Storage Capacity of a Multilayer Neural Network with Binary Weights

1991 Europhys. Lett. 14 107

(http://iopscience.iop.org/0295-5075/14/2/003)

EUROPHYSICS LETTERS 15 January 1991

Europhys. Lett., 14 (2), pp. 107-112 (1991)

Storage Capacity of a Multilayer Neural Network with Binary Weights.

E. BARKAI and I. KANTER

Department of Physics, Bar-Ilan University - Ramat Gan 52100, Israel

(received 11 October 1990; accepted 13 November 1990)

PACS. 05.50 – Lattice theory and statistics; Ising problems.
PACS. 87.10 – General, theoretical, and mathematical biophysics (inc. logic of biosystems, quantum biology, and relevant aspects of thermodynamics, information theory, cybernetics, and bionics).

Abstract. – Statistical mechanics is applied to estimate the maximal capacity per weight (α_c) of a two-layer feed-forward network with binary weights, functioning as a parity machine of the hidden units. For $K \geq 2$ hidden units, the maximal theoretical capacity is achieved, $\alpha_c = 1$, and the average overlap between different solutions is zero. These results agree with the simulations. At finite temperature one-step replica symmetry breaking solution is found, which appears to be exact.

Much work has been done in the recent past on the statistical nature of a one-layer network, known as a perceptron [1, 2]. However, the computational power of such one-layer network is limited, since it cannot solve nonseparable problems. Furthermore, having in mind applications to biological systems, multilayer networks may play an important role. Hence, quantitative estimation of the capability of *multilayer* architectures is very interesting. In particular, it is interesting to know how much the information capacity *per synapse* (*i.e.* per weight) is larger in multilayer systems than in one-layer perceptron.

In this letter a two-layer feed-forward network is studied using statistical mechanics approach. The architecture of the network consists of N binary input units, one hidden layer with K continuous or discrete units and a single binary output unit. The input units are divided into K equal disjoint sets, each one of them consists of N/K units. The l's hidden unit is connected only to the i-th input via a binary weight, $J_i = \pm 1$, such that $N(l-1)/K < i \leq Nl/K$. Therefore, each one of the input units is connected only to one hidden unit, *i.e.* the receptive fields of the hidden units are nonoverlapping.

The configuration of the input is denoted by $\{s_i\}$ $i = 1, \ldots, N$ with $s_i = \pm 1$. The state of the l-th hidden unit is equal to its induced local field

$$h_l = \sum_{i=(N(l-1)/K)+1}^{Nl/K} J_i s_i \equiv \boldsymbol{J}_l \cdot \boldsymbol{s}_l , \tag{1}$$

where J_i is a binary weight which can take the values ± 1, $\boldsymbol{J}_l$ and $\boldsymbol{s}_l$ are vectors of rank N/K. The output unit, o, of the network is just the sign of the product of the K hidden units

$$o = \operatorname{sgn}\left(\prod_{l=1}^{K} h_l\right). \tag{2}$$

It is now clear that the functioning of this network is independent of the exact nature of the hidden units, *i.e.* continuous or discrete. This network is known as a parity machine network [3] with binary weights, since the output is the parity of the internal representation of the hidden units. It is important to note that this architecture is a two-layer network only where the output unit is a Sigma-Pi unit. A version of this model with continuous weights was studied in ref. [4].

Like in the study of the one-layer perceptron, the task of the network is a mapping of a random input on a random output. More precisely, the μ-th pattern consists of $\xi_i^\mu = \pm 1$ with equal probability, where $i = 1, \ldots, N$, $\mu = 1, \ldots, P$ and α is defined to be P/N. The desired output of the μ-th pattern is $y^\mu = \pm 1$ with equal probability. The question is to calculate, as a function of K, the maximal number of patterns P_c that can be tought to the network in the limit $N \to \infty$. It is also important to understand the statistical nature of the solutions in the phase space of the weights.

The case $K = 1$ is exactly the perceptron with binary weights. This problem has been addressed by Gardner and Derrida [5-7], who found that within the replica-symmetric (RS) ansatz $\alpha_c = 4/\pi$, but showed that RS must be broken. Krauth and Mezard [8] found a solution with one-step replica symmetry breaking [9, 10] (RSB) which seems to be exact. The critical capacity is ~ 0.83, where the zero-temperature entropy vanishes.

Let us first summarize the main results of this work where the details are given below.

a) For $K \geq 2$, $\alpha_c = 1$. Unlike the case of continuous weights [4], where α_c scales with $\log_2 K$, in the binary case the maximal capacity is independent of K and equals the bound of the maximal theoretical embedding information.

b) For $\alpha < 1$, the entropy of the system is equal to $(N - P)\ln 2$ and the order parameter q (which will be defined below) is equal to zero.

c) At finite temperature, there is a critical line independent of K, $\alpha_c(T_c)$, where the entropy of the system vanishes. For $T > T_c$ the system is in the paramagnetic phase characterized by $q = 0$, where below it a glassy phase appears. The function order parameter, $q(x)$, of this glassy phase is characterized by one step of RSB, where $q_0 = 0$, $q_1 = 1$ and the breakpoint $m = T/T_c$.

The derivation of these results are given below, however, let us first indicate that the upper bound for the capacity of the system is $\alpha_c = 1$. This upper bound is obtained by using the annealed approximation or by a simple argument of information theory, where each weight contains exactly one bit of information. Nevertheless, the exact maximal capacity and the nature of the system are investigated using the method suggested by Gardner and Derrida [5, 6]. For each set of coupling the energy is defined as the number of patterns which are not embedded:

$$E(J) = \sum_{\mu=1}^{P} \Theta\left(y^\mu \prod_{l=1}^{k} J_l \cdot \xi_l^\mu / \sqrt{N}\right). \tag{3}$$

Following ref. [6], for a given set of patterns we introduce the partition function at temperature $T = 1/\beta$

$$Z = \mathrm{Tr}_{\{J_i = \pm 1\}} \exp\left[-\beta E(J)\right]. \tag{4}$$

In the zero temperature limit the partition function, eq. (4), is exactly the volume, V, in the weights space occupied by the networks which achieve the desired output. The patterns can be learnt if and only if the zero-temperature internal energy vanishes. In this limit the entropy of the system per weight is $\ln(Z)/N$. Since the entropy is a nonincreasing function of α, the maximal capacity is defined as the lowest α such that the entropy vanishes.

In the computation of extensive thermodynamic quantities one concentrates on $\ln Z$. This quantity is averaged over the distribution of the quenched random patterns $\{\xi_l^\mu, y^\mu\}$, using the replica method. In the calculations one introduces a set of order parameters

$$q_l^{\alpha\beta} = \frac{K}{N} \boldsymbol{J}_l^\alpha \cdot \boldsymbol{J}_l^\beta, \tag{5}$$

where $|q_l^{\alpha\beta}| < 1$. The physical meaning of $q_l^{\alpha\beta}$ is the overlap between the weights belonging to the l-th hidden unit in two replicas, α and β. Note that because the receptive fields of two different hidden units are nonoverlapping, there is *a priori* only one type of order parameter, $q_l^{\alpha\beta}$. Furthermore, $q_l^{\alpha\beta} = q_{l'}^{\alpha\beta}$, since after the average over the disorder all the hidden units are equivalent.

Under the RS assumption, $q_l^{\alpha\beta} = (1-q)\,\delta_{\alpha\beta} + q$, one can show that

$$G_{\mathrm{RS}} = \langle\!\langle \ln Z \rangle\!\rangle = \mathrm{ext}_{\{q,\phi\}} \left\{ -\frac{1}{2}\phi(1-q) + \int_{-\infty}^{\infty} \mathrm{D}z \ln\left(2\cosh\left(\sqrt{\phi}\,z\right)\right) + \right.$$
$$\left. + \alpha \int_{-\infty}^{\infty} \prod_l \mathrm{D}t_l \ln\left\{ \exp[-\beta] + (1-\exp[-\beta]) \left[\mathrm{Tr}_{\{\tau_l=\pm 1\}} \prod_l H(Q\tau_l t_l)\, \Theta\left(\prod_l \tau_l\right) \right]\right]\right\}\right\}, \tag{6}$$

where the symbol $\langle\!\langle \ldots \rangle\!\rangle$ stands for the average over the patterns, $\alpha = P/N$, $Q = (q/(1-q))^{1/2}$, $\mathrm{D}t = \exp[-t^2/2]\,\mathrm{d}t/(2\pi)^{1/2}$, $H(x) = \int_x^\infty \mathrm{D}x$ and q and ϕ are determined by the saddle point (SP) equations.

As a function of α one can distinguish between two regimes. *a*) The regime $0 < \alpha < \alpha_0$ is characterized by $q = 0$. For the case $K = 2$, for instance, $\alpha_0 = \dfrac{\pi^2(1+\exp[-\beta])^2}{8(1-\exp[-\beta])^2}$, where it is clear that α_0 is an increasing function of K. *b*) At α_0 the system undergoes a transition to a glassy phase which is characterized by $q > 0$. The order parameter q is an increasing function of α and goes to 1 for the case $K = 2$, for instance, at $\alpha \sim 3.5$. One can also verify that the RS solution is stable at least up to α_0.

The upper bound for the maximal capacity, $\alpha_c = 1$, indicates that the stability analysis or the capacity where $q \to 1$ are not good criteria in this case to fix or to bound the maximal capacity. In contrast, the available volume of the solutions is a decreasing function of the capacity. Furthermore, since the volume cannot be less than one point in the phase space (at zero temperature), the capacity where the entropy vanishes is defined as the maximal capacity. This behaviour was recently observed in the binary perceptron case ($K = 1$), where the maximal capacity ($\alpha_c \sim 0.83$) is characterized by zero entropy, however, the order parameter q at the criticality is around 0.5 [8]. Therefore RS solution suggests that for any $K \geqslant 2$, the maximal capacity is $\alpha_c = 1$. In contrast with the case $K = 1$, where q is an increasing function of α, here q is a constant and equal to zero for any $\alpha \leqslant 1$. This result, $q = 0$, can explain why the annealed approximation gives the same behaviour as the RS solution.

The result that $q = 0$ seems to be surprising, since one expects that as α increases, the available volume as a solution decreases and correlations among different solutions are built. Note that the parity machine architecture is characterized by 2^{K-1} global symmetries, each one of them consists of the flip of the sign of weights which belong to even number of hidden units [4]. However, the glassy order parameter, q, does not relate to these global symmetries. In the following we will examine if RSB affect these results.

The averaged $\ln Z$ with one step of RSB is given by

$$\begin{aligned} G_{1\,\mathrm{step}}(\phi_0,\ \phi_1,\ q_0,\ q_1,\ m) = {} & \frac{1}{2}[\phi_1 q_1(1-m) - \phi_1 + m\phi_0 q_0] + \\ & + \frac{1}{m}\int_{-\infty}^{\infty} \mathrm{D}z \ln\left[\int_{-\infty}^{\infty} \mathrm{D}t\, 2^m \cosh^m(\sqrt{\Delta\phi}\, t + \sqrt{\phi_0}\, z)\right] + \\ & + \frac{\alpha}{m}\int_{-\infty}^{\infty} \prod_l \mathrm{D}t_l \ln\left\{\int_{-\infty}^{\infty} \prod_l \mathrm{D}z_l \left(\exp[-\beta] + (1-\exp[-\beta]) \cdot \right.\right. \\ & \left.\left. \cdot \left[\mathrm{Tr}_{\{\tau_l=\pm 1\}} \prod_l H\left[\tau_l\left(\sqrt{\frac{q_0}{1-q_1}}\, t_l + \sqrt{\frac{\Delta q}{1-q_1}}\, z_l\right)\right] \theta\left[\prod_l \tau_l\right]\right]\right)^m\right\}, \end{aligned} \tag{7}$$

where $\Delta\phi \equiv \phi_1 - \phi_0$ and $\Delta q \equiv q_1 - q_0$. In this parametrization the order parameter q_0 represents the overlap between the average solution in two different valleys:

$$q_0 = \frac{1}{N}\left\langle\!\!\left\langle \sum_i \langle J_i\rangle_\alpha \langle J_i\rangle_\beta \right\rangle\!\!\right\rangle, \tag{8}$$

q_1 represents the self-overlap within one valley

$$q_1 = \frac{1}{N}\left\langle\!\!\left\langle \sum_i \langle J_i\rangle_\alpha^2 \right\rangle\!\!\right\rangle, \tag{9}$$

and the breakpoint m relates to the Gibbs weights, P_α, of each one of the valleys

$$1 - m = \left\langle\!\!\left\langle \sum_\alpha P_\alpha^2 \right\rangle\!\!\right\rangle. \tag{10}$$

The five order parameters must be computed through the saddle point equations. In the zero-temperature limit and for $\alpha \leqslant 1$ we did not find any solution except the RS solution, $q_0 = q_1 = 0$. At finite temperature we find no RSB solution except in the limit $q_1 \to 1$. Following reference (7), the limit $q_1 = 1$ at finite temperature is studied. In this limit one can verify that $\phi_1 \to \infty$ and

$$G_{1\,\mathrm{step}}(q_0,\ \phi_0,\ 1,\ \infty,\ m,\ \beta) = \frac{1}{m} G_{\mathrm{RS}}(q_0,\ m^2\phi_0,\ \beta m), \tag{11}$$

where G_{RS} is given in eq. (6). The saddle point equations with respect to q_0 and ϕ_0 imply that $q_0 = q$ and $\phi_0 = \phi/m^2$, where q and ϕ are the RS order parameters at the inverse temperature βm. The saddle point with respect to m gives $S_{\mathrm{RS}} = S_{1\mathrm{step}} = 0$, where βm must be equal to $1/T_{\mathrm{c}}$, which is the inverse temperature where the RS entropy vanishes. This temperature as a function of α is given by

$$\alpha_{\mathrm{c}}(T) = -\ln(2)/\ln((1+\exp[-\beta])/2), \tag{12}$$

where as $T \to 0$, $\alpha_{\mathrm{c}} \to 1$. Therefore, the phase diagram within one step RSB is as follows: *a*) in the zero-temperature case, $q_0 = q_1 = 0$ for $\alpha < 1$. In this regime one can verify from eq. (11) that the entropy is given by $S = (N-P)\ln 2$ and vanishes as $\alpha \to 1$. The solution at $\alpha = 1$ is impossible, because of the essential singularity at this point. *b*) At finite temperature and at

$T > T_c$, the only solution is the paramagnetic solution characterized by $q_0 = q_1 = 0$. *c*) On the transition line, $T_c(\alpha)$, the entropy vanishes. At $T < T_c$, there exists one-step RSB solution which is defined by $m = T/T_c$, $q_0 = 0$ and $q_1 = 1$ and the free energy is independent of T and is equal to the RS free energy at T_c. The one-step solution is first order in the sense that the order parameter function, $q(x)$, is not continuous, but the free energy and its first derivative are still continuous at T_c. This behaviour is exactly identical to what was found in the random energy (RE) model, simplest spin glass [11, 12] and the Potts glass [13]. However, unlike the RE and the simplest spin glass models the energy levels are not independent random variables. This one can verify *only* from the low-temperature phase, from the structure of the last term in eq. (7). This behaviour is similar to the Potts glass in the large p limit (where p is the number of Potts states) [13] and to a highly diluted RE model [14]. In both cases there are correlations among the energy levels, which however are not strong enough to change the solution in the mean-field limit. As in ref. [11, 12], the physical interpretation of the order parameter is that in the thermodynamic limit the average self-overlap within one pure state is 1, where the overlap between the average solution in two different pure states is zero. In the thermodynamic limit, the phase space contains many pure states, separated by huge barriers. The inextensive difference among the energies of the pure states is responsible for the spectra of weights, P_α, where the mean square of the weights is given by $1 - m = 1 - T/T_c$ (see eq. (10)). On the critical line, $T_c(\alpha)$, the entropy per weight vanishes in the leading order, however, it seems that below it the entropy is still finite and only a few pure states become dominant, as in the case of the RE model.

In order to check whether the one-step solution is exact, a solution with two steps RSB is examined. In this parametrization the partition function is constructed from eight order parameters q_0, q_1, q_2, ϕ_0, ϕ_1, ϕ_2, m_1 and m_2. In the limit $q_2 = 1$ and $\phi_2 \to \infty$ one can show that

$$G_{2\text{steps}}(q_0, \phi_0, q_1, \phi_1, 1, \infty, m_1, m_2, \beta) = \frac{1}{m_2} G_{1\text{step}}(q_0, m_2^2\phi_0, q_1, m_2^2\phi_1, m_1/m_2, \beta m_2), \tag{13}$$

which is the same equality previously obtained for the case $K = 1$ [8]. Optimization of $G_{2\text{steps}}$ with respect to m_1 is equivalent to optimization of the one-step solution with respect to m. However, such solution was not found except where $q_1 = 1$ and $q_0 = 0$, which means that the two-step solution is degenerate with the one-step solution. Hence, the one-step solution appears to be exact. Let us only comment now that there exists another one-step solution where the transition for the case $K = 2$, for instance, occurs at $\alpha = \pi^2(1 + \exp[-\beta])^2/8(1 - \exp[-\beta])^2$, where $q_1 \to 0$. However, this cannot be a physical transition, since it occurs in lower temperatures than the previously discussed transition.

The theory was checked by exhaustive search simulations. In these simulations all the 2^N possible configurations in the phase space were examined, where N is the size of the input. More precisely, using the properties of the global symmetries, only 2^{N-K+1} configurations have to be checked. The size of the examined systems in the simulations was up to $N = 26$ with $K = 2$ and 3. The main results of the simulations at a zero temperature are: *a*) in fig. 1 the fraction of samples with solutions is plotted *vs.* α. Each point is averaged over at least 50 samples. The maximal capacity is fixed at α where the fraction of samples with solutions is equal to 1/2. It is clear from fig. 1 that α_c is an increasing function of N. Furthermore, for any examined N, the fraction of solutions for $P = N - 1$ is greater than 1/2. These results strongly indicate that $\alpha_c \to 1$. *b*) In fig. 2 the logarithm of the volume which is available as a solution is plotted *vs.* P for $K = 2$ and $N = 22$. At zero temperature the theory predicts that for $\alpha < 1$, $q = 0$ and $\log V$ is a straight line with a slope equal to $-\log 2 \simeq -0.3$. It is clear from fig. 2 that up to $P = 19$, $\log V$ scales linearly with P with a slope ~ -0.29. The deviations from a linear curve for $P > 19$ are due to strong fluctuations in finite systems, as the volume goes to zero. This result agrees with the theoretical prediction for the entropy

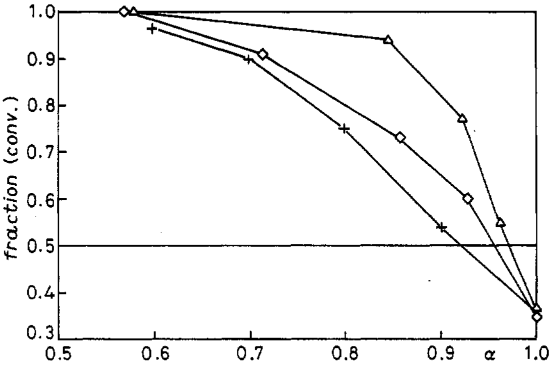

Fig. 1.

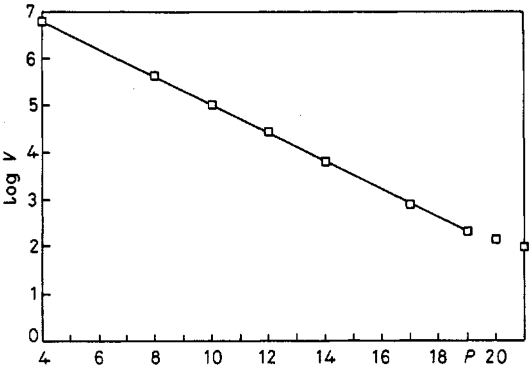

Fig. 2.

Fig. 1. – The fraction of samples with solutions *vs.* α in an exhaustive search, where $K = 2$ and $N = 10$, 14 and 26 with +, ◇ and △, respectively.

Fig. 2. – $\log V$ as a function of P, for $K = 2$ and $N = 22$.

with $q = 0$, for $\alpha < 1$. *c*) Preliminary results of simulations at finite temperatures indicate a strong first-order transition from the paramagnetic phase to the glassy phase, where $T_c(\alpha)$ is close to the theoretical prediction.

Finally, we would like to comment that a direct measurement of q is difficult, since the definition of configurations which belong to one valley and barriers between different valleys are not well defined in finite systems. Nevertheless, a remarkable insight on the structure of the solutions in the phase space was found where the average overlap between each pair of solutions was calculated as a function of P. It is clear that for $K > 1$ solutions which can be obtained by global symmetries should be deleted, otherwise the overlap by definition is zero. In the case $K = 1$ and $N = 21$ this overlap increases as a function of α up to ~ 0.5 near $\alpha \sim 0.8$ [8]. In contrast, in the case $K = 2$ and $N = 22$ this overlap was less than 0.08 for any $P < 21$. Note that this measure is not exactly equivalent to the physical meaning of the order parameter q_1 (see eq. (9)).

* * *

Discussions with D. HANSEL are acknowledged. This research of IK is supported by the Israel Ministry of Science and Development.

REFERENCES

[1] MINSKY M. L. and PAPER S., *Perceptron* (MIT Press, Cambridge, Mass.) 1969.
[2] COVER T., *IEEE Trans. Electron. Comput.*, **14** (1965) 326.
[3] MITCHISON G. J. and DURBIN R. N., *Biol. Cybern.*, **60** (1989) 345.
[4] BARKAI E., HANSEL D. and KANTER I., preprint (1990).
[5] GARDNER E., *J. Phys. A*, **21** (1988) 257.
[6] GARDNER E. and DERRIDA D., *J. Phys. A*, **21** (1988) 271.
[7] See, for instance, the memorial volume of GARDNER E., *J. Phys. A*, **22** (1989).
[8] KRAUTH W. and MEZARD M., *J. Phys. (Paris)*, **50** (1989) 3057.
[9] PARISI G., *J. Phys. A*, **13** (1980) 1101.
[10] For a review, see MEZARD M., PARISI G. and VIRASORO M. A., *Spin Glass Theory and Beyond* (World Scientific, Singapore) 1987.
[11] DERRIDA B., *Phys. Rev. B*, **24** (1981) 2613.
[12] GROSS D. J. and MEZARD M., *Nucl. Phys. B*, **240** (1984) 431.
[13] GROSS D. J., KANTER I. and SOMPOLINSKY H., *Phys. Rev. Lett.*, **55** (1985) 304.
[14] KANTER I., *J. Phys. C*, **20** (1987) L-257.