Generalization in a Large Committee Machine

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

March 6, 1992

Abstract

We study generalization in a committee machine with non–overlapping receptive fields trained to implement a function of the same structure. Using the replica method, we calculate the generalization error in the limit of a large number of hidden units. For continuous weights the generalization error falls off asymptotically inversely proportional to α, the number of training examples per weight. For binary weights we find a discontinuous transition from poor to perfect generalization followed by a wide region of metastability. Broken replica symmetry is found within this region at low temperatures. The first–order transition occurs at a lower and the metastability limit at a higher value of α than in the simple perceptron.

PACS. 87.10, 02.50, 64.60C

A good part of the utility of neural networks lies in their ability to learn a function from examples. Therefore there has been a good deal of theoretical work on calculating the *generalization ability* of different networks, i.e. their ability to give the correct output for an input not used in training. Most of these calculations have dealt with single–layer nets. Extensions to networks with a hidden layer include a model with small hidden receptive fields[1], some general results on networks whose outputs are continuous functions of their inputs[2, 3], and a calculation for a so–called *committee machine*[4] learning a function which could be implemented by a simple perceptron (i.e. one with no hidden units) in the high–temperature (i.e. high–noise) limit[5]. In this letter we make a step toward a more general understanding of nets with a single hidden layer by solving the 'tree' version of the committee machine in the limit of a large number of committee members (hidden units). In the present problem, the function to be learned is matched to the machine, i.e. the 'teacher' and 'student' are both tree committee machines.

For networks with a single output and 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 flipping the signs of all input–hidden weights to units which lead on to the output through negative connections. In the restricted version of the model which we solve, the receptive fields of the hidden units do not overlap. Furthermore we study a large–net limit in which the total number of inputs (N) and the number of hidden units (K) both go to infinity, but with $K \ll N$.

Our specific task is to calculate the generalization error ϵ_g (the average error on an example not in the training set) as a function of the ratio $\alpha = p/N$ of the training set size to the total number of weights in the network.

The 'capacity' problem (i.e. the number of random input–output associations a network is capable of learning) has already been solved for this machine[6, 7], and the present calculation of the generalization error is done in a similar way, starting from an expression for the volume in weight space in which the network gives the correct output[8]. As in the corresponding calculations for simple perceptrons[2, 9, 10], we weight the volume in weight space with a thermal factor $\exp(-\beta E)$, where E is the total error on the training examples. The resulting statistical mechanical problem, formulated in replica formalism and averaged over all possible sets of training examples, is then solved, assuming replica symmetry, for both continuous and binary input–hidden weights.

Like the calculations for simple perceptrons, our theory has two order parameters: q, the pattern average of the square of the average input–hidden weight vector, and R, the average overlap between this weight vector and a corresponding one for the teacher.

In the (zero–temperature) capacity calculations of Barkai et al it is found that the tree committee machine admits a striking simplification in the large–K limit. The expression for the 'free energy' to be minimized with respect to q is exactly the same as that for the simple perceptron except that in one term (the 'energy' term) q must be replaced by $f(q) = (2/\pi)\sin^{-1} q$. Here we show that a similar simplification also occurs in the generalization problem, also for nonzero temperature. The replacement of q by $f(q)$ is also valid here, and we find that R has to be replaced by $f(R)$, both in the energy term in the expression for the free energy and in the formula for the generalization error as a function of R.

The derivation of this basic result is somewhat long and will be presented in full elsewhere. Here we only have space for a brief sketch. We start with the training energy

$$E = \sum_\mu E_\mu = \sum_\mu \theta(-h^\mu_{\mathsf{u}} h^\mu_{\mathsf{w}}), \tag{1}$$

where μ labels training inputs, and

$$h^\mu_{\mathsf{w}} = \frac{1}{\sqrt{K}} \sum_i \mathrm{sgn}\left(\sqrt{\frac{K}{N}} \sum_j w_{ij}\xi^\mu_{ij}\right) \tag{2}$$

is the net input to the output unit of the student, and correspondingly for the teacher with weights u_{ij}. The convention is that i labels hidden units and j the input units: ξ_{ij} is the jth input to branch i. There are K/N values of j for each i, i.e. N inputs in all. We write the Boltzmann weight in the form

$$\exp(-\beta E) = \prod_\mu [\theta(h^\mu_{\mathsf{u}} h^\mu_{\mathsf{w}}) + e^{-\beta}\theta(-h^\mu_{\mathsf{u}} h^\mu_{\mathsf{w}})] \tag{3}$$

The partition function Z is then obtained by integrating (3) over the *a priori* weight density $\rho_0(\mathsf{w})$. For continuous weights this is a product of delta–functions constraining each $\mathbf{w}_i$ to length $\sqrt{N/K}$. For binary weights we just get a sum over each $w_{ij} = \pm 1$.

To average $\ln Z$ over the training patterns, we use replicas in the standard way. Employing the identities $\theta(ab) = \theta(a)\theta(b) + \theta(-a)\theta(-b)$ and $\theta^p(x) = \theta(x)$ $(p \geq 1)$, we are led to a product of p identical single–pattern averages

$$\langle\langle \theta(h_{\mathsf{u}}) \prod_{\alpha=1}^{n} \phi(h_{\mathsf{w}^\alpha}) + \theta(-h_{\mathsf{u}}) \prod_{\alpha=1}^{n} \phi(-h_{\mathsf{w}^\alpha}) \rangle\rangle \tag{4}$$

to evaluate (one for each training pattern), with $\phi(h) = \theta(h) + e^{-\beta}\theta(-h)$. To carry out these averages, we introduce internal representations σ_i^α, writing

$$\phi(h_{\mathsf{w}^\alpha}) = \sum_{\{\sigma_i^\alpha\}} \left[\theta\left(\frac{1}{\sqrt{K}} \sum_i \sigma_i^\alpha \right) + e^{-\beta}\theta\left(\frac{-1}{\sqrt{K}} \sum_i \sigma_i^\alpha \right) \right] \times \theta\left(\sigma_i^\alpha \sqrt{\frac{K}{N}} \sum_j w_{ij}^\alpha \xi_{ij} \right) \tag{5}$$

and similarly for the teacher $\theta(h_{\mathsf{u}})$, with another set of internal representation variables τ_i. For each hidden unit i we then employ the Gardner representation of the input–hidden θ–functions, average over the input patterns, and introduce the order parameters $q_i^{\alpha\beta} = K/N \sum_j w_{ij}^\alpha w_{ij}^\beta$ and $R_i^\alpha = K/N \sum_j w_{ij}^\alpha u_{ij}$. We assume 'translational invariance' i.e. that $q_i^{\alpha\beta}$ and R_i^α are independent of i.

The next part of the calculation follows exactly that for the simple perceptron [10] and leads to an expression for each i of the form

$$\int_0^\infty D\mu_i \int_{-\infty}^{\infty} Dt_i \prod_\alpha H\left[\sigma_i^\alpha \left(\frac{t_i\sqrt{q - R^2} - \mu_i R \tau_i}{\sqrt{1-q}} \right) \right] \tag{6}$$

(with $H(x) = (2\pi)^{-1/2} \int_x^\infty e^{-u^2/2} du$). Introducing Gardner representations for the hidden–output factors in (5) (and the corresponding one for the teacher) then gives us a trace

$$\sum_{\{\sigma_i^\alpha\}} \sum_{\{\tau_i\}} e^{-iy\tau_i/\sqrt{K}} \prod_\alpha \left\{ e^{-ix^\alpha \sigma_i^\alpha/\sqrt{K}} H\left[\sigma_i^\alpha \left(\frac{t_i\sqrt{q - R^2} - \mu_i R \tau_i}{\sqrt{1-q}} \right) \right] \right\} \tag{7}$$

to be done over the σ_i^α's and τ_i. (Here x^α and y are integration variables in the representations of the step functions.) We expand the result in $1/K$ to obtain integrals of the form

$$f_p(q, R) = \int_0^\infty D\mu \int_{-\infty}^{\infty} Dt \left[\hat{H}\left(\frac{t\sqrt{q - R^2} - \mu R}{\sqrt{1-q}} \right) \right]^p \tag{8}$$

for $p = 1, 2$, where $\hat{H}(x) \equiv 1 - 2H(x)$. We find that $f_1(q, R) = -\pi^{-1}\sin^{-1} R$ (independent of q), and $f_2(q, R) = \pi^{-1}\sin^{-1} q$ (independent of R).

Doing the integrals over the variables introduced in the representations of the hidden–output step functions is now just like what is done for the simple perceptron, except that

q is replaced by $f(q) = 2f_2(q) = (2/\pi)\sin^{-1} q$ and R is replaced by $f(R) = -2f_1(R) = (2/\pi)\sin^{-1} R$.

The foregoing summarizes the calculation of the 'energy' term $G_1(q, R)$ in the final expression for the free energy. That of the 'entropy' term G_2 is exactly the same (with $q_i = q$ and $R_i = R$) as in the simple perceptron, giving

$$G_2 = -\tfrac{1}{2}(1-q)\hat{q} - R\hat{R} + \int Dz \ln\cosh\left(\sqrt{\hat{q}}z + \hat{R}\right) \tag{9}$$

for binary weights and

$$G_2(q, R) = \tfrac{1}{2}\left[\frac{q - R^2}{1-q} + \ln(1-q)\right] \tag{10}$$

(after eliminating the auxiliary variables $\hat{q}$ and $\hat{R}$) for continuous weights. In either case, we must find the saddle point of $G = G_1 + G_2$ with respect to R, q, $\hat{R}$ and $\hat{q}$. The saddle point equations are the same as in the simple perceptron case up to a factor

$$\frac{df(R)}{dR} = \frac{2}{\pi}\frac{1}{\sqrt{1-R^2}} \tag{11}$$

in the equation $\partial G/\partial R = 0$ and the corresponding factor df/dq in $\partial G/\partial q = 0$.

From the saddle point order parameter values we must calculate the generalization error. The calculation is again analogous to that for the simple perceptron. There an overlap of R between student and teacher weight vectors (for the same input vector) led via simple geometry to $\epsilon_g = \pi^{-1}\cos^{-1} R$. Here we have identical weights for the hidden–output layer but an effective overlap of hidden–layer internal representations equal to $(2/\pi)\sin^{-1} R$, so we obtain

$$\epsilon_g = \pi^{-1}\cos^{-1}[\frac{2}{\pi}\sin^{-1} R], \tag{12}$$

i.e. once again the replacement of R in the expression for the simple perceptron by $(2/\pi)\sin^{-1} R$.

For a network with continuous weights, the solution of the saddle point equations yields an algebraically decreasing generalization error. There is no phase transition at any value of α or T. For $T = 0$ there are easy relations between the overlaps and their conjugate parameters, $R = q$ and $\hat{R} = \hat{q}$, respectively. This simplifies the saddle point equations and the asymptotic form of the generalization error in powers of $1/\alpha$ can be easily obtained as

$$\epsilon_g = 2\sqrt{2}\left(\int Dt \frac{e^{-t^2/2}}{H(t)}\right)^{-1}\frac{1}{\alpha} + \mathcal{O}\left(\frac{1}{\alpha^2}\right) = \frac{1.25}{\alpha} + \mathcal{O}\left(\frac{1}{\alpha^2}\right). \tag{13}$$

This is just twice the ϵ_g found for the simple perceptron in this limit.

Contrary to the previous case, the committee machine with binary weights exhibits a phase transition at all temperatures from poor to perfect generalization. At zero temperature, the situation is again simplified by the relations $R = q$ and $\hat{R} = \hat{q}$. The corresponding generalization error as function of α is shown in figure 1. The saddle point equations have

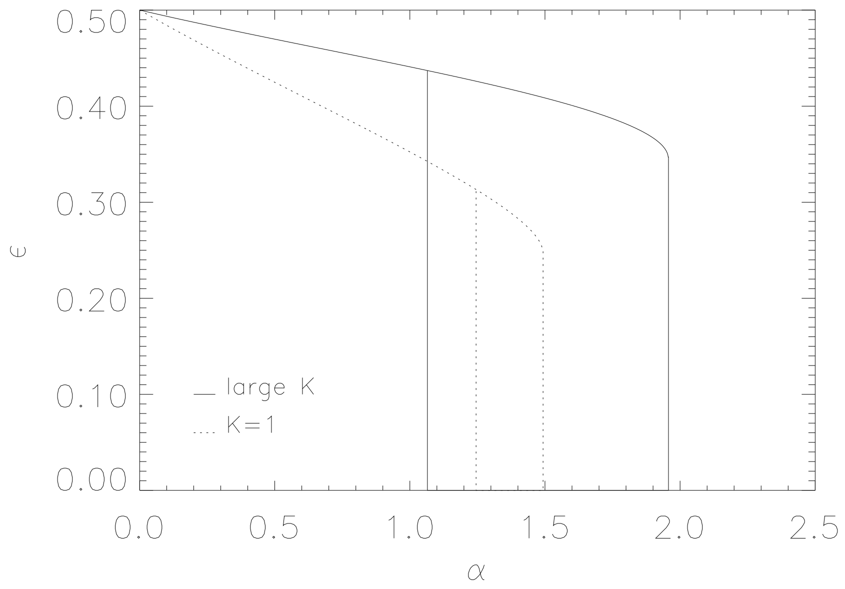

Figure 1: Zero temperature learning curves of the tree committee machine for large–K (solid lines) and $K = 1$ (dotted lines).

a solution with $R = 1$ and zero free energy for all values of α. If the number of training examples is small, however, a second solution with $R < 1$ has a lower free energy and therefore corresponds to the equilibrium state. This is the region of poor generalization. At a critical value $\alpha_c = 1.07$ of the load parameter the $R = 1$ saddle point becomes the global minimum of the free energy and the system exhibits a first order phase transition to the state of perfect generalization. In the region $1.07 < \alpha < 1.96$ the $R < 1$ solution remains metastable; it disappears at the spinodal point $\alpha = 1.96$.

At finite temperature we find the same qualitative picture, and the complete replica–symmetric (RS) phase diagram is shown in figure 2. The solid line corresponds to the phase transition to perfect generalization, and in the region between the solid and the dashed lines the $R < 1$ state of poor generalization is metastable.

We compare these results to the corresponding phase diagram for the simple perceptron as given in [11]. The large–K committee machine shows a qualitative behaviour similar to the simple perceptron ($K = 1$). The locations of the thermodynamic transitions and the spinodal line, however, are different. In the large–K limit at any fixed temperature the transition to perfect generalization occurs earlier. On the other hand the spinodal point, where the metastable $R < 1$ solution disappears, moves to higher values of α. This results in a wider region of metastability. Figure 1 shows furthermore that the generalization error of the $R < 1$ state in the large–K committee machine is higher than in the simple perceptron.

We can understand this difference in the following way. The architecture of the committee machine is more complicated and allows for different internal representations of each input to output mapping. Therefore it is less likely to find the correct generalization from a limited set of training examples. In thermodynamic terms this ambiguity in the internal

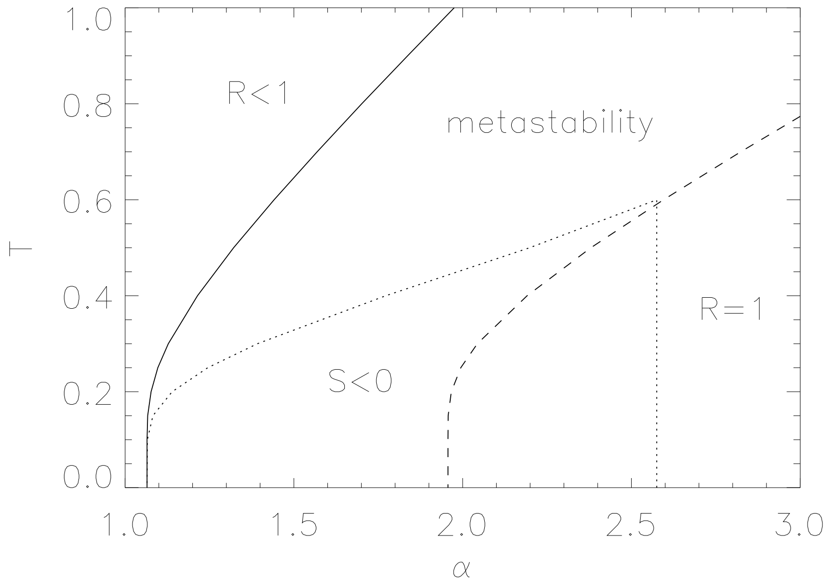

Figure 2: Replica–symmetric phase diagram of the large–K tree committee machine with binary weights. The solid line marks the thermodynamic transition from the state of poor generalization and $R < 1$ to perfect generalization and $R = 1$. In the region between the solid and the dashed lines the $R < 1$ state is metastable. The dotted line marks the line of zero entropy. Below this line the replica–symmetric solution is not correct.

representations implies a larger free energy for the large–K system at any fixed values of α and T. Therefore the free energy of the $R < 1$ state in the large–K system becomes positive earlier, explaining the shifting of the phase transition towards smaller values of α.

The case of general finite K is rather more involved, but we studied the annealed approximation for $K = 3$ which indicates a rather smooth K–dependence for $1 < K < \infty$.

In order to check the validity of the RS solution, we evaluated the thermodynamic entropy $S = \beta^2 dF/d\beta$ of the $R < 1$ state, and the dotted line in figure 2 shows the location of vanishing entropy. Since a negative entropy is unphysical for a binary system, the RS solution cannot be valid in the region below this line. One can construct a broken–replica–symmetry description of a metastable spin glass phase similar to that for the simple perceptron in [11]. This will be described elsewhere.

In summary, we studied the generalization properties of a large committee machine with non–overlapping receptive fields in the replica–symmetric approximation. The network was trained to implement a target rule given by another two–layer network of the same architecture. For a network with continuous weights we found a smoothly decreasing generalization error as function of the number of training examples. The asymptotic form of the learning curves follows a $1/\alpha$ law, which has been found in a variety of other models. A network with binary weights exhibits a first order phase transition from poor to perfect generalization, reflecting the fact that the target rule is realizable by the network. The thermodynamic transition is followed by a broad region of metastability with RSB at low temperatures.

It would be very desirable to extend this analysis to fully connected architectures, which have previously been studied in the context of zero temperature learning from random examples [6, 7, 12] and in the high–T limit implementing a linearly separable problem [5]. Giving up the symmetry between the student network and the target rule would be a further extension, allowing for unrealizable rules, which have been of recent interest in simple perceptron learning [2, 13].

We would like to thank S. Solla, K. Bauer, A. Krogh, and G. Pöppel for helpful 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] H. Sompolinsky, N. Tishby, Europhys.Lett. **13**, 567 (1990).

[2] H.S. Seung, H. Sompolinsky, and N. Tishby, submitted to Phys.Rev. A.

[3] A. Krogh, J. Hertz, *Advances in Neural Information Processing Systems,* Morgan Kaufmann, to be published 1992

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

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

[6] E. Barkai, D. Hansel, and H. Sompolinsky, preprint 1991

[7] A. Engel, H.M. Köhler, F. Tschepke, H. Vollmayr, and A. Zippelius, Phys. Rev. A, to be published

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

[9] E. Gardner, B. Derrida, J.Phys. A **21**, 271 (1988).

[10] G. Györgyi and N. Tishby in *Neural Networks and Spin Glasses*, edited by K. Thuemann and R. Köberle (World Scientific, Singapore, 1990).

[11] H. Sompolinsky, N. Tishby, and H.S. Seung, Phys.Rev.Lett. **65**, 1683 (1990).

[12] M. Griniasty, T. Grossman, preprint 1991

[13] T. Watkin, A. Rau, Phys. Rev. A, to be published