Associative memory with high information content

J. Buhmann, R. Divko, and K. Schulten
Physik-Department, Technische Universität München, James-Franck-Strasse, 8046 Garching bei München, Federal Republic of Germany
(Received 11 January 1988; revised manuscript received 19 October 1988)

An extension of Hopfield's model with adaptive threshold and inhibitory interactions yields a network capable of nearly optimal storage of patterns of low activity. A replica symmetric solution of the mean-field equations is presented for the noise-free case. The following properties are demonstrated: For a low level of activity a the storage capacity increases as $-(a \ln a)^{-1}$; up to 0.38 bits per synapse can be stored; spurious states can be suppressed; the network is not opinionated, i.e., it can categorize inputs as not similar enough to patterns stored.

The Hopfield dynamic spin model[1] with symmetric, nonlocal heterogeneous spin-spin interactions presents an extreme abstraction of biological neural networks. It can be analyzed by methods of statistical mechanics of spin-glasses[2] and by techniques of coding theory.[3] Such analysis has provided most important information on storage capacity, role of noise, and recall performance. Cognitive science and artificial intelligence are beginning to embrace dynamical models like neural networks for information processing, since these models provide some understanding of collective computation in systems with a large number of simple elements, and since they share some properties of biological networks, such as fault tolerance, massive parallelism without a system-wide clock cycle, and high connectivity.

Recent work on spinoidal networks focuses on two important aspects: (i) storage of sequences of patterns and (ii) optimization of storage capacity and recall behavior. In this paper we will be concerned with the latter, although our approach is also relevant to the former aspect.[4,5] One approach to optimal storage and recall involves iterative learning algorithms for the interaction matrix of a network.[6,7] This approach certainly leads to the best results, albeit requiring a computationally intensive learning phase. We want to propose an alternative method for optimal storage based on a simple relationship between patterns stored and the interaction matrix. The network model proposed is most suitable for storage of correlated patterns of small activity level a, i.e., so-called biased or "sparcely-coded" patterns which involve only a small fraction of neurons firing. We have provided our network with an adaptive neural threshold and global inhibition between neurons. These features bear most important properties: high (close to optimal, see below) storage capacity, no spurious states, a special state for "no recognition." By adjusting the threshold one can further choose between effective storage and good associativity. Storage capacity of the network, as judged by criteria derived by Gardner,[7] is optimal for small activity levels a. Recall quality of stored patterns is nearly optimal as well. The network bridges the gap between matrix memory[8] and Hopfield models; it can be interpreted as a generalization of the original Hopfield model[1] for an arbitrary level of activity a.

There exist several other proposals for effective, associative storage of sparsely coded patterns.[8,9] The model of Moopenn *et al.*[9] involves a storage prescription with local inhibition, where the space of possible patterns is restricted to diluted patterns which differ from biased patterns. Supression of spurious states is also observed in that model. Another approach to storing biased patterns in Hopfield models had been investigated in Ref. 10, however the assumed dynamics did not yield high information content.

The network we propose is composed of N neurons described by dynamic variables $\mathbf{S}=\{S_i\}_{i=1}^N$. Neuron i is either firing ($S_i=1$) or quiet ($S_i=0$). The variables are updated asynchronously according to a probabilistic rule. The rule is based on a molecular field $h_i=\sum_k W_{ik}S_k$ which represents the interaction of neuron i with all other neurons. With probability $f_i=\{1+\exp[-(h_i-U)/T]\}^{-1}$ neuron i fires at time $t+\Delta t$, otherwise it is quiet. The parameters U and T are the threshold potential and the network temperature. The patterns $\xi^\nu=\{\xi_i^\nu\}_{i=1}^N$ to be stored are chosen according to the distribution $P(\xi_i^\nu)=a\delta(\xi_i^\nu-1)+(1-a)\delta(\xi_i^\nu)$ $\forall \nu,i$ with a representing the fraction of firing neurons. For $a\neq\frac{1}{2}$ the patterns are biased. We have recently suggested a network model to store and recall sequences of M biased patterns.[4] Since pattern sequences appear to comprise a most important data format for information processing by networks, low-activity pattern storage deserves special attention. Observations on biological neural systems in the state of normal function also reveal low levels of neural activity.[11]

The synaptic connections between neurons are chosen according to a hypothesis of Hebb which postulates *assemblies* of cooperating neurons, i.e., of neurons with mutual excitatory interactions, and competition between such assemblies, i.e., inhibitory interactions between neurons belonging to different assemblies. In the present model the interaction matrix is defined by the rule

$$W_{ik}=\frac{1}{a(1-a)N}\sum_{\nu=1}^{p}(\xi_i^\nu-a)(\xi_k^\nu-a)-\frac{\gamma}{aN},\quad i\neq k\ . \qquad (1)$$

The first term in (1) describes excitatory interactions and favors firing patterns $\mathbf{S}=\xi^\nu$. The sets $\mathcal{M}^\nu=\{i|\xi_i^\nu=1\}$ correspond to Hebb's *neural assemblies* of cooperating

neurons. The second term in (1) introduces global inhibition between all neurons of the network, in particular also between neurons belonging to different $\mathcal{M}^\nu$.

The interactions W_{ik} in (1) between two neurons i,k are symmetric, i.e., $W_{ik}=W_{ki}$; self-interaction is forbidden ($W_{ii}=0$). Accordingly, an energy function $\mathcal{H}=-\frac{1}{2}\sum_{i\neq k} W_{ik}S_iS_k+U\sum_i S_i$ exists. In the case of a finite number p of stored patterns a standard analysis[2] reveals that the network state is described by two parameters

$$m^\nu=\frac{1}{aN}\left\langle\!\left\langle \sum_i (\xi_i^\nu-a)\langle S_i\rangle \right\rangle\!\right\rangle$$

and

$$x=\frac{1}{aN}\left\langle\!\left\langle \sum_i \langle S_i\rangle \right\rangle\!\right\rangle .$$

x measures the total activity in the network and m^ν essentially the overlap of the network state $\mathbf{S}$ with the pattern state ξ^ν. $\langle\langle\ \rangle\rangle$ denotes an average over quenched random variables ξ^ν and $\langle\ \rangle$ thermal averaging. The sum

$$m^\nu+ax=\frac{1}{aN}\left\langle\!\left\langle \sum_i \xi_i^\nu\langle S_i\rangle \right\rangle\!\right\rangle$$

counts the number of active neurons in a network state $\mathbf{S}$ which are also active in pattern ν. The analysis proves that at $T=0$ and for the parameter range defined by $U+\gamma<1-a$ all states exhibiting macroscopic overlap with one pattern are stable states.

Network states which one seeks to avoid during associative recall are those which involve macroscopic overlaps with more than one pattern. In the Hopfield model[2] such so-called spurious states become stable in the order of three, five, etc., overlaps when temperature is lowered. In the present model the most stable of such unwanted states are the ones exhibiting overlaps with two patterns. A stability of these states is particularly ruinous for recall of pattern sequences.[4] In the limit of small a values ($a\ll 0.1$) for sufficiently strong inhibition [$\gamma>\gamma_c=(1-U)/2$] such network states are always unstable.

The standard analysis also showed that for positive threshold $U>0$ the network state with no activity ($m^\nu=0\ \forall\nu,\ x=0$) is always an equilibrium state. This state emerges asymptotically if the initial state of the network does not exhibit significant overlap with any of the patterns stored. It plays the important role of an indicator for no recognition. This qualifies our model, in contrast, e.g., to the Hopfield model, as *not opinionated* since it does not force every initial state to be recognized as one of the stored patterns.

We have studied network dynamics in more detail for the situation that two patterns, say 1 and 2, do not overlap with any of the other patterns stored. When the network is in a state of overlap with solely these two patterns the dynamics depends only on m^1,m^2 and allows a characterization of the basins of attraction for pattern 1, pattern 2, and the state of no recognition. We found that the basins of attraction are well behaved, i.e., sufficiently large, of equal size,[5] and become small for large values of U and γ.[12] This behavior actually holds in the whole pattern space and not only in the subspace of patterns 1 and 2.

We now consider the storage capacity $\alpha_c=p_{\max}/N$ of the network and assume an infinite network with an infinite number of stored patterns, albeit finite α. A most straightforward consideration of the signal-to-noise ratio for a single spin can give insight into the properties of the network. Let the network be in a state $\mathbf{S}\approx\xi^1$ corresponding to recall of pattern ξ^1. This state should be stabilized by the local fields h_i defined above. h_i can be decomposed in a signal part h_i^s and a remaining (random) part h_i^r, the latter originating from perturbations due to other stored patterns, i.e., $h_i=h_i^s+h_i^r$. The signal part together with the threshold takes on the two values $h_i^s-U=1-a-\gamma-U$ for $\xi_i^1=1$ and $h_i^s-U=-a-\gamma-U$ for $\xi_i^1=0$. The overlap of pattern 1 with the infinitely many patterns $\nu>1$ results in Gaussian noise with zero mean and variance $\langle\langle(h_i^r)^2\rangle\rangle=\alpha a$. The resulting signal-to-noise ratios are then $\rho_1=(1-a-\gamma-U)/\sqrt{\alpha a}$ and $\rho_0=(a+\gamma+U)/\sqrt{\alpha a}$ for neurons i with $\xi_i^1=1$ and $\xi_i^1=0$, respectively. Recall would be optimal in the case $\rho_1=\rho_0$. This can be achieved through adaptation of the threshold to the optimal value $U_{\text{opt}}=\frac{1}{2}-a-\gamma$. The resulting optimal signal-to-noise ratio is $\rho_{\text{opt}}=1/\sqrt{4\alpha a}$ which shows that threshold adaptation or adaptive inhibition yield a recall behavior which improves with decreasing a. The latter behavior is opposite to that of the network investigated in Ref. 10. A rough estimate of the storage capacity can be based on the assumption that ρ_{opt} does not exceed a critical value. This yields the prediction $\alpha_c\sim a^{-1}$.

A more-detailed analysis bears logarithmic corrections to this prediction, i.e., $\alpha_c\sim a^{-1}f(1/\ln a)$. For this analysis which follows Amit *et al.*[2] we assume that the network state has macroscopic overlap with a finite number s of patterns. The microscopic overlap with the remaining $p-s$ patterns is then the origin of additional noise. The partition function of the network can be evaluated with the replica technique (see also Ref. 13). The network is characterized by order parameters m^ν and x defined above as well as by ($\beta=a\beta$)

$$q=\frac{1}{aN}\left\langle\!\left\langle \sum_i \langle S_i\rangle^2 \right\rangle\!\right\rangle , \tag{2}$$

$$r=\frac{1}{\alpha(1-a)}\sum_{\mu>s}\left[\frac{1}{aN}\sum_i(\xi_i^\nu-a)\langle S_i\rangle\right]^2 , \tag{3}$$

$$y=\frac{1}{\alpha(1-a)}\sum_{\mu>s}\left\langle\left[\frac{1}{aN}\sum_i(\xi_i^\nu-a)S_i\right]^2\right\rangle -\frac{2}{\alpha\beta}\left[\gamma x+U+\frac{\alpha}{2}\right] . \tag{4}$$

The parameter q corresponds to the Edwards-Anderson parameter in spin-glass theories. r and y characterize the mean and thermal fluctuations of the overlap between the thermodynamic state and patterns $\nu>s$ which are *not* condensed. The free-energy density f for our network model is

$$f=\frac{a}{2(1-a)}\sum_{\nu}(m^{\nu})^2+\frac{a\gamma}{2}x^2$$
$$+a\left[U+\frac{\alpha}{2}\right]x-\frac{a\alpha\bar{\beta}}{2}(rq-xy)$$
$$+\frac{\alpha}{2\beta}\left[\ln(1-C)-\frac{\bar{\beta}q}{1-C}\right]$$
$$-\frac{1}{\beta}\int\mathcal{D}z\langle\langle\ln[1+\exp(\beta\Phi)]\rangle\rangle$$

with

$$\mathcal{D}z=\frac{dz}{\sqrt{2\pi}}\exp\left[-\frac{z^2}{2}\right],$$
$$C=\bar{\beta}(x-q),$$

and

$$\Phi=\sum_{\nu}\frac{\xi^{\nu}-a}{1-a}m^{\nu}+\frac{\alpha\bar{\beta}}{2}(y-r)+\sqrt{a\alpha r}z\ .$$

From this expression one can derive the mean-field equations for the order parameters.

The network storage capacity can be characterized from a study of the simple case that the network in the zero-temperature limit has macroscopic overlap with only one pattern, i.e., $m^{\nu}=m\delta_{\nu,1}$. The essential order parameters (r and y yield algebraic expressions and are inserted) are solutions of the mean-field equations

$$m=\frac{1-a}{2}[\mathrm{erfc}(-\Phi_1)-\mathrm{erfc}(-\Phi_0)]\ , \tag{5a}$$

$$x=\tfrac{1}{2}\mathrm{erfc}(-\Phi_1)+\frac{1-a}{2a}\mathrm{erfc}(-\Phi_0)\ , \tag{5b}$$

$$C=\frac{1-C}{\sqrt{2\pi\alpha ax}}[a\ \exp(-\Phi_1^2)+(1-a)\exp(-\Phi_0^2)] \tag{5c}$$

with

$$\mathrm{erfc}(x)=\frac{2}{\sqrt{\pi}}\int_x^{\infty}\exp(-t^2)dt\ ,$$
$$\Phi_0=\frac{1-C}{\sqrt{2\alpha ax}}\left[-\frac{a}{1-a}m-U-\gamma x+\frac{\alpha}{2}\frac{C}{1-C}\right]$$

and

$$\Phi_1=\frac{1-C}{\sqrt{2\alpha ax}}\left[m-U-\gamma x+\frac{\alpha}{2}\frac{C}{1-C}\right].$$

The order parameters resulting from numerical solutions of Eqs. (5) for different storage density α are presented in Fig. 1. The results show that $m+ax$ and x at low storage density approach unity, i.e., recall is precise. Increase of α first reduces m and x slightly, i.e., the recall state adopts a small, but increasing fraction of false bits. Further increase of storage density results in a sudden destabilization of the solutions of Eqs. (5). The density for which the breakdown occurs yields the storage capacity α_c of the network. The value of α_c depends on a and U. The breakdown is preceded by an increase of

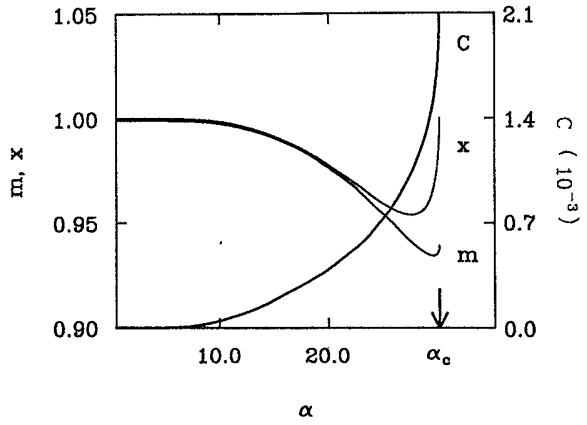

FIG. 1 Dependence of the mean-field parameters m, x, and C on storage density α ($a=10^{-3}$, $U=0.7$, $\alpha_c=30.16$).

m and x due to increasing overlaps with noncondensed patterns. For $\alpha>\alpha_c$ the network relaxes to a state with high activity and vanishing overlap with any stored pattern ν, i.e., $x\gg1$, $m^{\nu}=0\ \forall\nu$, or to the state of no activity ($m^{\nu}=0\ \forall\nu$, $x=0$).

The entropy at zero temperature $S(\alpha)=(\alpha/2)[\ln(1-C)+C/(1-C)]$ remains negative for all values of α. For $U=0.7$, $\gamma=0$, and $a=10^{-3}$ one determines $S(\alpha)\geq-3.31\times10^{-5}$. This small value implies that the effect of replica symmetry breaking should be small for our network model, as in the Hopfield model.[2]

The storage properties of a network with $U=0.7$ and $\gamma=0$ are presented in Fig. 2. To focus on the correction to $\alpha_c\sim a^{-1}$ the axes $a\alpha$ and $-1/\ln a$ are chosen. The area below the solid curve (labeled $a\alpha_c$) corresponds to networks with satisfactory recall behavior, the area above to networks incapable of recall. The phase boundary corresponds to the storage capacity multiplied by a. This line would be constant if $\alpha_c\sim a^{-1}$ would hold exactly. The phase boundary at small a values is linear. This most remarkable behavior coincides with the upper bound

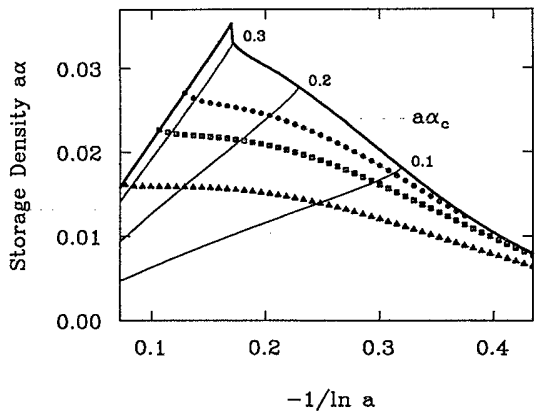

FIG. 2 Storage capacity α_c multiplied by a (bold line) as a function of $-(\ln a)^{-1}$ ($U=0.7$, $\gamma=0.0$). The thin lines denote (α,a) values with constant information (0.1,0.2,0.3 bits) stored per synapse. The symbols indicate contour lines of (α,a) values for which networks recall with 0.95 (○), 0.97 (□), and 0.99 (△) accuracy.

$\alpha_c \sim -(a \ln a)^{-1}$ derived by Gardner.[14] This suggests that storage in our network for small a is nearly optimal.

In order to test this supposition we consider the information content of the network which is obtained by calculating the entropy of the pattern states $\mathbf{S}=\xi^\nu$ diminished by the information deficit due to recall errors. The entropy per neuron of a network state with $m^\nu=m\delta_{\nu,1}$ is

$$S_N = -a\bar{m}\ln(a\bar{m}) - a(1-\bar{m})\ln(a-a\bar{m}) - a(x-\bar{m})\ln(ax-a\bar{m}) - [1-a(1+x-\bar{m})]\ln[1-a(1+x-\bar{m})]$$

with $\bar{m}=m^\nu+ax$. The entropy per neuron of a pure pattern ξ^ν is $S_P=-a\ln a-(1-a)\ln(1-a)$, the entropy loss due to recall errors is $\Delta S=S_N-S_P$, and the information content per synapse is $I(\alpha,a)=\alpha(S_P-\Delta S)/\ln 2$ (in bits). In Fig. 2 we have presented the contour lines for $I(\alpha,a)=0.1,0.2,0.3$. The results show that in the vicinity of the linear phase boundary assumed to present optimal storage more than 0.3 bits per synapse are stored. We have found that the information stored per synapse could be as large as 0.38 for $U=0.75$, $a=5.6\times10^{-7}$. This value is close to the bounds determined for asymmetric networks in the limit $a \ll 1$ by Gardner.[14,8]

The proposed network does not recall with complete precision. In order to judge its storage behavior the recall error measured by the deviation of $m^\nu+ax$ from the value one needs to be investigated. This information is also included in Fig. 2. The areas below the lines designated by circles, squares, and triangles correspond to a and α values for which networks recall with more than 95%, 97%, and 99% of correct bits, respectively. The results show that the price for better recall precision is a reduction in storage density.

To judge the efficiency of the associative memory one needs to solve the mean-field theory of networks with clipped synapses, i.e., synapses with strength $W_{ik}\in\{W_-,W_+\}$. In this case, not only the stored information, but also the information necessary to build the network can be evaluated. Our network with clipped synapses can be treated similar to the clipped Hopfield model[15] and yields a reduction of information content due to an additional noise source originating from the synaptic nonlinearity. We have found that after clipping, optimal storage at $U=0.75$, $a=10^{-6}$, is reduced from $I=0.38$ to $I=0.28$.

Finally we would like to emphasize again that the proposed network provides a nearly optimal associative memory for strongly biased patterns, i.e., for sparse coding. Since many interesting information processing tasks involve sparse coding the network appears to be a most promising candidate for practical applications.

This work has been supported by the German Ministry for Science and Technology (Grant No. ITR-8800-G9). J.B. is supported by a NATO grant (DAAD 300/402/513/9).

[1]J. J. Hopfield, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982); **81**, 3088 (1984).
[2]D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. A **32**, 1007 (1985); Phys. Rev. Lett. **55**, 1530 (1985).
[3]R. J. McEliece, E. C. Posner, E. R. Rodemich, and S. S. Venkatesh, IEEE Trans. Inf. Theory **IT-33**, 461 (1987).
[4]J. Buhmann and K. Schulten, Europhys. Lett. **4**, 1205 (1987); in *Neural Computers*, edited by R. Eckmiller and C. v.d. Malsburg (Springer, Berlin, 1987).
[5]J. Buhmann, thesis, Technische Universität München, 1988.
[6]W. Krauth and M. Mezard, J. Phys. A **20**, L745 (1987); G. Pöppel and U. Krey, Europhys. Lett. **4**, 979 (1987).
[7]E. J. Gardner, J. Phys. A **21**, 257 (1988).
[8]D. J. Willshaw, O. P. Buneman, and H. C. Longuet-Higgins, Nature **222**, 960 (1969); G. Palm, Biol. Cybern. **36**, 19 (1980).
[9]A. Moopenn, J. Lambe, and P. Thakoor, IEEE Trans. Syst. Man. Cybern. **SMC-17**, 325 (1987).
[10]D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. A **35**, 2293 (1987).
[11]M. Abeles, *Local Cortical Circuits* (Springer, Berlin, 1982).
[12]For large values of U ($U>0.7$) a noisy pattern representing the initial state with less than 30% wrong bits is associatively restored to the proper pattern state.
[13]A. D. Bruce, E. J. Gardner, and D. J. Wallace, J. Phys. A **20**, 2909 (1987); after completion of our calculations we took notice of this paper where the case $a=0.5$, $U=0$, $\gamma=0$, is solved.
[14]Gardner (Ref. 7) yields an information content per synapse $I=1/(2\ln 2)=0.72$ (0.36) for asymmetric (symmetric) networks.
[15]H. Sompolinsky, Phys. Rev. A **34**, 2571 (1986).