EUROPHYSICS LETTERS 1 November 1992

Europhys. Lett., **20** (5), pp. 471-476 (1992)

Memorization without Generalization in a Multilayered Neural Network.

D. HANSEL, G. MATO and C. MEUNIER

Centre de Physique Theorique, Ecole Polytechnique - 91128 Palaiseau Cedex, France

(received 27 April 1992; accepted in final form 20 August 1992)

PACS. 87.10 – General, theoretical, and mathematical biophysics (inc. logic of biosystems, quantum biology and relevant aspects of thermodynamics, information theory, cybernetics and bionics).

PACS. 05.90 – Other topics in statistical physics and thermodynamics.

Abstract. – The supervised learning of a rule that can be realized by a multilayer network (the teacher) functioning as a parity machine with $K = 2$ hidden units and nonoverlapping receptive fields is studied. The student network is supposed to have the same architecture as the teacher. Application of statistical mechanics shows that when the number of examples is smaller than a critical value P_* the trained network is unable to generalize the rule from the examples. Numerical simulations exhibiting this phenomenon are discussed.

Statistical mechanics has been shown to be a natural and powerful tool for studying generalization abilities of large neural networks. A general framework has been set recently by [1-4]. It has been applied to analyse and to classify the properties of toy models for supervised learning of one-layer perceptrons [4-6].

Less is known about the statistical mechanics of supervised learning in *multilayer* feedforward networks. It has been shown that for smooth multilayer feedforward networks the generalization error approaches asymptotically its minimal value as the inverse of the number of examples [4]. The same form for the asymptotic generalization error has been found in the learning of a nonoverlapping Committee machine [7]. However, for nonsmooth networks no general result is known and the shape of the generalization curve may depend on the details of the problem. Moreover, properties of learning in multilayer networks appear to be much richer due to possible symmetries of the student architecture and/or of the task to be learnt. Spontaneous breaking of these symmetries when the size of the training set is enlarged may therefore lead to various kinds of phase transitions (see, for instance, in [8]).

In this work we study a simple multilayered network, namely a parity machine. We concentrate mainly on an architecture with $K = 2$ hidden units trained to implement a realizable target rule extracted from a teacher with the same architecture. Our main result is the finding of a phase in the training process where the student network is memorizing the examples correctly without being able to generalize to new examples.

The nonoverlapping parity machine architecture consists of N binary input units, one hidden layer with K binary units and a single binary output unit. The input units are divided into K disjoint sets, consisting of N/K units. The k-th hidden unit is connected to the i-th

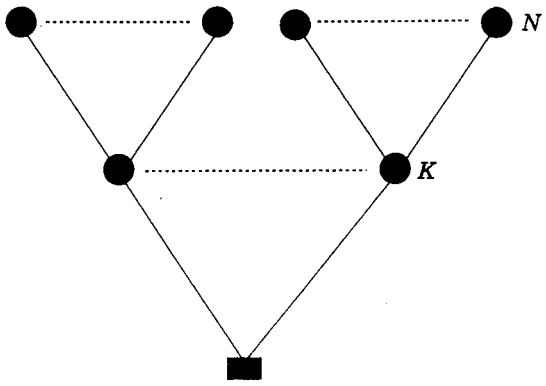

Fig. 1. – The nonoverlapping parity machine with N input units and K hidden units.

input via the weight J_i only for i such that $N(k-1)/K < i \leqslant Nk/K$ ($k = 1, \ldots, K$). Therefore, each input unit is connected to only one hidden unit, *i.e.* the receptive fields of the hidden units are nonoverlapping (see fig. 1).

The input state is denoted by S_i, $i = 1, \ldots, N$ with $S_i = \pm 1$. The state of the k-th hidden unit is equal to the sign of its induced local field

$$\sigma_k = \operatorname{sign}\left[\sum_{i=N(k-1)/K+1}^{Nk/K} J_i S_i\right] \equiv \boldsymbol{J}\cdot\boldsymbol{S}\,. \tag{1}$$

The output unit, o, of the network is just the sign of the product of the K hidden units

$$o = \prod_{l=1}^{K} \sigma_l\,. \tag{2}$$

In other words the output of the network is the parity of the internal representation of the hidden units for a given set of weights and a state of the input layer.

The target function to be learnt by the student network is denoted by $\pi_0(\boldsymbol{S})$ and corresponds to the classification of input performed by a teacher network having a nonoverlapping parity machine architecture with weights $\boldsymbol{J}^0$. The goal of the training is to adjust the weights J_i defining the student so that its output $\pi(\boldsymbol{J}, \boldsymbol{S})$ best approximates the target rule $\pi_0(\boldsymbol{S})$.

The inputs $\boldsymbol{S}$ are chosen at random from the entire input space according to some *a priori* measure $\mathrm{d}\mu(\boldsymbol{S})$ and the performance of the trained network on an input $\boldsymbol{S}$ is quantified by an error function $\varepsilon(\boldsymbol{J}, \boldsymbol{S})$ which measures the deviation of the network output from the target rule. The supervised training is performed by presenting the student with a set of P examples $(\boldsymbol{S}^l, \pi_0(\boldsymbol{S}^l))$, $l = 1, \ldots, P$ and the goal is to construct a student network that performs well on *all* the inputs, *i.e.* that minimizes the generalization error:

$$\varepsilon_g(\boldsymbol{J}) = \int \mathrm{d}\mu(\boldsymbol{S})\, \varepsilon(\boldsymbol{J}; \boldsymbol{S})\,. \tag{3}$$

The examples are used to construct a training energy:

$$E(\boldsymbol{J}) = \sum_{l=1}^{P} \varepsilon(\boldsymbol{J}; \boldsymbol{S}^l)\,. \tag{4}$$

In the statistical-mechanics framework one supposes a Gibbs distribution of networks: $P(\boldsymbol{J}) = (1/Z)\exp[-\beta E(\boldsymbol{J})]$ [4, 5, 9]. Such a distribution has been shown to occur under

general assumptions on the learning procedure [4, 9]. The normalization factor Z is the partition function

$$Z = \int \mathrm{d}\mu(\boldsymbol{J}) \exp\left[-\beta E(\boldsymbol{J})\right], \tag{5}$$

where $\mathrm{d}\mu(\boldsymbol{J})$ is an *a priori* measure on the network space and $T = 1/\beta$ is a temperature.

As shown in [4] the proper thermodynamic limit is achieved when P and N go to infinity with $\alpha = P/N$ constant. At equilibrium, the average training and generalization error are, respectively,

$$\varepsilon_t(T, \alpha) = \frac{1}{N\alpha}[\langle E(\boldsymbol{J})\rangle] \tag{6}$$

and

$$\varepsilon_g = [\langle \varepsilon(\boldsymbol{J})\rangle] \tag{7}$$

where $[...]$ denotes averaging over the training set and $\langle ...\rangle$ denotes the thermal average over the Gibbs distribution. The important thermodynamic quantities are the free energy $F = -T[\ln Z]$ and the entropy $S = -\int \mathrm{d}\mu(\boldsymbol{J})[P(\boldsymbol{J}) \ln P(\boldsymbol{J})]$.

In the following the error function is chosen as

$$\varepsilon(\boldsymbol{J}, \boldsymbol{S}) = \theta(-\pi_0(\boldsymbol{S})\pi(\boldsymbol{J}, \boldsymbol{S})). \tag{8}$$

It is zero if the input $\boldsymbol{S}$ is correctly classified by the student network and one otherwise. Therefore the training error is simply the number of the examples in the training set that are misclassified by the student and the training energy is minimized by networks that are classifying all the examples correctly.

The following analysis will be limited to $K = 2$ hidden units. For a given training set the network space displays a symmetry property. Indeed, it is clear that two networks $\boldsymbol{J}$ and $\boldsymbol{J}'$ related by the reversal symmetry $\boldsymbol{J}' = -\boldsymbol{J}'$ have exactly the same training energy. This is a consequence of the architecture of the student network working as a parity machine and this is true for any task that such a network is trained with. We will see that this simple reversal symmetry has a strong consequence on the learning curve of the student: if the size of the training set is too small the student network is memorizing the examples of the training set without being able to extract any regularity or rule from these examples.

Two different situations are studied in the following: 1) both the teacher and the student have continuous and normalized weights $\left(i.e.\ \ \mathrm{d}\mu(\boldsymbol{J}) = \mathrm{d}\boldsymbol{J}\,\delta\left(\sum_{i=1}^{N/2} J_i^2 - 1\right)\delta\left(\sum_{i=N/2+1}^{N} J_i^2 - 1\right)\right)$, 2) both the teacher and the student have binary ± 1 weights.

First we present our main results for continuous networks at zero temperature.

Computing the free energy by using the replica trick and a replica symmetric ansatz (RS), one finds that it depends on two types of order parameters, namely the overlaps of the student with the teacher

$$R_k = \left[\left\langle \sum_{i=N(k-1)/2+1}^{Nk/2} J_i J_i^0 \right\rangle\right] \tag{9}$$

and the Edwards-Anderson order parameter

$$q_k = \left[\left\langle \sum_{i=N(k-1)/2+1}^{Nk/2} J_i \right\rangle^2\right] \tag{10}$$

($k = 1, 2$). These order parameters are determined by stationarity conditions (saddle-point equations) of the free energy. In particular one finds that they do not depend on the index k: $R_k = R$, $q_k = q$, for all k's. This is a consequence of the nonoverlapping receptive fields of the

network. Once the saddle-point equations are solved, one can calculate the generalization error, which in the large-N limit is a function of R alone, namely

$$\varepsilon_g = \frac{2}{\pi}\arccos R - \frac{2}{\pi^2}(\arccos R)^2 . \tag{11}$$

The saddle-point equations possess one paramagnetic solution with $R = q = 0$. This solution *exists for all* α's and corresponds to a generalization error $\varepsilon_g = 0.5$, *i.e.* the student network is memorizing the training set but is unable to generalize. However for $\alpha > \alpha_* = \pi^2/8$ another solution appears. For this solution, R starts from zero continuously at α_* and increases monotonously reaching 1 asymptotically for $\alpha \to \infty$ as: $1 - R \propto (1/\alpha)^2$. This corresponds to an asymptotic generalization error: $\varepsilon_g \propto 1/\alpha$. Note that this is an example of a nonsmooth multilayered network displaying a $1/\alpha$ asymptotics for ε_g.

The existence of the «rote memorization» phase at values of $\alpha < \alpha_*$ is a consequence of the inversion symmetry of the network space. For $\alpha < \alpha_*$ the two networks, with weights $\boldsymbol{J}$ and $\boldsymbol{J}' = -\boldsymbol{J}$ have the same training energy and are in the same ergodic component. As a consequence the thermal averages of R and q are zero. At α_* the symmetry is broken and $\boldsymbol{J}$ and $\boldsymbol{J}'$ no longer belong to the same ergodic component. This transition gives rise to a nonzero value for R and q. Calculating the Hessian matrix at the saddle point shows that the replica symmetric solution with $q = R \neq 0$ is locally stable. On the other hand, the paramagnetic solution with $q = R = 0$ loses stability at α_*.

At that point it is useful to present results from numerical simulations. We have simulated networks of different sizes: $N = 100$, 200, 400 and 800. The training was performed with the Least Action Algorithm (LAA). A description of this learning strategy can be found in [10, 11]. The training set was chosen at random and the training procedure was initialized from a randomly chosen initial condition. The number of samples of training sets (N_{samp}) was taken between 500 (large N or α) and 10000 (small N or α).

We have measured the overlap R of each of the N_{samp} trained networks (with 0 training error) with the teacher and we have plotted the histogram of this quantity over the N_{samp} realizations. Two regimes were found depending on the size of the training set. At small value of α ($\alpha < 3.0$) the histogram has a Gaussian shape well peaked around zero, and the variance of the histogram, $\sigma(N, \alpha)$, scales like $\sigma^2(N, \alpha) \sim 1/N$. (For instance for $\alpha = 0.1$ one finds: $\sigma^2(N, \alpha) = 1.08(\pm 0.01)(1/N) - 5(\pm 7)\cdot 10^{-5}$.) This shows that in the thermodynamic limit the overlap with the teacher is 0. For a given size, increasing α makes the Gaussian histogram broader. This is followed eventually by the splitting of the distribution into two well-separated peaks. These two peaks are located around two symmetrical values $\pm R_{\text{peak}}(\alpha)$ and we have checked that their width decreases with N roughly like $1/\sqrt{N}$.

This behaviour, displaying a phase where the trained network is unable to generalize, is similar to the prediction of the RS solution described above. However the minimal value of α for which the trained network generalizes is significantly larger than the prediction of this solution (by a factor of 2). This is not totally surprising since there is no guarantee that the LAA is constructing networks according to the Gibbs measure. Nevertheless, it is remarkable that the values of R_{peak} are very close to the statistical-mechanics prediction above $\alpha \approx 4$ as can be seen from fig. 2.

These facts suggest that for $1.23 < \alpha < 4$, apart from the set of networks described by the RS solution, another set exists and that starting the LAA from random initial conditions leads to networks belonging to this set.

In order to interpret this behaviour we have looked for solutions of the saddle-point equations that break the replica symmetry [12]. We have studied in details the one-step replica-symmetry-breaking ansatz. We found that: 1) such a solution starts to exist at $\alpha = \alpha_*$ coexisting with the RS solution; 2) the only solution we have found at that level of RSB, for

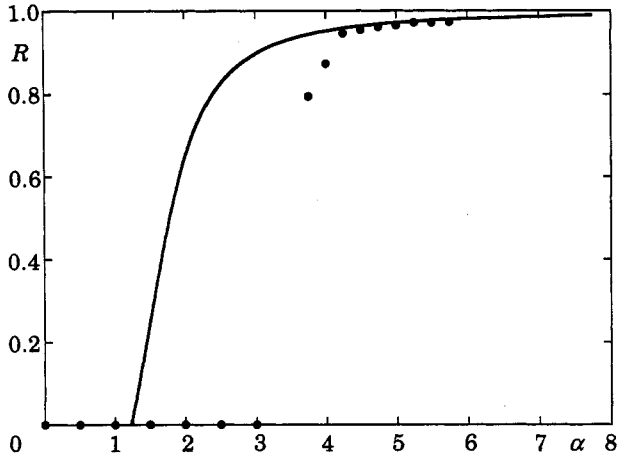

Fig. 2. – The overlap R *vs.* α for the RS symmetric solution (full line), and for the numerical simulation (dots).

$\alpha > \alpha_*$, has a vanishing overlap R; 3) as a consequence the student is behaving like if it were learning a random mapping and thus this solution does not exist for $\alpha > \alpha_c^{(1)}$, where $\alpha_c^{(1)}$ is the one-step RSB estimate of the capacity for the nonoverlapping $K = 2$-parity machine [13]; 4) in the whole domain of existence the volume of this one step RSB solution is lower than the volume of the RS solution. As in the case of the random mapping problem the replica symmetry has to be broken beyond the first step. At all orders of the RSB one can find a solution that satisfies (1) to (3), eventually leading to the stable solution of the random mapping problem. We conjecture that this solution is also locally stable when it is considered as a solution of the saddle-point equations for the generalization problem. As a consequence, the RSB phase should extend up to the capacity of the $K = 2$-parity machine, *i.e.* up to $\alpha = \alpha_c \simeq 3.8$ [13] and in this phase the student network is unable to generalize. This phase is not the thermodynamically stable phase: the branch of the RS solution for which the student generalizes is more stable.

This conjecture is compatible with the following interpretation of the simulations: for $\alpha < 3.2$ and for random initial conditions the LAA is always leading to a network that belongs to this part of the solution space even if it has lower volume than the RS solution. This effect is essentially a consequence of a *metastability*. The space of RSB solutions is given by a set of disjoint subspaces and the typical overlap between two of these subspaces is 0. Starting the algorithm with random initial conditions (that are quasi-orthogonal to the teacher) it converges to one of these subspaces rather than reaching the RS solutions and the statistical properties of the networks it constructs are exactly the same as for the learning of a random mapping. Actually, for values of α close to 3 we have found a strong slowing-down of the training. The number of epochs needed to learn the training set perfectly diverges like $1/(\alpha_0 - \alpha)^\gamma$, where $\alpha_0 = 3.1 \pm 0.1$ and $\gamma = 2.2 \pm 0.2$. Note that these values correspond to the fit found in the learning of a random mapping [13]. For $\alpha > 3.2$ the algorithm is not able to reach these solutions, even when they exist (the effective capacity of the LAA is close to 3.2), and for $\alpha > 4$ these solutions disappear. In that regime the algorithm only finds the part of the space corresponding to the RS solution.

When the teacher and the student have binary weights ± 1, the phase where there is no generalization still exists. For this system we considered the high-temperature approxi-

mation. In this way one can evaluate the average free energy without introducing replicas [14].

This free energy has always a local minimum at $R = \pm 1$, with value $f(\pm 1) = 0$. There is another solution of the saddle-point equation $\partial f/\partial R = 0$. This equation can be solved to obtain R as a function of $\tilde{\alpha} = \alpha/T$, and then ε_g. For $\tilde{\alpha} < 2.33$ the system does not generalize at all, and after this value is reached the generalization energy falls to zero. The free energy of this solution before the jump is given by $\beta f^0 = \tilde{\alpha}/2 - \ln 2$. Therefore, for $\tilde{\alpha} \geqslant 2 \ln 2 \approx 1.38$ it has a greater free energy than the perfect generalization solution, and is only metastable. Monte Carlo simulations were performed to verify these results. In particular metastability was checked by choosing the teacher configuration as the initial condition for $\tilde{\alpha} > 1.38$.

The remarkable fact that there is a critical value of α below which the student learns the examples perfectly but does not generalize at all is a consequence of the particular symmetry of the parity machine. We have also found such «rote memorization» phase for $K > 2$-parity machine. Such a «delay» to generalization has also been recently found in other architectures displaying similar symmetries [15]. An interesting question is whether it is possible to find a training strategy that would allow us to extend the generalization phase to values of α that are smaller than α_*. A natural idea would be to use a querying procedure to build the training set [16, 17]. This question is under current investigation.

* * *

Many useful discussions with E. BARKAI, Dr. S. SEUNG and Prof. H. SOMPOLINSKY are acknowledged. GM acknowledges the hospitality of the Centre de Physique Theorique in the Ecole Polytechnique.

REFERENCES

[1] GARDNER E., *J. Phys. A*, **21**(1988) 257.
[2] GARDNER E. and DERRIDA B., *J. Phys. A*, **21** (1988) 271.
[3] GARDNER E. and DERRIDA B., *J. Phys. A*, **22** (1989) 1983.
[4] SEUNG H. S., SOMPOLINSKY H. and TISHBY N., *Phys. Rev. A*, **45** (1992) 6056.
[5] HANSEL D. and SOMPOLINSKY H., *Europhys. Lett.*, **11** (1990) 687.
[6] GYORGYI G. and TISHBY N., *Statistical theory of learning a rule*, in *Neural Networks and Spin Glasses*, edited by W. K. THEUMANN and R. KOBERLE (World Scientific, Singapore) 1990, p. 3.
[7] MATO G. and PARGA N., *Generalizing Properties of Multilayered Neural Networks*, to appear in *J. Phys. A* (1992).
[8] SCHWARZE H., OPPER M. and KINZEL W., *Generalization in a Two-Layer Neural Network*, preprint (1991).
[9] TISHBY N., LEVIN E. and SOLLA S., *Proceedings of the IEEE - Special Issue on Neural Networks*, Vol. 2 (1989) 403.
[10] MITCHISON G. J. and DURBIN R. M., *Biol. Cybern.*, **60** (1989) 345.
[11] BARKAI E., HANSEL D. and SOMPOLINSKY H., *Phys. Rev. A*, **45** (1992) 4146.
[12] MEZARD M., PARISI G. and VIRASORO M., *Spin Glass Theory and Beyond* (World Scientific, Singapore) 1987.
[13] BARKAI E., HANSEL D. and KANTER I., *Phys. Rev. Lett.*, **65** (1990) 2312.
[14] SOMPOLINSKY H., TISHBY N. and SEUNG H. S., *Phys. Rev. Lett.*, **65** (1990) 1683.
[15] BARKAI E., unpublished.
[16] BAUM E., *IEEE Trans. in Neural Network* (1990).
[17] OPPER M. A., SEUNG H. S. and SOMPOLINSLY H., Query by Committee, Preprint Racah Institute and Center for Neural Computation (1992).