J. Phys. A: Math. Gen. 25 (1992) 5055–5069. Printed in the UK

Short- and long-range connections in autoassociative memory

Dominic O'Kane† and Alessandro Treves‡

† Theoretical Physics, University of Oxford, 1 Keble Road, Oxford OX1 3NP, UK
‡ Dept of Experimental Psychology, South Parks Road, Oxford OX1 3UD, UK

Received 16 December 1991, in final form 22 May 1992

Abstract. We consider memory retrieval in a network of M modules. A module consists of N neuronal units, each of which is connected to all $N-1$ other units within the same module, and to L units distributed randomly throughout all the other modules. Both short- and long-range connections are symmetric. The units are threshold-linear with a continuous positive output. Each module can retrieve one of D local activity patterns, or 'features', stored on the corresponding short-range connections. Furthermore, P global activity patterns, each consisting of combinations of M local features, are stored on the dilute long-range connections. When $M \gg 1$ the long-range connections endow the network with attractor states correlated with a single global pattern, and we study its storage capacity within a mean-field approach. If $P = D$, and each feature appears in only one pattern, our model reduces to an intermediate case between fully connected and highly dilute architectures, whose capacities we recover in the appropriate limits. As P/D takes larger (integer) values, the maximum P grows, but it remains asymptotically proportional to N rather than to $L + N - 1$ (the total number of connections per unit). The maximum amount of retrievable information per synapse, on the other hand, decreases. Moreover, as P/D grows, retrieval attractors have to compete with a 'memory glass' state, involving the retrieval of spurious combinations of features, whose existence and stability we describe analytically. We suggest implications for neocortical memory functions.

1. Introduction

Ever since Marr [7], local excitatory connections among pyramidal cells (the recurrent collaterals) have often been hypothesized to provide patches of neocortex with the properties of an autoassociative memory. The role of the collaterals, much as in many formal network models, would be to retrieve one of a set of several local activity patterns on the basis of a partial cue. However, some neurobiologists, e.g. Braitenberg [5], tend to regard the *whole* of neocortex, or at least its so-called association areas, as a kind of associative memory device, in which some of the long-range cortico–cortical connections, alongside short-range ones§, would also implement the storage and retrieval of *global* activity patterns. This view inevitably poses the problem of the storage capacity that such a device would have, and whether it would be consistent with broad constraints of plausibility and efficiency.

We address this question by considering a formal model which includes, as its crucial aspect, the operation of both short- and long-range connections in memory retrieval. This involves both a specific architecture, or connectivity, and a particular

§ Note that there tend to be about as many long-range connections as there are short-range ($\sim 10^4$) [1].

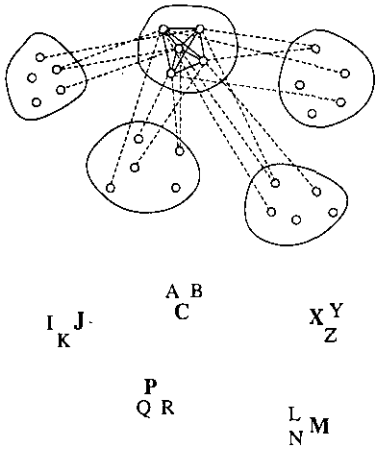
0305-4470/92/195055+15$04.50 © 1992 IOP Publishing Ltd

organization of the memory patterns (which can be conceived as resulting from a separate storage process). The remaining ingredients of the model, such as the type of processing units, the representation of inhibition and so on, are taken as in [12, 14], where an appropriate framework has been suggested for discussing issues of memory capacity in a cortically plausible context which includes both sparse coding and inhibitory control of excitatory activity.

The model is introduced in section 2, and the capacity analysis presented in section 3. The main results are briefly discussed in the last section, while some of the formal derivations are included in the appendices.

2. A modular model

Without necessarily subscribing to the view that cortices can be chopped up into discrete modular elements, we consider a network in which units are grouped into M modules, each containing N units. In each module, units receive both what we shall call short- and long-range connections. Short-range are those inputs which come from the other $N-1$ units within the same module, while long-range connections originate at L units distributed randomly, and differently for each receiving unit, throughout the remaining modules (see figure 1). All connections are taken to be symmetric.



p \ m	1	2	3	4	5
1	**J**	**P**	**C**	**M**	**X**
2	I	Q	C	M	Z
3	K	P	B	L	Y
4	J	R	A	L	Z
5	K	R	B	N	X
6	I	Q	A	N	Y

Figure 1. Schematic representation of the architecture of the network (top) (only the connections relative to one module are drawn for clarity) and of the corresponding memory organization (middle). Full lines represent short-range connections while the broken lines denote long-range connections. The table at the bottom gives an example of how features could combine into patterns when $\mu = 2$. Boldface letters denote one particular pattern being retrieved.

P patterns are stored on both the short- and long-range connections. However each of these (global, network) patterns is made up of M *features*, one per module, each drawn from a repertoire of D features stored in that module. We assume that the distribution of firing rates η representing each feature d within a module m (occurring during a learning phase not described here, and which would result in Hebbian-like synaptic modifications) is given independently for each unit i (and d, m) by some probability distribution $P_\eta(\eta_{i_m}^d)$. Note that $i = 1, \ldots, N$, $d = 1, \ldots, D$ and $m = 1, \ldots, M$. A global pattern, labelled p (with $p = 1, \ldots, P$), is then a random combination $\{d_1^p, \ldots, d_m^p, \ldots, d_M^p\}$. For simplicity, we shall assume that $P/D \equiv \mu$ is integral, and that features are assigned to patterns by randomly partitioning the P patterns in each module into D groups of μ elements (there are $P!/(\mu!)^D$ possible such assignments).

Note that for $\mu > 1$ different patterns will share some common features, but that any pair will on average be represented by the same feature in only a fraction $1/D$ of the modules. One need not worry about setting an upper limit on μ, as the calculation will show the model to be viable for only relatively low values of μ.

The total number of connections to a neurone is given by $C = L + N - 1$, and we define

$$\gamma = \frac{L}{C} \tag{1}$$

as the fraction of long-range connections.

As in [14] we impose that the probability distribution P_η satisfy $\langle\eta\rangle = \langle\eta^2\rangle = a$. The variable a thus defined is a measure of the *sparse coding*, while $T_0 \equiv (\langle\eta^2\rangle - \langle\eta\rangle^2)/a^2 = (1-a)/a$ sets the natural scale for the inverse of the gain, to be defined below. When necessary for explicit calculations, we shall take a simple binary form for P_η, since it has been shown [14] that more realistic distributions produce similar results.

The short- and long-range connection strengths are given by 'Hebbian' covariance rules [14] of the form

$$\begin{aligned} J_{i_m j_m}^{\text{short}} &= \frac{\mu}{C} \sum_{d=1}^{D} \left(\frac{\eta_{i_m}^d}{a} - 1\right)\left(\frac{\eta_{j_m}^d}{a} - 1\right) \\ J_{i_m j_n}^{\text{long}} &= \frac{c_{i_m j_n}}{C} \sum_{p=1}^{P} \left(\frac{\eta_{i_m}^p}{a} - 1\right)\left(\frac{\eta_{j_n}^p}{a} - 1\right) \end{aligned} \tag{2}$$

(each feature is repeated μ times while storing memories within a module). The variable $c_{i_m j_n}$ reflects the dilution and is 1 with probability q and 0 with probability $(1-q)$ where

$$q = \frac{L}{N(M-1)}\,. \tag{3}$$

Strong dilution is imposed on the long-range connections by letting $q \ll 1$. We take the equilibrium states of the net to be equivalent to those obtained through a dynamics with random asynchronous updating of the form

$$V_{i_m}(\tau + \delta\tau) = \begin{cases} 0 & \text{if } h_{i_m}(\tau) < T_{\text{thr}} \\ g(h_{i_m}(\tau) - T_{\text{thr}}) & \text{if } h_{i_m}(\tau) > T_{\text{thr}} \end{cases}$$

where g is the *gain* of each unit, $T_{\rm thr}$ its threshold, $h_{i_m}(\tau)$ is the local field at unit i_m at time τ and is given by

$$h_{i_m}(\tau) = \sum_{j_m(\neq i_m)} J^{\rm short}_{i_m j_m} V_{j_m} + \sum_{n(\neq m)} \sum_{j_n} J^{\rm long}_{i_m j_n} V_{j_n} + b(x) + s^\sigma \frac{\eta^\sigma_{i_m}}{a} \tag{4}$$

where $b(x)$ expresses the feedback control of the mean activity, $x = \langle V \rangle$, by inhibitory interneurones which are assumed to react very quickly to the average excitatory activity x. The form of $b(x)$ is irrelevant to a determination of the pattern storage capacity and its integral is denoted by $B(x)$ such that

$$B(x) = \int^x {\rm d}x' b(x') \,. \tag{5}$$

Note that the positive output V has a threshold linear dependence on the field h_{i_m}, which has been argued to approximate the current-to-frequency transduction of typical pyramidal cells. The last symmetry-breaking term represents a persistent external stimulus used as a cue for retrieving a particular pattern σ.

3. Attractor states and storage capacity

The attractor states of the system [2] are studied by looking for the minima of the free energy. The first two terms in the system Hamiltonian

$$H = -\sum_m \sum_{(i_m, j_m)} J^{\rm short}_{i_m j_m} V_{i_m} V_{j_m} - \sum_{(m,n)} \sum_{i_m, j_n} J^{\rm long}_{i_m j_n} V_{i_m} V_{j_n} - MNB(x)$$
$$- s^\sigma \sum_{m,i} \frac{\eta^\sigma_{i_m}}{a} V_{i_m} \tag{6}$$

represent the contributions of the short- and long-range connections. The significance of the other terms has been explained above. To average over the quenched random patterns in the free energy, one adopts [3] the replica trick based on the identity

$$\langle \ln Z \rangle_\eta = \lim_{n \to 0} \frac{\langle Z^n \rangle_\eta - 1}{n} \tag{7}$$

which results in the generation of an n-replica thermodynamic partition function

$$\langle Z^n \rangle_\eta = \left\langle {\rm Tr}_{\{V^\gamma_{i_m}\}} \exp\left(-\beta \sum_{\gamma=1}^{n} H^\gamma \right) \right\rangle_\eta \,. \tag{8}$$

The average over the dilution (appendix A) is straightforward (equivalent to the approach first used by Sompolinsky [11] in which the dilution is treated as a random synaptic noise subject to a Gaussian distribution) as long as $L \gg 1$ and $D \to \infty$. To allow for the existence of states which have a finite correlation with no more than one pattern, we also have to assume that $M \gg 1$.

The states of the system can be characterized by macroscopic order parameters. Labelling γ, δ individual replicas, the order parameters are the overlap of the state of the whole network with the one pattern being retrieved

$$\hat{x}^{\gamma\sigma} = \frac{1}{MN} \sum_{m,i_m} \left(\frac{\eta_{i_m}^{\sigma} - a}{a} \right) V_{i_m}^{\gamma} \tag{9}$$

the overlap with feature d within module m

$$\hat{x}_m^{\gamma d} = \frac{1}{N} \sum_{i_m} \left(\frac{\eta_{i_m}^{d} - a}{a} \right) V_{i_m}^{\gamma} \tag{10}$$

the mean network activity

$$x^{\gamma} = \frac{1}{MN} \sum_{m,i_m} V_{i_m}^{\gamma} \tag{11}$$

and the mean square activity in a module

$$y_m^{\gamma\delta} = \frac{1}{N} \sum_{i_m} V_{i_m}^{\gamma} V_{i_m}^{\delta} \,. \tag{12}$$

Since it was shown [13] that replica symmetry breaking is less significant for a system of continuous output units subject to a global activity constraint than for a system of binary, locally constrained units [10], we are justified in assuming a replica symmetry ansatz. With such an ansatz the final form of the free energy is (appendix B)

$$\begin{aligned} f = & -\frac{\mu(1-\gamma)}{2M} \sum_m (\hat{x}_m^1)^2 - \frac{1}{M} \sum_m \hat{t}_m^1 \hat{x}_m^1 - \hat{t}^{\sigma} \hat{x}^{\sigma} - \frac{\gamma}{2} (\hat{x}^{\sigma})^2 - tx \\ & + \frac{PT_o(1-\gamma)}{2NM} \sum_m y_{mo} + \frac{\beta\Delta^2\gamma^2}{4} (y_0^2 - y_1^2) - \frac{1}{M} \sum_m (r_{m0} y_{m0} - r_{m1} y_{m1}) \\ & - \frac{\beta\Delta^2\gamma^2}{2M} \left(y_0 \sum_m y_{m0} - y_1 \sum_m y_{m1} \right) - s^{\sigma}(x + \hat{x}^{\sigma}) - B(x) \\ & - \frac{1}{\beta M} \sum_m \left\langle\!\!\left\langle \int_{-\infty}^{\infty} \mathrm{D}z \ln \mathrm{Tr}(h, h_2) \right\rangle\!\!\right\rangle_{\eta} \\ & + \frac{D}{2\beta MN} \sum_m \left(\ln[1 - \beta T_o \mu(1-\gamma)(y_{m0} - y_{m1})] \right. \\ & \left. - \frac{\beta T_o \mu(1-\gamma) y_{m1}}{[1 - \beta T_o \mu(1-\gamma)(y_{m0} - y_{m1})]} \right) \end{aligned} \tag{13}$$

where

$$\begin{aligned} & h = -t - (\hat{t}^{\sigma} + \hat{t}_m^1) \left(\frac{\eta^{\sigma} - a}{a} \right) - z \sqrt{\frac{-2r_{m1}}{\beta}} \qquad h_2 = r_{m1} - r_{m0} \\ & \mathrm{Tr}(h, h_2) = \mathrm{Tr}_{\{V\}} \exp(\beta h V + \beta h_2 V^2) \qquad \mathrm{D}z = \frac{\mathrm{d}z}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2} \,. \end{aligned} \tag{14}$$

The order parameters characterizing the equilibrium states of the system at an inverse temperature β are arrived at by minimizing the free energy with respect to the variables of integration, yielding the saddle-point equations (appendix C). We then use these equations to derive the storage capacity of the network. We define

$$\psi = \beta T_o(y_0 - y_1) \qquad \rho = \frac{1}{T_o[\gamma + \mu(1-\gamma)]}\sqrt{\frac{-2r_{m1}}{\beta}} \tag{15}$$

and consider only the zero temperature limit in which

$$\lim_{\beta\to\infty}\frac{1}{\beta}\ln \mathrm{Tr}(h, h_2) = \begin{cases} 0 & \text{if } h < T_{\mathrm{thr}} \\ \dfrac{g'}{2}(h - T_{\mathrm{thr}})^2 & \text{if } h > T_{\mathrm{thr}} \end{cases}$$

where $1/g' = 1/g - 2h_2$. Note that $y_0 - y_1$ scales as $1/\beta$ signifying the fact that the system 'freezes' into a ground state in the low-temperature limit. Substituting the saddle point equations for $t, \hat{t}^\sigma$ and $\hat{t}^1_m$ into the equations for h and letting for simplicity $s^\sigma \to 0$ gives

$$h = [\gamma + \mu(1-\gamma)]\left[\tilde{b}(x) + \hat{x}^\sigma\left(\frac{\eta^\sigma}{a} - 1\right) - z\rho T_o\right] \tag{16}$$

where $\tilde{b}(x)$ is the rescaled function $b(x)$.

Parameter ρ measures the amplitude of the slow noise due to the multitude of uncondensed patterns. We define the following two signal to noise ratios:

$$w = [\tilde{b}(x) - \hat{x}^\sigma - \tilde{T}_{\mathrm{thr}}]/\rho T_o \quad \text{(uniform)} \qquad v = [\hat{x}^\sigma]/\rho T_o \quad \text{(pattern specific)}. \tag{17}$$

We write the final equations using the pattern averages defined in appendix D. Of the two equations which have to be satisfied on the v, w plane, one yields the capacity α_c as the maximum value of

$$\alpha = \frac{P}{C} \tag{18}$$

for which there are still solutions to

$$A_2^2 = \alpha A_3\left(\frac{\mu(1-\gamma)}{[\gamma + \mu(1-\gamma)\frac{A_1}{A_2}]^2} + \frac{\gamma}{[\gamma + (1-\gamma)\mu]^2}\right). \tag{19}$$

and, when solved together with the first, the other equation determines the range allowed for the gain, g:

$$A_2[\gamma + \mu(1-\gamma)] = \frac{1}{gT_o} - \alpha\left(\frac{\gamma(1-\frac{A_1}{A_2})}{[\gamma + \mu(1-\gamma)]} + \frac{[\gamma + \mu(1-\gamma)]}{[\gamma + \mu(1-\gamma)\frac{A_1}{A_2}]} - 1\right). \tag{20}$$

Once P_η has been chosen (19) can be solved numerically to give α_c. For simplicity we choose a binary P_η, such that $\eta = 1$ with probability a and $\eta = 0$ with probability $(1-a)$, and begin with the case $\mu = 1$, i.e. when there are as many patterns as there

are features in each module. In this case, as $\gamma \to 0$ (short-range connectivity only) the equations reduce to those for a fully connected network of N neurones storing D patterns (having M such networks—uncoupled—does not affect their capacity), while for $\gamma \to 1$ (long-range connectivity only) they reduce to those of a highly dilute [6] network storing P patterns [14]. For $0 < \gamma < 1$, α_c interpolates smoothly, at any given a, between the two limits. This is illustrated in figure 2, where α_c is plotted against a for $\gamma = 0, 0.5$ and 1. In the dependence on a (approx $\alpha_c \propto 1/[a \ln(1/a)]$) the figure reproduces the well known effects of sparse coding. As cortically realistic values of the sparse coding parameter are very difficult to determine, here and in some of the following figures we display the results for all values of a, so as to make the conclusions valid irrespective of whatever particular range might apply in a given system.

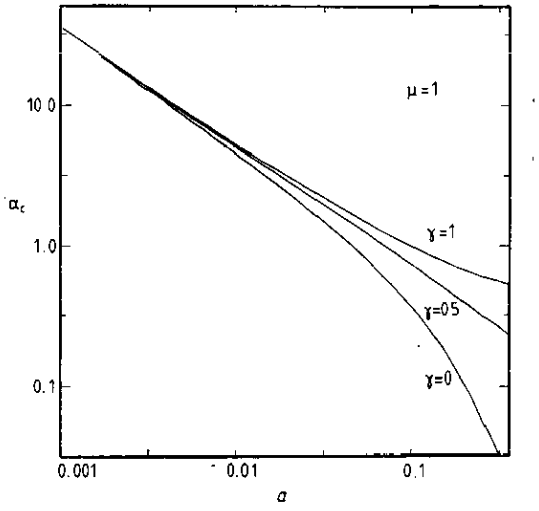


Figure 2. Storage capacity α_c as a function of sparseness of coding a for $\gamma = 0.0$, 0.5 and 1.0.

The decomposition of the patterns into features becomes relevant when $\mu > 1$, and it is again convenient to consider first the two connectivity limits. When $\gamma = 0$ one is left with the equation for a fully connected network, with α replaced by $\alpha/\mu \equiv D/C$. This simply reflects the fact that each of the modules is limited by its own capacity α_c but, without coupling between the modules, there is no limit on P as such. When $\gamma = 1$, μ drops out of the capacity equation. This signifies that in this dilute connectivity limit the fact that a feature may belong, locally, to several patterns, has no effect because of the connections being all long-range.

The interesting result is that for intermediate γ one again obtains intermediate capacities. In particular, as μ grows, α_c also grows—asymptotically as $\mu(1-\gamma)$. Once again it is the limit on D/N, the capacity of the modules, which constrains the capacity of the network.

This is misleading though, unless proper account is taken of another type of attractor state, which appears with the modular organization. It is the state in which the features that are retrieved in each module do not combine into any global pattern, so that none of the pattern overlaps is finite (for $M \to \infty$). This state could be termed a 'memory glass' as the network is, as it were, frozen into a disordered configuration, not of single unit states, but of memory fragments. It can be seen that, whereas there

are no thermodynamic solutions corresponding to k-mixtures of network patterns (with $k > 1$), there is a solution corresponding to one feature overlap being non-zero within each module, but all patterns overlaps vanishing (as $1/D$).

The 'storage capacity' equation for such a solution can be derived from the saddle point equations already found. It turns out to be

$$A_2^2 = \alpha A_3 \left(\frac{A_2^2}{\mu(1-\gamma)A_1^2} + \frac{\gamma}{\mu^2(1-\gamma)^2} \right) . \tag{21}$$

Analysing such a solution, one finds a stability line of the form

$$\gamma A_2 = A_1[\mu(1-\gamma) + \gamma] \tag{22}$$

beyond which the memory glass state is unstable to the emergence of a global pattern—it has become a saddle point of the free energy—and thus the network evolves into a retrieval state.

For very low values of μ, the value of α at which the memory glass solution becomes unstable, according to (22), falls well below the critical loading of the retrieval state. For slightly larger values of μ, the memory glass solution disappears, when α grows and crosses the limit expressed in (21), before ever reaching the instability line of (22). For larger μ, however, the difference between the α-value at which the memory glass solution disappears, and α_c, at which retrieval states disappear, becomes narrow, and with increasing μ the two limit have the same asymptotic behaviour. Hence at large μ most of the α range useful for retrieval, up to α_c, is 'infested' by the memory glass state. While the rate at which the two capacity limits approach each other depends on the fraction γ of long-range connections (figure 3), it is only for $\gamma \to 1$ that they are still substantially far apart for $\mu > 1$. As the P memory patterns represent but a negligible proportion of the D^M possible combinations of features, it is likely that in such a situation any kind of dynamics would result in the memory glass basin of attraction dominating over the tiny retrieval basins. Admittedly, the presence of a persistent external stimulus (which has been taken to vanish in the final steps of our calculation) would enhance the capacity of retrieval states [12], and thus provide some relief; but then it is the external stimulus, not the long-range connections, which succeeds in linking up the features retrieved locally into a meaningful combination, the global pattern.

The question of the storage capacity is better understood by also considering the *information* capacity [4] i.e. the total amount of information [9] that can be retrieved from the neurons per synapse. For our network of continuous response units this is defined as in [12], and depends on three factors: the information stored in each activity pattern as determined by P_η, the fraction of it that can be retrieved (in the presence of interference effects due to the extensive memory loading), and finally the number of independent activity distributions which can simultaneously be in storage. The only proviso is that in the present case the independent activity patterns are those characterizing the local features, rather than the global network patterns; hence the total information per synapse is obtained by multiplying the retrievable information per unit present in a feature by α/μ, rather than by α:

$$I = \frac{\alpha}{\mu \ln 2} \left\langle \ln \left(\frac{\text{Prob}(\eta, V)}{\text{Prob}(\eta)\text{Prob}(V)} \right) \right\rangle_\eta . \tag{23}$$

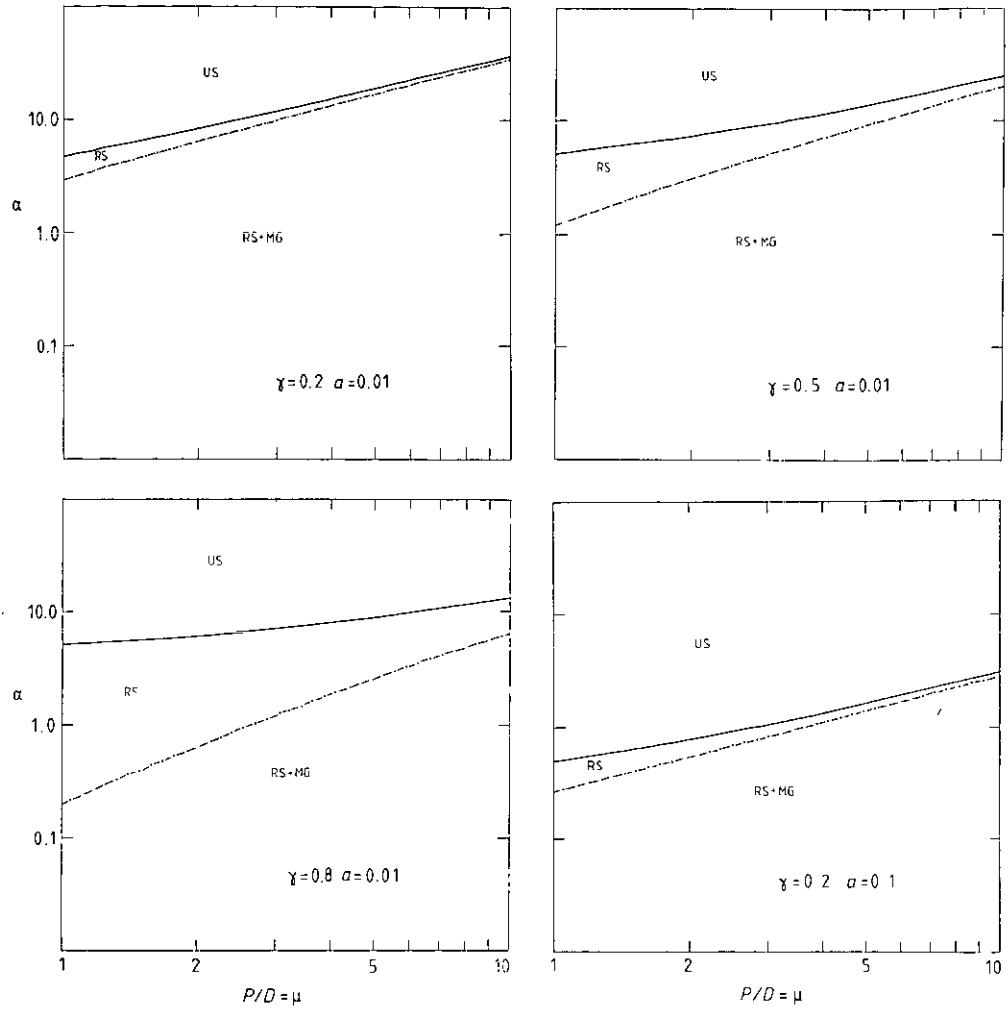


Figure 3. Comparison between the storage capacity α_c for the retrieval state (RS), and the upper limit on the existence of the memory glass state (MG) as a function of μ (for the sake of clarity the lines are drawn as if μ were a continuous variable)—in the right-hand diagram overleaf one can also distinguish the stability line, above which the memory glass becomes an unstable saddle point. (US) is the uniform state in which there is no retrieval, either locally or globally.

I_m, the retrievable information maximised over g and α, has been calculated to be around 0.1–0.3 bits per synapse for several different cases of autoassociative memory, with a mild dependence on the sparse coding parameter a. We calculate I_m for our case and the results are plotted in figure 4 for $\gamma = 0.5$ and $\mu = 1, 2, 3, 10, 100$. Although the maximum I_m occurs for $\alpha < \alpha_c$, and its dependence on α is more complex than simply multiplicative, the asymptotic behaviour with μ is determined by the fact that the $1/\mu$ factor cancels out the main μ dependence, in α_c, and $I_m(\mu)$ becomes essentially constant. This occurs after an initial decrease, for low values of μ, as shown in figure 4.

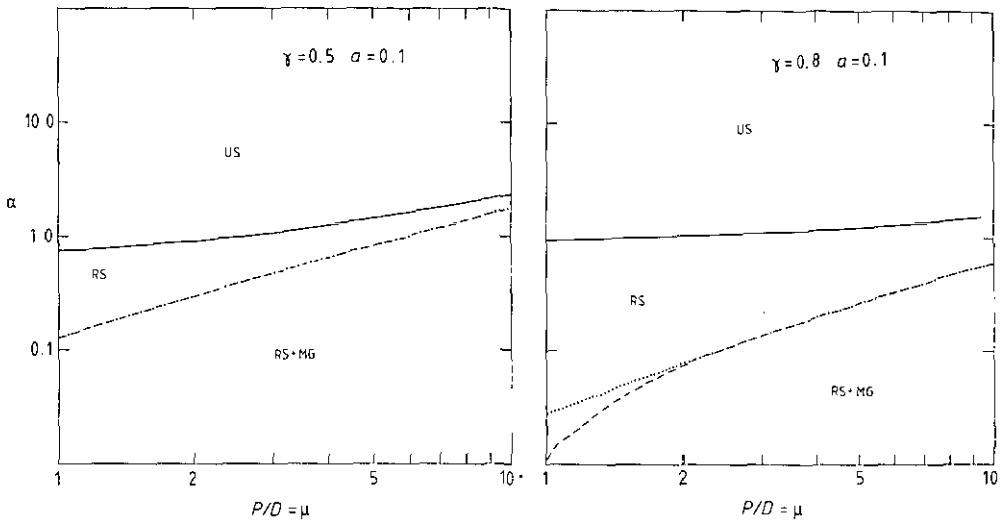


Figure 3. (Continued)

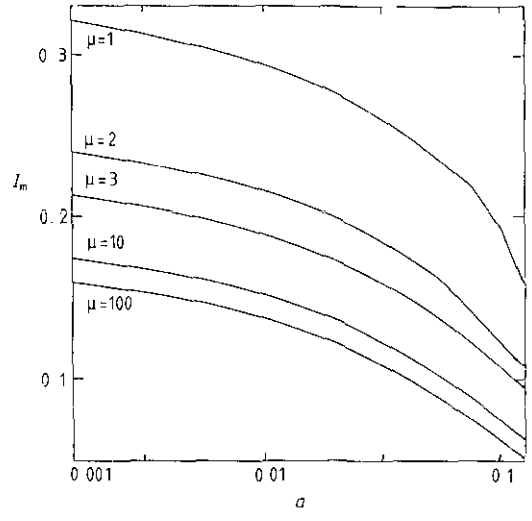


Figure 4. Maximum retrievable information I_m, in bits per synapse, versus a, for $\mu = 1$, 2, 3, 10 and 100 with $\gamma = 0.5$.

4. Discussion

The results we obtain are hardly surprising. Essentially, the number of patterns that can be stored depends on the number of connections per unit and on the sparseness of the coding as in simpler autoassociative memory models, while the total retrievable information is of the order of a fraction of a bit per synapse times the total number of synapses on which it is stored. Even going into quantitative details, no strange effect emerges. When $\mu = 1$, i.e. the global patterns are simply collections of non-repeating local features, one is left with an ordinary autoassociative memory whose connectivity is partly dense (locally) and partly dilute (globally). Its storage capacity interpolates, as the fraction γ of long-range connections increases, between the lower limit of a fully connected network and the higher limit of an extremely diluted one.

When the same local feature appears in more and more global patterns, i.e. for increasing μ, the local components of the signal and the noise gain importance, and eventually dominate, over the long-range components. As a result, what constrains the capacity is the (local) need to retrieve one of D features using only $N-1=(1-\gamma)C$ connections per unit. However, this does not allow a free ride on μ, i.e. to increase P at will, because soon the long-range connections become unable to link up the local features into previously stored global patterns. This effect is manifested in the emergence of a memory glass state corresponding to the network lapsing into a spurious combination of features.

The conclusion of the capacity analysis is that the present network is *not* a viable model, however simplistic, for the organization of memory in neocortex. Whatever type of memory ability one might want to consider, and whichever way one might want to estimate the storage capacity of a real brain, clearly having the number of memory items scaling with the number of connections per unit, as is the case here, rather than increasing with the size of the system, is wholly implausible from a biological point of view.

If the model adequately defines the quantitative constraints implicit in a certain theoretical view, then there may be several reasons why the theoretical view is itself inadequate. One reason could be that memories are not processed in neocortex in the form of activity patterns, coarsely spread on a temporal scale of tens of milliseconds, but of temporally finer elements, defined for example by the synchronicity of spike emission by selected groups of cells [1]. Another perhaps more economical reason could lie in the long-range projections of neocortical cells being very different from the sort of random, uniformly distributed connections considered here. A third could be associated with the process of memory storage being very different in neocortex from the autoassociative mechanism expressed by our 'Hebbian' learning of statistically independent patterns, which may be more immediately relevant, instead, to hippocampal processing [8]. However one regards these issues, the analysis of our rather crude model indicates that a simple-minded autoassociative memory approach is probably unable to offer clues useful to understanding memory processing on a cortical scale.

Acknowledgments

We thank D Sherrington, T Watkin and E T Rolls for useful discussions. DO'K is supported by a DENI Research Studentship and A T by an EEC BRAIN grant.

Appendix A. Dilution

We take the average as follows.

$$\prod_{(m,n)}\prod_{i_m,j_n}\left\langle\!\!\left\langle \exp\left[\beta\gamma\frac{c_{i_m j_n}}{L}\sum_{\gamma,p}\left(\frac{\eta_{i_m}^p-a}{a}\right)\left(\frac{\eta_{j_n}^p-a}{a}\right)V_{i_m}^\gamma V_{j_n}^\gamma\right]\right\rangle\!\!\right\rangle_c$$

$$=\prod_{(m,n)}\prod_{i_m,j_n}\exp\ln\left[\!\!\left[1+\frac{L}{MN}\right.\right.$$

$$\left.\left.\times\left\{\exp\left[\frac{\beta\gamma}{L}\sum_{\gamma,p}\left(\frac{\eta_{i_m}^p-a}{a}\right)\left(\frac{\eta_{j_n}^p-a}{a}\right)V_{i_m}^\gamma V_{j_n}^\gamma\right]-1\right\}\right]\!\!\right]. \tag{24}$$

An expansion may be performed with respect to the argument of the second exponent. This can then be expanded in the argument of the logarithm. Provided $L \gg 1$, only three terms scale extensively. The first of these is the fully connected term, then there are two corrections. The dilution term becomes

$$\exp\left(\frac{\beta^2\Delta^2\gamma^2}{2NM}\sum_{\gamma,\delta}\sum_{(m,n)}\sum_{i,j}V_{i_m}^{\gamma}V_{j_n}^{\gamma}V_{i_m}^{\delta}V_{j_n}^{\delta}\right) \tag{25}$$

where Δ comes from a self-average over network patterns. However, one must be careful since the averaging over the dilution has introduced a coupling between the activity of a given unit in two different patterns. These patterns may share, in the module of that unit, the same feature—a possibility which must be accounted for in the self-average. One finds that

$$\Delta^2 = \frac{P(1-\frac{L}{MN})T_o^2}{L} \to \frac{\alpha T_o^2}{\gamma}. \tag{26}$$

Appendix B. Free energy

Keeping only extensively scaling terms and having averaged over the dilution, the n-replica partition function becomes

$$\begin{aligned}\langle Z^n\rangle_c = \mathrm{Tr}_{\{V_{i_m}^{\gamma}\}}\exp\beta N\sum_{\gamma,d,m}\frac{\mu(1-\gamma)}{2}\left[\frac{1}{N}\sum_{i_m}\left(\frac{\eta_{i_m}^{d}-a}{a}\right)V_{i_m}^{\gamma}\right]^2\\ \times\exp\frac{\beta\gamma NM}{2}\sum_{\gamma}\left[\frac{1}{MN}\sum_{m,i_m}\left(\frac{\eta_{i_m}^{\sigma}-a}{a}\right)V_{i_m}^{\gamma}\right]^2\\ \times\exp\left[\frac{\beta^2\Delta^2\gamma^2NM}{4}\sum_{\gamma,\delta}\left(\frac{1}{MN}\sum_{m,i_m}V_{i_m}^{\gamma}V_{i_m}^{\delta}\right)^2\right.\\ \left.-\frac{\beta PT_o(1-\gamma)}{2N}\sum_{\gamma,m,i_m}(V_{i_m}^{\gamma})^2\right]\\ \times\exp\beta\sum_{\gamma}\left[s^{\sigma}\sum_{m,i_m}\frac{\eta_{i_m}^{\sigma}}{a}V_{i_m}^{\gamma}+MNB\left(\frac{1}{MN}\sum_{m,i_m}V_{i_m}^{\gamma}\right)\right].\end{aligned} \tag{27}$$

One can perform the average with respect to η^d for the uncondensed patterns i.e. $d > 1$ by expanding to second order—higher-order terms do not scale extensively—then averaging and re-exponentiating the result. It is then possible to integrate out the order parameters relating to the uncondensed patterns; $\hat{x}_m^{\gamma d}$ and $\hat{t}_m^{d\gamma}$ for $d > 1$.

$$\begin{aligned}\langle\langle Z^n\rangle_c\rangle_\eta = \left(\frac{\beta N}{2\pi}\right)^{n(\frac{n+1}{2}+M)}\left(\frac{\beta NM}{2\pi}\right)^{n(2+\frac{n}{2})}\\ \times\int_{-\infty}^{\infty}\mathrm{d}\hat{x}_m^{1\gamma}\mathrm{d}\hat{t}_m^{1\gamma}\mathrm{d}x^{\gamma}\mathrm{d}t^{\gamma}\mathrm{d}\hat{x}^{\sigma\gamma}\mathrm{d}\hat{t}^{\sigma\gamma}\mathrm{d}y_m^{\gamma\delta}\mathrm{d}r_m^{\gamma\delta}\mathrm{d}y^{\gamma\delta}\times\end{aligned}$$

$$\times \exp \beta N \sum_{\gamma} \left(\frac{\mu(1-\gamma)}{2} \sum_{m} (\hat{x}_m^{1\gamma})^2 + \frac{M}{2} (\hat{x}^{\sigma\gamma})^2 \right.$$

$$\left. + \mathrm{i} \sum_{m} \hat{t}_m^{1\gamma} \hat{x}_m^{1\gamma} + \mathrm{i} M \hat{t}^{\sigma\gamma} \hat{x}^{\sigma\gamma} + \mathrm{i} M t^{\gamma} x^{\gamma} \right)$$

$$\times \exp \beta N \left(- \frac{P T_o (1-\gamma)}{2N} \sum_{\gamma, m} y_m^{\gamma\gamma} - \frac{\beta \Delta^2 \gamma^2 M}{4} \sum_{\gamma, \delta} (y^{\gamma\delta})^2 \right.$$

$$\left. + \mathrm{i} \sum_{(\gamma,\delta) m} r_m^{\gamma\delta} y_m^{\gamma\delta} + \frac{\beta \Delta^2 \gamma^2}{2} \sum_{\delta, \gamma} y^{\gamma\delta} \sum_{m} y_m^{\gamma\delta} \right)$$

$$\times \exp \left(- \frac{1}{2} \sum_{m, d} \mathrm{Tr}_{\gamma} \ln(\mathbf{1} - \beta T_o \mu (1-\gamma) \mathbf{Y_m}) \right.$$

$$\left. - \mathrm{i} \beta M N \sum_{\gamma} \left[s^{\sigma} (x^{\gamma} + \hat{x}_m^{1\gamma}) + B(x^{\gamma}) \right] \right)$$

$$\times \mathrm{Tr}_{\{V_{i_m}^{\gamma}\}} \exp -\mathrm{i} \beta \sum_{m, i} \left\{ \sum_{(\gamma,\delta)} r_m^{\gamma\delta} V_{im}^{\gamma} V_{i_m}^{\delta} \right.$$

$$\left. + \sum_{\gamma} \left[t^{\gamma} + \left(\hat{t}_m^{1\gamma} + \hat{t}^{\sigma\gamma} \right) \left(\frac{\eta_{i_m}^{\sigma} - a}{a} \right) \right] V_{i_m}^{\gamma} \right\}. \tag{28}$$

Rotating variables $\mathrm{i}\hat{t}_m^{1\gamma} \to \hat{t}_m^{1\gamma}$, $\mathrm{i} r_m^{\gamma\delta} \to r_m^{\gamma\delta}$, $\mathrm{i} t^{\gamma} \to t^{\gamma}$ and $\mathrm{i}\hat{t}^{\sigma} \to \hat{t}^{\sigma}$ we can then equate this to $\exp -\beta N M n f$ at the saddle point as $N \to \infty$. We assume a replica symmetry ansatz. The trace over neurones can be simplified to a single site trace using the Hubbard–Stratonovitch identity and simplifies further in the $n \to 0$ limit.

Appendix C. Saddle-point equations

$$\hat{t}_m^1 = -\mu(1-\gamma)\hat{x}_m^1 \qquad \hat{x}_m^1 = \left\langle\!\!\left\langle \left(\frac{\eta^{\sigma} - a}{a} \right) \frac{\mathrm{d}}{\mathrm{d}h} \frac{1}{\beta} \ln \mathrm{Tr}(h, h_2) \right\rangle\!\!\right\rangle_{\eta}$$

$$\hat{x}^{\sigma} = \frac{1}{M} \sum_{m} \hat{x}_m^1 \qquad \hat{t}^{\sigma} = -\gamma \hat{x}^{\sigma} - s^{\sigma} \qquad t = -b(x) - s^{\sigma}$$

$$x = \frac{1}{M} \sum_{m} \left\langle\!\!\left\langle \int \mathrm{D}z \frac{\mathrm{d}}{\mathrm{d}h} \frac{1}{\beta} \ln \mathrm{Tr}(h, h_2) \right\rangle\!\!\right\rangle_{\eta}$$

$$y_{m0} = \left\langle\!\!\left\langle \int \mathrm{D}z \frac{\mathrm{d}}{\mathrm{d}h_2} \frac{1}{\beta} \ln \mathrm{Tr}(h, h_2) \right\rangle\!\!\right\rangle_{\eta}$$

$$y_{m1} = \left\langle\!\!\left\langle \int \mathrm{D}z \frac{\mathrm{d}}{\mathrm{d}h_2} \frac{1}{\beta} \ln \mathrm{Tr}(h, h_2) \right\rangle\!\!\right\rangle_{\eta} - \frac{1}{\beta} \left\langle\!\!\left\langle \int \mathrm{D}z \frac{\mathrm{d}^2}{\mathrm{d}h^2} \frac{1}{\beta} \ln \mathrm{Tr}(h, h_2) \right\rangle\!\!\right\rangle_{\eta} \tag{29}$$

$$y_0 = \frac{1}{M} \sum_{m} y_{m0} \qquad \text{and} \qquad y_1 = \frac{1}{M} \sum_{m} y_{m1}$$

$$r_{m0} = -\frac{DT_o}{2N}\frac{\mu(1-\gamma)[1-\mu(1-\gamma)\beta T_o(y_{m0}-2y_{m1})]}{[1-\mu(1-\gamma)\beta T_o(y_{m0}-y_{m1})]^2} + \frac{PT_o(1-\gamma)}{2N} - \frac{\beta\Delta^2\gamma^2 y_0}{2}$$

$$r_{m1} = -\frac{DT_o}{2N}\frac{[\mu(1-\gamma)]^2\beta T_o y_{m1}}{[1-\mu(1-\gamma)\beta T_o(y_{m0}-y_{m1})]^2} - \frac{\beta\Delta^2\gamma^2 y_1}{2}.$$

The variables defined in the text become

$$\psi = T_o g' \left\langle\!\!\left\langle \int_{h>thr} \mathrm{D}z \right\rangle\!\!\right\rangle_\eta$$

$$\rho^2 = \frac{D(\mu(1-\gamma))^2 y_1}{N[\gamma+\mu(1-\gamma)]^2[1-\mu(1-\gamma)\psi]^2} + \frac{\Delta^2\gamma^2 y_1}{T_o^2[\gamma+\mu(1-\gamma)]^2} \tag{30}$$

$$h_2 = \frac{DT_o\mu(1-\gamma)}{2N[1-\mu(1-\gamma)\psi]} + \frac{\Delta^2\gamma^2\psi}{2T_o} - \frac{PT_o(1-\gamma)}{2N}$$

Appendix D. Pattern averages

$$A_1(w,v) = \frac{1}{vT_o}\left\langle\!\!\left\langle \left(\frac{\eta^\sigma - a}{a}\right)\int^+ \mathrm{D}z\left(w + v\frac{\eta}{a} - z\right)\right\rangle\!\!\right\rangle_\eta - \left\langle\!\!\left\langle \int^+ \mathrm{D}z \right\rangle\!\!\right\rangle_\eta$$

$$A_2(w,v) = \frac{1}{vT_o}\left\langle\!\!\left\langle \left(\frac{\eta^\sigma - a}{a}\right)\int^+ \mathrm{D}z\left(w + v\frac{\eta}{a} - z\right)\right\rangle\!\!\right\rangle_\eta \tag{31}$$

$$A_3(w,v) = \left\langle\!\!\left\langle \int^+ \mathrm{D}z\left(w + v\frac{\eta}{a} - z\right)^2\right\rangle\!\!\right\rangle_\eta$$

where the range of integration in each is over $w + v(\eta/a) - z > 0$.

Appendix E. The memory glass solution

Beginning at the saddle-point equations we impose $\hat{x}^\sigma = 0$, and $\hat{x}^1_m$ non-zero for only a negligible fraction of modules to give

$$h = b(x) + s^\sigma + [\mu(1-\gamma)]\left(\frac{\eta - a}{a}\right)\hat{x}^1_m - zT_o\rho\mu(1-\gamma) \tag{32}$$

and, defining the pattern averages as before, we find that

$$A_2\mu(1-\gamma) = \frac{1}{g'T_o} \tag{33}$$

allowing us to write (21). To derive the stability we look for the zero-valued eigenvalues of the Hessian matrix. This is quite straightforward and results in (22).

References

[1] Abeles M 1991 *Corticonics* (Cambridge: Cambridge University Press)
[2] Amit D J 1989 *Modelling Brain Function* (Cambridge: Cambridge University Press)
[3] Amit D J, Gutfreund H and Sompolinsky H 1987 *Ann. Phys., NY* **173** 30–67
[4] Amit D J, Gutfreund H and Sompolinsky H 1987 *Phys. Rev.* A **35** 2293–303
[5] Braitenberg V and Schütz A 1991 *Anatomy of the Cortex: Statistics and Geometry* (Berlin: Springer)
[6] Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167–73
[7] Marr D 1970 *Proc. R. Soc.* B **176** 161–234; 1971 *Phil. Trans. R. Soc.* B **262** 23–81
[8] Rolls E T 1989 *Neural Models of Plasticity* ed J H Byrne and W O Berry 240–65 (New York: Academic)
[9] Shannon C E and Weaver W 1949 *The Mathematical Theory of Communication* (Urbana, IL: University of Illinois Press)
[10] Sherrington D and Kirkpatrick S 1975 *Phys. Rev. Lett.* **35** 1792
[11] Sompolinsky H 1987 *Phys. Rev.* A **34** 2571
[12] Treves A 1990 *Phys. Rev.* A **42** 2418–30
[13] Treves A 1991 *J. Phys. A: Math. Gen.* **24** 2645–54
[14] Treves A and Rolls E T 1991 *Network* **2** 371–97