Statistical Mechanics of a Multilayered Neural Network

E. Barkai,[1] D. Hansel,[2,a] and I. Kanter[1]
[1]*Department of Physics, Bar-Ilan University, Ramat Gan 52100, Israel*
[2]*Racah Institute of Physics, Hebrew University, Jerusalem, Israel 91904*
(Received 11 July 1990)

Statistical mechanics is applied to estimate the maximal information capacity per synapse (α_c) of a multilayered feedforward neural network, functioning as a parity machine. For a large number of hidden units, K, the replica-symmetric solution overestimates dramatically the capacity, $\alpha_c \propto K^2$. However, a one-step replica-symmetry breaking gives $\alpha_c \sim \ln K/\ln 2$, which coincides with a theoretical upper bound. It is suggested that this asymptotic behavior is exact. Results for finite K are also discussed.

PACS numbers: 87.10.+e, 05.50.+q, 64.60.Cn

Analogies between networks of formal neurons and random magnetic spin systems have suggested the use of statistical mechanics in the study of properties of neural networks.[1] Among these systems, perceptron networks play a central role.[2] The prototype of this class of architectures is the one-layered perceptron,[2] consisting of one input layer of N binary units and one binary output unit. Gardner, in her pioneering work, has demonstrated that a statistical-mechanics approach can be helpful for studying properties of this simplest perceptron. She was able to rederive the already known result of the information maximal capacity,[3] introducing a general framework allowing a systematic study of this kind of system.[4,5] This approach has been recently applied to study various properties of the one-layer perceptron.[6]

However, as is well known, the computational power of such a 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 larger the information capacity *per synapse* (i.e., per weight) is in multilayer systems compared with the one-layer perceptron.

The problem of multilayered networks has been addressed by Baum[7] and Mitchison and Durbin,[8] where bounds for the information capacity were obtained using geometrical methods. In this work, a two-layer feedforward network is studied using the statistical-mechanics approach. The architecture of the network consists of N binary input units, one hidden layer with K continuous units (because of the particular internal representation of the studied problem, see below, the results are independent of the exact nature of the hidden units, i.e., continuous or discrete) and a single binary output unit. The input units are divided into K disjoint sets, each one of them consisting of N/K units. The lth hidden unit is connected only to the ith input via the weight J_i 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 (see Fig. 1).

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

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

where $\mathbf{J}_l$ and $\mathbf{s}_l$ 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

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

This network is known as a parity machine,[8] since the output is the parity of the internal representation of the hidden units for a given set of weights and a state of the input layer. (Strictly speaking, this architecture is a two-layered network only where the output unit is a Σ-Π unit.)

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 μth pattern consists of

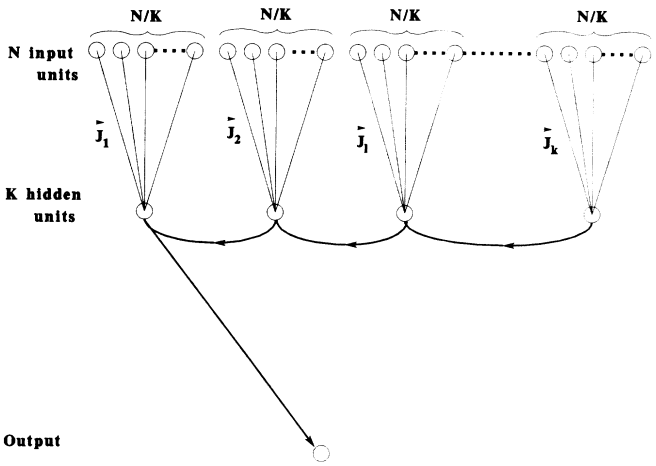

FIG. 1. The architecture of a parity machine with nonoverlapping receptive fields with N input units and K hidden units.

$\xi_i^\mu = \pm 1$ with equal probability, where $i = 1, \ldots, N$ and $\mu = 1, \ldots, P$. The desired output of the μth 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 taught 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. For the case $K = 1$, the network is exactly the perceptron problem and $P_c = 2N$.

Following Gardner's method, the problem is formulated in the statistical framework as follows. For a given realization of the P patterns, the normalized fractional volume in the weight space occupied by the networks which achieve the desired output is given by

$$V = \int_{-\infty}^{\infty} \prod_{i=1}^{N} dJ_i \prod_l \delta(\mathbf{J}_l \cdot \mathbf{J}_l - N/K) \prod_\mu \Theta\left(y^\mu \prod_l \mathbf{J}_l \cdot \xi_l^\mu\right). \tag{3}$$

In the computation one concentrates on $\ln V$ which in the thermodynamic limit is an extensive quantity. This quantity is averaged over the quenched distribution of the random input and output patterns, $\{\xi_i^\mu, y^\mu\}$, using the replica method.[9] In the calculations one introduces a set of order parameters

$$q_l^{\alpha\beta} = (K/N)\mathbf{J}_l^\alpha \cdot \mathbf{J}_l^\beta, \tag{4}$$

where $|q_l^{\alpha\beta}| < 1$. The physical meaning of $q_l^{\alpha\beta}$ is the overlap between the weights belonging to the lth 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. The case of a fully connected parity machine, where *a priori* $\ln V$ depends on more than one type of order parameter, was recently investigated by Mezard and Patarnello[10] within the replica-symmetric (RS) ansatz.

Under the RS assumption, $q_l^{\alpha\beta} = (1-q)\delta_{\alpha\beta} + q$ (q is independent of l, since after the average over the disorder, all the hidden units are equivalent), one can show that

$$\langle\langle \ln V \rangle\rangle = \mathrm{ext}_{\{q\}} \left\{ \tfrac{1}{2} \ln(1-q) + \frac{q}{2(1-q)} + \alpha \int_{-\infty}^{\infty} \prod_l Dt_l \ln\left[\mathrm{Tr}_{\{\tau_l = \pm 1\}} \prod_l H(Q\tau_l t_l) \Theta\left(\prod_l \tau_l\right)\right]\right\}, \tag{5}$$

where $\alpha = P/N$, $Q = [q/(1-q)]^{1/2}$, $Dt = e^{-t^2/2} dt/(2\pi)^{1/2}$, $H(x) = \int_x^\infty Dx$, and q is determined by the saddle-point (SP) equation. For each number of hidden units, K, one can distinguish between four regimes of α: (a) The region $0 < \alpha < \alpha_0$ is characterized by $q = 0$. For the cases $K = 2$, 3, and 4, $\alpha_0 = \pi^2/8$, 6.5, and ~ 12, respectively. (b) For $K = 2$ the system undergoes a second-order phase transition at α_0, where near α_0, $q \sim 4/\pi^2(\alpha - \alpha_0)$. For $K > 2$ the transition at α_0 is a first-order phase transition. At the transition for $K = 3$ and 4, numerical solutions of the SP equation give $q \simeq 0.51$ and 0.85, respectively. (c) For $\alpha_0 < \alpha < \alpha_c$, q increases with α up to the maximal capacity α_c where $q = 1$. Quantitatively, the critical capacity is fixed by the following equation:

$$\alpha_c^{-1}(K) = \int_{-\infty}^{\infty} Dy\, y^2. \tag{6}$$

For the case $K = 2$, $\alpha_c \simeq 5.5$ and for large K, $\alpha_c \propto K^2$. (d) The RS solution is locally stable for $K = 2$ up to α_0, whereas for $K > 2$ the RS solution is stable up to α_c. The details of the derivation of these results and the full dependence of α as a function of K and q will be given elsewhere.[11]

Result (a) is remarkable, since in general it is expected that as α increases correlations are built among different solutions. This is indeed the case of the one-layer perceptron, where q is positive for any finite α. However, the parity-machine network is different from the one-layer perceptron in two aspects. The first difference is that each pattern has 2^{K-1} legal internal representations of the hidden units, which gives $2^{P(K-1)}$ legal internal representations for all the patterns. The second difference is that the state of the output unit is invariant under 2^{K-1} global symmetries. Each global symmetry consists of the transformation $J_l \to -J_l$ for an *even* number of hidden units. Note that the freedom in the choice of the internal representation is common to all multilayered networks. However, the discussed global symmetries characterize especially the parity-machine network. Hence, for small α the fractional volume of the solution is enlarged, in comparison to a simple perceptron, by the free choice of the internal representation. For $\alpha < \alpha_0$ the different solutions in the weight space are connected, including regions related to each other by global symmetries, which leads to the existence of a paramagnetic phase with $q = 0$. For large enough α, $\alpha > \alpha_0$, the ergodicity is broken and a transition to a spin-glass phase occurs.

For $K > 3$ the system undergoes a first-order phase transition which can be understood by the symmetry of the cost function [see Eq. (3)]. This quantity can be expressed in the form

$$\prod_\mu \Theta\left(y^\mu \sum_{i_1, \ldots, i_K} \mathbf{J}_{i_1} \cdots \mathbf{J}_{i_K} \xi_{i_1}^\mu \cdots \xi_{i_K}^\mu\right),$$

where J_{i_l} is a weight connected to the lth hidden unit. For $K > 3$ the symmetry of the discussed system reminds one of the symmetry of Hamiltonian systems with multispin (soft) interactions where a first-order transition is found.[12]

The replica-symmetric prediction that $\alpha_c \propto K^2$ for large K is certainly wrong, since an upper bound of $\ln K/\ln 2$ for α_c is obtained by a straightforward generalization of the method used by Mitchison and Durbin[8] to our

nonoverlapping architecture. The failure to give an estimate for the maximal capacity compatible with this upper bound is due to a strong replica-symmetry-breaking (RSB) effect. In the following, the solution within a one-step RSB is discussed.

The one-step RSB solution is defined by three order parameters, m, q_0, and q_1. For $0<x<m$, $q(x)=q_0$ and for $m<x\leq 1$, $q(x)=q_1$. The averaged $\ln V/N$ in this framework is given by

$$\langle\langle \ln V\rangle\rangle = \mathrm{ext}_{\{q_0,q_1,m\}}\left[\frac{1+(m-1)\Delta q}{2A}+\frac{m-1}{2m}\ln(1-q_1)+\frac{\ln A}{2m}+\frac{\alpha}{m}\int_{-\infty}^{\infty}\prod Dz_l \ln\langle X^m\rangle\right], \tag{7}$$

where $\Delta q = q_1 - q_0$, $A = 1-q_1+m\Delta q$, and in general $\langle f(X)\rangle$ is defined by

$$\langle f(X)\rangle = \int_{-\infty}^{\infty}\prod_l Dt_l f\left[\mathrm{Tr}_{\{\tau_l=\pm 1\}}\prod_l H\left(\tau_l\left\{\left[\frac{q_0}{1-q_1}\right]^{1/2} z_l+\left[\frac{\Delta q}{1-q_1}\right]^{1/2} t_l\right\}\right)\Theta\left(\prod_l \tau_l\right)\right]. \tag{8}$$

The three order parameters are determined by the three SP equations with respect to q_0, q_1, and m.

One first examines the properties of the solution to these three equations near the transition to the RSB phase for the cases $K=2$ and 3 and then the critical capacity will be studied as a function of K.

In the case of $K=2$, the RS solution is unstable for $\alpha>\alpha_0=\pi^2/8$ and a transition to the RSB phase is expected. A perturbative expansion of SP equations around α_0 gives the result that the system undergoes a second-order phase transition to a RSB phase which bifurcates continuously from the RS solution. This phase is characterized by $q_0=0$, and the first terms of the expansion of q_1 and m in the vicinity of α_0 are

$$q_1 = 4/\pi^2(\alpha-\alpha_0)+16/3\pi^4(72/\pi^2-1)(\alpha-\alpha_0)^2$$

and $m=4/\pi^2(1+16/\pi^2)(\alpha-\alpha_0)$. Note that q_1 differs from the RS solution only at the second order in $\alpha-\alpha_0$. As α increases above α_0, the numerical solutions of the three SP equations indicate that $q_0=0$ (which is always a solution), q_1 increases to 1, and m increases up to some maximal value and then decreases to zero. The behavior of m is similar to the one-step RSB solution of Ising spin glass.[13]

In the case of $K=3$, the transition to the RSB phase occurs at $\alpha_{\mathrm{RSB}}\simeq 3.2$, whereas the RS solution with $q\neq 0$ appears above $\alpha_0\simeq 6.15$. At the transition one can find numerically that $m=1$ and $q_1\simeq 0.93$, where $q_0=0$ is a solution for any α. This transition is first order in the sense that q_1 is discontinuous, but the averaged $\ln V$ (free energy) is continuous at α_{RSB}, since $\int q(x)dx \propto \alpha-\alpha_{\mathrm{RSB}}$. A similar behavior was found at the transition to the glassy low-temperature phase of Potts glass systems[14] and spin glasses with multispin interactions.[15,16] As was explained above, the symmetry of the examined systems is similar to those systems and such a discontinuous transition is expected. Since the RS solution is stable around α_{RSB}, the preferred phase is the one which minimizes $\ln V$. The comparison between these two phases was carried out and indeed gives the result that the preferred phase above α_{RSB} is the one-step solution. As α increases above α_{RSB}, $q_0=0$, $q_1\to 1$, and m decreases to zero. The full curves of m and q_1 as a function of α and K will be given elsewhere.

Let us concentrate now on the limit $\alpha\to\alpha_c$, where $q_1\to 1$, $m\to 0$, and $m/(1-q_1)\equiv c$, where c is of $O(1)$. This scaling of m and $1-q_1$ is found in various spin-glass systems and was confirmed numerically in our problem for the cases $K=2$ and 3 by solving numerically the SP equations near α_c. The averaged $\ln V$ in this limit is given by

$$\langle\langle \ln V\rangle\rangle = \frac{1}{2m}\left[m\ln(m/c)+\ln(1+c)-2\alpha\ln 2+2\alpha\ln\left(1+2K\int_0^{\infty}\frac{dt}{\sqrt{2\pi}}e^{-(t^2/2)(1+c)}[2H(t)]^{K-1}\right)\right]. \tag{9}$$

The extremum of this equation with respect to c and m gives, for $K=2$ and 3, $\alpha_c\simeq 4.06$ and 5, respectively (corresponding to $c\simeq 14.3$ and 67.2). The result of the critical capacity for K up to 50 is given in Fig. 2, which indicates that for large K, α_c scales with $\ln K$. It is remarkable that this *one-step* calculation is compatible with the bound of Ref. 8.

In the large-K limit (but $K/N\to 0$), Eq. (9) gives

$$\langle\langle \ln V\rangle\rangle = \frac{1}{2m}[m\ln(m/c)+\ln(1+c)-2\alpha\ln 2+2\alpha\ln(1+K/\sqrt{c})] \tag{10}$$

and indeed one can obtain that in the leading order the critical capacity for large K is given by

$$\alpha_c(K\to\infty)=\ln K/\ln 2, \tag{11}$$

whereas $c\simeq(K\ln K/\ln 2)^2$.

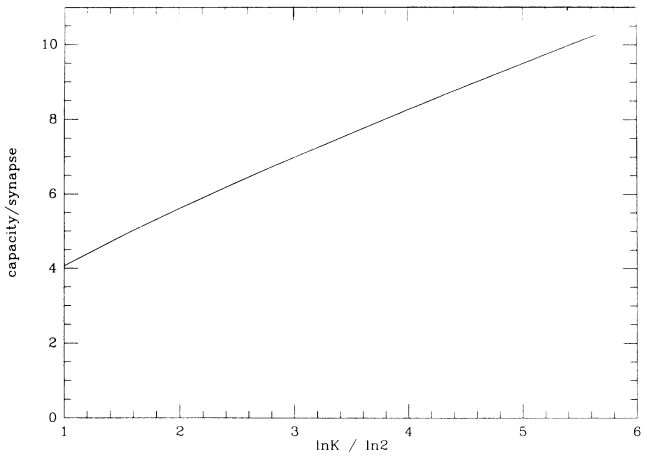

FIG. 2. The maximal capacity per synapse as a function of $\ln K/\ln 2$.

This result coincides with the upper bound for α_c one can obtain by a generalization of the geometric method used in Ref. 8. An important question is whether the one-step solution is exact in this limit or whether it is just a good approximation, where the exact α_c is less than $\ln K/\ln 2$. The stability analysis of the one-step solution is not necessarily a good criterion, since the transition as a function of α is in some sense a first-order transition. Hence, the procedure to calculate the exact value of α_c is to minimize $\langle\langle \ln V/N \rangle\rangle$ in the general framework of Parisi's solution, where α_c is fixed for each parametrization. This goal certainly deserves further research. Nevertheless, we think that Eq. (11) is an exact result for the following reasons. At the transition to the RSB phase, the state in each one of the valleys is almost frozen even for finite K, since $q_1 \simeq 0.93$ for $K=3$. It is expected that the self-overlap q_1 is an increasing function of K, and in the large-K limit, $q_1 \to 1$. Furthermore, some similarities between this problem in the large-K limit and the random-energy model and the simplest spin glass[15,16] suggest that the one-step solution is exact. Finally, it is reasonable to think that for the fully connected parity machine the capacity per synapse is also given by (11) and that the SP solution has only one nonzero order parameter, q_1.

Simulations in the case $K=2$ and with N up to 1000 were carried out using the least-action algorithm,[8] where the results were averaged over at least fifty samples. It was found that $t_{\rm av}^{-1/2} \propto \alpha$, for $\alpha > 1.5$, where $t_{\rm av}$ is the average convergence time. This scaling of convergence time was also obtained for the perceptron case. A finite-size analysis of α_c, where $t_{\rm av}$ diverges, suggested that $\alpha_c \simeq 3.2$. It was also found that the algorithm is inefficient to learn more than one-half of the samples at the value ~ 3.2. The difference between this result and the one-step solution where $\alpha_c \simeq 4$ could be a consequence of the fact that the exact form of the order parameter is a more structured function than a one step. Another source of this discrepancy is the fact that the measured α_c is the maximal capacity of this particular algorithm.

The existence of the paramagnetic phase and the nature of the transition to the spin-glass phase have been confirmed in preliminary simulations for $K=2$ and 3. The more structured RSB phase is currently under analytical investigation.

We would like to thank Professor H. Sompolinsky for many useful discussions and encouragement. One of us (D.H.) thanks Professor S. Amari for bringing the paper of Mitchison and Durbin to his attention. The research of I.K. is supported by the Israel Ministry of Science and Development.

[a]Member of CNRS, on leave from the Centre de Physique Théorique, Ecole Polytechnique, 91128 Palaiseau, France.

[1]For a review, see D. J. Amit, *Modeling Brain Function* (Cambridge Univ. Press, New York, 1989).

[2]M. L. Minsky and S. Papert, *Perceptron* (MIT Press, Cambridge, 1969).

[3]T. Cover, IEEE Trans. Electron. Comput. **14**, 326 (1965).

[4]E. Gardner, J. Phys. A **21**, 257 (1988).

[5]E. Gardner and D. Derrida, J. Phys. A **21**, 271 (1988).

[6]See, for instance, the memorial volume of Elisabeth Gardner, J. Phys. A **22** (1989).

[7]E. Baum, J. Complexity **4**, 193 (1988).

[8]G. J. Mitchison and R. N. Durbin, Biol. Cybern. **60**, 345 (1989).

[9]S. Kirkpatrick and D. Sherrington, Phys. Rev. B **17**, 4384 (1978).

[10]M. Mezard and S. Patarnello (unpublished).

[11]E. Barkai, D. Hansel, and I. Kanter (unpublished).

[12]E. Gardner, Nucl. Phys. **B257**, 747 (1985).

[13]G. Parisi, J. Phys. A **13**, 1101 (1980).

[14]D. J. Gross, I. Kanter, and H. Sompolinsky, Phys. Rev. Lett. **55**, 304 (1985).

[15]B. Derrida, Phys. Rev. B **24**, 2613 (1981).

[16]D. J. Gross and M. Mezard, Nucl. Phys. **B240**, 431 (1984).