Information Theory of a Multilayer Neural Network with Discrete Weights

EUROPHYSICS LETTERS 7 January 1992
Europhys. Lett., 17 (2), pp. 181-186 (1992) *published in November 1991*

Information Theory of a Multilayer Neural Network with Discrete Weights.

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

(received 26 March 1991; accepted in final form 30 September 1991)

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 discrete weights of depth l, functioning as a parity machine of the K hidden units. For each K and $l \leqslant l_0(K)$, the maximal theoretical capacity $\alpha_c = \log_2(2l)$ is achieved, the capacity per bit is 1, the average overlap between different solutions is zero and $l_0(K) \propto \log K$ for large K. At finite temperature, a one-step replica symmetry-breaking solution is found to be exact for $l \leqslant l_0(K)$.

In the recent past, statistical-mechanical methods have been applied to investigate the properties of neural networks. Among these systems, the class of multilayer networks plays an important role [1]. The prototype of this class of architecture is the one-layer perceptron [1], consisting of one input layer of N binary units and one-binary output unit. Various statistical mechanical properties of the one-layer perceptron such as the maximal capacity and the generalization ability of the network have been recently investigated by using the pioneering work of Elizabeth Gardner [2, 3].

However, the computational capability of a one-layer perceptron is limited, since it cannot solve nonseparable problems. Furthermore, multilayer networks may play an important role in the functioning of biological systems. In particular, it is important to examine the advantages of multilayer networks over the perceptron with respect to quantities such as storage capacity per weight and the effect of synaptic depth (the resolution of the synaptic strength). Furthermore, the applicability of neural networks to biology and to the construction of real devices requires the understanding of the interplay between the synaptic depth and the properties of the network. The effect of the synaptic depth on the system is crucial, since the implementation of a deeper synaptic depth is much more difficult and expensive in real neural network circuits [4]. Hence, an optimal depth should be defined and predicted for each task of the network.

In this letter a two-layer feed-forward network is studied using a statistical-mechanical 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 consisting of N/K units. The j's hidden unit is connected only to the i-th input via a weight of depth l, $J_i = \pm 1, \pm 2, \ldots, \pm l$, such that $N(j-1)/K < i \leqslant Nj/K$. The configuration of the input is denoted by $\{s_i\}$, $i = 1, \ldots, N$, with

$s_i = \pm 1$. The state of the j-th hidden unit is equal to its induced local field

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

where J_i is a weight of depth l and $\boldsymbol{J}_j$ and $\boldsymbol{s}_j$ are vectors of rank N/K. The output unit, o, is just the sign of the product of the K hidden units $o = \mathrm{sgn}\left(\prod_{l=1}^{K} h_l\right)$. Hence, the output is a parity of the internal representation of the hidden units and is independent of the exact nature of the hidden units, *i.e.* continuous or discrete.

As 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 $\xi_i^\mu = \pm 1$ and the output unit $y^\mu = \pm 1$ with equal probability, where $i = 1, \ldots, N$, $\mu = 1, \ldots, P$ and α is defined as P/N. The question is to calculate as a function of K, the maximal number of patterns P_c that can be taught to the network in the thermodynamic limit.

The case $K = l = 1$ is exactly the perceptron with binary weights. This problem has been raised by Gardner and Derrida [3], who have given the replica-symmetric (RS) solution and an upper limit for its validity, $\alpha_c = 4/\pi$. Krauth and Mezard [5] recognized that the disappearance of the entropy defines the critical capacity in that model, and found from the RS solution $\alpha_c \simeq 0.83$ [6], where the zero-temperature entropy vanishes. There the RS ansatz gives way to a solution with one-step replica symmetry breaking (RSB), which appears to be the globally stable one. This case with $l > 1$ was examined under RS assumption in ref. [7]. The case of binary weights with $K \geqslant 2$ was recently investigated by Barkai and Kanter [8], who found that, for any $K \geqslant 2$, $\alpha_c = 1$.

The main results of this work are: *a*) For any $K \geqslant 2$, there is a critical synaptic depth l_0 such that for $l \leqslant l_0$, $\alpha_c = \log_2(2l)$, which is the maximal theoretical capacity where the capacity per bit is 1. *b*) For $l \leqslant l_0$ and $\alpha < \log_2(2l)$, the entropy of the system is equal to $(N \log_2(2l) - P) \ln 2$ and the order parameter q (which will be defined below) is zero. *c*) For $l > l_0$, one-step RSB appears at α_0, where the overlap q_1 jumps to a positive value which is a decreasing function of l. At the maximal capacity the entropy vanishes and $q_1 < 1$, but in a two-step solution $q_2 = 1$. For large l, the capacity per bit decreases logarithmically to zero. *d*) Our results strongly indicate that $l_0 \propto \log K$, for large K. *e*) At finite temperature and $l \leqslant l_0$, there is a critical line, $\alpha_c(T_c, l)$, where the entropy of the system vanishes. For $T > T_c$, $q = 0$, where below it one-step RSB appears, $q_0 = 0$, $q_1 = 1$ and $m = T/T_c$.

Let us first indicate that the upper bound for the capacity of the system is $\alpha_c = \log_2(2l)$. One can obtain this upper bound by using the annealed approximation or by the following simple argument of information theory. Each weight is an integer belonging to the range $[-l, l]$, which can be represented by $\log_2(2l)$ bits. Hence, each weight contains exactly $\log_2(2l)$ bits of information. Nevertheless, the exact maximal capacity and the nature of the system are investigated using the method suggested by Gardner and Derrida [3]. For each set of coupling the energy is defined as the number of patterns which are not embedded

$$E(J) = \sum_{\mu=1}^{P} \Theta\left(-y^\mu \prod_{l=1}^{K} \boldsymbol{J}_l \cdot \xi_l^\mu / \sqrt{N}\right). \tag{2}$$

Note that the symmetry of the energy function, eq. (2), is the same as Hamiltonian systems with multi spin interactions. This observation suggests that a second-order phase transition and a first-order phase transition is expected for $K = 2$ and $K > 2$, respectively, which is indeed verified below. The partition function of eq. (2) in the zero-temperature limit is exactly the volume, V, which achieves the desired output and the entropy per weight is $\ln(Z)/N$. The patterns can be learnt if and only if the zero-temperature internal energy vanishes. 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 $\{\xi_i^\mu, y^\mu\}$, using the replica method. In the calculations one introduces a set of order parameters

$$q_l^{\alpha\beta} = \frac{K}{N} \boldsymbol{J}_l^\alpha \cdot \boldsymbol{J}_l^\beta , \tag{3}$$

where $|q_l^{\alpha\beta}| \leqslant 1$. The physical meaning of $q_l^{\alpha\beta}$ is the overlap between the weights belonging to the l-th hidden unit in two different replicas, α and β. Under the RS assumption one can show that

$$G_{\mathrm{RS}} = \langle\!\langle \ln Z \rangle\!\rangle = \mathrm{ext}_{\{q,\phi\}} \Bigg\{ -\frac{1}{2}\phi(1-q) + \int_{-\infty}^{\infty} \mathrm{D}z \ln\left(2 \sum_{j=1}^{l} \exp[(\phi(q-1)/2)(j^2-1)] \cosh(j\sqrt{\phi}\, z) \right) + \\ + \alpha \int_{-\infty}^{\infty} \prod_l \mathrm{D}t_l \ln \left\{ \exp[-\beta] + (1-\exp[-\beta]) \left[\mathrm{Tr}_{\{\tau_l = \pm 1\}} \prod_l H(Q\tau_l t_l)\, \Theta\left(\prod_l \tau_l\right) \right] \right\} \Bigg\} , \tag{4}$$

where the symbol $\langle\!\langle \ldots \rangle\!\rangle$ stands for the average over the patterns, $Q = (q/(1-q))^{1/2}$, $\mathrm{D}t = \exp[-t^2/2]\,\mathrm{d}t/(2\pi)^{1/2}$, $H(x) = \int_x^\infty \mathrm{D}x$ and q and ϕ are determined by the saddle point (SP) equations. Note that the diagonal order parameter, $q^{\alpha\alpha}$, disappears in eq. (4) after the rescaling $q^{\alpha\beta} \to q^{\alpha\beta} q^{\alpha\alpha}$.

At $T = 0$ and 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, α_0 is independent of l and equal to $\pi^2/8$. For the case $K = 3$ and $l = 2$, $\alpha_0 \sim 5.7$, where as $l \to \infty$ the result of the continuous case, $\alpha_0 = 6.5$, is expected. *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 $\alpha_c \sim 3.53$.

The upper bound for the maximal capacity, $\alpha_c = \log_2(2l)$, 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 volume of the solutions is a decreasing function of the capacity. Therefore, α where the entropy vanishes is defined as the maximal capacity. This behaviour was recently observed in the binary perceptron case ($K = 1$) [5] and in the discussed architecture with binary weights, $\alpha_c \sim 0.83$ and 1, respectively. Therefore, RS solution suggests that for $l < 2^{\alpha_c(K)-1}$, the maximal capacity is $\alpha_c = \log_2(2l)$, where $\alpha_c(K) \propto K^2$ for large K. This result violates the upper bound and furthermore the maximal capacity is bounded from above by the capacity of the network with continuous weights, where $\alpha_c \sim \log_2 K$ for large K [9]. Hence, a dramatic effect of RSB is expected, where an interplay between the number of hidden units K and the synaptic depth l occurs.

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

$$G_{1\,\mathrm{step}}(\phi_0, \phi_1, q_0, q_1, m) = \frac{1}{2}C + \frac{1}{m} \int_{-\infty}^{\infty} \mathrm{D}z \ln\left[\int_{-\infty}^{\infty} \mathrm{D}t \cdot \left(\sum_{j=1}^{l} \exp[C(j^2-1)/2]\, 2\cosh(j(\sqrt{\Delta\phi}\, t) + \sqrt{\phi_0}\, z) \right)^m \right] + \\ + \frac{\alpha}{m} \int_{-\infty}^{\infty} \prod_l \mathrm{D}t_l \ln \left\{ \int_{-\infty}^{\infty} \prod_l \mathrm{d}z_l \left(\exp[-\beta] + (1-\exp[-\beta]) \left[\mathrm{Tr}_{\{\tau_l = \pm 1\}} \prod_l H\left[\tau_l \left(\sqrt{\frac{q_0}{1-q_1}}\, t_l + \sqrt{\frac{\Delta q}{1-q_1}}\, z_l \right) \right] \Theta\left[\prod_l \tau l \right] \right] \right)^m \right\} , \tag{5}$$

where $C \equiv \phi_1 q_1(1-m) - \phi_1 + m\phi_0 q_0$, $\Delta\phi \equiv \phi_1 - \phi_0$ 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_1 represents the self-overlap within one valley and m relates to the Gibbs weights P_α, $1 - m = \langle\langle \sum_\alpha P_\alpha^2 \rangle\rangle$. The five order parameters must be computed through the SP equations. In the cases of binary weights and continuous weights it was found that $q_0 = \phi_0 = 0$ [8, 9] and therefore it obviously holds in the intermediate case of arbitrary depth. In this case and for $K = 2$, a perturbation expansion of the SP equations around α_0 gives that the system undergoes a second-order phase transition to a RSB phase, where at the transition q_1 and $m \to 0$. At the critical capacity for the case $l = 2$, for instance, the entropy vanishes at $\alpha_c \sim 1.97$, where $q_c \sim 0.86$. Hence, it is clear that $\alpha_c(l)$ is an increasing function of l, from 1.97 at $l = 2$ up to 4.06 as $l \to \infty$. For the cases $K \geq 3$, a first-order transition as a function of α is expected, where q_1 is discontinuous, but the free energy is still continuous, since at the transition $m \to 1$. The solutions of the SP equations in this limit for the case $K = 3$, 4, 5 and $2 \leq l \leq 14$ are carried out numerically and are summarized in tables I-III.

Tables I-III indicate that $\alpha_0(K, l) > \log_2(2l)$ for $l \leq l_0$, where $l_0 = 3$, 5, and 12 for $K = 3$, 4 and 5, respectively. Hence, for $l \leq l_0(K)$ the maximal theoretical capacity is achieved, $\log_2(2l)$, the average overlap between different solutions is zero, and the entropy decreases linearly with α, $S = N(\log_2(2l) - \alpha)\ln 2$.

TABLE I. – *Results for continuous weights.*

K	α_0	α_c	$(\alpha_0 - \alpha_c)/\alpha_c$
1	0	2.00	1.0
2	1.28	4.06	0.69
3	3.18	5.0	0.36
4	3.95	5.61	0.29
5	5.12	6.06	0.15

TABLE II. – *Results for* $K = 3$.

l	q	α_0	$\log_2(2l)$
2	0.998	2.22	2.0
3	0.948	2.63	2.58
4	0.939	2.73	3.0
∞	0.929	3.18	

TABLE III. – *Results for* $K = 4$ *and* 5.

K	l	α_0	$\log_2(2l)$
4	2	2.57	2.00
4	4	3.35	3.00
4	5	3.44	3.32
4	6	3.49	3.58
5	7	4.56	3.80
5	9	4.61	4.17
5	12	4.64	4.58
5	14	4.66	4.81

For $l > l_0$, a transition to one-step RSB phase occurs at α_0, since $\alpha_0 < \log_2(2l)$. At the transition q_1 jumps discontinuously and increases as a function of α up to α_c, where the entropy vanishes. For the cases $K = 2$, $l = 2$, 3 and $K = 3$, $l = 5$, for instance, $\alpha_c \sim 1.97$, 2.47, 3.28 and $q_c \sim 0.86$, 0.895, 0.96, respectively. It is clear that $\alpha_c(l)$ is bounded from above by the maximal capacity of the continuous case, α_c(cont). Nevertheless, the exact calculation of α_c for $l > l_0$ is much more difficult, since the effect of two-step RSB is not negligible for the continuous case [10], and for the discrete case at finite temperature as will be explained below. It is also physically expected that as the entropy vanishes there is a complete freezing in each one of the valleys and therefore the upper plateau of $q(x)$ is 1. Nevertheless, our results indicate that it is constructive to consider besides the *capacity per synapse* also the *capacity per bit* of information stored in the synapses. For $l \leqslant l_0$, the capacity per embedded bit is 1 and q and S are given by a simple form. In contrast, for $l > l_0$ the capacity per embedded bit is less than or equal to 1 and q and S are not simple functions of α. Note that for $\log_2(2l) > \alpha$(cont), the capacity per bit is *less* than 1 and decreases logarithmically as a function of l towards zero as $l \to \infty$.

An important question is what is the interplay between l_0 and K. A direct investigation of the nature of α_0 as a function of K is a difficult numerical and theoretical task. Nevertheless, a preliminary question is whether α_0(cont) scales as α_c with $\log_2 K$ for large K. In table I the results for α_0 and α_c for $K = 2$, 3, 4 and 5 indicate that the fraction of the capacity where RSB occurs, $(\alpha_c - \alpha_0)/\alpha_c$, decreases with K. It is strongly suggested that in the leading order as $K \to \infty$, $(\alpha_c - \alpha_0)/\alpha_c$ decreases to zero (or to a small constant). Therefore, α_0(cont) is also proportional to $\log_2 K$ for large K. On the other hand, if the solution's points for a fixed l are distributed homogeneously in the solution's space of the continuous case, it is expected that l_0 is fixed by the maximal integer which obeys the inequality, $\log_2(2l) \leqslant \alpha_0$(cont). For small K and l, the resolution of the weights is rough and fluctuations in the homogeneous distribution drives the quantity $\log_2(2l_0)/\alpha_0$(cont) below 1. However as K increases, it is expected that this quantity goes to 1. One can verify from tables I-III that this is indeed our case where $\log_2(2l_0)/\alpha_0(\text{cont}) = 0.70$, 0.84, 0.87 and 0.90 for $K = 2$, 3, 4 and 5, respectively. In conclusion, our results suggest that l_0 also scales with $\log_2 K$. For large K, the capacity per bit is 1 up to $\log_2 K$, where above it the capacity decreases logarithmically with l.

Another conclusion from table II is that for a fixed K, $q_1(\alpha_0)$ is a decreasing function of l. This behaviour of q_1 can be explained in the following way. Assume that the solution's points for finite l are distributed homogeneously in the solution's space of the continuous case. Since the entropy $\propto \log_2(2l)$, as l decreases, the number of solution's points decrease exponentially and the distance between them grows (*i.e.* the energy barrier grows). Therefore, it is expected that at the transition to the RSB phase, each valley of the continuous case splits into many disconnected small valleys, and the average overlap within each one of the valleys increases as l becomes smaller. Tables I-III also indicate that for a fixed l, α_0 increases with K. This fact could be explained, since as K increases the number of legal internal representations for all the patterns is $2^{P(K-1)}$. This picture suggests that valleys which are disconnected for a given K are connected for larger K, due to the additional freedom in the internal representation.

In the case $\alpha_0(l) > \log_2(2l)$ the entropy vanishes at $\log_2(2l)$. Nevertheless, it is interesting to know the behaviour at finite temperature for $\alpha > \log_2(2l)$. In this case we did not find any solution except the RS solution, $q_0 = q_1 = 0$ and a one-step RSB solution in the limit $q_1 \to 1$. Following ref. [5], the limit $q_1 = 1$ at finite temperature is studied. In this limit one can verify that $\phi_1 \to \infty$ and

$$G_{1\,\text{step}}(q_0, \phi_0, 1, \infty, m, \beta) = \frac{1}{m} G_{\text{RS}}(q_0, m^2\phi_0, \beta m), \tag{6}$$

where G_{RS} is given in eq. (4). The SP 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 SP with respect to m gives $S_{RS} = S_{1step} = 0$, where βm must be equal to $1/T_c$, which is the inverse temperature where the RS entropy vanishes. This temperature as a function of α is given by $\alpha_c(T) = -\ln(2l)/\ln((1+\exp[-\beta])/2)$, where as $T \to 0$, $\alpha_c \to \log_2(2l)$. For $T > T_c$, the only solution is the paramagnetic solution, $q_0 = q_1 = 0$. On the transition line, $T_c(\alpha)$, the entropy vanishes. At $T < T_c$, there exists a 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 [11]. In order to check whether the one-step solution is exact, a solution with two-step RSB is examined and gives in the limit $q_2 = 1$ and $\phi_2 \to \infty$, $G_{2steps}(q_0, \phi_0, q_1, \phi_1, 1, \infty, m_1, m_2, \beta) = G_{1steps}(q_0, m_2^2\phi_0, q_1, m_2^2\phi_1, m_1/m_2, \beta m_2)/m_2$. An optimization of G_{2steps} with respect to m_1 is equivalent to an optimization of the one-step solution with respect to m, and hence the two-step solution coincides with the one-step solution in the region $\alpha < \alpha_0(K, l)$ and $l \leq l_0(K)$. For $\alpha < \alpha_0(K, l)$ the trivial RS solution $q = 0$ holds. For $\alpha > \alpha_0(K, l)$ and $l > l_0(K)$, there is a solution with $q_2 = 1$, $m_2 = T/T_c$ and q_1 and m_1 are given by the one-step solution. This solution is more physically sensible than the one-step solution, since as the entropy vanishes the upper plateau $q_2 = 1$.

Finally, we would like to comment that preliminary results also indicate that the behaviour of the system depends only on the number of degrees of freedom per weight. For instance, the case $K = 3$, $J = \pm 1, \pm 5$ gives $\alpha_0 \sim 2.11$. On the other hand, the case $J = \pm 1, \pm 10$ gives $\alpha_0 \sim 1.96$, which is due to strong fluctuations at small K, since for $K = 4$ it gives $\alpha_c \sim 2.27$, which is greater than 2. This behaviour is expected, since one can transform the new discrete values into the previous ones by remembering that each ± 10 is actually represented by ± 2.

REFERENCES

[1] MINSKY M. L. and PAPERT S., *Perceptron* (MIT Press, Cambridge, Mass.) 1969.

[2] GARDNER E., *J. Phys. A*, 21 (1988) 257.

[3] GARDNER E. and DERRIDA D., *J. Phys. A*, 21 (1988) 271.

[4] BOSER B. and SOLLA S. A., private communication.

[5] KRAUTH W. and MEZARD M., *J. Phys. (Paris)*, 50 (1989) 3057.

[6] For review, see MEZARD M., PARISI G. and VIRASORO M. A., *Spin Glass Theory and Beyond* (World Scientific, Singapore) 1987.

[7] GUTFREUND H. and STEIN Y., *J. Phys. A*, 23 (1990) 2613.

[8] BARKAI E. and KANTER I., *Europhys. Lett.*, 14 (1991) 107.

[9] BARKAI E., HANSEL D. and KANTER I., *Phys. Rev. Lett.*, 65 (1990) 2321.

[10] KANTER I., unpublished.

[11] DERRIDA B., *Phys. Rev. B*, 24 (1981) 2613.