Vol. 28 (1997) ACTA PHYSICA POLONICA B No 7

STORAGE CAPACITY OF MULTI-LAYERED NEURAL NETWORKS WITH BINARY WEIGHTS*

W. TARKOWSKI** AND J.L. VAN HEMMEN

Physik-Department der Technischen Universität München
D-85747 Garching bei München, Germany

(Received October 17, 1996)

Using statistical physics methods we investigate two-layered perceptrons which consist of N binary input neurons, K hidden units and a single output node. Four basic types of such networks are considered: the so-called *Committee*, *Parity*, and *AND Machines* which make a decision based on a majority, parity, and the logical AND rules, respectively (for these cases the weights that connect hidden units and output node are taken to be equal to one), and the *General Machine* where one allows all the synaptic couplings to vary. For these kinds of network we examine two types of architecture: fully connected and tree-connected ones (with overlapping and non-overlapping receptive fields, respectively). All the above mentioned machines have binary weights. Our basic interest is focused on the storage capabilities of such networks which realize $p = \alpha N$ random, unbiased dichotomies (α denotes the so-called storage ratio). The analysis is done using the annealed approximation and is valid for all values of K. The critical (maximal) storage capacity of the fully connected *Committee Machine* reads $\alpha_c = K$, while in the case of the tree-structure one gets $\alpha_c = 1$, independently of K. The results obtained for the *Parity Machine* are exactly the same as those for the *Committee* network. The optimal storage of the *AND Machine* depends on the distribution of the outputs for the patterns. These associations are studied in detail. We have found also that the capacity of the *General Machines* remains the same as compared to systems with fixed weights between intermediate layer and the output node. Some of the findings (especially these concerning the storage capacity of the *Parity Machine*) are in a good agreement with known numerical results.

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

* The paper is partially supported by the KBN Grants Nos. 2-P302-045-06 and 1-P106-004-06.

** Permanent address: Centrum Fizyki Teoretycznej, Polska Akademia Nauk, Al. Lotników 32/46, 02-668 Warsaw, Poland.

1. Introduction

One of the most important problems in the theory of neural networks concerns the storage capacity. For several years many attempts have been made in order to investigate properties of networks with synaptic couplings (weights, or strengths) which are constructed according to a definite rule, such as Hebb's rule, the pseudo-inverse one, and many others — see Refs. [1, 2] for a detailed review. A new and powerful line of research in the field of learning a rule has been introduced by Gardner [3], who proposed that one should look for the optimal configuration(s) of couplings in the space of all possible network weights.

Gardner's original idea has been extended widely since then; see, for example, Refs. [4–10]. However, most of the results which were obtained using this approach concern only a simple perceptron that consists of a single layer of input units connected directly with output node(s). It is well-known that computational capabilities of such a neuronal structure are very limited. A single-layer perceptron cannot realize, for example, the so-called nonlinearly-separable rules such as, for instance, the exclusive alternative — XOR one [11]. Its memory capacity seems also to be relatively small in comparison with numerical results obtained for more complicated systems. It then seems very sensible and strongly recommended to extend Gardner's method in order to investigate multi-layered, at least two-layered, networks. Such structures that contain, besides input and output units, also the so-called hidden nodes can solve problems that are *not* linearly separable, and have better computational capabilities than a simple perceptron. This is due to the freedom in the choice of the internal representations, which one does not fix *a priori*.

The adaption of Gardner's method to multi-layered networks has in fact been done in a recent series of papers [12–20]. They contain both numerical and theoretical results. The latter have been obtained by the aid of the so-called replica method [21]. In order to evaluate the quantity $\langle\langle \ln \mathcal{Z} \rangle\rangle$, one estimates the average over the memorized data-patterns distribution of the n-th power of the partition function $\frac{1}{n} \ln \langle\langle \mathcal{Z}^n \rangle\rangle$ where n is a natural number, and takes the $n \to 0$ limit afterwards. Most of the theoretical results have been obtained after quite complicated calculations and take into account the replica-symmetry-breaking effects as well. They are, however, obtained only for some limited cases as, for instance, small values of K (as $K = 3$), or $K \to \infty$ and for simplest network architectures (as the tree-one). This suggests that one should look for another way to estimate the storage properties of the investigated systems, even if it would allow us to get only approximate results.

In this paper, we try to solve the problem of the maximal capacity in multi-layered networks using, among others, a generalization of the method developed in Ref. [6]. It consists of a straightforward evaluation of Gardner's integral (denoted in the literature by V [3]) without any replica method impact. Instead, we employ in the present work the annealed approximation [21]. We investigate four well-known types of two-layered perceptron: the so-called *Committee*, *Parity*, and *AND Machines* which make a decision based on a majority, parity, and the logical AND rule, respectively (for these cases the weights that connect hidden units and output node are taken to be equal to one), and the *General Machine*, where one allows all the synaptic couplings to vary. For these networks two kinds of architecture are considered: a fully connected structure (with "overlapping" receptive fields) where every hidden neuron is coupled to all the inputs (Fig. 1), and a tree-structure system (with "non-overlapping" receptive fields) where different hidden units do not share the input nodes (Fig. 2). We consider networks with binary weights only. All our results have been obtained for the general case of *any* number of hidden (K) units. The theoretical predictions for the investigated machines (especially for the *Parity Machine*) are in a good (or even very good) agreement with numerical findings presented in the literature; see, for instance, Refs. [13, 17]. We should, however, add that our annealed-approximation results concerning *Committee* and *Parity Machines* may be obtained using other, well-known and simpler methods, than the one presented in this paper. Nevertheless, our approach seems to be quite generic and could be applied to each kind of network with any activation rule, architecture, or couplings. Therefore, we present it in detail.

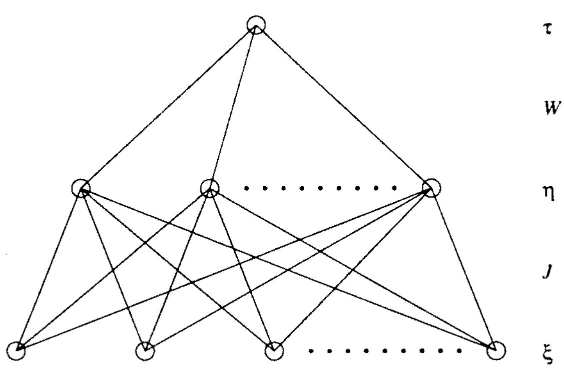

Fig. 1. Two-layered perceptron with fully connected architecture. All the input nodes $\{\xi_j\}$ are coupled, through the couplings $\{J_{kj}\}$, with all the hidden-layer neurons, $\{\eta_k\}$. These, in turn, are connected to the output τ by means of the synaptic weights $\{W_k\}$.

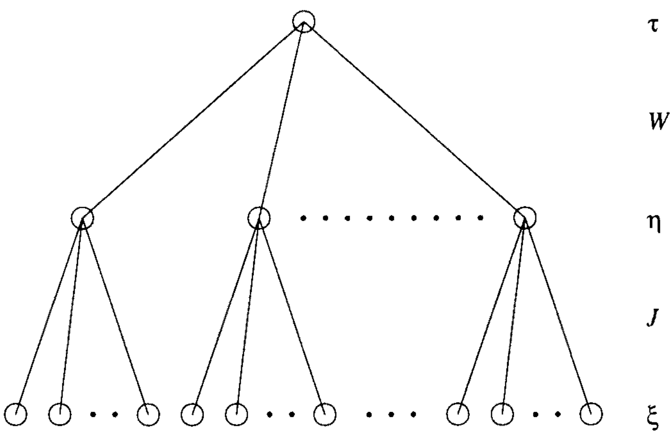

Fig. 2. Tree-connected two-layered perceptron. The distinct groups of input neurons $\{\xi_j\}$ are connected, through the couplings $\{J_{kj}\}$, to the distinct hidden nodes, $\{\eta_k\}$. Every hidden neuron is coupled, by the synaptic strengths $\{W_k\}$, with the output τ.

The paper is organized as follows: In Section 2 we present a short but general description of the statistical approach which is adopted in this work. In Sections 3, 4, 5 and 6 the *Committee*, *Parity*, *AND* and *General Machines* are, respectively, considered in detail. Section 7 contains some final conclusions and remarks.

2. The model

In the present paper we investigate networks which consist of N binary inputs $\xi_j = \pm 1$, with $1 \le j \le N$, connected by a set of binary neuronal couplings $\{J_{kj} = \pm 1\}$ with K hidden units $\eta_k = \pm 1$, $1 \le k \le K$. The synaptic signal is then sent to the single output node $\tau = \pm 1$ with the help of synaptic weights $\{W_k = \pm 1\}$. Usually, the strengths connecting the intermediate layer with the neuron τ are fixed to be equal to one (with the exception of the case of the *General Machine*, where one allows these couplings to vary), and the state of the output is determined by performing a specified Boolean function of the states of the hidden units. Here we do not define in detail the architectures or possible dynamical equations of the systems as this will be done in the next Sections.

The problem of learning in such a structure may be formulated as follows: Having a certain number of input-output relations $\left\{\xi_j^\mu\right\} \to \{\tau^\mu\}$, for $1 \le \mu \le p$ input patterns with prescribed output τ^μ and $1 \le j \le N$ input neurons, can one find a set of synaptic connections (to be called a "solution") for a network which performs each pair of the input-output associations $\left\{\xi_j^\mu\right\} \to \{\tau^\mu\}$ correctly? We try to answer the question of how many patterns $\left\{\xi_j^\mu, \tau^\mu\right\}$ can be maximally memorized in a network which has a specified architecture and follows a definite updating rule that determines dynamics of the system. In another words, we want to know what is the *number* of memorized dichotomies for which one can find a set of synaptic couplings that can perfectly implement all the rules $\left\{\xi_j^\mu\right\} \to \{\tau^\mu\}$. One should add that a large number of learning algorithms has already been constructed; see Refs. [1, 2]. These, then, assume that the existence problem has been solved. Existence of solutions is a corollary of the present work. The problem of learning multi-layered perceptrons is, in general, a very hard task, which takes much computer time and memory. In fact, no learning algorithm has been proven *to converge* to the desired solution, if the latter exists; there are many local minima and metastable states in the configuration space, *etc.*

In order to solve the problem of determining the maximal storage capacity, we use statistical mechanics methods. In so doing, we define the cost, or energy function

$$\mathcal{E}\left(\{J_{kj}\},\{W_k\},\left\{\xi_j^\mu\right\}\right) = \sum_{\mu=1}^{p} \Theta\left[-\tau\left(\{J_{kj}\},\{W_k\},\left\{\xi_j^\mu\right\}\right)\tau^\mu\right] , \tag{1}$$

which corresponds to the number of ill-memorized patterns. The partition function reads then

$$\mathcal{Z} = \mathrm{Tr}_{\left\{J_{kj}\right\},\{W_k\}} \exp\left[-\beta\mathcal{E}\left(\{J_{kj}\},\{W_k\},\left\{\xi_j^\mu\right\}\right)\right] , \tag{2}$$

where $\beta = 1/T$ is the noise measure, and T denotes the temperature. Note that the trace $\mathrm{Tr}_{\left\{J_{kj}\right\},\{W_k\}}$ should be performed while taking into account the binary character of synaptic connections $\{J_{kj}\}$ and $\{W_k\}$.

The free energy of the model can then be determine as follows:

$$\mathcal{F} = -T\langle\!\langle \ln \mathcal{Z} \rangle\!\rangle , \tag{3}$$

where $\langle\!\langle\cdot\rangle\!\rangle$ denotes the quenched average over the probability distribution of the patterns $\left\{\xi_j^\mu\right\}$,

$$\Pr\left(\xi_j^\mu\right) = \frac{1}{2}\delta\left(\xi_j^\mu + 1\right) + \frac{1}{2}\delta\left(\xi_j^\mu - 1\right) , \tag{4}$$

for all j and μ. We perform our calculations for statistically independent, unbiased and uncorrelated patterns $\left\{\xi_j^\mu\right\}$, which are, in addition, also independent of $\{\tau^\mu\}$. The probability distribution of the "output part" of the data, $\{\tau^\mu\}$, has not any influence on the final result — apart from the case of the *AND Machine*; see Section 5.

The entropy function can be obtained by differentiation,

$$\mathcal{S} = -\frac{\partial \mathcal{F}}{\partial T} \; . \tag{5}$$

We are, in practice, interested in the noiseless limit $T = 0$, or $\beta \to \infty$. For such a case one gets

$$\mathcal{S} = \left\langle\!\!\left\langle \ln \left\{ \mathrm{Tr}_{\{J_{kj}\},\{W_k\}} \prod_{\mu=1}^{p} \Theta\left[\tau\left(\{J_{kj}\},\{W_k\},\left\{\xi_j^\mu\right\}\right)\tau^\mu\right] \right\} \right\rangle\!\!\right\rangle \; . \tag{6}$$

Note that the limit $\beta \to \infty$ should, in fact, be taken *after* performing the average $\langle\!\langle\cdot\rangle\!\rangle$. It is easy to see that this procedure gives us exactly Eq. (6).

The average over the logarithm function is especially hard to evaluate. In order to simplify our calculations we then make use of the annealed approximation [21] and assume

$$\mathcal{S} = \ln \left\langle\!\!\left\langle \mathrm{Tr}_{\{J_{kj}\},\{W_k\}} \prod_{\mu=1}^{p} \Theta\left[\tau\left(\{J_{kj}\},\{W_k\},\left\{\xi_j^\mu\right\}\right)\tau^\mu\right] \right\rangle\!\!\right\rangle \; . \tag{7}$$

We will show that this approach provides quite reasonable results for the presently studied cases.

Further procedures depend on the particular form of the entropy function for each of the investigated systems and will be described in detail in the following Sections.

3. Committee Machine

3.1. Fully connected architecture

We consider the fully connected *Committee Machine* which follows the two-stage (inputs — hidden units — output) dynamics of the form

$$\eta_k = \mathrm{sgn}\left[\sum_{j=1}^{N} J_{kj}\xi_j\right] \qquad \text{for each } k = 1,\ldots,K \tag{8}$$

with the output

$$\tau = \mathrm{sgn}\left[\sum_{k=1}^{K} \eta_k\right] \tag{9}$$

representing a majority rule.

The entropy function (7) for such a case reads

$$\mathcal{S} = \ln\left\langle\!\!\left\langle \sum_{\{J_{kj}=\pm1\}} \prod_{\mu=1}^{p} \Theta\left[\tau^{\mu}\sum_{k=1}^{K}\operatorname{sgn}\left(\sum_{j=1}^{N} J_{kj}\xi_j^{\mu}\right)\right]\right\rangle\!\!\right\rangle \equiv \ln\langle\!\langle\mathcal{A}\rangle\!\rangle \;. \tag{10}$$

One straightforwardly evaluates the quantity $\mathcal{A}$ in a manner similar to the one described in Ref. [6]. First, we observe that the transformation $\xi_j^{\mu} \to \tau^{\mu}\xi_j^{\mu}$ (for each j, μ) in the case of the patterns distribution given in Eq. (4) will not change the properties of the model, and, at the end, of the final results. In further calculations we then simply omit $\{\tau^{\mu}\}$.

In order to evaluate the entropy function (10) we introduce the new variables

$$x_{k\mu} = \sum_j \frac{J_{kj}\xi_j^{\mu}}{\sqrt{N}} \tag{11}$$

for each k and μ, which gives

$$\mathcal{A} = \sum_{\{J_{ij}=\pm1\}} \int_{-\infty}^{+\infty} \prod_{k,\mu} dx_{k\mu} \int_{-\infty}^{+\infty} \prod_{k,\mu} \frac{d\lambda_{k\mu}}{2\pi} \prod_{\mu} \Theta\left(\sum_k \operatorname{sgn} x_{k\mu}\right) \times \exp\left[i\sum_{k,\mu} x_{k\mu}\lambda_{k\mu} - i\sum_{k,j} J_{kj}\sum_{\mu}\frac{\lambda_{k\mu}\xi_j^{\mu}}{\sqrt{N}}\right] \;. \tag{12}$$

The sum over $\{J_{kj} = \pm 1\}$, and then the Gaussian integral over $\{\lambda_{k\mu}\}$ in Eq. (12) can be performed exactly. As a result we obtain expression

$$\mathcal{A} = \frac{2^{KN}}{\pi^{\frac{\alpha KN}{2}}} \frac{1}{(\det M_{\mu\mu'})^{\frac{K}{2}}} \int_{-\infty}^{+\infty} \prod_{k,\mu} dx_{k\mu} \prod_{\mu} \Theta\left(\sum_k \operatorname{sgn} x_{k\mu}\right) \times \exp\left[-\sum_{k,\mu,\mu'} x_{k\mu} M^{-1}_{\mu\mu'} x_{k\mu'}\right] \;, \tag{13}$$

where $\alpha \equiv p/N$ denotes the so-called storage ratio and the overlap matrix

$$M_{\mu\mu'} = \frac{1}{N}\sum_j \xi_j^{\mu}\xi_j^{\mu'} \tag{14}$$

is a non-negative definite one.

In order to derive expression (13) we have made an approximation which, actually, has no influence on the final result. We have dropped the terms higher than second-order from the series expansion of the cosine function which was obtained after taking the trace over all possible binary couplings configurations; *c.f.* Refs. [3–5]. An easy analysis shows that the sum of these clipped higher-order terms is, on the average, much less than the last which remains — the second-order element. This already holds for smallest values of N. For example, if $N = 2$, the relative error in the final result, which is caused by cutting of the mentioned above series, can be estimated to be not greater than a few percent. For growing N the error tends, of course, to zero. It is then obvious that in such a case the storage capacity, which takes the value of the natural number closest to $p = \alpha_c N$, remains without any change after this approximation. For instance, if $p = 1.51$ or $p = 2.49$, the number of memorized patterns is for both cases, on the average, equal to 2. Similar considerations are true for all the following Sections, and hence we will not repeat them any further.

Next we should consider the Heavyside Θ functions which appear in Eq. (13). Signum functions appear in their arguments. The quantities $\{x_{k\mu}\}$ in these expressions can take all possible values. It is obvious that either for $x_{k\mu} \in (-\infty, 0)$ or for $x_{k\mu} \in (0, +\infty)$, the function sgn $(x_{k\mu})$ remains constant (either -1 or $+1$, respectively). On the other hand, the multiple integral $\int_{-\infty}^{+\infty} \prod_{k,\mu} dx_{k\mu}$ can be expressed as the sum of $2^{\alpha K N}$ partial ones with $x_{k\mu}$ belonging to the interval $(-\infty, 0)$ or to $(0, +\infty)$. We have to estimate how many of these integrals take a nonzero (identical) value, which occurs for $\Theta(\cdot) = 1$. This means that we are to investigate, for each μ, the condition

$$\sum_k \operatorname{sgn} x_{k\mu} \geq 0 \ . \tag{15}$$

It is easy to calculate the number of nonzero partial integrals for every μ. A straightforward combinatorics gives for odd K

$$\Sigma_{1a} = \binom{K}{\frac{K+1}{2}} + \binom{K}{\frac{K+1}{2}+1} + \binom{K}{\frac{K+1}{2}+2} + \ldots + \binom{K}{K} , \tag{16}$$

and for even K

$$\Sigma_{1b} = \frac{1}{2}\binom{K}{\frac{K}{2}} + \binom{K}{\frac{K}{2}+1} + \binom{K}{\frac{K}{2}+2} + \ldots + \binom{K}{K} . \tag{17}$$

The prefactor 1/2 in front of $\binom{K}{K/2}$ in (17) reflects the for the network somewhat confusing situation where half of the hidden units has a plus, the other half — a minus sign. The output node can then take both values ± 1 with probability 1/2. Of course, for such a case $\sum_k \operatorname{sgn} x_{k\mu} = 0$ but

$\Theta(0) = 1/2$.[1] Both sums Σ_{1a} and Σ_{1b} equal 2^{K-1}. We now return to our original problem, evaluating (13).

The entropy function (10) is given by

$$\mathcal{S} = \ln\left\langle\!\!\left\langle \frac{2^{KN}}{\pi^{\frac{\alpha KN}{2}}} \frac{1}{\left(\det M_{\mu\mu'}\right)^{\frac{K}{2}}} \int_{-\infty}^{+\infty} \prod_{k,\mu} dx_{k\mu} \prod_{\mu} \Theta\left(\sum_k \operatorname{sgn} x_{k\mu}\right) \right.\right. \\ \left.\left. \times \exp\left[-\sum_{k,\mu,\mu'} x_{k\mu} M^{-1}_{\mu\mu'} x_{k\mu'}\right]\right\rangle\!\!\right\rangle . \qquad (18)$$

The integrals over $\{x_{k\mu}\}$ can be written as a sum over $2^{\alpha KN}$ partial ones with the variables $\{x_{k\mu}\}$ belonging either to the interval $(-\infty, 0)$ or $(0, +\infty)$, for each k, μ. However, because of the above combinatorial investigations, only $1/2^{\alpha N}$ (since $2^{\alpha(K-1)N}/2^{\alpha KN} \equiv 1/2^{\alpha N}$) of the terms of the sum of these partial integrals are nonzero. We shall now prove that

$$\left\langle\!\!\left\langle (\det M_{\mu\mu'})^{-\frac{K}{2}} \int_{-\infty}^{+\infty} \prod_{k,\mu} dx_{k\mu}\, 2^{\alpha N} \prod_{\mu} \Theta\left(\sum_k \operatorname{sgn} x_{k\mu}\right) \right.\right. \\ \left.\left. \times \exp\left[-\sum_{k,\mu,\mu'} x_{k\mu} M^{-1}_{\mu\mu'} x_{k\mu'}\right]\right\rangle\!\!\right\rangle \\ = \left\langle\!\!\left\langle (\det M_{\mu\mu'})^{-\frac{K}{2}} \int_{-\infty}^{+\infty} \prod_{k,\mu} dx_{k\mu} \exp\left[-\sum_{k,\mu,\mu'} x_{k\mu} M^{-1}_{\mu\mu'} x_{k\mu'}\right]\right\rangle\!\!\right\rangle, \qquad (19)$$

for any value of K and N. To this end we evaluate the right-hand side of Eq. (19), assuming the pattern $\left\{\xi^{\mu}_{j}\right\}$ distribution given by expression (4),

$$\left\langle\!\!\left\langle (\det M_{\mu\mu'})^{-\frac{K}{2}} \int_{-\infty}^{+\infty} \prod_{k,\mu} dx_{k\mu} \exp\left[-\sum_{k,\mu,\mu'} x_{k\mu} M^{-1}_{\mu\mu'} x_{k\mu'}\right]\right\rangle\!\!\right\rangle \\ = \left\langle\!\!\left\langle \pi^{-\frac{KN}{2}(1+\alpha)} \int \prod_{j,k} da_{kj} \exp\left[-\sum_{j,k} a^2_{kj}\right]\right.\right.$$

---

[1] A consistent definition requires $\Theta(0) = \int_{-\infty}^{+\infty} \Theta(x)\delta(x)\,dx = [\Theta(x)\Theta(x)]_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} \delta(x)\Theta(x)\,dx = 1 - \Theta(0)$.

$$\times \int \prod_{k,\mu} db_{k\mu} \int \prod_{k,\mu} dx_{k\mu} \exp\left[-2i\sum_{j,k} a_{kj} \sum_{\mu} \frac{b_{k\mu}\xi_j^\mu}{\sqrt{N}} + 2i\sum_{k,\mu} b_{k\mu}x_{k\mu}\right]\Bigg\rangle\Bigg\rangle$$

$$= \pi^{-\frac{KN}{2}(1+\alpha)} \int \prod_{j,k} da_{kj} \exp\left[-\sum_{j,k} a_{kj}^2\right]$$

$$\times \sum_{\{x_{k\mu}\}} \int \prod_{k,\mu} db_{k\mu} \int_{\{x_{k\mu}\in(-\infty,0)\,\mathrm{or}\,(0,+\infty)\}} \prod_{k,\mu} dx_{k\mu}$$

$$\times \exp\left[2i\sum_{k,\mu} b_{k\mu}x_{k\mu}\right] \prod_{j,\mu} \cos\left(2\sum_{k} \frac{a_{kj}b_{k\mu}}{\sqrt{N}}\right), \quad (20)$$

where the sum $\sum_{\{x_{k\mu}\}}$ is with respect to all $2^{\alpha KN}$ nonzero multiple integrals $\int_{\{x_{k\mu}\in(-\infty,0)\,\mathrm{or}\,(0,+\infty)\}} \prod_{k,\mu} dx_{k\mu}$. For any value of the index $k = k_0$, and for all $\mu = 1, \ldots, \alpha N$, in any elements of the sum $\sum_{\{x_{k\mu}\}}$, we readily verify that the following transformation

$$a_{k_0 j} \to -a_{k_0 j} \qquad \text{for each } j = 1, \ldots, N, \quad (21)$$

$$b_{k_0\mu} \to -b_{k_0\mu}\,, \quad (22)$$

$$x_{k_0\mu} \to -x_{k_0\mu}\,, \quad (23)$$

leaves expression (20) without any change — with only one exception: For every value of μ, in appropriate terms of the sum $\sum_{\{x_{k\mu}\}}$, the integration limits of $\{x_{k_0\mu}\}$ are altered: either $(-\infty,0) \to (0,+\infty)$ or $(0,+\infty) \to (-\infty,0)$. It is easy to verify that another transformation (for fixed $\mu = \mu_0$, with respect to every value of $k = 1, \ldots, K$),

$$b_{k\mu_0} \to -b_{k\mu_0}\,, \quad (24)$$

$$x_{k\mu_0} \to -x_{k\mu_0}\,, \quad (25)$$

changes the integration intervals of $\{x_{k\mu_0}\}$ ($k = 1, \ldots, K$) while keeping the right-hand side of Eq. (20) constant. We should stress that the transformations (21)–(23) and (24)–(25) may be, of course, performed for each component of the sum $\sum_{\{x_{k\mu}\}}$ in Eq. (20) separately. It is now easy to see that the averaged values of integrals over different "octants" of the $\{x_{k\mu}\}$ space — see Eqs. (19) and (20) — are all equal to one another. This ends the proof of Eq. (19).

One can now return to Eq. (18). The entropy function may be rewritten

$$\mathcal{S} = \ln\left\langle\!\!\left\langle \frac{2^{KN}\,(\det M_{\mu\mu'})^{-\frac{K}{2}}}{2^{\alpha N}\pi^{\frac{\alpha K N}{2}}}\left[\int_{-\infty}^{+\infty}\prod_{\mu} dx_\mu \exp\left(-\sum_{\mu,\mu'} x_\mu M^{-1}_{\mu\mu'}x_{\mu'}\right)\right]^K\right\rangle\!\!\right\rangle$$
$$= KN\ln 2 - \alpha N\ln 2\,. \quad (26)$$

On the other hand, in the maximum capacity limit, we expect only one set of couplings to exist, and the critical-storage condition $\mathcal{S} = \ln 1 = 0$ will give the capacity ratio (per input)

$$\alpha_c = K\,. \quad (27)$$

As our calculations are done within the annealed approximation, the above result should be thought as an upper bound on the real ("quenched") critical capacity.

3.2. Tree-connected architecture

We now turn to investigating the tree-connected *Committee Machine* which is characterized by the dynamics

$$\eta_k = \operatorname{sgn}\left[\sum_{j=[N(k-1)/K]+1}^{Nk/K} J_{kj}\xi_j\right] \qquad \text{for every } k = 1,\ldots,K, \quad (28)$$

$$\tau = \operatorname{sgn}\left[\sum_{k=1}^{K}\eta_k\right]\,, \quad (29)$$

where we assume that N/K is a natural number.

For our problem the synaptic connections $\{J_{kj}\}$, $j = [N(k-1)/K]+1,\ldots,Nk/K$ with $k = 1,\ldots,K$, take binary values $J_{kj} = \pm 1$ for each k and j. The entropy function for this problem reads

$$\mathcal{S} = \ln\left\langle\!\!\left\langle \sum_{\{J_{kj}=\pm 1\}}\prod_{\mu=1}^{p}\Theta\left[\tau^\mu\sum_{k=1}^{K}\operatorname{sgn}\left(\sum_{j=[N(k-1)/K]+1}^{Nk/K} J_{kj}\xi_j^\mu\right)\right]\right\rangle\!\!\right\rangle \equiv \ln\langle\!\langle\mathcal{A}\rangle\!\rangle\,. \quad (30)$$

This leads to

$$\mathcal{A} = \frac{2^N}{\pi^{\frac{\alpha K N}{2}}}\frac{1}{(\det M_{\mu\mu'})^{\frac{K}{2}}}\int_{-\infty}^{+\infty}\prod_{k,\mu} dx_{k\mu}\prod_{\mu}\Theta\left(\sum_k \operatorname{sgn} x_{k\mu}\right)$$
$$\times\exp\left[-\sum_{k,\mu,\mu'} x_{k\mu}M^{-1}_{\mu\mu'}x_{k\mu'}\right]\,. \quad (31)$$

We note that here the index j can take, for each value of k, only N/K values (instead of N, as in the case of a fully connected system).

After straightforward calculations similar to those in Section 3.1 we arrive at

$$\mathcal{S} = N \ln 2 - \alpha N \ln 2 \ , \tag{32}$$

which gives (with the maximal-storage condition $\mathcal{S} = 0$) the critical capacity

$$\alpha_c = 1 \ . \tag{33}$$

This is independent of K and N. The result (33) is, apparently, the same as that already obtained within the annealed approximation, and remains somewhat in contradiction with both numerical [22] and theoretical results — obtained using the replica method — known from the literature [5, 17]. In the limit $N \to \infty$, namely, one gets for $K = 1$, which is the simple perceptron case, $\alpha_c \cong 0.833$ [5]. If $K = 3$, $\alpha_c \cong 0.92$ [15, 17]. When K is sufficiently large, $\alpha_c \cong 0.95$ [17]. On the other hand, for small values of N, computer simulation findings [22] indicate that the value of the storage ratio is very close to 1. It is easy to check that, actually, for $N = 1, 2, 3, 4, 5$, $\alpha_c = 1$ (we can simply consider all the possible configurations of couplings and binary unbiased patterns). One then should stress that all these results are quite close to the findings obtained using a relatively simple method described in this paper.

4. Parity Machine

4.1. Fully connected architecture

The fully connected *Parity Machine* is driven by the following dynamics,

$$\eta_k = \operatorname{sgn}\left[\sum_{j=1}^{N} J_{kj}\xi_j\right] \qquad \text{for each } k = 1, \ldots, K, \tag{34}$$

$$\tau = \operatorname{sgn}\left[\prod_{k=1}^{K} \eta_k\right] \ . \tag{35}$$

The evaluation of the entropy function

$$\mathcal{S} = \ln\left\langle\!\!\left\langle \sum_{\{J_{kj}=\pm 1\}} \prod_{\mu=1}^{p} \Theta\left[\tau^\mu \prod_{k=1}^{K} \operatorname{sgn}\left(\sum_{j=1}^{N} J_{kj}\xi_j^\mu\right)\right]\right\rangle\!\!\right\rangle \equiv \ln\langle\!\langle \mathcal{A} \rangle\!\rangle \tag{36}$$

is very similar as in Section 3.1. Thus we find

$$\mathcal{A} = \frac{2^{KN}}{\pi^{\frac{\alpha KN}{2}}} \frac{1}{(\det M_{\mu\mu'})^{\frac{K}{2}}} \int\limits_{-\infty}^{+\infty} \prod_{k,\mu} dx_{k\mu} \prod_{\mu} \Theta\left(\tau^{\mu} \prod_{k} \operatorname{sgn} x_{k\mu}\right) \times \exp\left[-\sum_{k,\mu,\mu'} x_{k\mu} M^{-1}_{\mu\mu'} x_{k\mu'}\right] , \tag{37}$$

where the overlap matrix is given by

$$M_{\mu\mu'} = \tau^{\mu}\tau^{\mu'} \frac{1}{N} \sum_{j} \xi_j^{\mu} \xi_j^{\mu'} . \tag{38}$$

As one sees, all the calculations which we must perform are very similar to those of Section 3.1. The only difference is that, because we cannot use the transformation $\left\{\xi_j^{\mu}\right\} \rightarrow \left\{\tau^{\mu}\xi_j^{\mu}\right\}$, the condition (15) should be

$$\tau^{\mu} \sum_{k} \operatorname{sgn} x_{k\mu} \geq 0 , \tag{39}$$

whatever μ. The analogues of sums Σ_{1a} and Σ_{1b} may be evaluated easily according to the sign of τ^{μ}. For $\tau^{\mu} = +1$, odd values of K imply

$$\Sigma_{2a} = \binom{K}{0} + \binom{K}{2} + \binom{K}{4} + \ldots + \binom{K}{K-1} , \tag{40}$$

and for even K one obtains

$$\Sigma_{2b} = \binom{K}{0} + \binom{K}{2} + \binom{K}{4} + \ldots + \binom{K}{K} . \tag{41}$$

If $\tau^{\mu} = -1$ and K is an odd number,

$$\Sigma_{3a} = \binom{K}{1} + \binom{K}{3} + \binom{K}{5} + \ldots + \binom{K}{K} , \tag{42}$$

whereas for even K

$$\Sigma_{3b} = \binom{K}{1} + \binom{K}{3} + \binom{K}{5} + \ldots + \binom{K}{K-1} . \tag{43}$$

Nicely, all the above sums equal 2^{K-1}. Similar considerations as in Section 3.1 (while taking into account the presence of the variables $\{\tau^\mu\}$ here) lead us to the conclusion

$$\left\langle\!\!\left\langle (\det M_{\mu\mu'})^{-\frac{K}{2}} \int_{-\infty}^{+\infty} \prod_{k,\mu} dx_{k\mu} \prod_{\mu} \Theta\left(\tau^\mu \prod_k \operatorname{sgn} x_{k\mu}\right) \times \exp\left[-\sum_{k,\mu,\mu'} x_{k\mu} M^{-1}_{\mu\mu'} x_{k\mu'}\right]\right\rangle\!\!\right\rangle$$

$$= \left\langle\!\!\left\langle (\det M_{\mu\mu'})^{-\frac{K}{2}} \int_{-\infty}^{+\infty} \prod_{k,\mu} dx_{k\mu} \left(\frac{1}{2}\right)^{\alpha N} \exp\left[-\sum_{k,\mu,\mu'} x_{k\mu} M^{-1}_{\mu\mu'} x_{k\mu'}\right]\right\rangle\!\!\right\rangle . \quad (44)$$

Further considerations, which are identical to those in Section 3.1, give us the same result as for the *Committee Machine*. We then end up with

$$\alpha_c = K \ , \quad (45)$$

whatever N.

4.2. Tree-connected structure

In the case of a tree-connected structure the dynamics of a network takes the form

$$\eta_k = \operatorname{sgn}\left[\sum_{j=[N(k-1)/K]+1}^{Nk/K} J_{kj}\xi_j\right] \qquad \text{for every } k = 1, \ldots, K, \quad (46)$$

$$\tau = \operatorname{sgn}\left[\prod_{k=1}^{K} \eta_k\right] , \quad (47)$$

where N/K is an integer. The case of binary couplings can now be handled straightforwardly.

The discrete weights imply the following form of the entropy function

$$\mathcal{S} = \ln\left\langle\!\!\left\langle \sum_{\{J_{kj}=\pm 1\}} \prod_{\mu=1}^{p} \Theta\left[\tau^\mu \sum_{k=1}^{K} \operatorname{sgn}\left(\sum_{j=[N(k-1)/K]+1}^{Nk/K} J_{kj}\xi_j^\mu\right)\right]\right\rangle\!\!\right\rangle \equiv \ln \langle\langle \mathcal{A} \rangle\rangle \ , \quad (48)$$

and here

$$\mathcal{A} = \frac{2^N}{\pi^{\frac{\alpha K N}{2}}} \frac{1}{(\det M_{\mu\mu'})^{\frac{K}{2}}} \int_{-\infty}^{+\infty} \prod_{k,\mu} dx_{k\mu} \prod_{\mu} \Theta\left(\tau^{\mu} \prod_{k} \operatorname{sgn} x_{k\mu}\right) \times \exp\left[- \sum_{k,\mu,\mu'} x_{k\mu} M^{-1}_{\mu\mu'} x_{k\mu'}\right] . \tag{49}$$

The storage ratio reads

$$\alpha_c = 1 , \tag{50}$$

for all values of K and N. This result agrees with both theoretical and numerical findings of Ref. [13], where the tree-connected *Parity Machine* with binary connections was studied in the thermodynamic limit $N \to \infty$ using the replica method. One then could expect that the result of Section 4.1 for the fully connected *Parity Machine*, see Eq. (45), is also exact.

5. AND Machine

We now consider the so-called *AND Machine*. The output unit in such a network takes the value $+1$ when every hidden node is active ($+1$), and -1 in all the other cases. Below we study both the fully connected and the tree-connected structure.

5.1. Fully connected structure

The dynamics of a fully connected network can be defined by

$$\eta_k = \operatorname{sgn}\left[\sum_{j=1}^{N} J_{kj} \xi_j\right] \qquad \text{for each } k = 1, \ldots, K, \tag{51}$$

$$\tau = \operatorname{sgn}\left[\sum_{k=1}^{K} \eta_k + 1 - K\right] . \tag{52}$$

Taking $\{J_{kj} = \pm 1\}$ one can write the entropy function,

$$\mathcal{S} = \ln\left\langle\!\!\left\langle \sum_{\{J_{kj}=\pm 1\}} \prod_{\mu=1}^{p} \Theta\left[\tau^{\mu} \sum_{k=1}^{K} \operatorname{sgn}\left(\sum_{j=1}^{N} J_{kj} \xi_j^{\mu}\right) + \tau^{\mu} - K\tau^{\mu}\right]\right\rangle\!\!\right\rangle \equiv \ln\langle\!\langle \mathcal{A} \rangle\!\rangle . \tag{53}$$

Proceeding similarly as in previous Sections we obtain

$$\mathcal{A} = \frac{2^{KN}}{\pi^{\frac{\alpha KN}{2}}} \frac{1}{(\det M_{\mu\mu'})^{\frac{K}{2}}} \int_{-\infty}^{+\infty} \prod_{k,\mu} dx_{k\mu} \prod_{\mu} \Theta\Big(\sum_k \operatorname{sgn} x_{k\mu} + \tau^\mu - K\tau^\mu \Big) \times \exp\Big[- \sum_{k,\mu,\mu'} x_{k\mu} M^{-1}_{\mu\mu'} x_{k\mu'} \Big] , \quad (54)$$

with the matrix $\{M_{\mu\mu'}\}$ given by expression (14).

Here the distribution of the "output-part" of the patterns, $\{\tau^\mu\}$, plays an important role. We note, however, that the transformation $\{\xi_j^\mu\} \to \{\tau^\mu \xi_j^\mu\}$ can be done. Let us assume that the data $\{\tau^\mu\}$ are uncorrelated but biased,

$$\Pr(\tau^\mu) = \frac{1-m}{2}\delta(\tau^\mu + 1) + \frac{1+m}{2}\delta(\tau^\mu - 1) \quad (55)$$

for each μ, with $-1 \leq m \leq 1$. The average $\langle\langle\cdot\rangle\rangle$ is then with respect to the distributions of both $\{\xi_j^\mu\}$ and $\{\tau^\mu\}$.

Preliminary calculations in this case are similar as in Section 3.1, but instead of the condition (15) we have here

$$\sum_k \operatorname{sgn} x_{k\mu} + \tau^\mu - K\tau^\mu \geq 0 . \quad (56)$$

Further combinatorial arguments depend on the signs of $\{\tau^\mu\}$. For the case $\tau^\mu = +1$ the analogue of expressions (16), (17) and (40)–(43) is simply $\Sigma_{4a} = 1$, whereas, if $\tau^\mu = -1$, $\Sigma_{4b} = 2^K - 1$.

The quantities $\{\tau^\mu\}$ do not depend on the index k and one can perform the average over $\{\tau^\mu\}$ in an easy manner. Straightforward calculations lead to the final expression for the entropy function,

$$\begin{aligned} \mathcal{S} &= \ln 2^{KN} + \ln \left[\frac{1^{\alpha\frac{1+m}{2}N} \left(2^K - 1\right)^{\alpha\frac{1-m}{2}N}}{2^{\alpha KN}} \right] \\ &= KN \ln 2 + \alpha N \left[\frac{1+m}{2} \ln\left(\frac{1}{2^K}\right) + \frac{1-m}{2} \ln\left(1 - \frac{1}{2^K}\right) \right] , \end{aligned} \quad (57)$$

so that

$$\alpha_c = \frac{2K \ln 2}{2 \ln 2^K - (1-m)\ln(2^K - 1)} . \quad (58)$$

5.2. Tree-connected structure

The dynamics of a tree-connected structure can be written

$$\eta_k = \mathrm{sgn}\left[\sum_{j=[N(k-1)/K]+1}^{Nk/K} J_{kj}\xi_j\right] \qquad \text{for every } k = 1,\ldots,K, \tag{59}$$

$$\tau = \mathrm{sgn}\left[\sum_{k=1}^{K} \eta_k + 1 - K\right] , \tag{60}$$

with N/K being an integer.

The entropy function takes here the form

$$\mathcal{S} = \ln\left\langle\!\left\langle \sum_{\{J_{kj}=\pm 1\}} \prod_{\mu=1}^{p} \Theta\left[\tau^\mu \sum_{k=1}^{K} \mathrm{sgn}\left(\sum_{j=[N(k-1)/K]+1}^{Nk/K} J_{kj}\xi_j^\mu\right) + \tau^\mu - K\tau^\mu\right]\right\rangle\!\right\rangle$$

$$\equiv \ln\langle\!\langle \mathcal{A}\rangle\!\rangle \tag{61}$$

and, after some calculations, we arrive at

$$\mathcal{A} = \frac{2^N}{\pi^{\frac{\alpha K N}{2}}} \frac{1}{(\det M_{\mu\mu'})^{\frac{K}{2}}} \int_{-\infty}^{+\infty} \prod_{k,\mu} dx_{k\mu} \prod_{\mu} \Theta\left(\sum_k \mathrm{sgn}\, x_{k\mu} + \tau^\mu - K\tau^\mu\right)$$

$$\times \exp\left[-\sum_{k,\mu,\mu'} x_{k\mu} M_{\mu\mu'}^{-1} x_{k\mu'}\right] . \tag{62}$$

The final result reads

$$\alpha_c = \frac{2\ln 2}{2\ln 2^K - (1-m)\ln(2^K - 1)} . \tag{63}$$

The capacities per adaptable synapse for fully connected and tree-structured networks are then identical; see Eqs. (58) and (63). The same holds obviously also for the *Committee* and *Parity Machines* studied in Sections 3 and 4. Note that if $m = +1$, we have $\eta_k^\mu = +1$ for all k, μ and the storage ratio,

$$\alpha_c = \frac{1}{K} , \tag{64}$$

decreases with K. This can be interpreted by saying that it becomes increasingly difficult for all the hidden units to assume the value $+1$ as K increases.

6. General Machine

In the case of the *General Machine* one allows all the synaptic couplings to vary, including the $\{W_k\}$ which connect hidden units with the output.

The dynamics of the fully and tree-connected *General-Committee Machines* can be described by

$$\eta_k = \operatorname{sgn}\left[\sum_{j=1}^{N} J_{kj}\xi_j\right] \qquad \text{for each } k = 1, \ldots, K, \tag{65}$$

$$\tau = \operatorname{sgn}\left[\sum_{k=1}^{K} W_k\eta_k\right], \tag{66}$$

and

$$\eta_k = \operatorname{sgn}\left[\sum_{j=[N(k-1)/K]+1}^{Nk/K} J_{kj}\xi_j\right] \qquad \text{for every } k = 1, \ldots, K, \tag{67}$$

$$\tau = \operatorname{sgn}\left[\sum_{k=1}^{K} W_k\eta_k\right], \tag{68}$$

respectively. The fully and tree-structured *General-Parity Machines* are driven by the updating rules,

$$\eta_k = \operatorname{sgn}\left[\sum_{j=1}^{N} J_{kj}\xi_j\right] \qquad \text{for each } k = 1, \ldots, K, \tag{69}$$

$$\tau = \operatorname{sgn}\left[\prod_{k=1}^{K} W_k\eta_k\right], \tag{70}$$

as well as

$$\eta_k = \operatorname{sgn}\left[\sum_{j=[N(k-1)/K]+1}^{Nk/K} J_{kj}\xi_j\right] \qquad \text{for every } k = 1, \ldots, K, \tag{71}$$

$$\tau = \operatorname{sgn}\left[\prod_{k=1}^{K} W_k\eta_k\right]. \tag{72}$$

In the case of the fully and tree-connected *General-AND Machines* the dynamics takes the form

$$\eta_k = \operatorname{sgn}\left[\sum_{j=1}^{N} J_{kj}\xi_j\right] \qquad \text{for each } k = 1, \ldots, K, \tag{73}$$

$$\tau = \mathrm{sgn}\left[\sum_{k=1}^{K} W_k \eta_k + 1 - K\right] , \tag{74}$$

and

$$\eta_k = \mathrm{sgn}\left[\sum_{j=[N(k-1)/K]+1}^{Nk/K} J_{kj}\xi_j\right] \qquad \text{for every } k = 1, \ldots, K, \tag{75}$$

$$\tau = \mathrm{sgn}\left[\sum_{k=1}^{K} W_k \eta_k + 1 - K\right] , \tag{76}$$

respectively.

The basic form of the partition function (2) for the fully connected *General-Committee Machine* with dynamics given by (65) and (66) and binary strengths $\{J_{kj} = \pm 1\}$, $\{W_k = \pm 1\}$ reads

$$\mathcal{Z} = \sum_{\{J_{kj}=\pm 1\},\{W_k=\pm 1\}} \exp\left\{-\beta \sum_{\mu=1}^{p} \Theta\left[-\tau^\mu \sum_{k=1}^{K} W_k \,\mathrm{sgn} \sum_{j=1}^{N} J_{kj}\xi_j^\mu\right]\right\} . \tag{77}$$

It is obvious that here one can introduce the auxiliary variables $T_{kj} = W_k J_{kj}$ for each k and j. The partition function can then be written

$$\mathcal{Z} = 2^K \sum_{\{T_{kj}=\pm 1\}} \exp\left\{-\beta \sum_{\mu=1}^{p} \Theta\left[-\tau^\mu \sum_{k=1}^{K} \mathrm{sgn} \sum_{j=1}^{N} T_{kj}\xi_j^\mu\right]\right\} . \tag{78}$$

The annealed entropy in the noiseless limit $\beta \to \infty$ is thus given by

$$\mathcal{S} = K \ln 2 + \ln\left\langle\left\langle \sum_{\{J_{kj}=\pm 1\}} \prod_{\mu=1}^{p} \Theta\left[\tau^\mu \sum_{k=1}^{K} \mathrm{sgn}\left(\sum_{j=1}^{N} J_{kj}\xi_j^\mu\right)\right]\right\rangle\right\rangle . \tag{79}$$

The second term in above expression is exactly the same as the right-hand side of Eq. (10); see Section 3.1. In order to evaluate the entropy in a second manner (as the Boltzmann function) we should take into account the following degeneracy: For any value(s) of $k = k_0$ and any solution $\{W_{k_0},\{J_{k_0 j}\}\}$ also the configuration $\{-W_{k_0},\{-J_{k_0 j}\}\}$ can perfectly remember the same set of data. In the saturation limit we then end up with

$$\mathcal{S} = \ln 2^K . \tag{80}$$

Thus the storage capacity takes the value

$$\alpha_c = K . \tag{81}$$

We can also investigate all the other machines and architectures — for the proper dynamics equations see the beginning of this Section — with adaptive weights which connect the hidden layer with output unit. In such cases one has to add to all the entropy functions given by expressions (30), (36), (48), (53) and (61) the term $K \ln 2$, but also the critical-storage condition $\mathcal{S} = 0$ is changed to be $\mathcal{S} = K \ln 2$. We obtain that the maximal capacity of the fully connected *General-Parity Machine* is also given by (81). The storage ratio for the tree-structured *General-Committee* and *Parity Machines* reads

$$\alpha_c = 1 \; . \tag{82}$$

The fully connected *General-AND Machine* can store per input

$$\alpha_c = \frac{2K \ln 2}{2 \ln 2^K - (1-m) \ln(2^K - 1)} \; , \tag{83}$$

and for this network with the tree-architecture one has

$$\alpha_c = \frac{2 \ln 2}{2 \ln 2^K - (1-m) \ln(2^K - 1)} \; . \tag{84}$$

All the above results are valid for any value of K. We conclude that machines in which the couplings connecting the intermediate layer with the output node are allowed to vary, cannot store more patterns than networks with fixed *a priori* values of the weights $\{W_k\}$. This general conclusion also holds for the calculations done with the "quenched" entropy, Eq. (6).

We can generalize the results of this Section in an easy way to more-than-two-layered networks which can have any architecture and any kind of connections (continuous or two-valued) between input and the first hidden layer, but are tree-structured, with binary weights, between the first intermediate layer, any number of further hidden ones, and the output node. The methods to handle such a case are very similar to those described in this Section. One now has to introduce "super-weights" — $\{T_{kj}\}$ — for the whole network. Our calculations demand here a slightly more complicated combinatorics than that in Sections 3, 4 and 5 but give the same result as in the previous parts of this work. The only difference for the final result is caused by the fact that in the expression for the partition function (78) one now encounters, instead of the term 2^K, the factor 2^R where R is the total number of all the hidden connections between the first intermediate layer and the output unit; see also Ref. [15]. On the other hand, in the saturation limit, the Boltzmann entropy should have the form $\mathcal{S} = R \ln 2$. The storage capacity ratio, checked against the results of parts 3, 4 and 5 of this paper, thus remains unchanged.

7. Conclusions

In this work we have investigated four basic types of two-layered network with two kinds of architectures as well as with binary synaptic couplings. Our interest was focused on the storage capabilities of these associative machines. In order to estimate their maximal memory capacity we have adopted a statistical mechanics approach. It is worth stressing that the calculations done in this paper are valid for any number of the hidden units.

We have compared our results with both theoretical and numerical findings which are known from the literature. In some cases ($K \to \infty$, networks with tree-architectures, and especially the *Parity Machine* and structures with small numbers of input neurons N) a good (or even very good) agreement with computer simulations has been found. In the limit $K \to \infty$, our results are also very close to the theoretical findings obtained within much more complicated, from the mathematical point of view, replica method (we should stress that there does not exist, to our knowledge, any attempt to apply replica method to the case of fully connected networks). The results obtained for the *Parity Machines* seem to even be exact (for any value of K). We would like, however, to add that our annealed-approximation results concerning *Committee* and *Parity Machines* may be obtained using other, well-known and somewhat simpler methods, than the one presented in this paper. Nevertheless, our approach seems to be quite generic and could be applied to each kind of network with any activation rule, architecture, or couplings. Therefore, we have presented it in detail.

At the end of this paper, we would like to attract reader's attention to two points:

1. All of our results have been obtained using the annealed approximation; see Eq. (7). One can conclude that this approach may sometimes give better (closer to reality) results than the replica method, especially if the replica-symmetry-breaking effects are hard to incorporate. It is also well-known [23] that in some cases the logarithm of specified random functions (of the form very similar to the ones investigated in this work) is a self-averaging quantity.

2. We might wish to try a further attempt to study networks with more than two layers. After Section 6 we may, however, conclude that the storage capacity of many-layered networks need not be much greater than that of two-layered perceptrons or, strictly speaking, may be close to this value. In fact, the connections between inputs and the first hidden layer play the most important — decoding role for sending synaptic signals by associative machines. Other (further) connections mainly provide the execution of the proper Boolean function which has to be performed by the network.

One of us (W.T.) thanks the *Alexander von Humboldt-Stiftung* for making his research stay in Germany possible. W.T. would also like to gratefully acknowledge Professor Maciej Lewenstein for many enlightening discussions.

REFERENCES

[1] E. Domany, J.L. van Hemmen, K. Schulten, editors, *Models of Neural Networks*. Springer–Verlag, Berlin 1991.

[2] D.J. Amit, *Modeling Brain Function: The World of Attractor Neural Networks*. Cambridge University Press, Cambridge 1989.

[3] E. Gardner, *J. Phys.* **A21**, 257 (1988).

[4] E. Gardner, B. Derrida, *J. Phys.* **A21**, 271 (1988).

[5] W. Krauth, M. Mézard, *J. Phys. France* **50**, 3057 (1989).

[6] W. Tarkowski, M. Komarnicki, M. Lewenstein, *J. Phys.* **A24**, 4197 (1991).

[7] H.S. Seung, H. Sompolinsky, N. Tishby, *Phys. Rev.* **A45**, 6056 (1992).

[8] T. L. H. Watkin, A. Rau, M. Biehl, *Rev. Mod. Phys.* **65**, 499 (1993).

[9] W. Tarkowski, M. Lewenstein *J. Phys.* **A26**, 2453 (1993).

[10] W. Tarkowski, M. Lewenstein, *J. Phys.* **A26**, 3669 (1993).

[11] D.E. Rumelhart, J.L. McClelland, editors, *Parallel Distributed Processing: Explorations in Microstructure of Cognition*, volume I and II, MIT Press, Cambridge, Mass. 1986.

[12] E. Barkai, D. Hansel, I. Kanter, *Phys. Rev. Lett.* **65**, 2312 (1990).

[13] E. Barkai, I. Kanter, *Europhys. Lett.* **14**, 107 (1991).

[14] M. Biehl, M. Opper, *Phys. Rev.* **A44**, 6888 (1991).

[15] I. Kocher, R. Monasson, *J. Phys.* **A25**, 367 (1992).

[16] I. Kanter, *Europhys. Lett.* **17**, 181 (1992).

[17] E. Barkai, D. Hansel, H. Sompolinsky, *Phys. Rev.* **A45**, 4146 (1992).

[18] A. Engel, H.M. Köhler, F. Tschepke, H. Vollmayr, A. Zippelius. *Phys. Rev.* **A45**, 7590 (1992).

[19] M. Griniasty, T. Grossman, *Phys. Rev.* **A45**, 8924 (1992).

[20] E. Eisenstein, I. Kanter, *Europhys. Lett.* **21**, 501 (1993).

[21] M. Mézard, G. Parisi, M. Virasoro, editors, *Spin Glass Theory and Beyond*, World Scientific, Singapore 1987.

[22] W. Krauth, M. Opper, *J. Phys.* **A22**, L519 (1989).

[23] W. Tarkowski, M. Lewenstein, *J. Phys.* **A25**, 6251 (1992).