Multilayer Neural Networks with Extensively Many Hidden Units

Michal Rosen-Zvi,[1] Andreas Engel,[2] and Ido Kanter[1]

[1]*Minerva Center and Department of Physics, Bar-Ilan University, Ramat-Gan, 52900 Israel*

[2]*Institut für Theoretische Physik, Otto-von-Guericke Universität, PSF 4120, 39016 Magdeburg, Germany*

(Received 21 March 2001; revised manuscript received 23 May 2001; published 25 July 2001)

The information processing abilities of a multilayer neural network with a number of hidden units scaling as the input dimension are studied using statistical mechanics methods. The mapping from the input layer to the hidden units is performed by general symmetric Boolean functions, whereas the hidden layer is connected to the output by either discrete or continuous couplings. Introducing an overlap in the space of Boolean functions as order parameter, the storage capacity is found to scale with the logarithm of the number of implementable Boolean functions. The generalization behavior is smooth for continuous couplings and shows a discontinuous transition to perfect generalization for discrete ones.

DOI: 10.1103/PhysRevLett.87.078101 PACS numbers: 84.35.+i

Statistical mechanics investigations of artificial neural networks continue to play a stimulating role in the scientific dialogue between disciplines as diverse as neurophysiology, mathematical statistics, computer science, and information theory. In particular, the study of feed-forward neural networks pioneered by Gardner [1] has revealed a variety of interesting results on how these systems may learn different tasks of information processing from examples (for a review, see [2]). Of particular importance in this respect are multilayer networks (MLN) which are able to implement any function between input and output [3]. This makes them attractive candidates for many practical applications.

Although it is well known that very many hidden units are needed in order to realize the computational power of MLN, statistical mechanics studies have so far been mostly restricted to systems with very few hidden units compared to the number of inputs [4]. In the present Letter we overcome this limitation and study the storage and generalization abilities of a tree MLN in which the size of the hidden layer scales in the same way as the input dimension.

We consider a MLN with N binary hidden units $\tau_i = \pm 1$, $i = 1, \ldots, N$ feeding a binary output $\sigma = \mathrm{sgn}(\sum_i J_i \tau_i)$ through a coupling vector $\mathbf{J} = J_1, \ldots, J_N$. The hidden units are determined via Boolean functions $\tau_i = B_i(\mathbf{S}_i)$ by disjoint sets of inputs $\mathbf{S}_i = S_{i1}, \ldots, S_{iL}$ containing L elements each. We are interested in the limit $N \to \infty$ with L remaining constant. In order to keep the connection with neural network architectures we restrict ourselves to *symmetric* Boolean functions characterized by $B_i(-\mathbf{S}_i) = -B_i(\mathbf{S}_i)$. There are $2^{2^{L-1}}$ such functions with L inputs.

In order to investigate the storage and generalization properties of the network it is necessary to quantify the flexibility in the input-output mapping when adapting the Boolean functions B_i and the couplings J_i. The statistical mechanics approach to this problem (for a detailed discussion, see [2]) considers a set of αLN *random* inputs $\boldsymbol{\xi}_i^\mu$, $\mu = 1, \ldots, \alpha LN$, the components $\xi_{i1}^\mu, \ldots, \xi_{iL}^\mu$ of which are independent, identically distributed random variables with zero mean and unit variance. One then determines the *typical* fraction of the phase space spanned by those Boolean functions and couplings that are able to map all inputs on outputs $\sigma^\mu = 1$. This typical fraction is given by $\exp(Ns)$ with the *quenched entropy*

$$s = \lim_{N\to\infty} \frac{1}{N} \left\langle\!\!\left\langle \int d\mu(\mathbf{J}) \operatorname*{Tr}_{\{B_i\}} \prod_{\mu=1}^{\alpha LN} \theta\left(\sum_i J_i B_i(\boldsymbol{\xi}_i^\mu) \right) \right\rangle\!\!\right\rangle_{\{\boldsymbol{\xi}_i^\mu\}} . \tag{1}$$

Here $d\mu(\mathbf{J})$ is the proper measure in the space of couplings $\mathbf{J}$ and the trace denotes the sum over all symmetric Boolean functions. The product in the integrand is nonzero only if the arguments of all the θ functions are positive and hence projects out the admissible part of the phase space.

The double angle in (1) stands for the average over the input distribution. To perform this average the replica method relying on the identity $\langle\!\langle \ln x \rangle\!\rangle = \lim_{n\to 0}[(\langle\!\langle x^n \rangle\!\rangle - 1)/n]$ may be used. Analytic progress can be made by employing standard techniques [2] and introducing an overlap between two solutions in the combined space of couplings $\mathbf{J}$ and Boolean functions B_i of the form

$$q^{ab} = \frac{1}{N} \sum_i J_i^a J_i^b \langle\!\langle B_i^a(\boldsymbol{\xi}) B_i^b(\boldsymbol{\xi}) \rangle\!\rangle_{\boldsymbol{\xi}} . \tag{2}$$

Exploiting the fact that the average in (2) is over a single, L-component vector $\boldsymbol{\xi}$ only and therefore involves only a finite number of terms and assuming replica symmetry, $q^{ab} = q$ for $a \neq b$, we find for the quenched entropy

$$s = \operatorname*{extr}_{q,\hat{q}} [G_C(q, \hat{q}) + G_S(\hat{q}) + \alpha L G_E(q)] . \tag{3}$$

The explicit expressions for the functions G_C, G_S, and G_E depend on L, the constraints on $\mathbf{J}$ and on whether the storage or the generalization problem is addressed.

Let us begin with the storage problem by asking for the storage capacity α_c defined as the maximal value of α for which the system can still realize all desired input-output mappings with probability 1. Performing the replica limit

with the number of replicas tending to zero characteristic for this problem we find

$$G_E(q) = \int Dt \ln H(Qt) \tag{4}$$

with the abbreviations $Dt = dt\, e^{-t^2/2}/\sqrt{2\pi}$, $H(x) = \int_x^\infty Dt$, and $Q = \sqrt{q/(1-q)}$. The expressions for G_C and G_S depend on the constraints on the coupling vector $\mathbf{J}$.

A particular simple case is given by Ising couplings $J_i = \pm 1$. From the symmetry of the Boolean functions it is clear that it is sufficient to consider $J_i = 1$ for all i. Consequently all flexibility of the network rests in the choice of the Boolean functions between input and the hidden layer and q is the sole overlap in the space of these Booleans. We then find $G_C = \hat{q}(1-q)/2$, where $\hat{q}$ denotes the conjugate order parameter to q. Moreover, using the identity

$$\operatorname*{Tr}_{\{B_i\}} \exp\left(\sqrt{\frac{\hat{q}}{2^{L-1}}} \sum_{\boldsymbol{\xi}} z_{\boldsymbol{\xi}} B_i(\boldsymbol{\xi})\right) = \prod_{\boldsymbol{\xi}} \left[2\cosh\left(\sqrt{\frac{\hat{q}}{2^{L-1}}}\, z_{\boldsymbol{\xi}}\right)\right] \tag{5}$$

with the sums and products over $\boldsymbol{\xi}$ running over all 2^{L-1} configurations of $\boldsymbol{\xi}$ with $\xi_1 = 1$ we get

$$G_S = 2^{L-1} \int Dz \ln\left[2\cosh\left(\sqrt{\frac{\hat{q}}{2^{L-1}}}\, z\right)\right]. \tag{6}$$

Under the transformations $\hat{q} \mapsto 2^{L-1}\hat{q}$ and $\alpha \mapsto 2^{L-1}\alpha/L$ the resulting expression for the entropy maps *exactly* on the result for the Ising perceptron corresponding to $L = 1$. We may therefore use the well-known results for this case [6]. The storage capacity is hence overestimated by the replica symmetric expression and the correct result

$$\alpha_c(L) = \alpha_c(1)\, 2^{L-1}/L \cong 0.83\, 2^{L-1}/L \tag{7}$$

is given by the value of α at which the entropy $s(\alpha)$ turns negative. The storage capacity is proportional to the *logarithm* of the number of implementable Boolean functions. This result is in accordance also with the rigorous upper bound $\alpha_c \le 2^{L-1}/L$ resulting from the annealed entropy $s^{\text{ann}} = (2^{L-1} - \alpha L)\ln 2$. As in the case of the Ising perceptron this bound is related to information theory. The full specification of the network with all $J_i = 1$ requires $N\, 2^{L-1}$ bits of information necessary to pin down the N Boolean functions B_i. Therefore the machine cannot store more than $N\, 2^{L-1}$ bits and α_c cannot exceed $2^{L-1}/L$.

Figure 1 compares the analytical result $\alpha_c(L=3) \cong 1.11$ with numerical simulations using exact enumerations. Even for the small sizes accessible to this numerical technique we find a steepening of the transition with increasing N and a crossing point of the curves close to the theoretical prediction.

It is possible to generalize the above analysis to the case of discrete couplings with finite synaptic depth l of the form $J_i = \pm 1/l, \pm 2/l, \ldots, \pm 1$ by building on the analysis of the analogous case for the perceptron [7,8]. In this case the additional order parameter $\bar{q} = \sum_i (J_i^a)^2/N$, and its conjugate $\hat{\bar{q}}$, must be introduced. For G_E we again find (4) where now $Q = \sqrt{q/(\bar{q}-q)}$. Moreover, $G_C = -\hat{\bar{q}}\bar{q} + \hat{q}q/2$ and

$$G_S = \int \prod_{\boldsymbol{\xi}} Dz_{\boldsymbol{\xi}} \ln \operatorname*{Tr}_J \exp\left[-\left(\frac{\hat{q}}{2} - \hat{\bar{q}}\right)J^2\right] \prod_{\boldsymbol{\xi}} 2\cosh\left(J\sqrt{\frac{\hat{q}}{2^{L-1}}}\, z_{\boldsymbol{\xi}}\right),$$

with Tr_J denoting the trace over the $2l$ possible values of the couplings J_i. Using these results we have numerically calculated the storage capacity $\alpha_c(l)$ for the simplest case $L = 3$ as a function of the synaptic depth l. The results are shown in Fig. 2. Starting from $\alpha_c \cong 1.11$ for the Ising case, $l = 1$, the capacity increases monotonically with l. It is rather difficult to compare these analytical findings with numerical simulations since the effects of the finite synaptic depth do not show up at the small values of N accessible to exact enumerations [9].

To complete the analysis of the storage properties we analyze the case of continuous couplings $\mathbf{J}$ between hidden and output layers. It is convenient to eliminate the additional order parameter k necessary in this case to enforce the normalization $\mathbf{J}^2 = N$ by introducing $\hat{Q} = \hat{q}/(k + \hat{q})$. Within replica symmetry the quenched entropy s is then again of the form (3) with $G_C = \ln[1 - \hat{Q}/(1+Q^2)]/2$, G_E given by (4), and the extremum is now with respect to Q and $\hat{Q}$. Moreover

$$G_S = \int \prod_{\boldsymbol{\xi}} Dz_{\boldsymbol{\xi}} \ln \operatorname*{Tr}_B \exp\left\{\frac{\hat{Q}}{2^L}\left[\sum_{\boldsymbol{\xi}} z_{\boldsymbol{\xi}} B(\boldsymbol{\xi})\right]^2\right\}. \tag{8}$$

The storage capacity α_c can be obtained from these expressions in the limits $Q \to \infty$, $\hat{Q} \to \infty$ corresponding to $q \to 1$. This limit indicates that different solutions of the storage problem may at most differ in a nonextensive number of components J_i and Boolean functions B_i. We find $G_E \sim -Q^2/4$ and $G_S \sim \hat{Q}[1/2 + (2^{L-1} - 1)/\pi]$ giving rise to

$$\alpha_c^{RS} = \frac{2 + \frac{4}{\pi}(2^{L-1} - 1)}{L}. \tag{9}$$

For $L = 3$ this yields $\alpha_c^{RS} = 2/3 + 4/\pi \cong 1.94$ which is included as a horizontal line in Fig. 2. From the scaling of the order parameters describing the case of finite synaptic depth we expect that the results for the storage capacity should converge to the value obtained for continuous

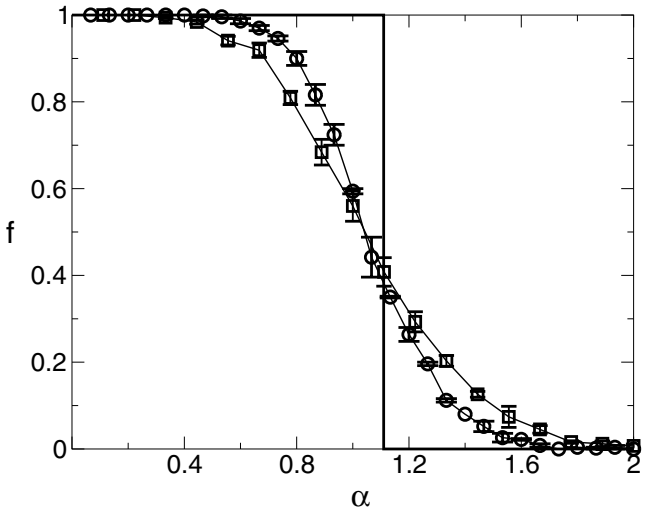

FIG. 1. Fraction f of $3\alpha N$ random input-output mappings implementable by a MLN with $3N$ inputs and N hidden units as a function of α for $N = 3$ (squares) and $N = 5$ (circles). The couplings between hidden units and output are fixed to $J_i = 1$ for all i and enumerations are performed over all combinations of symmetric Boolean functions B_i between input and the hidden layer. For every value of α, 200 input realizations were averaged over. The solid line gives the analytical result describing the limit $N \to \infty$.

couplings if $l \to \infty$. From Fig. 2 it can be seen that this asymptotic approach is likely to be very slow. This is similar to MLN with a few hidden units [10].

It is possible to derive an upper bound for α_c as has been done for MLN with a finite number of hidden units [11]. For a *fixed* set $\boldsymbol{\xi}_i^\mu, \mu = 1, \ldots, \alpha LN$ of inputs we can generate at most $2^{N(2^{L-1}-1)}$ different configurations τ_i of hidden units by using different combinations of Boolean functions B_i. Since the mapping from the hidden layer to

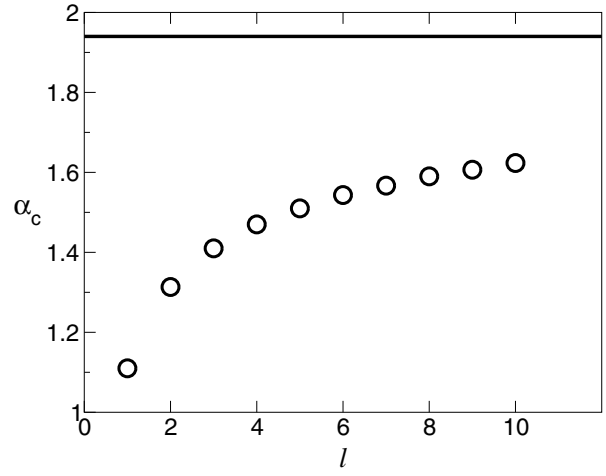

FIG. 2. Storage capacity of a MLN with $N \to \infty$ hidden units and $3N$ inputs with the couplings J_i between the hidden layer and output taking $2l$ discrete values. The inputs are mapped to the hidden layer by symmetric Boolean functions B_i. The line gives the result for continuous couplings J_i.

the output is performed by a perceptron, each hidden configuration gives rise to the desired output with probability $C(p, N)/2^p$, where $C(p, N)$ is the number of dichotomies calculated in [12]. Neglecting correlations between the different hidden configurations and using $p = \alpha LN$ we find for $N \to \infty$ that the solution α^{MD} of the equation

$$(2^{L-1} - 1)\ln 2 = (\alpha L - 1)\ln(\alpha L - 1) - \alpha L \ln\frac{\alpha L}{2} \tag{10}$$

represents an upper bound to the storage capacity α_c. For $L = 1$ we get $\alpha^{MD} = 2$ as expected, for $L = 3$ we find $\alpha^{MD} \cong 2.394$, and for large L the asymptotic behavior is $\alpha^{MD} \sim 2^{L-1}/L + \ln 2$. Contrary to what is known for MLN with a finite number of hidden units [4], the replica symmetric result (9) for the storage capacity is below this upper bound for $L = 3$ and shows the same scaling for large L.

Nevertheless we expect that the replica symmetric result (9) overestimates the storage capacity as is known for other types of MLN [4]. For large storage ratios α the solution space is likely to be disconnected, and it becomes impossible to determine its volume by introducing a single overlap scale (2) only. An improved characterization of the situation can be obtained using the well-known technique of replica symmetry breaking [13]. This problem will be addressed in future work [14].

Finally let us elucidate the generalization problem, i.e., the ability of the network to infer a rule from examples. To this end we consider as usual two networks of the same type with the couplings and Boolean function of one of them (the "teacher") fixed at random. The other network (the "student") receives a set of randomly chosen inputs $\boldsymbol{\xi}_i^\mu, \mu = 1, \ldots, \alpha LN$ together with the corresponding outputs σ_T^μ generated by the teacher. The task for the student is to imitate the teacher as closely as possible. The success in doing so is quantified by the generalization error ε defined as the probability that a *newly* chosen random input is classified differently by teacher and student.

The statistical mechanics analysis of the generalization problem builds again on the expression (1) for the quenched entropy with the number of replicas now tending to 1 rather than to 0 [2,15]. A nice feature of this limit is that replica symmetry is known to be stable. The order parameter q defined in (2) now gives the typical overlap between teacher and student and determines the generalization error ε in a simple way. In the present situation we have the standard relation $\varepsilon = (\arccos q)/\pi$. Moreover, (4) is replaced by

$$G_E(q) = 2\int Dt\, H(Qt)\ln H(Qt). \tag{11}$$

For Ising couplings, $J_i = \pm 1$, the problem can again be mapped exactly on the Ising perceptron. Correspondingly there is a *discontinuous* transition to perfect learning [16]; i.e., $\varepsilon = 0$ for $\alpha > \alpha_d$ with $\alpha_d = 1.24\ 2^{L-1}/L$. This

transition occurs when all Boolean functions of the student "lock" onto the corresponding input-hidden mappings of the teacher.

In the case of continuous couplings we find instead $G_C = Q^2\hat{Q}/[(1 + Q^2)(1 - \hat{Q})]$ and

$$G_S = \frac{1}{2}\ln(1 - \hat{Q}) + \sqrt{1 - \hat{Q}} \int \prod_{\boldsymbol{\xi}} Dz_{\boldsymbol{\xi}}\, g(z_{\boldsymbol{\xi}}) \ln g(z_{\boldsymbol{\xi}}), \quad (12)$$

where

$$g(z_{\boldsymbol{\xi}}) = \operatorname*{Tr}_B \exp\left[\frac{\hat{Q}}{2^L}\left(\sum_{\boldsymbol{\xi}} z_{\boldsymbol{\xi}} B(\boldsymbol{\xi})\right)^2\right]. \quad (13)$$

For small α this gives rise to $\varepsilon \sim 1/2 - \alpha L/(\pi^2 2^{L-1})$ which coincides with the result for the perceptron for $L = 1$, as it should. With increasing L the initial decay of the generalization error becomes slower, reflecting the increasing complexity and storage abilities of the network. There is no retarded learning because of the nonzero correlation between hidden units and output [17]. There is also no discontinuous drop of the generalization error at some intermediate value of α which would occur if an extensive fraction of student Boolean functions locked onto their teacher counterparts. Finally, for large α we find $\varepsilon \sim 0.625/(L\alpha)$. This is identical to the asymptotic behavior of a perceptron with N inputs and would hence occur if all Boolean functions between teacher and student were matched and only the hidden to output couplings had still to be adapted. For continuous couplings we therefore find a smooth generalization behavior characterized by a gradual freezing of the degrees of freedom associated with the Boolean functions and an asymptotic decay dominated by the fine-tuning of the student couplings between hidden layer and output.

In conclusion, we have quantitatively characterized the storage and generalization abilities of a multilayer neural network with a number of hidden units scaling as the input dimension. The mapping from the input to the hidden layer is realized by symmetric Boolean functions with L inputs. The storage capacity is found to be proportional to the logarithm of the number of these Boolean functions divided by L. The more conventional case in which the hidden units are the outputs of perceptrons with couplings $\mathbf{w}_i$ can also be analyzed [14]. In this case the identity (5) is to be replaced by somewhat long and unwieldy expressions resulting from the restriction of the trace to linear separable Boolean functions. Anticipating that the above scaling of the storage capacity persists and observing that the logarithm of the number of Boolean functions which can be implemented by a perceptron with L inputs is $O(L^2)$, we arrive at the interesting conclusion that the number of stored input-output relations *per weight* of the network is proportional to L. This implies that doubling the number of couplings in the network would increase the storage capacity by a factor of 2. This makes the proposed architecture superior to MLN with few ($K \ll N$) hidden units in which the storage capacity is known to increase at most logarithmically with the number of weights.

We have benefited from discussions with Wolfgang Kinzel, Robert Urbanczik, Peter Reimann, and Stephan Mertens. We thank the Max-Planck-Institut für Physik Komplexer Systeme in Dresden, where this work was completed, for hospitality, and the GIF for support.

[1] E. Gardner, J. Phys. A **21**, 257 (1988); E. Gardner and B. Derrida, J. Phys. A **21**, 271 (1988).
[2] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, 2001).
[3] G. Cybenko, Math. Control Signals Syst. **2**, 303 (1989).
[4] E. Barkai, D. Hansel, and I. Kanter, Phys. Rev. Lett. **65**, 2312 (1990); E. Barkai, D. Hansel, and H. Sompolinsky, Phys. Rev. A **45**, 4146 (1992); A. Engel, H. M. Köhler, F. Tschepke, H. Vollmayr, and A. Zippelius, Phys. Rev. A **45**, 7590 (1992); H. Schwarze and J. Hertz, Europhys. Lett. **20**, 375 (1992); R. Monasson and R. Zecchina, Phys. Rev. Lett. **75**, 2432 (1995); R. Urbanczik, Europhys. Lett. **35**, 553 (1996); one of the rare exceptions is [5].
[5] A. Bethge, R. Kühn, and H. Horner, J. Phys. A **27**, 1929 (1994).
[6] W. Krauth and M. Mezard, J. Phys. (Paris) **50**, 3057 (1989).
[7] H. Gutfreund and Y. Stein, J. Phys. A **23**, 2613 (1990).
[8] I. Kanter, Europhys. Lett. **17**, 181 (1992).
[9] A. Priel, M. Blatt, T. Grossman, E. Domany, and I. Kanter, Phys. Rev. E **50**, 577 (1994).
[10] R. Urbanczik, Europhys. Lett. **26**, 233 (1994).
[11] G. J. Mitchison and R. M. Durbin, Biol. Cybernet. **60**, 345 (1989).
[12] T. M. Cover, IEEE Trans. Electron. Comput. **14**, 326 (1965).
[13] M. Mezard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[14] M. Rosen-Zvi, A. Engel, and I. Kanter (to be published).
[15] M. Opper and D. Haussler, Phys. Rev. Lett. **66**, 2677 (1991).
[16] G. Györgyi, Phys. Rev. A **41**, 7097 (1990).
[17] M. Biehl and A. Mietzner, J. Phys. A **27**, 1885 (1994); B. Schottky, J. Phys. A **28**, 4515 (1995); C. Van den Broeck and P. Reimann, Phys. Rev. Lett. **76**, 2188 (1996).