Numerical Study of Back-Propagation Learning Algorithms for Multilayer Networks

1993 Europhys. Lett. 21 501

(http://iopscience.iop.org/0295-5075/21/4/020)

EUROPHYSICS LETTERS 1 February 1993

Europhys. Lett., **21** (4), pp. 501-506 (1993)

Numerical Study of Back-Propagation Learning Algorithms for Multilayer Networks.

E. EISENSTEIN and I. KANTER

Department of Physics, Bar-Ilan University - Ramat Gan 52900, Israel

(received 17 June 1992; accepted in final form 18 November 1992)

PACS. 87.10 – General theoretical and mathematical biophysics.
PACS. 02.50 – Probability theory, stochastic processes, and statistics.
PACS. 05.20 – Statistical mechanics.

Abstract. – A back-propagation learning algorithm is examined numerically for feedforward multilayer networks with one-hidden-layer functions as a parity machine or as a committee machine of the internal representation of the hidden units. It is found that the maximal known theoretical capacity is saturated and that the convergent time is not exponential with the size of the system. The results also indicate the possibility of a replica-symmetry-breaking phase with the lack of local minima.

There has recently been a growing interest in the theory and application of computing structure based on the architecture of neural networks [1]. An important class of architectures studied is the feedforward layered structure functioning as an input-output device. The prototype of this class of architectures is the one-layer perceptron consisting of N-binary-input units which are connected with N continuous weights to a one-binary-output unit [2]. The basic task of the perceptron is to embed pairs of input-output relations. Various statistical mechanical properties such as the maximal capacity, the maximal number of input-output relations which can be stored in the perceptron, have been recently investigated by using the method developed by Gardner [3].

A non-trivial extension of this method to multilayer networks (MLN) with one hidden layer was recently examined analytically for some fixed mappings between the hidden units and the output unit. The first solvable case was the parity machine (PM) where the output is the parity of the internal representation of the hidden units [4]. Recently, the AND machine [5] and the committee machine [6, 7], where the output is equal to the AND or to the majority among the hidden units, respectively, were also solved analytically in some limits. These MLN are important for applications, since networks with one hidden layer can solve nonseparable problems which cannot be implemented in the architecture of the perceptron.

The analytical method suggested by Gardner [8] is aimed at finding, for instance, the maximal capacity of the network. This method leads mostly, in the case of MLN, to a complicated replica-symmetry-breaking (RSB) phase [4, 6, 7] which can only be approximated analytically. Furthermore, proof of the existence of a solution in the phase space of the weights which fulfils the task of the network does not necessarily indicate that a solution can be obtained in a finite time. A constructive way of finding a solution for the task of the network is the learning algorithm. For the architecture of the perceptron many learning algorithms are known, and saturate the maximal theoretical capacity [8]. For the

interesting case of MLN, the efficiency of the learning algorithms is in doubt. The first class of known algorithms is based on heuristic methods such as the least-action algorithm (LAA) [9]. The achievable maximal capacity of this algorithm is less than the predicted theoretical capacity [4-7]. In the second class, the algorithm is an efficient search in the space of the internal representations of the hidden units, but the learning time may be unbounded [10]. For both classes, the learning algorithms are not associated with an energy function. Hence, any relation between the performance of the algorithms and the phase diagram predicted by theoretical methods is impossible. Moreover, if a solution in the phase space exists but is surrounded by many local minima, then the possibility of finding any algorithm which saturates the maximal capacity is in question. In such a case, the notions maximal theoretical capacity and maximal capacity that can be achieved by any learning algorithms are different. Furthermore, it is interesting to verify whether the maximal achievable capacity is related to the features of the phase which characterizes the system up to this particular capacity. Note that in all feedforward architectures which were examined, the region where a convergent learning algorithm can be found analytically or numerically is characterized by a replica-symmetric (RS) phase. This behaviour was found in the case of the perceptron [8] and in all the examined MLN [4-7]. In contrast, no learning algorithm with bounded learning time is known analytically or even numerically for the whole region with the structure of RSB phase [4-7]. These results may suggest a relation between the phase diagram or the structure of the energy landscape and the maximal capacity which can be achieved by a learning algorithm. In order to examine these questions, a careful analysis of a back-propagation (BP) learning algorithm, which is associated with an energy function, is studied in this work.

The architectures to be examined in this letter consist 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 equal disjoint sets (blocks). The l-th hidden unit is connected only to the elements of the l-th set via the continuous weights $\{J_i\}$, $N(l-1)/K < i \leqslant Nl/K$.

The configuration of the input is defined by $\{s_i\}$, $i = 1, ..., N$ with $s_i = \pm 1$. The state of the l-th hidden unit, σ_l, is defined as the sign of the induced local field

$$\sigma_l = \operatorname{sgn}(h_l)\,, \tag{1}$$

where

$$h_l = \sum J_i s_i \Big/ \sqrt{\sum J^2} \equiv \boldsymbol{J}_l \cdot \boldsymbol{s}_l\,. \tag{2}$$

The summation in eq. (2) is over the l-th set and $\boldsymbol{J}_l$ and $\boldsymbol{s}_l$ are vectors of rank N/K. The output unit of the network, o, is a fixed Boolean function, f, of the internal representation of the hidden units

$$o = f[\sigma_1, \ldots, \sigma_K]\,. \tag{3}$$

The task of the network is a mapping of a set of P random input patterns, $\{\xi_i^\mu\}$, $i = 1, ..., N$, $\mu = 1, ..., P$ and $\xi_i^\mu = \pm 1$ with equal probability, onto a set of P random binary outputs $y^\mu = \pm 1$ with equal probability. The goal is to calculate, as a function of K, the maximal number, P_c, of input-output pairs that can be stored in the network. Since the maximal capacity is of $O(N)$, it is convenient to introduce the symbol $\alpha_c \equiv P_c/N$.

The first analytically solvable case of an MLN with one hidden layer is the parity machine (PM) [4], $o = \operatorname{sgn}\left[\prod_{l=1}^{K} \sigma_l\right]$. The main results for the PM [4] within one step of RSB are that up to $\alpha_0(K)$ the system is in the paramagnetic phase. At $\alpha_0(K)$ the system undergoes either a second-order phase transition ($K = 2$) or a first-order phase transition ($K > 2$) to a glassy phase. The maximal capacity, α_c, is obtained where the entropy vanishes and scales with

$\log_2 K$. In particular, it is found that for $K = 2$ and 3, $\alpha_0 \simeq 1.27$ and 3.2 and $\alpha_c \simeq 4.06$ and 5, respectively. The second case which was examined analytically is the committee machine (CM), $o = \mathrm{sgn}\left[\sum_{l=1}^{K} \sigma_l\right]$. The results within one step of RSB were examined in detail only for the case $K = 3$ where a transition to an RSB phase occurs at $\alpha_0 \simeq 1.67$ and $\alpha_c \simeq 3.0$ [6,7]. The third case which was recently examined [5] only within the RS ansatz is the AND machine [5]. For $K = 3$, $\alpha_c \simeq 3.66$ and for $K > 3$, α_c is bounded from above by 4.

In all of these cases of MLN, stochastic algorithms [4-7] converge well in the RS regions and even seem to saturate the maximal capacity for the AND machine where the RS phase seems to be globally stable up to α_c [5]. Nevertheless, in cases where an RSB phase appears for large α, all the known algorithms give maximal capacity which is $20 \div 50$ percent less than the analytically predicted maximal capacity. The source for these deviations could be either approximations in the scheme of RSB or the quality of the algorithms which for some technical reasons or fundamental reasons, such as the existence of local minima, cannot converge in the whole RSB phase.

In the following, simulations of a BP learning algorithm for the case of the PM are presented, where an extension to the CM is discussed later. The energy function examined below for the PM is defined by

$$E = -\sum_{\mu=1}^{P} X^{\mu}[1 - \mathrm{tgh}(X^{\mu})]\,\theta(-X^{\mu}), \tag{4}$$

where

$$X^{\mu} = g y^{\mu}\left[\prod_{l=1}^{K} h_l^{\mu}\right]. \tag{5}$$

The parameter g is the gain factor of the energy function and h_l^{μ} is defined in eq. (2). In the limit $g \to \infty$, the term in the brackets of eq. (4) is proportional to the well-discussed step function energy [3]. The minimization of E with respect to the weights, $\{J_i\}$, is usually performed through sequential iterative updates using some form of gradient descent

$$\boldsymbol{J}^{t+1} = \boldsymbol{J}^{t} - \lambda \nabla E, \tag{6}$$

where $\boldsymbol{J}$ is a vector of rank N, ∇E is the gradient of E with respect to $\boldsymbol{J}$ and λ is used to adjust the size of the updating step. It is clear that a pattern is embedded in the network if its X^{μ} is positive. Hence, the energy function is positive except in the ground state, where all the patterns are embedded and $E = 0$. In the simulations the weights are updated sequentially following eq. (6), which is one of the versions of the BP algorithm [11,12].

The algorithm is terminated when no changes occur in all the weights. There are only two possibilities for such a situation: *a*) $E = 0$ and a solution is found, *b*) the algorithm is stuck in one of the local minima of the energy function E.

Let us prove now that in some particular limits local minima do not exist. A necessary condition for a local minimum is that

$$\frac{\partial E}{\partial J_i} = -\sum_{\mu} \frac{\partial X^{\mu}}{\partial J_i} a^{\mu} = 0, \qquad \forall i, \tag{7}$$

where $a^{\mu} = 1 - \mathrm{tgh}(X^{\mu}) - X^{\mu}/\cosh^2(X^{\mu})$, the summation is only over patterns which are not embedded and

$$\frac{\partial X^{\mu}}{\partial J_i} = \frac{X^{\mu}}{\sqrt{\sum J^2}}\left(\xi_i^{\mu}/h_1^{\mu} - J_i/\sqrt{\sum J^2}\right), \tag{8}$$

where J_i belongs, for instance, to the first block. Using the equalities, eq. (7), one can show that within one block

$$\frac{\partial^2 E}{\partial J_i \, \partial J_j} = \sum_{\mu} (\delta_{ij} - J_i J_j) K_{\mu} + (\xi_i^{\mu}/h_1^{\mu} - J_i)(\xi_j^{\mu}/h_1^{\mu} - J_j) M_{\mu}, \tag{9}$$

where $K^{\mu} = X^{\mu} a^{\mu}$ and $M^{\mu} = 2(1 - X^{\mu} \operatorname{tgh} X^{\mu})(X^{\mu})^2/\cosh^2(X^{\mu})$. It is now straightforward to show that within the subspace of only two weights (other weights are fixed), the necessary condition for a local minimum is given in the leading order where $h_i^{\mu} = O(1)$, $\forall \mu$ by

$$\sum_{\mu} K^{\mu} + M^{\mu}(1 \pm \xi_i^{\mu} \xi_j^{\mu})/(h_1^{\mu})^2 > 0. \tag{10}$$

However, there are with probability one in the finite P limit some pairs of input units which obey $\xi_i^{\mu} = \xi_j^{\mu}$, $\forall \mu$. Since $K^{\mu} < 0$ the inequality equation (10) is violated in these pairs and local minima do not exist. In other extreme limits where h_i^{μ} is not of $O(1)$, it is trivial to verify that local minima do not exist. The extension of this result to the general case of finite α limit is difficult. Nevertheless, in the case of sign-constrained wieights (each weight has a fixed sign, for instance, positive) equalities (7) can be written as

$$\sum_{\mu} b^{\mu} \xi_i^{\mu} = J_i \sum_{\mu} a^{\mu} X^{\mu} \equiv A J_i < 0, \tag{11}$$

where $b^{\mu} \equiv X^{\mu} a^{\mu}/h_1^{\mu}$. Since $\{\xi_i^{\mu}\}$ are random variables, eq. (11) stands for a set of N random linear equalities with P parameters, $\{a^{\mu}\}$, which in the optimal case are not correlated. It is clear that the number of such random equalities which can be solved simultaneously is bounded from above by the number of the inequalities, $\sum b^{\mu} \xi_i^{\mu} < 0$, which can be solved simultaneously. Since this number is well known to be twice the number of the parameters, it implies that at least for $\alpha < 1/2K$ local minima do not exist. Nevertheless, the simulations suggest that either local minima do not exist for any α or a small number of them exist but with small basin of attractions.

Before the results of the simulations are presented, let us now discuss some of the properties which characterized the energy function and the algorithm, eqs. (4)-(6). The normalization of the weights in the dynamical process is necessary, since as the values of the weights become large, the derivatives of the energy function become small, the convergent time to a solution can be immense. Another comment is that there is a critical size of the step, $\lambda_{\max}$, which depends on the architecture of the network and on g, where only for $\lambda < \lambda_{\max}$ a solution can be obtained [13]. However, as λ becomes smaller, the convergent time to a solution increases. Therefore, we carried out heavy simulations to fix the values of g and λ which minimize the convergent time. The last comment is that there is an advantage for the prefactor X^{μ} in the energy function, eq. (4), since the derivative is finite even for large $-X^{\mu}$ but does not vanish for small X^{μ}.

In the simulations we concentrate on architecture with 2 and 3 hidden units, where the size of each block, N/K, is between 5 and 15. The results for $K = 2$ and 3 are presented in fig. 1, where each point was averaged over $20 \div 80$ samples. The maximal capacity is defined as the capacity where only in one-half of the samples a solution is found.

The results of the simulations indicate that α_c scales with $1/N$ and is equal in the thermodynamic limit to $\simeq 3.9$ and 4.9 for $K = 2$ and 3, respectively. These maximal capacities are very close to the prediction of the calculations within the framework of one-step RSB, where $\alpha_c = 4.06$ and 5.0, respectively. Furthermore, for $K = 2$ and within two steps of RSB $\alpha_c \simeq 3.9$ [14]. It is important to note that the LAA does not saturate the maximal capacity and gives for these cases 3.2 and 3.6 [14]. The convergent time, t_{con}, as a function of α for a typical

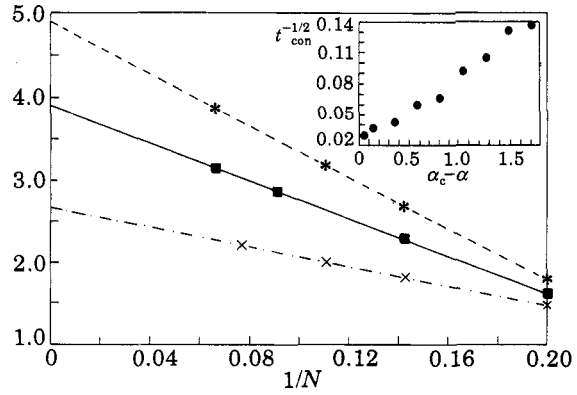

Fig. 1. – Results of the simulations for the PM with $K = 2$ (solid line) and 3 (dashed line) and for the CM with $K = 3$ (dotted line). The insert is $t_{con}^{-1/2}$ *vs.* $\alpha_c - \alpha$ for the PM with $N = 11$ and $K = 2$.

case, $K = 2$ and $N = 11$, is presented in the insert of fig. 1 and indicates that

$$t_{con} = A(N)/[\alpha_c(N) - \alpha]^2 , \tag{12}$$

where $A(N)$ is a constant. In order to verify whether $A(N)$ diverges with N, we carried out simulations for N up to 500 and fixed $(\alpha_c(N) - \alpha)$ to be equal to one. The results indicate that $A(N)$ is almost independent of N and hence the convergent time is not exponential with N. Note also that the deviation of the results from the prediction of one step RSB is less than 4 percent, which seems a reasonable difference due to corrections of a more structured functional order parameter. Note that in almost all cases of analytical solvable spin-glass (SG) systems the correction of one step RSB is a few percent in comparison to the RS solution, but the correction of two-step RSB is negligible in comparison to the first correction [15]. The reason why the corrections of two steps are not negligible in the discussed systems is a result of the wrong order of the maximal capacity in the RS solution, $\alpha_c \propto K^2$. Hence, one-step solution is the first approximation in the right order as the role that RS plays in the case of SG systems. In conclusion, the analytical maximal capacity can be achieved by the algorithm in a time which is not exponential with the size of the system.

The extension of the BP algorithm to the case of the CM, for instance, can be done in the following way:

$$E = \sum_{\mu=1}^{P} \Theta\left[-y^\mu \left(\sum_{l=1}^{K} \operatorname{sgn}(h_l^\mu) \right) \right] \sum_{l=1}^{K} (h_l^\mu)^2 \Theta(-h_l^\mu y^\mu) . \tag{13}$$

The first theta-function indicates that only patterns which are not embedded contribute to the energy function. The second theta-function indicates that the system prefers to flip the sign of hidden units which are antiparallel to the desired output. The results of the simulations are presented in fig. 1 and indicate that $\alpha_c \simeq 2.7$, where the result of one-step RSB gives $\simeq 3.0$. Note that the LAA gives in this case $\alpha_c \simeq 2.4$.

Finally, note that in almost all SG systems the transition to the glassy phase can be found by the RS ansatz, but the globally stable glassy phase is characterized by RSB [15]. This RSB phase is characterized by many ground states which are separated by huge energy barriers and many local minima. This structure of the energy landscape is responsible for the huge relaxation times which are observed, for instance, in simulations. A globally stable RS phase exists only in some exceptional cases [16], however, the structure of their energy

landscape was not examined analytically or even numerically. In contrast, the statistical nature of all the examined MLN exhibits a region where the RS phase is globally stable. The physical meaning of such a solution is that the phase space contains only one connected region (except global symmetries) which fulfils the task of the network. This behaviour is similar to ferromagnetic (FM) systems, and might indicate that, as for the FM case, the region of the RS phase is also characterized by the lack of local minima. Hence, a BP learning algorithm can converge in this region. The main question is the general structure of the RSB phase, and in particular in the case of MLN. There are two options, both of them surprising. In the first option, the theoretically predicted α_c can be achieved by a BP algorithm as is found in the above-mentioned simulations. This result indicates that, unlike SG systems, there is a possibility of an RSB phase which consists of many disconnected regions which fulfil the task of the network, but with the lack of local minima. In this framework, the maximal achievable capacity is the same as the maximal theoretical capacity and there is no remarkable relation between the phase diagram and the behaviour of the algorithm as a function of α. Nevertheless, it is possible to explain the results in a second surprising way. In the original step function energy [3], which is discontinuous as a function of the weights, the maximal capacity is indeed α_0 and the algorithm converges only in the RS phase. In contrast, the energy function equation (4) is continuous, since the gain factor g is finite. This smoothing of the energy function is responsible for the lack of local minima. Indeed, preliminary simulations on the step function energy indicate that the maximal capacity is much lower than α_c which might indicate that the α_c is indeed α_0. This behaviour indicates a strong relationship between the thermodynamic properties of the system and the efficiency of the algorithm. Note nevertheless that the investigation of the step function energy is difficult, since the energy function is flat in the region where the number of embedded patterns is fixed and any simple local algorithm is the same as random walks in these flats.

* * *

Discussions and comments on the manuscript of E. DOMANY and T. GROSSMAN are gratefully acknowledged.

REFERENCES

[1] See, for example HERTZ J., KROGH A. and PALMER R., *Introduction to the Theory of Neural Computation* (Addison-Wesley Press) 1991.
[2] MINSKY M. L. and PAPERT S., *Perceptron* (MIT Press, Cambridge) 1969.
[3] See, for instance, the memorial volume of GARDNER E., *J. Phys. A*, **22** (1989).
[4] BARKAI E., HANSEL D. and KANTER I., *Phys. Rev. Lett.*, **65** (1990) 2312.
[5] GRINIASTY M. and GROSSMAN T., to be published in *Phys. Rev. A*.
[6] ENGEL A., KOHLER H. M., TSCHEPKE F., VOLLMAYR H. and ZIPPELIUS A., preprint (1992).
[7] BARKAI E., HANSEL D. and SOMPOLINSKY H., *Phys. Rev. A*, **45** (1992) 4146.
[8] GARDNER E., *J. Phys. A*, **21** (1988) 257; GARDNER E. and DERRIDA D., *J. Phys. A*, **21** (1988) 271.
[9] MITCHISON G. J. and DURBIN R. N., *Biol. Cyber.*, **60** (1989) 345.
[10] NABUTOVSKY D., GROSSMAN T. and DOMANY E., *Complex System*, **4** (1990) 519.
[11] RUMELHART D. E., HINTON G. E. and WILLIAMS R. J., *Nature*, **323** (1986) 533.
[12] LE CUN Y., *Proc. Cognit.*, **85** (1985) 599.
[13] LE CUN Y., KANTER I. and SOLLA S. A., *Phys. Rev. Lett.*, **66** (1991) 2396.
[14] EINSTEIN E. and KANTER I., unpublished.
[15] For a review, see MÉZARD M., PARISI G. and VIRASORO M. A., *Spin Glass Theory and Beyond* (World Scientific, Singapore) 1987.
[16] KOSTERLITZ J. M., THOULESS D. J. and JONES R. C., *Phys. Rev. Lett.*, **36** (1976) 1217.