of the individual hidden units in the student network.

We thank M. Biehl and 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.

# References

[1] T. Watkin, A. Rau, and M. Biehl, preprint Oxford University, 1992

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

[3] E. Gardner, J.Phys. A **21**, 257 (1988).

[4] J.S. Denker *et al.*, Complex Systems **1**, 877 (1987).

[5] R. Hecht–Nielsen, in *IEEE First International Conference on Neural Networks,* ed. by M. Caudill and C. Butler (IEEE, New York, 1987), Vol. II.

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

[7] H. Sompolinsky, N. Tishby, Europhys.Lett. **13**, 567 (1990).

[8] N.J. Nilsson *Learning Machines,* McGraw–Hill, New York 1965.

[9] H. Schwarze, J. Hertz, Europhys. Lett. (to be published).

[10] H. Schwarze, M. Opper, and W. Kinzel, preprint 1991

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

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