Quasi-optimized memorization and retrieval dynamics in sparsely connected neural network models

Karl E. Kürten

Classification
Physics Abstracts
87.30 — 75.10H — 64.60

Quasi-optimized memorization and retrieval dynamics in sparsely connected neural network models

Karl E. Kürten

Institut für Theoretische Physik, Universität zu Köln, D-5000 Köln, F.R.G. and HLRZ c/o KFA Jülich, D-5170 Jülich 1, F.R.G.

(*Reçu le 29 mars 1990, accepté le 4 avril 1990*)

Abstract. — Several network topologies with sparse connectivity suitable for information processing in neural network models are studied. A trial and error scheme for quasi-optimal neighbour search is shown to successfully overcome instabilities giving rise to chaotic behaviour. Moreover, computer simulations reveal that for random unbiased patterns the connectivity of the network can be adapted to the specific structure of the information the network is asked to capture such that it can be quasi-optimally stabilized. On the other hand, networks with random or purely nearest neighbour interactions are not competitive candidates for the realization of associative memories. It is shown further that sparsely connected network models with quasi-optimal neighbour search techniques substantially outperform their fully connected counterparts.

1. Introduction.

There has been a tremendous amount of interest in dynamical neural network models exhibiting collective phenomena associated with information processing and cognitive behaviour : storage and retrieval of memories, learning and forgetting, association and abstraction [1-3]. One detrimental assumption of some of the currently popular neural network models is full connectivity of the neuron-like elements. On the one hand, high connectivity is rarely found in nature. On the other, highly connected networks of respectable size, suitable for such cognitive tasks as pattern recognition and image processing, raise immense wiring problems in hardware realization and make real-time software simulations computationally expensive. In fact, it is well known that realistic biological networks are rather sparsely interconnected, and it is also well known that systems with *low* connectivity are often amenable to analytical study [3, 5].

2. Formalism.

The network model consists of N sparsely interconnected logical units only capable of the values $\sigma_i = 1$ and $\sigma_i = -1$, where each cell i is supposed to receive exactly K inputs from

other units of the network. The dynamical time evolution of the system is then formulated in terms of the local fields $h_i(t)$

$$h_i(t) = \sum_{j \neq i} c_{ij}\, \sigma_j(t)/|c|_i \quad i = 1, ..., N\,, \tag{1}$$

where j runs over the arbitrary K neighbours of cell i, not necessarily nearest neighbours, and the normalization factor $|c|_i$ is defined by

$$|c|_i = \left(\sum_{j \neq i} c_{ij}^2 \right)^{1/2} \quad i = 1, ..., N\,. \tag{2}$$

The interaction matrix C is assumed to be real and h_i can be interpreted as the net internal stimulus from cells synapsing onto cell i represented by the usual linear sum of weighted inputs. The actual binary state is then determined *via* the deterministic threshold rule

$$\sigma_i(t+1) = \text{sign}\ (h_i(t)) \quad i = 1, ..., N\,, \tag{3}$$

The network is now supposed to learn a series of $p = \alpha K$ random, unbiased configurations $\Pi^{(\mu)} = (\pi_1^{(\mu)}, ..., \pi_N^{(\mu)})$ $\mu = 1, ..., p$ with p being proportional to the connectivity parameter K. A necessary through minimal requirement is that the pN polarization parameters, $R_\nu^{(\mu)}$, defined as

$$R_i^{(\mu)} = \pi_i^{(\mu)} h_i(\pi_1^{(\pi)}, ..., \pi_N^{(\mu)}) \qquad i = 1, ..., N \quad \mu = 1, ..., p \tag{4}$$

are larger than zero, which guarantees that the patterns to be memorized are fixed points of the dynamics (1). In order to engrave the information as strongly as possible one has to adjust the couplings such that the polarization parameters in (4) satisfy the inequality

$$R_i^{(\mu)} > \kappa \quad i = 1, ..., N \quad \mu = 1, ..., p \tag{5}$$

with the quantity κ as large as possible. In the case of random, uncorrelated patterns and random connectivity the upper storage capacity α for given κ has been shown to satisfy [2]

$$\alpha = (p/K) = \left\{ \frac{1}{\sqrt{2\pi}} \int_{-\kappa}^{+\infty} (t+\kappa)^2 \times \exp\left(-\frac{1}{2} t^2\right) \mathrm{d}t \right\}^{-1} \tag{6}$$

in the thermodynamic limit. Equation (6) implies that the critical $\alpha_c = 2$ is reached, when κ vanishes. In view of technical applications, however, where the total number of cells N is finite and some patterns might be strongly correlated, one is faced with the problem to produce a set of c_{ij} so as to satisfy condition (5) with κ as large as possible. A straightforward generalization of the perceptron learning algorithm [6] suggests the following local dynamical learning rule

$$c_{ij}(t+1) = c_{ij}(t) + \frac{1}{K} \sum_{\mu=1}^{p} \theta(\kappa - R_i^{(\mu)})\, \pi_i^{(\mu)}\, \pi_j^{(\mu)}\,. \tag{7}$$

Thus, patterns well embedded have no impact on the modification of the synaptic efficiencies, whereas patterns not satisfying (5) become effective in modifying the couplings. If then, on average, the magnitude of the polarization parameters $R_i^{(\mu)}$ in (4) are maximized the sizes of the basins of attraction are also expected to be fairly close to their optimal magnitudes. Here it is useful to start from « tabula rasa » rather than from randomly chosen coupling coefficients.

3. Stability analysis.

Exact results can be derived in the thermodynamic limit for connectivity parameters $K \propto \text{Log}\ (N)$, if the connectivity as well as the patterns are chosen at random. The fractional Hamming distance between a spin configuration and an arbitrary stored pattern at time t can be given as a function of the distance at the previous time step and the polarization parameter κ related to the critical value of α by (6) [10, 5]

$$\begin{aligned} H_{t+1} &= \Phi[H_t, \kappa] \\ &= \frac{1}{\sqrt{2\pi}} \int_{\kappa(1/2-H)/\sqrt{H(1-H)}}^{+\infty} \exp\left(-\frac{1}{2}x^2\right) \\ &\quad \times \text{erf}\left[x \frac{\sqrt{H(1-H)}}{\frac{1}{2}-H}\right] \mathrm{d}x\,. \end{aligned} \tag{8}$$

Here, erf (z) denotes the error function defined as

$$\text{erf}\ (z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{1}{2}x^2\right) \mathrm{d}x\,. \tag{9}$$

The time evolution of $H(t)$ admits the natural fixed points $H \equiv 0$ and $H \equiv \frac{1}{2}$ and the existence of a phase transition from an ordered to a chaotic behaviour depends only on the magnitude of the slope of $\Phi(H)$ evaluated at the fixed points.

$$\frac{\mathrm{d}\Phi}{\mathrm{d}H} = \exp\left[-\frac{1}{2}\left(\frac{\kappa(1-2H)}{\sqrt{H(1-H)}}\right)^2\right] \left\{ \frac{\text{erf}\ (\kappa)\,\kappa}{(2\sqrt{H(1-H)})^2} + \frac{\exp\left(-\frac{1}{2}\kappa^2\right)}{\sqrt{2\pi}} \right\} / \sqrt{2\pi H(1-H)}\,. \tag{10}$$

Figure 1 shows that $\mathrm{d}\Phi/\mathrm{d}H_{|\,H\equiv 0}$ exhibits a singularity at $\kappa = 0$, which allows no recognition for $\alpha_c = 2$ since the fixed point $H \equiv 0$ is unstable. On the contrary, for $\kappa \neq 0$, $\mathrm{d}\Phi/\mathrm{d}H_{|\,H\equiv 0}$ and all higher derivatives remain strictly zero. Hence, the fixed point $H \equiv 0$ is always stable and any stored pattern can in principle be recovered if it shares a sufficiently large overlap with the initial condition. Moreover, as can be seen from figure 1, for $0 < \kappa < \kappa_G \approx 1.1$, $\mathrm{d}\Phi/\mathrm{d}H_{|\,H\equiv 1/2}$ does not exceed the value one which implies that in this parameter regime the fixed point $H \equiv \frac{1}{2}$ is also stable, giving rise to the existence of a strictly unstable fixed point H^*. Hence, the memory of initial conditions can be completely erased, since the time evolution of the overlap of slightly different configurations with a stored pattern might evolve to zero. According to equation (6) κ_G corresponds to a critical capacity $\alpha_G = 0.42$.

In contrast to random model networks [5] the prevalence of the ordered and chaotic phase not only depends on the network parameters, but also on the initial conditions. For κ above the critical κ_G the unstable fixed point H^* disappears and the fixed point $H \equiv \frac{1}{2}$ changes its stability from attraction to repulsion. Then, any initial configuration with a Hamming distance $H(t=0) < \frac{1}{2}$ with a stored pattern will lead to perfect recognition.

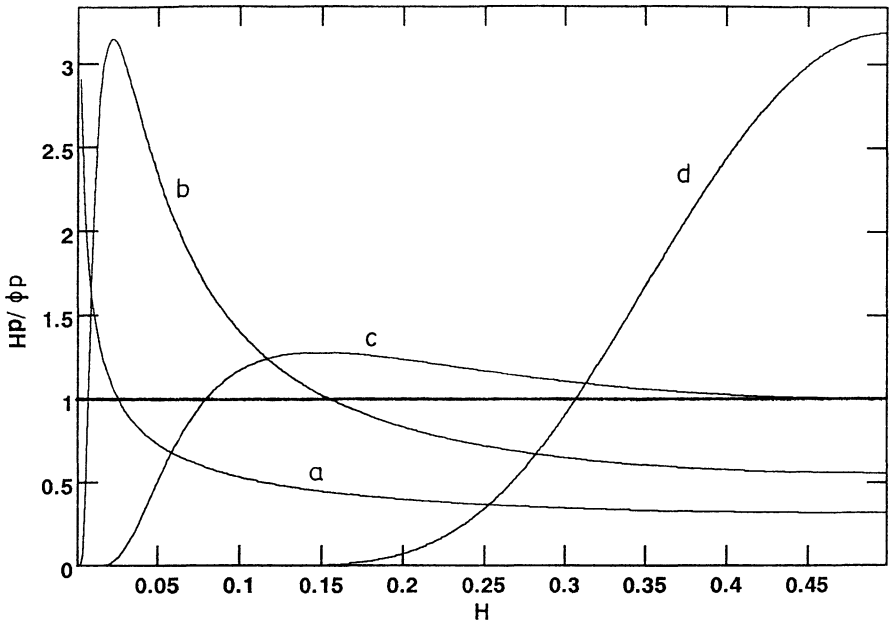

Fig. 1. — The slope of $\Phi(H)$ for a) $\kappa = 0$, b) $\kappa = 0.5$, c) $\kappa = 1.198$ and d) $\kappa = 4$.

For $H \ll \frac{1}{2}$ and $\kappa \neq 0$ equation (8) can be well approximated by

$$H_{t+1} = \text{erf}\ (\kappa)\left\{1 - \text{erf}\ \kappa\ \frac{\frac{1}{2} - H}{\sqrt{H(1-H)}}\right\}, \tag{11}$$

whereas for H close to the value $\frac{1}{2}$

$$H_{t+1} = \frac{1}{2} + 2\left(H - \frac{1}{2}\right) \times \left\{\text{erf}\ (\kappa)\ \kappa + \exp\left[-\frac{1}{2}\kappa^2\right] / \sqrt{2\pi}\right\} \tag{12}$$

is a good approximation. Equation (11) suggests rather fast local convergence to a stored pattern. Note that equation (8) is a scalar mean field equation in the thermodynamic limit and the two basins of attraction are well separated by the relative position of the unstable fixed point H^*. However, for a finite number of cells a small fractional Hamming distance does not guarantee convergence to a stored pattern, since judgement involves a measure in n-dimensional space, where, depending on the number of stored memories, the basins of attraction can be degraded by large crevices and holes close to the stored information. These holes allow errors to be incurred for small Hamming distances from the stored states and degrade the network performance substantially.

4. Computer simulations.

In this section we give some results of our computer simulations for three network models based on different topologies. We try to embed $p = \alpha K$ patterns within at most 25 learning

passes and stop the learning procedure (7) when the average of all polarization parameters reaches a near-optimal value. Before the retrieval phase the original patterns are degraded by random noise, and presented as initial conditions for the network to recall.

4.1 NEXT-NEIGHBOUR SQUARE LATTICE MODEL. — We first study networks with cells residing on a two-dimensional square lattice restricted to exclusively eight next-nearest neighbour interactions. Figure 2 shows that the fraction of bits not satisfying condition (5) even with $\kappa = 0$ grows with increasing α.

Thus, a portion of the patterns are not fixed points of the dynamics (1) and consequently cannot be recalled perfectly. Though for $\alpha = 0.375$ and $\alpha = 0.5$ figure 2 shows that the output-input ratio is essentially larger than one, the network is hardly capable of functioning as an associative memory. The results are in accord with Forrest's findings [7] who reports a somewhat smaller output-input ratio explainable through the restriction of only allowing clipped coupling coefficients, which reduces the storage capacity and the performance of the network. For $\alpha = 0.625$ the situation gets worse, where the retrieval rate is in fact less than the input rate. These results are independent of N for $N \geqslant 225$ and have been performed for $N = 225, N = 625$ and $N = 2\,500$ respectively, averaged over ten different initial configurations and some twenty independent systems.

It is obvious, therefore, that models with exclusively nearest-neighbour interactions are not at all designed to process *random* patterns.

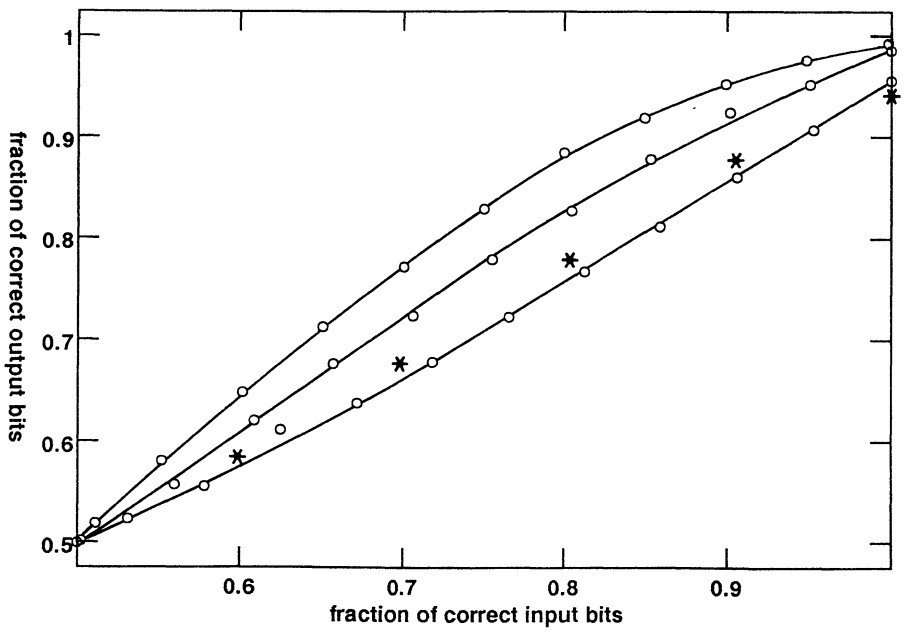

Fig. 2. — Mean fraction of correctly recalled bits as a function of the fraction of initially correct bis for $K = 8$ (square lattice), $\alpha = 0.375$, $\alpha = 0.5$ and $\alpha = 0.625$ (from above). Stars represent Forrest's results for $\alpha = 0.5$ based on clipped coupling coefficients ($|c_{ij}| = 1$).

4.2 RANDOM NEIGHBOUR LONG-RANGED MODEL. — In a next step we study a somewhat less restricted and more realistic class of networks, where the units choose their arbitrary neighbours at random.

As has been found for the next-nearest-neighbour model, computer simulations for $K = 8$ reveal that after the learning process a small portion of the patterns to be memorized are not fixed points of the dynamics (3), though this fraction decreases to zero with increasing N. Figure 3 shows the mean fraction of correctly retrieved spins as a function of the fraction of initially correct spins for $\alpha = 0.375$ and different total number of cells N. Note the strong size dependence clearly indicating that, for large N, an initial damage of less than fifty percent leads to an essentially perfect recall as predicted by Gardner's theoretical results [10] in section 3.

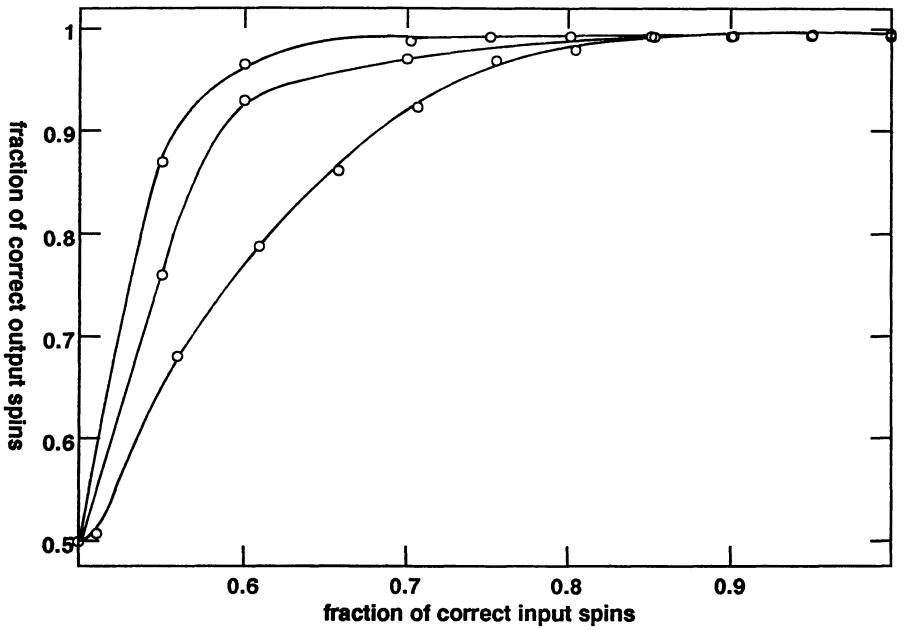

Fig. 3. — Mean fraction of correctly recalled spins as a function of the fraction of initially correct spins for $K = 8$ (random connectivity), $N = 225$, $N = 625$ and $N = 2\,500$ cells for $\alpha = 0.375$ (from below).

As to be expected the situation changes drastically for $\alpha = 0.5$. The stability of the fixed point $H \equiv \frac{1}{2}$ reduces substantially the magnitude of the basins of attraction of the individual attractors. We observe that with increasing degree of noise the average transient lengths for parameter values above the critical $\alpha_G = 0.42$ increases dramatically, being of the order of several magnitudes larger than below the critical α_G. Moreover, the system often does not lock into a fixed point after more than some five hundred time steps wandering around in phase space indicating that attractors corresponding to chaotic behaviour become dominant. Though the performance of random neighbour models is perfect for $\alpha = 0.375$ it performs very poorly and even breaks down for capacities α above the critical value α_G. The sudden appearance of long transients and failures in reaching a cyclic mode is quite reminiscent of damage spreading effects in networks with randomly chosen couplings, where above a critical parameter value disorder grows exponentially with the number of cells [8, 9] and chaotic dynamics prevails.

4.3 QUASI-OPTIMAL LONG-RANGED NEIGHBOUR MODEL. — One would like to have sparsely connected network models capable of an associative recall without error for sizes of the

storage capacity far beyond the critical capacity $\alpha_G = 0.42$. Furthermore, all the patterns to be memorized should be stable fixed points of the dynamics (1) and their basins of attraction should be as large as possible. As a first improvement, we adopt a trial and error scheme by choosing new random neighbours if the initial random neighbour choice fails to fulfill the embedding condition (5). The strategy is repeated until the network has found its quasi-optimal connections. Figure 4 shows the recall performance of such a network for $K = 8$ and $\alpha = 0.5$ for different numbers of cells N. The final connections have been obtained after a few trials on average, however, there are large fluctuations, since most of the polarization parameters satisfy (6) already after the initial random neighbour choice.

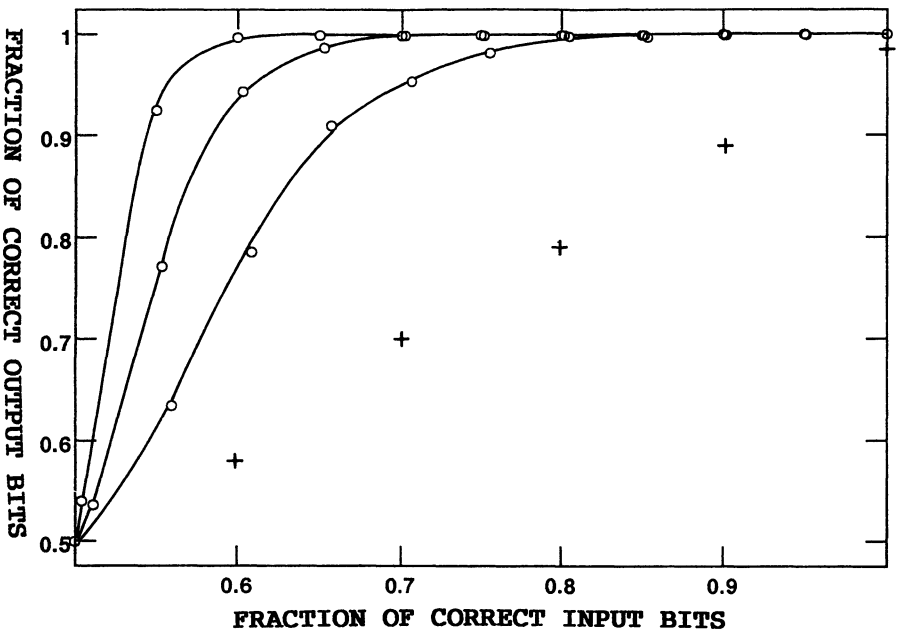

Fig. 4. — Mean fraction of correctly recalled bits as a function of the fraction of correct input bits for $K = 8$, $N = 225$, $N = 625$ and $N = 2\,500$ for $\alpha = 0.5$ (from above). Crosses show the same quantity for $N = 225$ cells *without* the trial and error strategy.

Note that the improved model for $\alpha = 0.5$ performs somewhat better than the one *without* quasi-optimization for $\alpha = 0.375$. Furthermore, computer simulations indicate that also for α far beyond 2 the trial and error scheme works successfully, though with increasing α the learning time increases rapidly due to larger amounts of trials until the embedding condition (5) is fulfilled. In figure 5 we compare the performance of these optimized sparsely connected networks to that with a fully connected model for $\alpha = 0.5$. Though in all networks the amount of stored information *in bits* is alike (3 200 bits), and all networks have the same number of 6 400 couplings it is quite clear from the figure that the sparsely connected networks outperform their fully connected counterparts.

Note, however, that for structured information, the information content is *not* the same, since with increasing N and decreasing number of connections K one can embed a smaller number of patterns and the diversity of the network capability shrinks substantially. On the other hand, the information content per bit is strongly dependent on the nature of the

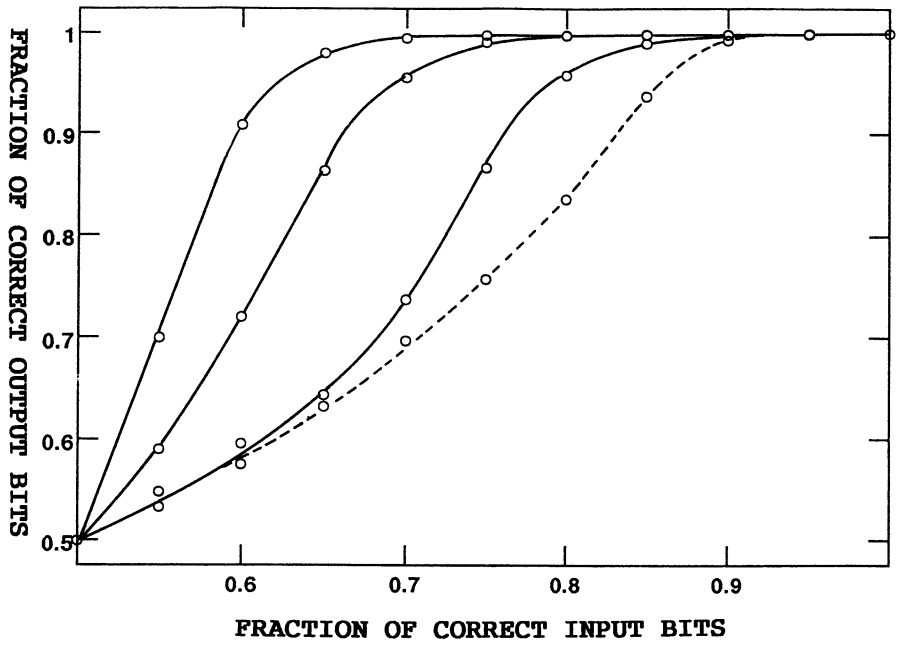

Fig. 5. — Mean fraction of correctly recalled spins as a function of the fraction of correct input spins for sparsely connected networks consisting of $N = 800$, $N = 400$ and $N = 200$ cells with $\alpha = 0.5$ and $K = 8$, $K = 16$ and $K = 32$ couplings, respectively (from the stop). The dashed line shows the same quantity for a fully connected network consisting of $N = 80$ cells.

patterns. Due to stronger correlations within the individual patterns the information content per bit decreases strongly with increasing resolution in the case of structured patterns. Thus, according to figure 5, depending on technical tasks, one can choose between memorizing a few particular patterns with relatively high resolution (K smaller) and a larger number of patterns with relatively low resolution (K larger).

Let us now study the performance of the network with optimized connectivity, where the ratio of the number of connections and the number of cells $\gamma = K/N$ remains constant. Figure 6 shows the retrieval performance of sparsely connected networks with $\gamma = 0.16$ compared to fully connected networks for different numbers of N and $\alpha = 0.5$.

As has been observed by Forrest [4] for the case $\gamma = 1$ the recall fraction exhibits a crossover effect for increasing N for $\gamma < 1$, too. As the number of cells N increases, there is a dramatic increase (decrease) in the level of errors, when the fraction of correct input bits is reduced below (increased above) a critical value. It is quite evident that the smaller the quantity γ the more the critical retrieval value is reduced so that the performance of the network improves.

Our network model is not restricted to the storage of random information but is also able to process artificially structured patterns [11]. Here, we first connect each cell with its eight next-nearest neighbours, while additional suitable long-ranged connections emerge from a trial and error scheme. The outcoming architecture is reminiscent of Braitenberg's picture [12] about the interconnectivity of cortical pyramidal cells making short-range connections with basal dendrites and long-range connections with apical dendrites. According to the physiological rule that low-efficacy synapses degenerate we suggest to apply a trial and error scheme by substituting bonds whose magnitudes are close to zero. Preliminary results reveal that this

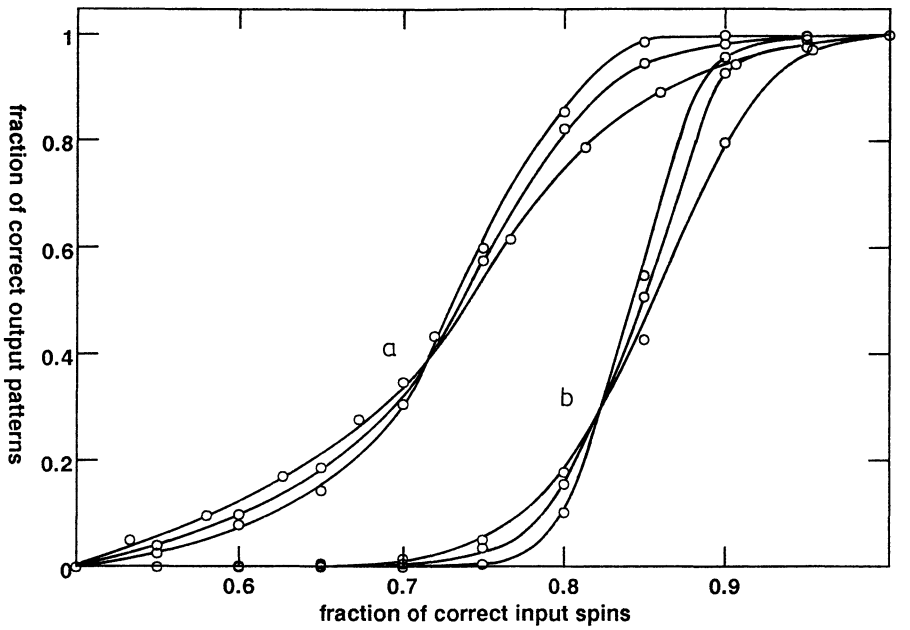

Fig. 6. — Mean fraction of correctly recalled patterns as a function of the fraction of initially correct spins for a) $N = 100$, $N = 200$ and $N = 300$ cells with $\gamma = 0.16$, and b) $N = 40$, $N = 80$ and $N = 120$ cells with $\gamma = 1$. The storage parameter α has been chosen as 0.5.

biologically motivated strategy is highly beneficial for improving the convergence to near optimal answers.

5. Conclusion.

We have shown that for randomly chosen information the performance of optimized sparsely connected networks is quite superior to that of short-ranged lattice models as well as to that of their fully-connected counterparts. It is quite evident that our strategy for solving the combinatorial optimization problem can be largely improved. Here, parallel genetic algorithms [13], in the spirit of the discrete Mendelian genetic model can be introduced with a view to get closer to near-optimal answers. The general philosophy is that the specific neighbour choice, as used here, and in future work also other network parameters such as degree of connectivity and thresholds ought to be adapted to the specific information the network has to capture and hence can be subject to modification during an optimization process.

Acknowledgments.

The author gratefully acknowledges financial support for this work by the German Federal Department of Research and Technology (BMFT) under grant Nr. ITR 8800K4. It is a pleasure to thank the von Seelen group for their hospitality at the Institute for Neuroinformatics in Bochum during several visits. The author appreciates many useful and stimulating discussions with J. W. Clark, J. Duarte, M. Husson, J. L. van Hemmen, U. Keller, G. Kohring, G. Senger and D. Stauffer.

References

[1] HOPFIELD J. J., *Proc. Natl. Acad. Sci.* **79** (1982) 2554.
[2] GARDNER E., *J. Phys. A* **21** (1987) 257.
[3] DERRIDA B., GARDNER E. and ZIPPELIUS A., *Europhys. Lett.* **4** (1987) 167.
[4] FORREST B. M., *J. Phys. A : Math. Gen.* **21** (1988) 245.
[5] KÜRTEN K. E., *Phys. Lett. A* **129** (1988) 157.
[6] MINSKY M. L. and PAPERT S., Perceptrons (Cambridge, MA : MIT Press) 1969.
[7] FORREST B. M., *J. Phys. France* **50** (1989) 2003.
[8] STAUFFER D., *Philos. Mag. B* **56** (1987) 901.
[9] KÜRTEN K. E., *J. Phys. France* **50** (1989) 2313.
[10] GARDNER E., *J. Phys. A* **22** (1989) 1969.
[11] KÜRTEN K. E., Parallel Processing in neural systems and computers, Eds. R. Eckmiller, G. Hartmann and G. Hauske (North-Holland) 1990, pp. 191-194.
[12] BRAITENBERG V., *Lect. Notes Biomath.* (Springer Verlag Berlin) **21** (1978) 171.
[13] MÜHLENBEIN H. and KINDERMANN J., Connectionism in Perspective, Eds. R. Pfeifer *et al.* (North-Holland) 1989, pp. 173-198.