Generalization properties of multilayered neural networks

1992 J. Phys. A: Math. Gen. 25 5047

(http://iopscience.iop.org/0305-4470/25/19/017)

J. Phys. A: Math. Gen. 25 (1992) 5047–5054. Printed in the UK

Generalization properties of multilayered neural networks

G Mato and N Parga

Centro Atómico Bariloche† and Instituto Balseiro‡, 8400 Bariloche, Argentina

Received 11 September 1991, in final form 23 April 1992

Abstract. Generalization properties of multilayered neural networks with binary couplings are studied in the high-temperature limit. The transition to the perfect generalization phase is evaluated for systems with an arbitrary number of layers. It is found that the thermodynamic transition occurs for a number of examples lower than for the perceptron, but the opposite occurs for the transition in which the poor generalization solution disappears. The generalization error is also decomposed according to the contributions coming from different numbers of hidden neurons that have a wrong sign in the internal field. This allows us to describe the generalization behaviour of the hidden neurons.

1. Introduction

Neural networks have been extensively studied using the methods of statistical mechanics [1]. Among the structures that have been considered are the multilayered networks. These are systems that implement a given mapping between the inputs and the outputs. The most simple of them is the perceptron, which is built with one input layer and one output unit; in more general architectures, one or more layers of hidden neurons are placed between them. Several properties of the perceptron are well understood. In particular, the maximum number of mappings that it can store has been evaluated [2–3].

Another interesting problem that has been addressed for the perceptron is the generalization ability, i.e. the capacity to predict a correct input–output relation from a set of examples [4, 5]. This problem can be considered in two ways. In the first the network is trained with a set of mappings (the training set) in which the output is chosen independently from the input. As there is a maximum number of such mappings that can be stored, the region beyond the critical capacity can be made accessible, only allowing errors in the input–output relation. This can be done by introducing a temperature parameter. Once these mappings are learnt (or the training error is the minimum possible at that temperature) the generalization error is evaluated over the rest of the possible mappings. Its behaviour with a number of examples gives rise to the generalization curve. It has been found that the introduction of a finite temperature can lower the value of the generalization error [6]. The training error is not zero, but the exploration of a greater portion of coupling space allows a better generalization performance to be found.

† Comisión Nacional de Energía Atómica.

‡ Comisión Nacional de Energía Atómica and Universidad Nacional de Cuyo.

In the second approach the mappings not obtained by choosing independently the inputs and the outputs σ, but these are given functions of the inputs s_i $(i=1,\ldots,N)$ and of a set of fixed couplings $\boldsymbol{J}^0$. Although the number of inputs in the training set $\{s\}_\mathrm{T}$ is arbitrarily large, the error can be always brought to zero at zero temperature because there is always a set of fixed couplings, the teacher $\boldsymbol{J}^0$, that implements the correct mapping. But if there are other networks yielding the correct mapping on the training set, inputs outside it can have a wrong output, giving rise to a non-zero generalization error. As the number of examples is increased the number of couplings compatible with them decreases, and the same happens with the generalization error. A first-order transition occurs, even at high temperature, for binary couplings [7].

The purpose of this work is to study the behaviour of the generalization error for the second of these approaches and for the architecture shown in figure 1. The couplings J_{ij} from neuron j in the input to neuron i in the first hidden layer are Ising variables $(J_{ij}=\pm 1)$. The others can be chosen equal to 1 because any coupling equal to -1 can be transformed in $+1$ through a gauge transformation.

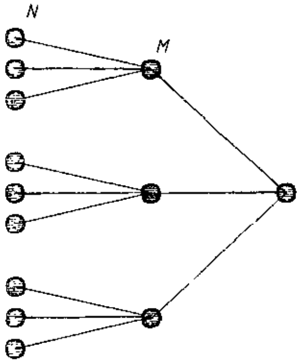

Figure 1. Non-overlapping architecture considered for the special case of one hidden layer.

Double-layered neural networks have also been considered from the point of view of maximal storage capacity [8-10]. The reader is referred to [8-10] for a discussion on the choice of the non-overlapping receptive field architecture we have adopted.

In the next section we study the behaviour of a network with only one hidden layer using the high-temperature and annealed approximations. Additional hidden layers are considered in section 3. Our conclusions are presented in section 4.

2. Networks with one hidden layer

The training energy for a set of couplings $\boldsymbol{J}$ is given by

$$E_\mathrm{T}(\boldsymbol{J},\boldsymbol{J}^0,\{s\}_\mathrm{T})=\sum_{\nu=1}^{p}\varepsilon(\boldsymbol{J},\boldsymbol{s}^\nu) \tag{1}$$

where $\{s\}_\mathrm{T}$ is the training set of size p, $\varepsilon(\boldsymbol{J},\boldsymbol{s}^\nu)=\theta(-\sigma_0^\nu\sigma^\nu)$, σ^ν $(\nu=1,\ldots,p=\alpha N)$ is the output of the network with couplings $\boldsymbol{J}$, σ_0^ν the output with couplings $\boldsymbol{J}^0$ and θ is the step function. The partition function is

$$Z=\sum_{J_{ij}=\pm 1}\exp(-\beta E_\mathrm{T}(\boldsymbol{J}^0,\boldsymbol{J},\{s\}_\mathrm{T})). \tag{2}$$

One should average $\ln Z$ over the input patterns using the replica method. Since for the perceptron the relevant properties are present even at small β [7] we will use a high-temperature approximation and the annealed approximation. In this case the average logarithm of the partition function can be written as

$$\langle \ln Z\rangle = \ln \sum_{J_{ij}=\pm 1} \exp(-\beta E(\boldsymbol{J}, \boldsymbol{J}^0)) \tag{3}$$

where the brackets $\langle \ldots \rangle$ denote the average over the input patterns $\{s^\nu\}$ with the distribution $P(s) = (\delta(s-1)+\delta(s+1))/2$ and $E(\boldsymbol{J}, \boldsymbol{J}^0)$ is the average training error:

$$E(\boldsymbol{J}, \boldsymbol{J}^0) = \int \mathrm{d}\boldsymbol{s}\, P(\boldsymbol{s}) E_{\mathrm{T}}(\boldsymbol{J}, \boldsymbol{J}^0, \{\boldsymbol{s}\}). \tag{4}$$

In this approximation the training and the generalization error per pattern are the same.

The average training error is a function of the overlaps between the couplings $\boldsymbol{J}$ and $\boldsymbol{J}^0$, $E(\boldsymbol{J}, \boldsymbol{J}^0) = E(m_i)$, where

$$m_i = \frac{1}{N/M} \sum_{j=1+(i-1)N/M}^{iN/M} J_{ij} J^0_{ij}. \tag{5}$$

From (3) the average free energy f can be written as

$$-\beta f = \int \prod_i \mathrm{d}m_i \exp(-\beta p \varepsilon(m_i) + S(m_i)) \tag{6}$$

where $\varepsilon(m_i) = E(m_i)/p$ is the average error per pattern and $S(m_i)$ is the entropy associated with the overlaps m_i:

$$S(m_i) = (N/M) \sum_{i=1}^{M} [-(1-m_i)/2 \ln(1-m_i) - (1+m_i)/2 \ln(1+m_i) + \ln 2]. \tag{7}$$

The integrals over m_i can be done through the saddle point method because both $p\varepsilon$ and S are extensive quantities.

First we consider a system with only one hidden layer with a finite number, M, of units. The average training energy per pattern can be written as

$$\varepsilon = \int_{-\infty}^{\infty} \mathrm{d}h_1\, \mathrm{d}h_2\, \theta(-h_1 h_2) \left\langle \delta\left(h_1 - M^{-1/2} \sum_i \eta_i^\nu\right) \delta\left(h_2 - M^{-1/2} \sum_i \eta_i^{0\nu}\right)\right\rangle \tag{8}$$

where η_i^ν and $\eta_i^{0\nu}$ are the hidden neuron representations of patterns $\{s^\nu\}$ for the current and the teacher networks, respectively. Defining the corresponding local field at these neurons as C_i^ν and $C_i^{0\nu}$, (8) can be written as

$$\begin{aligned}\varepsilon = \int_{-\infty}^{\infty} & \mathrm{d}h_1\, \mathrm{d}h_2\, \theta(-h_1 h_2) \prod_{i,\nu} \mathrm{d}C_i^\nu\, \mathrm{d}C_i^{0\nu} \\ & \times \delta\left(h_1 - M^{-1/2} \sum_i \mathrm{sgn}(C_i^\nu)\right) \delta\left(h_2 - M^{-1/2} \sum_i \mathrm{sgn}(C_i^{0\nu})\right) \\ & \times \left\langle \delta\left(C_i^\nu - (N/M)^{-1/2} \sum_j J_{ij} s_j\right) \delta\left(C_i^{0\nu} - (N/M)^{-1/2} \sum_j J^0_{ij} s_j\right)\right\rangle. \end{aligned} \tag{9}$$

Using the integral representation of the δ-function and averaging over the $\{s^\nu\}$ we have

$$\varepsilon(m_i)=\frac{1}{2\pi^2}\int_0^\infty \mathrm{d}h_1\,\mathrm{d}h_2\int_{-\infty}^{\infty}\mathrm{d}f_1\,\mathrm{d}f_2\exp[\mathrm{i}(f_1h_1+f_2h_2)]$$

$$\times\prod_i^M\{[1-\cos^{-1}(m_i)/\pi]\cos[(f_1-f_2)/M^{1/2}]$$

$$+\cos^{-1}(m_i)/\pi\cos[(f_1+f_2)/M^{1/2}]\}\tag{10}$$

which depends only on the overlaps m_i.

The contribution proportional to $\cos^{-1}(m_i)/\pi$ comes from the region in which $C_i^\nu C_i^{0\nu}<0$, i.e. hidden neuron i gives a non-zero contribution to the generalization error. The opposite happens for the term $1-\cos^{-1}(m_i)/\pi$. By expanding the product, the generalization error can be decomposed according to the number of hidden neurons which have not yet learnt the rule.

Assuming that in equilibrium all the overlaps m_i take the same value m, we have

$$\varepsilon(m)=\frac{1}{2}-\frac{1}{2\pi^2}\sum_{n=0,M-1,2}\binom{M}{n}\left(\frac{\Gamma[(M-n)/2]\Gamma[(n+1)/2]}{\Gamma[(M+1)/2]}\right)^2 a^{M-n}\tag{11}$$

where $a=1-2\cos^{-1}(m)/\pi$.

The overlap is given by the saddlepoint equation

$$m=\tanh(-\tilde{\alpha}\partial\varepsilon/\partial m)\tag{12}$$

where $\tilde{\alpha}=\alpha\beta$. There is always a local minimum of the free energy with $m=1$ and $\varepsilon=0$ that corresponds to perfect generalization.

Evaluating the solutions of (12) and inserting the values of the overlaps in (10) we obtain the training and generalization error for several values of M (see figure 2). We see that beyond $\tilde{\alpha}=\tilde{\alpha}_c$ the solution with $m<1$ disappears. This value increases with M. For low $\tilde{\alpha}$ the solution with $m<1$ has a lower free energy that the one with $m=1$. When $\tilde{\alpha}=\tilde{\alpha}_T$ both free energies are equal, showing the occurrence of a phase transition.

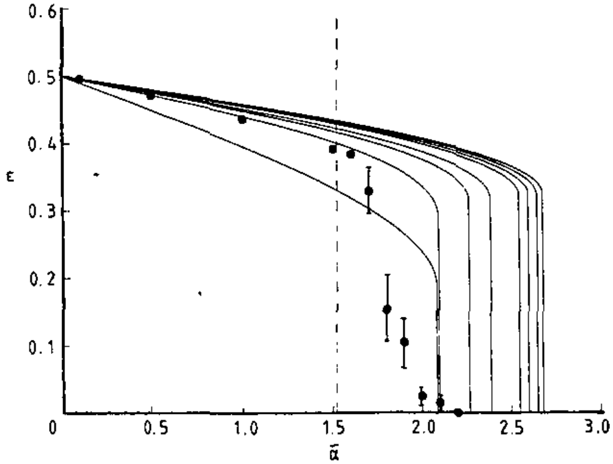

Figure 2. Full lines show the value of the generalization energy for $M=1, 3, 5, 7, 15, 25, 55$ and the asymptotic limit $M\gg1$ (from left to right). The points plotted are the results of Monte Carlo simulations for $N=93$, $M=3$ and $\beta=0.2$. The vertical broken line shows the thermodynamic transition for $M=3$.

The value of $\tilde{\alpha}_T$ decreases with M, for instance it is $\tilde{\alpha}_T \simeq 1.52$ for $M=3$ and $\tilde{\alpha}_T \simeq 1.48$ for $M=55$.

Monte Carlo simulations were done to verify these results. In figure 2 we can see the mean value of the generalization error per pattern for $M=3$, $N=93$ and $\beta=0.2$.

The simulation shows an intermediate behaviour between the thermodynamic transition and $\tilde{\alpha}_c$. This has also been observed in other systems [11, 12].

The assumption that $m_i=m$ has been verified in the simulation, beginning from a random initial condition on the couplings (i.e. $m_i \simeq 0$). The study of the free energy as a function of m_i using (7) and (9) reveals that there are local minima for which some of the m_i are equal to 1 while the others verify equations similar to (12). But these minima have a free energy that is higher than that of the soluton with $m_i=m$. Therefore, unless we begin with an initial condition of this type as will not be trapped in these minima.

Coming back to the equilibrium case with $m_i=m$, we can find that in the limit $M \gg 1$

$$\varepsilon(m)=\cos^{-1}[1-2\cos^{-1}(m)/\pi]/\pi. \qquad (13)$$

The value of the generalization error for this case is also displayed in figure 2. Comparing this result with the generalization error of the perceptron [7], $\varepsilon(m)=\cos^{-1}(m)/\pi$, it is interesting to see that the effect of the hidden layer is to 'renormalize' the overlap from m to $1-2\cos^{-1}(m)/\pi$.

Expanding the product over hidden neurons in (10) we can analyse the different contributions to the generalization error. As an example, let us consider the case $M=3$ (with $m_i=m$). The contribution to the error coming from one, two or three hidden neurons are

$$\varepsilon_1=\tfrac{3}{2}[1-\cos^{-1}(m)/\pi]^2\cos^{-1}(m)/\pi \qquad (14)$$

$$\varepsilon_2=\tfrac{3}{2}[1-\cos^{-1}(m)/\pi][\cos^{-1}(m)/\pi]^2 \qquad (15)$$

$$\varepsilon_3=[\cos^{-1}(m)/\pi]^3. \qquad (16)$$

It can be seen that the term in the expansion coming from perfect generalization in all the hidden neurons integrates to zero, as it should.

As shown in figure 3 the three contributions behave smoothly up to $\tilde{\alpha}=\tilde{\alpha}_c$, where all of them drop simultaneously to zero.

If perfect generalization was achieved by first eliminating errors from a single hidden neuron, ε_3 should drop for a value of $\tilde{\alpha}$ smaller than $\tilde{\alpha}_c$. This does not happen. This is probably a consequence of the fact that we are measuring generalization properties by looking only at the output neutron. Bad generalization in only two neurons, measured by ε_2, also has a systematic decay. On the other hand, as $\tilde{\alpha}$ increases ε_1 also increases, until a value of $\tilde{\alpha}$ near $\tilde{\alpha}_c$ is reached, showing that most of the errors correspond to bad performance of a single neuron.

In summary, although there are no partial first-order transitions associated with *perfect generalization of one or two neurons*, generalization is achieved by first eliminating errors coming from a large number of hidden neurons.

Another useful approach to the study of these systems is the annealed approximation. In this case we replace $\langle \ln Z\rangle$ in (3) by $\ln\langle Z\rangle$. It is easily shown that the annealed free energy is given by

$$-\beta f_A=\alpha\ln[1+(e^{-\beta}-1)\varepsilon_g]+S(m) \qquad (17)$$

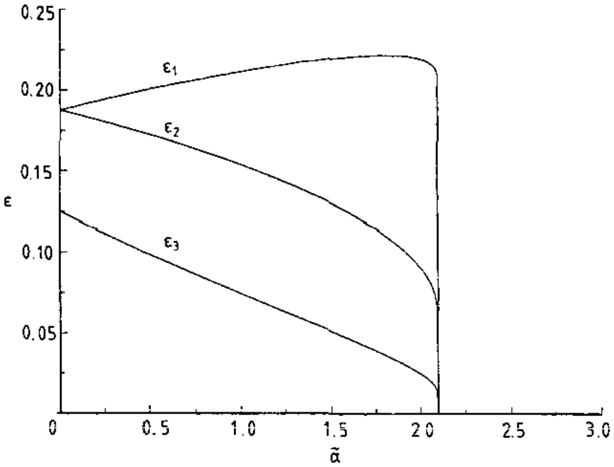

Figure 3. Different contributions to the generalization error for $M=3$ as a function of $\tilde{\alpha}$. ε_1, ε_2 and ε_3 are the contributions of one, two and three neurons with wrong signs in their internal field.

where ε_g is the generalization error. When $\beta \ll 1$ we obtain the same result as for the high-temperature limit, showing that in this limit the annealed approximation is exact.

The average training energy, given by $\varepsilon_t = -\partial(\beta f_A)/\partial\beta$, results in

$$\varepsilon_t = \frac{e^{-\beta}\varepsilon_g}{1+(e^{-\beta}-1)\varepsilon_g} \tag{18}$$

which is different from ε_g except for $\beta \to 0$. The generalization and training errors at $T=1$ for the annealed approximation are shown in figure 4 for $M=3$ hidden units.

At $\alpha > 2.75$ the only solution is the one with $m=1$ (perfect generalization) but there is a thermodynamic transition at $\alpha \simeq 2$.

The Monte Carlo simulation shows that the annealed approximation is good except in the region of perfect generalization, where it does not take into account the metastable states where the system is trapped.

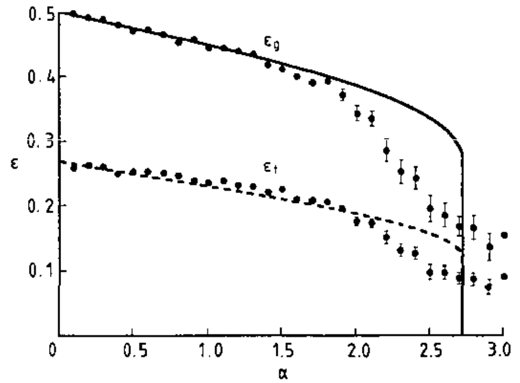

Figure 4. Full and broken lines show the value of the generalization and training energies as a function of α for $T=1$ and $M=3$. The points plotted are the results of Monte Carlo simulations for $N=75$ and 10 samples.

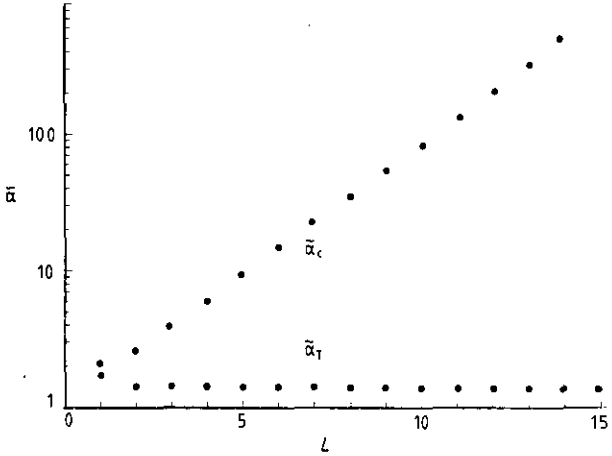

Figure 5. Critical values of $\tilde{\alpha}$ as a function of the number of layers L. $\tilde{\alpha}_c$ is where the poor generalization solution disappears and $\tilde{\alpha}_c$ is the thermodynamic transition.

3. Additional hidden layers

Using the ideas that lead to (9) one can easily find the expression for the generalization error of a system with L layers, with the restriction that the number of units in the ith layer must be much greater than the number of units in the $(i+1)$th layer.

The generalization energy of a system with L layers is a function of the correlation between the states of the output neuron when the couplings are the teacher $\boldsymbol{J}^0$ and the configuration $\boldsymbol{J}$. This correlation, denoted by c_L, can be written as a function of the correlations between the neurons in the previous layers. This procedure can be iterated until the first hidden layer:

$$c_1 \equiv \left\langle \operatorname{sgn}\left(N_{l-1}^{-1/2} \sum_j s_j^l \right) \operatorname{sgn}\left(N_{l-1}^{-1/2} \sum_j s_j^{0l} \right) \right\rangle = 1 - 2\cos^{-1}(c_{l-1})/\pi \qquad (19)$$

where N_l is the number of neurons in the lth layer and s_j^l (s_j^{0l}) is the state of the neuron j in the lth layer when the coupling connections are given by $\boldsymbol{J}$ $(\boldsymbol{J}^0)$.

When we arrive at the input layer we have that the correlations are given by

$$c_1 = 1 - 2\cos^{-1}(m)/\pi \qquad (20)$$

where m is the overlap between $\boldsymbol{J}$ and $\boldsymbol{J}^0$ defined in (5) (we assume that $m_i = m$).

In this way we obtain that the generalization error for a system with L layers is

$$\varepsilon^{(L)}(m) = \cos^{-1}(c_{L-1})/\pi \qquad (21)$$

where the correlations c_1 are given by (19) and (20).

Inserting (21) in (12) we find the overlap, and the free energy. In figure 5 we can see the behaviour of $\tilde{\alpha}_c$ and $\tilde{\alpha}_T$ as a function of the number of layers. As this number, L, tends to infinity, $\tilde{\alpha}_T$ tends to 1.38 while $\tilde{\alpha}_c$ grows exponentially: $\tilde{\alpha}_c \simeq (\pi/2)^L$.

4. Conclusions

We have studied the generalization properties of multilayered networks with binary couplings. Comparison with the perceptron reveals that as the number of layers increases, thermodynamic transition to perfect generalization has a lower value of $\tilde{\alpha}$, but the value where the poor generalization solution disappears increases. Monte Carlo simulations show an intermediate behaviour. In the region $\tilde{\alpha}_T \leqslant \tilde{\alpha} \leqslant \tilde{\alpha}_c$ the poor generalization solution is only a metastable state. Therefore, with a suitable annealing scheme it would be possible to arrive at the $m=1$ solution. In this case multilayered networks show better generalization properties than the perceptron because $\tilde{\alpha}_T$ is lower.

In this way we see that even in the high-temperature limit the system has a rich behaviour. The analysis at low temperatures requires the introduction of the replica technique for solving the full problem. However, following the steps of [13] it is easy to obtain a bound α_c at zero temperature. For the system with one hidden layer and in the limit $M \gg 1$, the result is $\alpha_c \leqslant 1.137$.

References

[1] Amit D J 1989 *Modelling Brain Function* (Cambridge: Cambridge University Press)

[2] Cover T M 1965 *IEEE Trans. Electron. Comput.* **EC-14** 326

[3] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257

[4] Denker J, Schwartz D, Wittner B, Solla S, Howard R, Jackel L and Hopfield J 1987 *Complex Systems* **1** 877

[5] Patarnello S and Carnevali P 1987 *Europhys. Lett.* **4** 503

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

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

[8] Barkai E, Hansel D and Kanter I 1990 *Phys. Rev. Lett.* **65** 2312
Barkai E and Kanter I 1991 *Europhys. Lett.* **14** 107

[9] Mato G, Moukarzel C and Parga N 1991 *Proc. Workshop on Neural Networks: From Biology to High Energy Physics (Elba, Italy, 1991)* ed O Benhar, C Bosio, P del Guidice and E Talbet (Pisa: ETS Editrice); 1992 *Phys. Rev.* A submitted

[10] Kocher I and Monasson R 1991 *Preprint LPTENS* 91/9

[11] Sompolinsky H and Tishby N 1990 *Europhys. Lett.* **13** 567

[12] Kocher I and Monasson R 1991 *Int. J. Neural Systems* **2** 115

[13] Gardner E and Derrida B 1989 *J. Phys. A: Math. Gen.* **22** 1983