Network 2 (1991) 371–397. Printed in the UK

What determines the capacity of autoassociative memories in the brain?

Alessandro Treves and Edmund T Rolls

Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford OX1 3UD, UK

Received 13 May 1991

Abstract. Threshold-linear (graded response) units approximate the real firing behaviour of pyramidal neurons in a simplified form, suited to the analytical study of large autoassociative networks. Here we extend previous results on threshold-linear networks to a much larger class of models, by considering different connectivities (including full feedback, highly diluted and multilayer feedforward architectures), different forms of Hebbian learning rules, and different distributions of firing rates (including realistic, continuous distributions of rates). This allows an evaluation of the main factors which may affect, in real cortical networks, the capacity for storage and retrieval of discrete firing patterns.

In each case a single equation is derived, which determines both α_c, the maximum number of retrievable patterns per synapse, and I_m, the maximum amount of retrievable information per synapse. It is shown that:

1. Non-specific effects, such as those usually ascribed to inhibition, or to neuromodulatory afferents, alter the overall response, but do not affect the capacity for retrieval.

2. The crucial parameter which affects α_c is a, the sparseness of the neural code. Results previously obtained with binary distributions of rates are shown to hold in general, namely that as $a \to 0$ (sparse coding) α_c grows proportionally to $(a \ln(1/a))^{-1}$, while I_m depends only weakly on a.

3. When the coding is sparse, α_c and I_m become independent of the connectivity, and in particular the relative disadvantages of fully connected feedback networks, which are prominent with non-sparse codes, disappear.

4. The precise form of the distribution of rates, and that of the learning rule used, turn out to have rather limited effects on α_c and I_m. It is to be noted, however, that when non-binary patterns are stored using nonlinear learning rules, such as those possibly modelling the action of NMDA receptors, information distortion occurs in retrieval, and results in a marked decrease of I_m.

These results may be applied to help understand the organization of a specific network in the hippocampus thought to operate as an autoassociative memory.

1. Introduction

Simplified formal models of associative memory neural networks [51,31,32,27,30,25] produce, when analysed with statistical techniques [6,4], potentially useful quantitative results. One needs, however, to know whether the results carry over to more realistic situations, before applying them profitably to study those networks in the brain which

0954-898X/91/020371+27$03.50

are presumed to be involved in memory. We attempt a step in this direction, focusing on *auto*associative memory, and on quantities measuring the capacity for retrieval. These quantities are less likely than others to depend drastically on the details of neuronal dynamics, but it has not been clear, so far, to what extent they depend on other features of real networks, and on the way those features are represented, simplified, or neglected altogether in formal network models. In particular:

- Experimentally observed distributions of firing rates are *continuously graded*, in contrast with the (essentially) *binary* distributions emerging in those models in which neurons are represented as binary units [34]. Does this discrepancy make the binary models irrelevant? What are the information capacities of networks with continuous distributions of rates?
- Which characteristics of the biophysical mechanism of *synaptic plasticity* are (computationally) crucial?
- What is the trade-off between *fully distributed* and *sparse* coding of memories in autoassociative nets? Is it of the type indicated by the simplest binary models?
- The same autoassociative memory functions can be implemented by very different patterns of connectivity. How do the various 'architectures' compare quantitatively? How important is the role of excitatory *feedback loops*? And in general, is it essential to model the detailed connectivity in order to estimate the memory capacity?

The approach we take is to study how the capacity varies within a class of models, in which we have introduced those elements of biological realism that are necessary in order to address the above questions; while retaining the simplicity which allows full analytical control. The way the compromise between biological fidelity and analytical tractability is worked out, especially at the level of modelling single neurons, is discussed in section 2. Section 3 proposes formal representations suited to evaluate the effects of different connectivities, of different memory codes, and of different learning mechanisms. The analytical results for the storage capacity of the network are presented in section 4, and their implications are discussed in section 5. The conclusions we arrive at will be put to actual use in a companion paper [49], in which we argue that they are instrumental in relating a neurobiological hypothesis (that a network in the hippocampus functions as an autoassociative memory [39]) to a series of experimentally testable predictions.

2. Real systems and models

2.1. Autoassociative systems in the brain

The models we consider are motivated by the attempt to better understand the organization of those systems in the brain, which have been hypothesized to function as autoassociative memories. These include the recurrent collaterals of the CA3 network in the hippocampus and the system of collateral connections between nearby pyramidal cells of neocortical association areas [31,32,38]. The hypothesis that the hippocampus includes, in its CA3 system, an autoassociative network important in episodic memory has been described elsewhere [38,40]. In the case of neocortical

collaterals, on the other hand, a possible autoassociative memory role should probably be set in the wider context of the much more complex, and yet largely to be understood, types of processing occurring in neocortex.

Both in CA3 and neocortex, the great majority of neurons are pyramidal cells [37], and the memory traces are considered to reside in the modification of the efficacies of the excitatory synapses between pyramidal cells. In fact, in formal studies of associative memory [51,22,30,27,25], it is pyramidal cells only which are included in the model networks. One element of realism we introduce, here, is an effective representation of a presumed regulatory role of inhibitory neurons, expressed as a non-specific modulation of the activity of pyramidal cells.

The information to be stored in the memory is considered to be coded as patterns of firing activities of pyramidal cells. The spatial resolution of the code is that of a single cell, and its temporal resolution, in the order of tens of ms, is taken to convey the firing rate, rather than the time of emission of individual spikes. This assumption does not exclude, of course, the possibility that much finer temporal encoding may be used as well in the brain, for example, earlier on in the sensory pathways [23]. In parallel, retrieving the stored information is taken to involve the evolution of each cell's rate in response to an ongoing, temporally rather coarse integration of many synaptic inputs, a process lasting in the order of a hundred ms or so. Again, this does not exclude that much faster processing may occur concurrently, or even that a substantial fraction of the same stored information be retrieved by the first few spikes [50].

In the case of CA3, it has been suggested that each distinct firing pattern codes an episodic memory [38], or the association of contextual information, which has already been processed extensively along various cortical pathways. Such information is given a unitary representation in terms of the firing pattern of a simple interconnected population, the CA3 pyramidal cells. In the models, a given set of units is taken to comprise a network, and its intrinsic connections store a discrete set of memory items. A crucial assumption is that the system is supposed to be able to retrieve each individual item, or firing pattern, when stimulated with an adequate partial cue. This occurs as the state of the network, defined as the list of instantaneous activity rates, dynamically evolves towards an attractor state [4]. The retrieval of each memory is thus implemented by a dynamical attractor. The present analysis is only directly concerned with this retrieval operation (whereas the storage of the information will be considered elsewhere [49]) and with the conditions that enable it, independently of the systems' level function subserved by such cued memory retrieval (a function which may be different between hippocampus and neocortex).

2.2. Neuronal current-to-frequency transduction

A crucial element involved in specifying the model is the (activation) function describing the input–output relation at the single neuron level. The input, here, is a real variable h representing the total synaptic current entering the cell, when it is subject to an asynchronous barrage of individual EPSPs and IPSPs. The output is a positive variable V, representing the instantaneous, or short-time averaged, firing rate of the cell. The relation of interest between the two is the one occurring in the long-time limit, when the cell has had time to adjust its output to a quasistationary average input. It is in this limiting 'static' condition, after transient effects have died out, that attractor states are characterized in the present analysis. For the attractor dynamics to sensibly represent memory retrieval, one has to require *a posteriori* that

cells approach their long-time limit behaviour within a psychological time scale of a few hundred ms.

The most direct experimental model of a macroscopically quasistationary asynchronous synaptic barrage is a long, constant current pulse injected into the cell body†. The resulting pattern of axon potentials has been studied *in vitro* for pyramidal cells from a variety of cortical regions. Data are usually presented as a series of current-to-frequency curves, each giving the instantaneous frequency, defined as the inverse nth inter-spike interval, as a function of injected current. For regular spiking pyramidal cells, in the long-time (large-n) limit one reaches an adapted firing state, given by a curve which rises from zero at a given threshold, and then is often approximately linear [24]. Sample values for the slope are 34 ± 18Hz nA^{-1} for CA1 pyramids in guinea pigs [29], 56 ± 14Hz nA^{-1} for layer 2/3 pyramids in cat visual cortex [33], and 56 ± 27Hz nA^{-1} for human neocortical pyramids [9]. The approximately linear range typically extends with no marked saturation effects throughout the region in which, as it appears from studies *in vivo*, cells normally operate (up to 10 Hz in CA3, 100 Hz in neocortex).

It seems, therefore, that a simple formal representation of regular spiking behaviour in the long-time limit could be effected by a threshold-linear transfer function,

$$V = \begin{cases} g(h - T_{hr}) & h > T_{hr} \\ 0 & h < T_{hr} \end{cases} \tag{1}$$

as shown in figure 1.

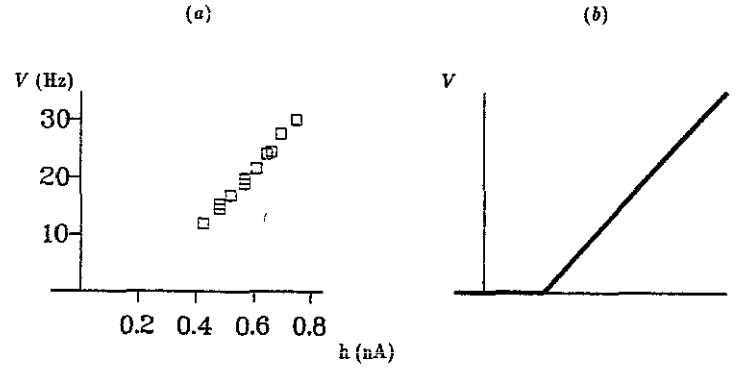


Figure 1. Threshold-linear dependence of the adapted firing rate on the integrated input current: (*a*) sample experimental data (from a layer 2/3 pyramidal cell in rat visual cortex [33]; (*b*) model.

Clearly, more sophistication would be required in order to model bursting behaviour or intrinsic oscillatory properties. While these are common phenomena, particularly in the hippocampus, it is possibly more important to take them properly into account while considering information storage rather than retrieval. On the other hand, the threshold-linear representation has the merit of modelling the two most

† Although it does not reproduce conductance changes affecting the cell *in vivo*, particularly those associated with the time course of inhibitory processes.

clearly present features of long-time adapted firing, namely the threshold effect and the graded response above threshold. It does not model saturation, which appears to occur at rates above those typical, in the adapted state, of 'normal' physiological regimes. It has the theoretical merit of its sheer simplicity, which allows extensive analytical understanding at the network level. Finally, it involves just two cell-dependent parameters, the gain g and threshold T_{hr}, corresponding to readily measurable characteristics of real cells. In the following network models, for simplicity g and T_{hr} will not be made to model cell-to-cell variability, but considered uniform throughout the population.

2.3. Integration of synaptic inputs

The total input to a given cell has been expressed as a current, yet the fundamental variables involved in synaptic action are ionic conductances. In relating the two, it is important to distinguish between different types of input. The type which, in these particular systems, is directly associated with memory retrieval is the one mediated by the modifiable synapses between pyramidal cells, where memory is thought to be stored. Neglecting active dendritic processes and considering each cell as electrotonically compact, this input can be approximately described by a linear summation of individual synaptic terms. The contribution of each term, when averaging over a short time rather than considering the effect of single spikes, is proportional to the firing rate V of the presynaptic cell, and the proportionality factor, or synaptic efficacy, is here denoted as J. Assuming that this type of input comes to a particular cell from a set of C other pyramidal cells, and that each makes a single synapse, a component of the integrated current will therefore be $\sum_{j=1}^{C} J_{ij} V_j$. It is useful, however, to separate out in each efficacy J the part which specifically encodes stored memory, that is the *modification* in the synaptic efficacy, J^c, due to associative learning. This leads to the expression

$$h_i = \sum_{j=1}^{C} J_{ij}^{c} V_j + (\text{non-specific terms}). \tag{2}$$

Note that while J is a positive number, the change J^c, and hence the relative contributions to h_i, can be positive or negative.

Additional specific inputs may come from outside, i.e. from cells which are not part of the autoassociative network under scrutiny. To leave open the possibility of modelling, for example, afferent stimulation which contributes in directing memory retrieval, it is useful to consider a further component Δh_i to the input current, which is assumed as externally assigned for each cell.

Non-specific inputs may come from a variety of sources, and in the model will all be lumped together into a third component, b_i, of h_i. The contributions subtracted above, from the non-learned baseline component of modifiable synaptic efficacies, will be pooled in the b term together with excitatory inputs mediated by non-modifiable synapses. In addition there will be inputs from a variety of inhibitory interneurons, non-specific neuromodulatory afferents, and so on. The effect of most of these inputs will be more complex, in real life, than just adding an additional term to the synaptic current; for example, they will affect the gain g of the adaptive firing curve. This situation is sometimes summarized by referring to subtractive and divisive inhibition.

Now, the individual firing of inhibitory neurons and the strengths of their synapses to pyramidal cells are of no interest in the model, inasmuch as they are assumed to

carry no specific information, and in fact are not explicitly represented. Their average firing level, and therefore their effect on pyramidal cells, will depend nevertheless on the firing activity of pyramidal cells. The approximation will be made to consider this effect, and that of the baseline excitatory efficacies as well, as dependent only on the average firing level of the same set of C cells that give inputs to a given cell, i.e. on the sum

$$X_i = \frac{1}{C}\sum_{j=1}^{C} V_j. \tag{3}$$

As a further approximation, for all cells the dependence on X will be given the same form, specified by a certain function $b(X)$ added to the input current.

The final form for the input current will therefore be

$$h_i = \sum_{j=1}^{C} J_{ij}^{c} V_j + \Delta h_i + b(X_i). \tag{4}$$

It should be noted that $b(X)$ may have a highly nonlinear form, modelling features of inhibitory action†; however, it is a special virtue of models using threshold-linear units [47] that the specific form is irrelevant to the performance of the associative memory, as analysed in the rest of this paper. As it turns out, it only contributes to determine the overall response level of pyramidal cells, but not the way the response is distributed in relation to the information stored in memory. It should finally be noted that, although different assumptions could be used, assuming uniformity in the form of these non-specific effects minimises noisy cell-to-cell fluctuations, and thus allows exploration of the conditions of optimal information retrieval.

3. Exploring different models

3.1. Architectures

In modelling the pattern of connectivity among principal cells, we study a number of limiting schemes, with the idea that from the comparison among the extremes it is possible to obtain information on general, less clearcut, cases. The first limit considered, model $\mathcal{M}1$, is a network of N units, in which each unit receives inputs, through modifiable synaptic links, from every other unit. The number of modifiable synapses onto each unit is thus $C = N - 1$. In addition, all units can receive inputs and transmit their output to and from both the surrounding environment and, through non-modifiable links, a population of interneurons. This *feedback* architecture is represented in the top left corner of figure 2. It is the one considered in conjuction with linear units by Kohonen [27], with binary-threshold units by Little [30] and Hopfield [25], and with binary synapses and diluted connectivity by Gardner-Medwin [22]. A common theme of these studies has been to show that memory traces can be stored in the efficacies of the synapses on the recurrent collaterals in such a way that the network is able to retrieve an activity pattern when presented with a partial cue

† Also the gain g might be modelled to be a complicated function of X.

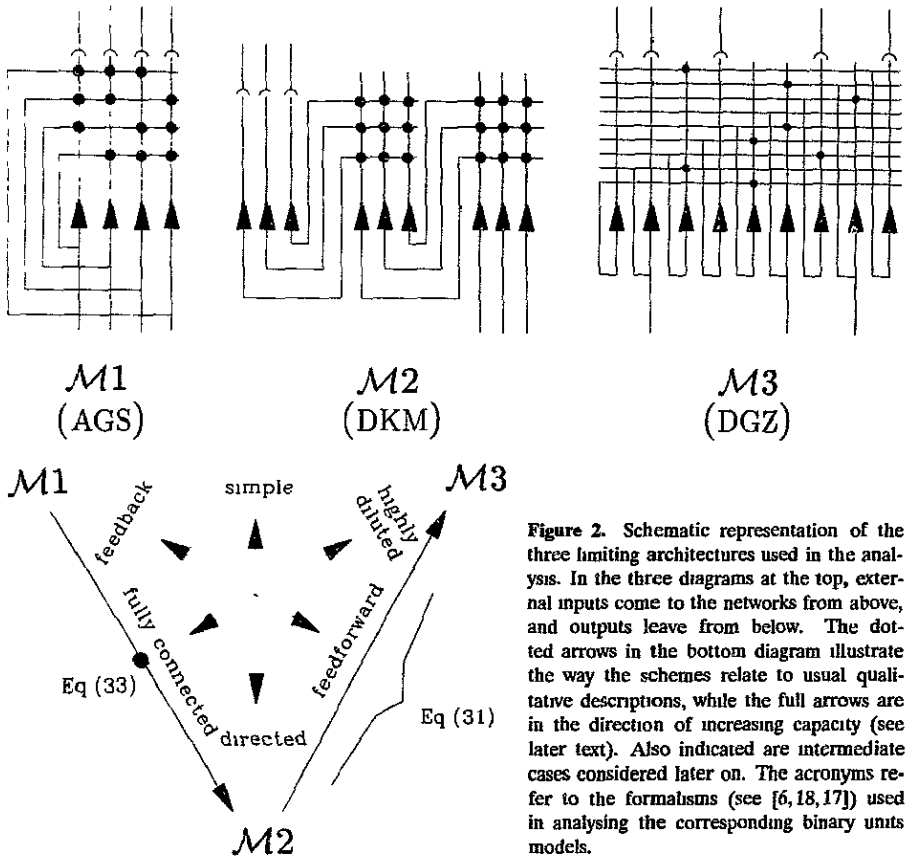


Figure 2. Schematic representation of the three limiting architectures used in the analysis. In the three diagrams at the top, external inputs come to the networks from above, and outputs leave from below. The dotted arrows in the bottom diagram illustrate the way the schemes relate to usual qualitative descriptions, while the full arrows are in the direction of increasing capacity (see later text). Also indicated are intermediate cases considered later on. The acronyms refer to the formalisms (see [6,18,17]) used in analysing the corresponding binary units models.

(pattern completion), and also in the presence of noise (noise tolerance) and when a substantial fraction of the network has been destroyed (robustness).

In another scheme, $\mathcal{M}2$, a series of layers each containing N units is considered. The inputs to the cells of one layer are assumed to be from the N cells of the preceding layers (see figure 2), hence $C = N$. External inputs affect the first layers (although they might be applied to subsequent layers as well) and then information is processed sequentially as in a multilayer perceptron until a final output is read off the last layer. This architecture has been adopted more often in artificial networks performing pattern recognition or classification, but it can be considered in the context of autoassociative memory as well [18]. In retrieving patterns from memory, a network of this type displays all the properties noted above, even if, if it were to be taken as a serious candidate to model biological circuitry, it would probably require additional elements in order to explain learning, i.e. to enable patterns to be stored in memory in the first place. The aim here is not to suggest this scheme as a realistic model, however, but only to use it as a useful benchmark in a comparison with $\mathcal{M}1$. The pattern of neuronal activities, which in $\mathcal{M}1$ is fed back along the recurrent col-

laterals onto itself, in $\mathcal{M}2$ is propagated forward, allowing in both cases an iterative reconstruction of the full pattern from a small cue.

In both $\mathcal{M}1$ and $\mathcal{M}2$ the connectivity is, in the sense of each scheme, complete. A third limiting case, $\mathcal{M}3$, is when the connectivity is very highly diluted, i.e. each unit receives inputs from a set of units which, although large, is actually a very small proportion of the total population of the system. Moreover, this set is chosen at random, independently of which sets happen to affect other units. In this situation it does not really matter whether the network is conceived of as a single layer of units interacting among themselves, or as a series of layers which supports a directional flow of information. Because of the sparse connectivity the system is feedforward, as the probability that the output of a cell will be propagated in a loop and eventually fed back into the cell itself is made negligible (mathematically, this can be imposed as an appropriate limiting relation between C and N†). The fact that this type of network, by allowing a simple statistical analysis, can be extremely useful in understanding the properties of autoassociative memory, which it retains, has been pointed out by Derrida *et al* [17].

These three architectures are considered here in order to provide some sort of boundary conditions, within which certain real systems may be enclosed. One may think, for example, of the hippocampal CA3 network in the rat, for which extensive anatomical data which bear on its computational function have been published [38,45,3]. There, if one considers the CA3 regions of both hippocampi in the two sides of the brain as forming a single network‡, $C \approx 1.2 \times 10^4$, $N \approx 6 \times 10^5$ (and as each CA3 cell projects out to subsequent stages this is also the size of the output representation), and the internal connectivity does not appear to be markedly directional, thereby not pointing to an underlying feedforward structure (within CA3 itself [26]). Which of these structural details affects function? We shall see that results like the inequalities (32)–(36) strongly suggest that C is the the only parameter that really matters, insofar as memory capacity is concerned.

$\mathcal{M}1$, $\mathcal{M}2$ and $\mathcal{M}3$ differ most substantially in the amount of hardware (neurons and synapses) they require to produce a given output representation, and also in the quality of information retrieval they allow. Thus, assuming both C and the size of the output representation to be fixed by external constraints, $\mathcal{M}2$ requires more neurons than $\mathcal{M}1$, by a factor essentially given by the number of iterations needed by the neural activities to converge to a pattern, and $\mathcal{M}3$ needs an unreasonably large amount of neurons, due to the low contact probability. Whether this can be counterbalanced by advantages may be clarified by the analysis that follows.

In all cases the analysis is simpler in the limit $C \to \infty$, which is relevant in a cortical context, as the synaptic inputs to principal cells are very many, and the collective action of many of them (maybe of the order of 10^2 [36]) is required to have a cell reach threshold. The formal treatment is easiest in the case $\mathcal{M}3$, of a very dilute connectivity, as the inputs to any given cell can be considered as uncorrelated [17]. In the multilayered case $\mathcal{M}2$, instead, one has to take into account the correlations stemming from the fact that any two cells in a given layer receive inputs from the same sets of cells, i.e. those in the previous layer. This can be accomodated in a

† Specifically, that as $C \to \infty$, as assumed later on, $\ln C / \ln N \to 0$ [28]

‡ In the rat, in fact, about half of the collaterals to a given CA3 cell appear to come from the contralateral side [3] In primates, by contrast the large majority of the connections appear to be ipsilateral [2], suggesting the need to regard the two CA3 systems in primates as distinct, if still interconnected, networks.

slightly more detailed statistical analysis ([18], and see the appendix). Finally, the fully connected feedback network $\mathcal{M}1$ [25] presents the additional complication that the correlations in the outputs of the N units are fed back in the inputs to the units themselves, leading to the need for a self-consistent treatment of the correlations. The resulting steady-state behaviour can be studied by adapting the type of free-energy analysis developed for spinglasses [6]. Whatever formalism is needed, it can be applied to the case of threshold-linear units [47,48].

3.2. Pattern statistics

Memory is retrieved from the network when neural activities, stimulated with a partial cue, evolve into a pattern strongly correlated with one of those which have been stored. What determined each stored distribution of activities in the first place, in the learning phase, is discussed, for a particular case, elsewhere [49]. To obtain a measure of the ability of the network to perform retrieval, a measure valid over most actual realizations of the stored patterns, one uses as a probe an *artificial* assignment of patterns of activity. This is done by assuming that p patterns (labelled $\mu = 1, \ldots, p$) are embedded with equal (RMS) strength in the synaptic efficacies. In each pattern each cell i is taken to code for independent information, i.e. its firing rate during learning η_i^μ is assigned as a random number, drawn independently, for all i and μ, from a probability distribution $P_\eta(\eta)$.

To study how the capacity of the network depends on the macroscopic features of the statistical distribution of patterns, one has to vary the parameters characterizing P_η. As η is a firing rate, $P_\eta(\eta) \geqslant 0$ only for $\eta \geqslant 0$, and $P_\eta(\eta) = 0$ otherwise. Moreover, $\int P_\eta \mathrm{d}\eta = 1$. Within these contraints, the first free parameter is the average firing rate,

$$a = \int P_\eta(\eta)\eta \mathrm{d}\eta \equiv \langle \eta \rangle_\eta \tag{5}$$

where $\langle \cdot \rangle_\eta$ denotes averages over P_η. This average pattern activity does not, however, affect memory encoding in the model (as the contribution of each pattern to the efficacies J^c is normalized, in the following, in units of a itself), nor it does affect retrieval (because, with a threshold-linear transfer function, the information retrieved does not depend on the absolute scale of the neuronal outputs). Therefore the average firing activity of the encoded patterns is irrelevant in determining the performance of the network, and the first relevant parameter is the average *square* activity. Imposing that also

$$a = \int P_\eta(\eta)\eta^2 \mathrm{d}\eta. \tag{6}$$

turns a into a parameter giving the ratio $\langle \eta \rangle_\eta^2 / \langle \eta^2 \rangle_\eta$, which in fact is a measure of the sparseness of the coding scheme†. This is evident when considering the specific example of a *binary* distribution

$$P_\eta(\eta) = (1 - a)\delta(\eta) + a\delta(\eta - 1) \tag{7}$$

† One could let $\langle \eta \rangle_\eta$ be an quantity independent of the sparseness a, and that, as mentioned, cancels out in the end. The present notation, however, is simpler when considering binary patterns, and is adopted here for consistency with previous literature.

with $\delta(x)$ being Dirac's function. With this distribution, a fraction $1 - a$ of the cells will be quiescent, on average, in any given pattern, while a fraction a will be active, firing at a rate set arbitrarily to 1. Again, the absolute scale of the firing rate will be influential in the network models considered here. Moreover, with the specific learning rules adopted in the next section to specify the strength of the synaptic efficacies, it will be clear that actually the rate of each individual cell could be rescaled independently of others, without affecting the performance of the net (this could be termed a *gauge* symmetry of the model).

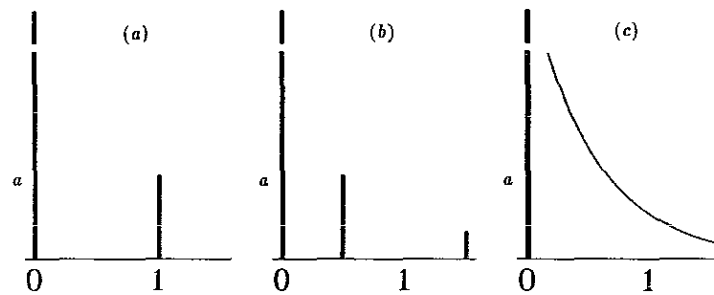


Figure 3. Examples of firing rates probability distributions used in the text: (*a*) binary (*b*) ternary (*c*) exponential

It is useful to quantify, through a, the sparseness of the coding scheme because that turns out to be the most crucial factor on which the performance of the models considered depends. The remaining features of P_η—the *structure* of the code—also affect performance, as will be discussed later on. One can consider various forms for P_η, all parametrized in terms of a. For instance a specific *ternary* choice that we have considered [47] is

$$P_\eta(\eta) = (1 - \tfrac{4}{3}a)\delta(\eta) + a\delta(\eta - \tfrac{1}{2}) + \tfrac{1}{3}a\delta(\eta - \tfrac{3}{2}). \tag{8}$$

As ternary distributions are the non-binary structures which can be specified with the least number of parameters, it has proven instructive [47] to systematically analyse the performance obtainable with ternary codes, and the case of equation (8) offers a good prototypical example of the features emerging with non-binary codes. Biologically more meaningful cases, however, are those of continuous distributions. Here the possibilities are infinite, but again one can select a representative example. One that is consistent with some recent experimental data [1] is a distribution with exponentially scarce high rates

$$P_\eta(\eta) = (1 - 2a)\delta(\eta) + 4a\exp(-2\eta). \tag{9}$$

This example on the one hand will demonstrate that autoassociative memories can function perfectly well with pattern distributions that are not bimodal or n-modal [1], and on the other will offer quantitative insight into some of the effects of a continuous distribution.

The parameter a measures the sparseness of the *stored* representation, the one used to code information in the learning phase. A different representation, whose degree of sparseness need not be the same, is generated by the retrieval process.

As the response of neuron i during retrieval is V_i, the sparseness of the retrieved representation can be quantified by

$$a_r = \frac{\langle V \rangle^2}{\langle V^2 \rangle} \tag{10}$$

where now the average $\langle \cdot \rangle$ is both over the random assignment of patterns and over the dynamical process of retrieval (which, for a deterministic process, essentially means over a set of possible initial conditions).

Finally, it has to be noted that the output of the network is assumed to be read out in terms of distributed, collective, 'macroscopic' measures. Thus, the degree to which the adapted state of an output set of N_o cells is correlated with the various stored patterns can be measured by the overlaps

$$x^\mu = \frac{1}{N_o} \sum_{i=1}^{N_o} \left\langle \frac{\eta_i^\mu}{a} V_i \right\rangle. \tag{11}$$

It is useful to define also the average activity

$$x = \frac{1}{N_o} \sum_{i=1}^{N_o} \langle V_i \rangle. \tag{12}$$

The scenario interpreted as memory retrieval is one [47] of a macroscopic correlation with just one learned pattern, i.e. one in which a single overlap is distinctly larger than the average activity, e.g. $x^1 \gg x$, while all other overlaps are scattered with small fluctuations around the average activity itself.

3.3. Learning rules

A *learning rule* is an expression that relates the modifiable component of the efficacy of the synaptic connection between two pyramidal cells to the firing activity of the two cells while storing in memory each of a set of patterns to be learned. This term, therefore, already implies the assumption that, to some approximation, the modifications due to synaptic plasticity are indeed expressible in terms of the two firing rates [12]. The change in the efficacy will be written, here, in the form

$$J_{ij}^c = \frac{1}{C} \sum_{\mu=1}^{p} F(\eta_i^\mu) G(\eta_j^\mu). \tag{13}$$

This expression represents a (linear) sum of contributions, each due to the learning of one pattern. Each contribution can be determined locally at the synapse, as it is the product of a factor which is a generic function F of the activity of the postsynaptic neuron times a factor depending (through, in general, a different function, G) on the activity of the presynaptic neuron. In the above expression there is no hierarchy nor ordering among the patterns: they are embedded with the same average strength. Further, the synaptic modification can be regarded as resulting from just one single presentation of the pattern (as well as from several), in contrast with other (i.e. error-correcting) rules which require an iterative procedure for learning.

It can be seen with simple signal-to-noise arguments that for the network to be able to retrieve a number of patterns p which, independently of the sparseness of the coding a, grows linearly with the number of modifiable synapses C, the average of the presynaptic factor over all the patterns must vanish†, i.e. $\langle G(\eta)\rangle_\eta = 0$. Moreover, the hypothesis that each arriving spike, when paired to suitable postsynaptic conditions, equally contributes to the modification of the efficacy, suggests a simple linear dependence of the factor G on the presynaptic rate η. If G is at most linear in η, the condition that $\langle G(\eta)\rangle_\eta = 0$ may be most simply enforced by having G proportional‡ to $\eta - \langle\eta\rangle$. Using $\langle\eta\rangle$ also as an overall normalization, one gets to the form

$$G(\eta) = \frac{\eta - \langle\eta\rangle}{\langle\eta\rangle} \equiv \frac{\eta}{a} - 1 \tag{14}$$

which will be adopted in the following, leading therefore to

$$J_{ij}^c = \frac{1}{C}\sum_{\mu=1}^{p} F(\eta_i^\mu)\left(\frac{\eta_j^\mu}{a} - 1\right). \tag{15}$$

It is useful to have a special notation for the variance of $G(\eta)$ over P_η, writing

$$T_0 \equiv \langle G^2(\eta)\rangle_\eta - \langle G(\eta)\rangle_\eta^2 \equiv \frac{1-a}{a}. \tag{16}$$

What remains free is the choice of the (postsynaptic) $F(\eta)$ factor. We consider here two specific examples.

In the case of feedback networks, the use of a form which results in the symmetry $J_{ij}^c = J_{ji}^c$ allows an analysis of the long-time behaviour of the neural activities based on the definition of a Lyapunov function [25, 14, 5]. This requires both that if there is a modifiable synapse from j to i then there is also one§ from i to j, and that

$$F(\eta) = \frac{\eta}{a} - 1 \tag{17}$$

which results, with equations (13)–(14), in a *covariance* rule for synaptic modification, i.e. the change in the efficacy is proportional to the covariance of pre- and postsynaptic activity [43]:

$$J_{ij}^c = \frac{1}{Ca^2}\sum_{\mu=1}^{p}(\eta_i^\mu - \langle\eta_i^\mu\rangle)(\eta_j^\mu - \langle\eta_j^\mu\rangle). \tag{18}$$

The average a_F and variance c_F of the postsynaptic factor over P_η are in this case $a_F = 0$ and $c_F = T_0$.

† This ensures that, when retrieving a pattern, the variance of the noise due to the storage of $p-1$ other patterns is proportional not to $p-1$, but just to $(p-1)/C$ [42, 52].

‡ Note that this implies that the sign of the efficacy change can reverse, for a given level of postsynaptic activity, depending on the level of presynaptic activity in the pattern.

§ This holds for the fully connected network $\mathcal{M}1$, but could also be imposed as an additional constraint in a network with diluted connectivity.

When the architecture of the network is such that its behaviour can be analysed with simpler methods, one can explore the effects of using alternative forms for the F factor. An interesting alternative would be to try to model some current neurobiological hypotheses on the mechanisms of long-term potentiation (LTP) based on the activation of NMDA-receptors. This activation seems to occur only when the postsynaptic membrane is very depolarized [16,15,20]. Such a feature might be modelled [10] by setting a threshold for synaptic modification (say, in learning a pattern) which is higher than the threshold above which the postsynaptic cell fires. If the pattern to be learned is binary, the nonlinearity due to this additional threshold is not expected to make much difference, as the modification occurs only at two discrete values of the postsynaptic rate, and the form of $F(\eta)$ in between the two values is irrelevant. New effects can already be seen, however, with ternary patterns. To be specific, having in mind the distributions of equations (7) and (8), let

$$F(\eta) = \frac{2\eta^2 - \eta}{a} \tag{19}$$

which implies that the modification occurs, both in the binary and in the ternary case considered, only for the highest postsynaptic firing rate. In the specific ternary case of equation (8) synapses onto neurons that fire at the intermediate rate while encoding the pattern are not affected. For the first two moments of F one has

$$c_F + a_F^2 = \begin{cases} 1/a & \text{binary} \\ 3/a & \text{ternary} \\ 7/a & \text{exponential} \end{cases} \tag{20}$$

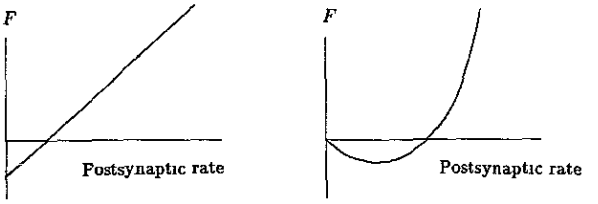
for this model NMDA rule, which will be used later on. Note that equation (19) implies a modification in the opposite direction for postsynaptic activations $\eta < 1/2$ (figure 4). This particular feature, which has been explicitly suggested with theoretical arguments [11], is included here merely for the sake of analytical simplicity (the simple nonlinearity in equation (19)). Similar nonlinearities due only to a threshold effect, and not including sign reversals, produce essentially the same results as derived for this particular case later on.

4. Capacity measures

4.1. Fixed point analyses

We are not concerned here with the *dynamics* of the evolution of the activity variables, but only with the stable *fixed points* of this evolution. More precisely, with the fixed points of the equations describing the evolution of *macroscopic* quantities, such as those defined at the end of subsection 3.2, quantities that monitor the collective activity of populations of cells. If processing in the system is distributed, it must be possible to read off its output in terms of quantities of this type—for associative memory retrieval, in particular, in terms of correlations with the stored patterns of activity (measured as in equation (11)).

Such fixed points are *not* associated with the stabilization of single-unit activity rates: these may still fluctuate, even when averaged over short times, once macroscopic quantities have reached a fixed point. Nor can one learn much from a calculation (when feasible) of the *time* it takes the model to reach a fixed point. As



Post-synaptic rate	Pre-synaptic rate					
	Covariance			NMDA		
	Zero	Low	High	Zero	Low	High
Zero	+	−	−−	0	0	0
Low	−	+	++	0	0	0
High	−−	++	++ ++	−−	++	++ ++

Figure 4. The two-sample learning rule considered. above, dependence of the postsynaptic factor on the firing rate, below, the possible modifications produced in the storing of the specific ternary pattern of equation (8).

noted in subsection 2.2 it is doubtful whether such a dynamical quantity would bear any meaningful resemblance to those characterizing the process in a real neural system. Third, even the existence or otherwise of fixed points associated with retrieval behaviour does not imply anything definite about whether even the model system will indeed retrieve or not. In the absence of a retrieval fixed point, the network might still be able to usefully extract large amounts of information encoded in the synaptic strengths, and in its presence it will in many situations fail to do so.

Still, as long as methods are lacking for studying realistic dynamics, the study of fixed points, which is based on the more reliable 'stationary' relations at the single-cell input–output level, is the best analytical tool available to give indications on how the behaviour of the network is affected by a variety of factors, such as those treated in this paper. *This* is the advantage of focusing on the *attractor* states in autoassociative memories [4]. Attractor retrieval fixed points cease to exist when the effective noise induced by memory loading goes beyond a certain limit. In the following, this limit on memory loading will be studied under the assumption that the loading itself is the dominant source of noise in the network, i.e. all other sources, be they fast- or slow-varying in time, are taken to be negligible and set to zero.

4.2. The equation for the storage capacity

The maximum on the number p of patterns that can be encoded on the synaptic strengths and individually retrieved is, when equation (14) holds, and in the limit $C \to \infty$, a number proportional to C. The maximum α_c of the proportionality factor $\alpha \equiv p/C$ can still span several orders of magnitude. A derivation of the fixed point equations, and of the resulting equation which yields α_c, is given for network $\mathcal{M}2$, in the appendix. In the two other cases it has been reported elsewhere, for network $\mathcal{M}1$ [47] and $\mathcal{M}3$ [48]; further novel examples will be mentioned in the text.

The equation for α_c is expressed in terms of averages that are functions of two

auxiliary variables v and w [47]. v quantifies the ratio of the signal specific to the pattern to be retrieved to the noise (due to memory loading), and w that of the uniform background signal, again to the noise.

The relevant averages are:

$$A_1(w,v) = \frac{1}{vT_0}\left\langle\left(\frac{\eta}{a}-1\right)\int^+ \mathrm{D}z\,\{w+v[1+F(\eta)]-z\}\right\rangle_\eta - \left\langle\int^+ \mathrm{D}z\right\rangle_\eta$$

$$A_2(w,v) = \frac{1}{vT_0}\left\langle\left(\frac{\eta}{a}-1\right)\int^+ \mathrm{D}z\,\{w+v[1+F(\eta)]-z\}\right\rangle_\eta \qquad (21)$$

$$A_3(w,v) = \left\langle\int^+ \mathrm{D}z\,\{w+v[1+F(\eta)]-z\}^2\right\rangle_\eta$$

where the averaging is over the distribution P_η and over a Gaussian variable z, and the z-average is carried out up to a threshold:

$$\int^+ \mathrm{D}z\,(\,) = \int_{-\infty}^{w+v[1+F(\eta)]} \frac{\mathrm{d}z}{\sqrt{2\pi}}\exp(-z^2/2)(\,). \qquad (22)$$

F is the postsynaptic factor introduced in equation (13).

The fully connected feedback model $\mathcal{M}1$ can be analysed [47] with the help of a suitably defined energy function (hence the constraint that $F(\eta)$ be given by equation (17)). This network, driven by its symmetric interaction, relaxes into an equilibrium state defined by a certain probability distribution in the space of all possible values of the set of neural activities. Some of the possible equilibrium states are associated with the retrieval of one of the patterns in memory. These retrieval states exist when a closed line in the v, w plane defined by the equation

$$A_1^2(v,w) - \alpha A_3(v,w) = 0 \qquad (\mathcal{M}1) \qquad (23)$$

is intersected by another line, whose position varies with the gain g of the model. As α increases, the perimeter of the line defined by equation (23) shrinks, until at $\alpha = \alpha_c$ the line reduces to a point and then disappears. Therefore, for $\alpha < \alpha_c$ retrieval states exist, for suitable values of the gain, while above this limit they do not, for *any* value of the gain†. α_c thus measures an optimal storage capacity, optimal in maximizing the number of stored patterns, and it brings out the functional dependence of this optimal number on the pattern statistics P_η.

In the multilayered structure $\mathcal{M}2$ (see figure 2), neural activities propagate in time from layer to layer. Correspondingly, macroscopic quantities can be considered at each time slice as defined on the population of units in the layer that was reached last. When these quantities reach a fixed point, no further processsing occurs in the following layers, which, macroscopically, appear to have the same distribution of

† Actually, one of the effects of feedback in this model is to *renormalize* the gain parameter g. The propagation of the correlations over feedback loops of different size makes single units respond as if they were endowed with an effective gain g' higher than the real one g. Summing over all loops yields an expression for the inverse effective gain $g'^{-1} = g^{-1} - \alpha T_0/(1-\psi)$ [47] or, when more correctly neglecting self-interactions, $g'^{-1} = g^{-1} - \alpha T_0\,\psi/(1-\psi)$, with $\psi = g'T_0(A_2 - A_1)$. This renormalization does not affect α_c.

activities. The analysis we present in the appendix shows that fixed points associated with retrieval exist provided that, similarly to the $\mathcal{M}1$ case, the equation

$$A_2^2(v,w) - \lambda_1[A_2(v,w) - A_1(v,w)]^2 - \alpha\lambda_2 A_3(v,w) = 0 \qquad (\mathcal{M}2) \tag{24}$$

is satisfied on a closed line in the v, w plane. Here it is not necessary to impose the symmetry condition $F(\eta) = G(\eta)$, and the analysis is valid for any form of $F(\eta)$. The parameters $\lambda_{1,2}$ equal 1 if equation (17) holds, and in general are defined as

$$\lambda_1 = \frac{(d_F - a_F)^2}{T_0^2} \qquad \lambda_2 = \frac{c_F + a_F^2}{T_0} \tag{25}$$

with

$$a_F = \langle F(\eta)\rangle_\eta \qquad c_F = \langle F^2(\eta)\rangle_\eta - \langle F(\eta)\rangle_\eta^2 \qquad d_F = \left\langle \frac{\eta}{a} F(\eta) \right\rangle_\eta . \tag{26}$$

In the highly diluted case ($\mathcal{M}3$), if a subset of the units is activated by an external stimulus the activation will spread in a sequence defined by the sparse connectivity, even if no layer structure was defined *a priori*. The equation yielding the optimal capacity reads [48]

$$A_2^2(v,w) - \alpha\lambda_2 A_3(v,w) = 0 \qquad (\mathcal{M}3) \tag{27}$$

and it gives α_c in the same fashion as above.

Up to now the network has been assumed to operate under no external influence, other than (uniform ones, and) the transient stimulus which determines the initial pattern of activation. A *persistent* and non-uniform external stimulus correlated with one or a few of the stored patterns can be modelled by setting in equation (4)

$$\Delta h_i = \sum_\mu s^\mu \frac{\eta_i^\mu}{a} \tag{28}$$

where the s^μ parametrize the strength of the patterns embedded in the ongoing afferent stimulation. This leads to higher optimal capacities, which can be derived within the same formalism ([47], [48]). Given that this resulting increase in capacity appears to be rather uniform over the variety of situations which are expressly studied in this paper, for simplicity results will be presented only in the case in which external stimuli are purely transient, $s^\mu = 0$. When considering a comparison with the performance of models studied in the literature (e.g. [22]) in which certain cells even have their output *clamped* by the afferent, is it useful to bear in mind two facts. While the results presented here refer to the case $s^\mu = 0$, which perhaps leads to capacity 'underestimation', they are also (a) derived for large systems and (b) valid as limits on the existence of retrieval states, which might be argued to lead to capacity 'overestimation'. This is because near these limits the basins of attraction of retrieval states will tend to be very small, allowing retrieval only when the initial network state already contains, by closely reproducing one of the stored patterns, essentially all the information to be retrieved, a situation obviously different from that of retrieval through completion of a partial cue.

It is also possible to derive capacity equations for network architectures which can be regarded as intermediate between the extreme cases $\mathcal{M}1$, $\mathcal{M}2$, $\mathcal{M}3$. An example is a multilayered network in which the connectivity from layer to layer is not complete, and each cell receives modifiable synapses from a subset of C cells randomly picked among the N ones of the previous layer. In this case α_c is determined by the equation

$$A_2^2\left[A_2^2-\lambda_1(A_2-A_1)\right]-\alpha\lambda_2 A_3\left[A_2^2-\left(1-\frac{C^2}{N^2}\right)\lambda_1(A_2-A_1)\right]=0 \tag{29}$$

where C/N is a measure of the degree of dilution in the connectivity. This formula can be applied to all networks intermediate between the highly diluted limit $\mathcal{M}3$, which is reached for $C/N \to 0$, and the multilayered structure $\mathcal{M}2$, whose capacity is given by the above equation when $C = N$.

Another case not treated before, and which may approach realistic connectivity patterns, is that of a network halfway between $\mathcal{M}1$ and $\mathcal{M}2$: in particular, one which is fully connected with modifiable reciprocal synapses within each one layer, and also, with unidirectional synapses, from one layer to the next. Activations spread from layer to layer, and the input from each preceding layer can be expressed in terms of macroscopic variables which have already reached a fixed point at the single-layer level. This is obtained by setting the input to a cell in layer l to

$$h_i^l=\sum_{j\neq i}J_{ij}^c V_j^l+b\left(\sum_{j\neq i}\frac{V_j^l}{N}\right)+\sum_k J_{ik}^c V_k^{l-1} \tag{30}$$

where the inputs through modifiable synapses come both from the neurons in the same layer, the V_j^l, and from those in the preceding one, the V_k^{l-1}. The J_{ij}^c are written in the symmetric form, with the F factor as in equation (17). The overall optimal capacity is determined by finding, through a free-energy calculation, the equations expressing equilibrium at the single-layer level, and then by considering global fixed points, at which macroscopic variables maintain their value from layer to layer†. The corresponding capacity equation is

$$A_1\frac{(A_1+A_2)}{2}-\alpha A_3=0. \tag{31}$$

The various equations derived so far can be used to order the different models on a capacity scale, as at the bottom of figure 2. In fact, it is easy to show analytically‡ that, under equal conditions (as defined by the choice of $F(\eta)$ and P_η), one has

$$\alpha_c(\mathcal{M}1)<\alpha_c(\text{equation (31)})<\alpha_c(\mathcal{M}2)<\alpha_c(\text{equation (29)})<\alpha_c(\mathcal{M}3). \tag{32}$$

Therefore the number of patterns that can be stored, for fixed C, in the various architectures considered is bounded by the two extreme cases of a fully connected

† Incidentally, this calculation answers a question raised in [8], indicating that such a sequence of networks would not improve considerably on the performance of a single, fully connected feedback net.

‡ To prove the string of inequalities in both equation (32) and equation (36) it suffices to check that the same relations hold between the left-hand sides of equations (23)–(31) for any given triplet v, w, α.

feedback network (lowest number of patterns) and of an extremely diluted connectivity (highest). Moreover, it will be shown in the following that in the neurobiologically interesting condition of *sparse coding* (such as it applies to the CA3 cells [35]) these two extremes approach, capacity-wise, each other.

Alongside the maximum number of firing patterns that can be stored, another meaningful measure of capacity for an associative network is the total amount of information [44], I, which can be stored in the synaptic efficacies *and* retrieved following stimulation by an appropriate cue. I can be defined [47] for arbitrary distributions P_η, generalizing similar expressions used with networks of binary units [7,21]. I can be computed for any $\alpha < \alpha_c$ and any appropriate value of the gain g by using the parameters v, w corresponding to the retrieval fixed point which exists for that choice of α, g. In terms of v and w, I is given by the expression (in bits per synapse)

$$I = \alpha\left\{\left\langle\int_0^\infty \frac{dt}{\sqrt{2\pi}} c_\eta^t(v,w)\log_2\left[\frac{c_\eta^t(v,w)}{\langle c_{\eta'}^t(v,w)\rangle_{\eta'}}\right]\right\rangle_\eta + \left\langle c_\eta^0(v,w)\log_2\left[\frac{c_\eta^0(v,w)}{\langle c_{\eta'}^0(v,w)\rangle_{\eta'}}\right]\right\rangle_\eta\right\} \tag{33}$$

with the Gaussian probability density c_η^t given by

$$c_\eta^t(v,w) = \exp\left(-\frac{\{w + v[1+F(\eta)] - t\}^2}{2}\right) \tag{34}$$

and the probability of zero response c_η^0 by

$$c_\eta^0(v,w) = \int_{w+v[1+F(\eta)]}^\infty \mathrm{Dz} \qquad (\equiv 1 + A_1 - A_2). \tag{35}$$

What will be considered here is only I_m the maximum of I over α ($< \alpha_c$) and g. This maximum occurs generally about halfway before α_c and is quite flat with respect to varying both α and g (in the range which still allows retrieval). It is this quantity which will be studied in the following in its dependence on the network's architecture, pattern statistics and learning rule.

It should be noted that, in parallel to equation (32), one can also show that

$$I_m(\mathcal{M}1) < I_m(\text{equation (31)}) < I_m(\mathcal{M}2) < I_m(\text{equation (29)}) < I_m(\mathcal{M}3). \tag{36}$$

4.3. Dependence on the coding scheme and on the learning rule

By computing the A averages one can evaluate α_c and I_m for a variety of situations, corresponding to different choices of P_η and $F(\eta)$. It is convenient to present the results as graphs where the quantities evaluated are plotted as functions of the sparse coding parameter a, which strongly affects them.

The task of understanding how different architectures affect the capacity is greatly simplified by the ordering relationships (32) and (36). A comparison could be made, for example, by plotting all the various curves $\alpha_c(a)$, for fixed P_η and assuming that $F(\eta)$ is given by the covariance rule, (17). One can start, however, with only the

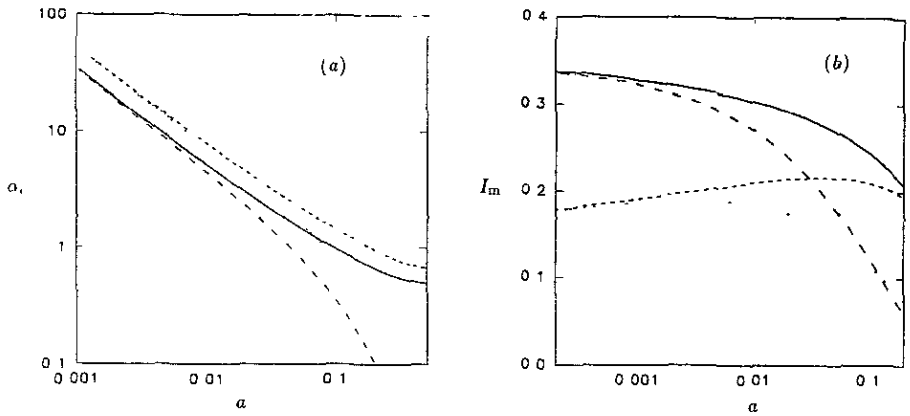


Figure 5. The storage capacity dependence on the connectivity: a fully connected feedback net (dash-dotted line for binary patterns and dotted line for exponential patterns) compared with an extremely diluted one (full line, binary; dashed line, exponential). (*a*) α_c, maximum number of stored pattern per modifiable synapses onto each neuron (*b*) I_m, maximum amount of retrievable information per neuron per modifiable synapses onto each neuron, in bits. a, is the sparse coding parameter.

curves for models $\mathcal{M}1$ and $\mathcal{M}3$, as shown in figure 5(*a*), as other cases fall in the middle, due to equation (32). Both the binary and the exponential forms for the pattern distributions are chosen in figure 5, to give two examples.

The first feature to notice is that α_c increases and eventually diverges as $a \to 0$, to first approximation as $(a \ln(1/a))^{-1}$. The same happens with binary neurons [13,46,19], and the reason is that while the signal pointing in the direction of one pattern grows, in our normalization, as a^{-1}, the root-mean-square noise due to memory loading grows only as $a^{-1/2}$. It should be noted that as soon as a is small enough to invalidate the assumption that aC is a very large number, the actual capacity will start to deviate from that predicted in the present theory.

The important observation to be made in figure 5(*a*) is that while $\alpha_c(a)$ is always higher for the highly diluted net, clearly the difference becomes irrelevant in the region of sparse coding.

The information content does not diverge as $a \to 0$ (figure 5(*b*). This is due to the fact that while α_c grows, the information contained in sparsely coded patterns decreases. Again, the difference between highly diluted and fully connected feedback nets (as well as that with all intermediate cases) disappears for low values of a.

Turning now to the dependence on the structure of the code, this can be studied by varying, for fixed architecture and learning rule, the probability distribution P_η. In [47] the performances obtainable with binary and ternary codes were contrasted in the $\mathcal{M}1$ case, and figure 6 shows the capacity of an $\mathcal{M}3$ network with binary (7), ternary (8) and exponential (9) pattern distributions. The comparison shows that the number of patterns that can be stored can be somewhat higher for non-binary distributions (of course, not for *any* non-binary distribution).

The way the structure of the code affects the amount of information retrievable from the memory, figure 6(*b*), is interesting. As mentioned above, three factors enter into I_m: the number of patterns stored (and figure 6(*a*) shows that $\alpha_c(a)$ is higher

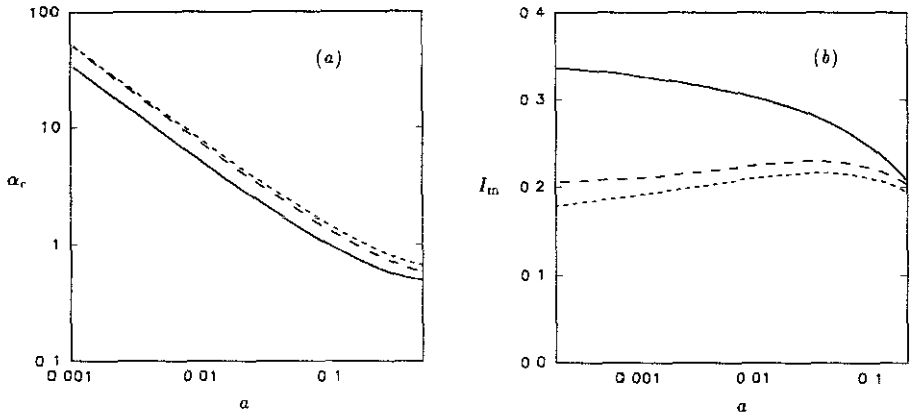


Figure 6. The effect of varying the structure of the code (*a*) $\alpha_c(a)$ and (*b*) $I_m(a)$, for binary (full line), ternary (long-dashed line) and exponential (short-dashed line) pattern distributions, for a highly diluted ($\mathcal{M}3$) net with a covariance learning rule

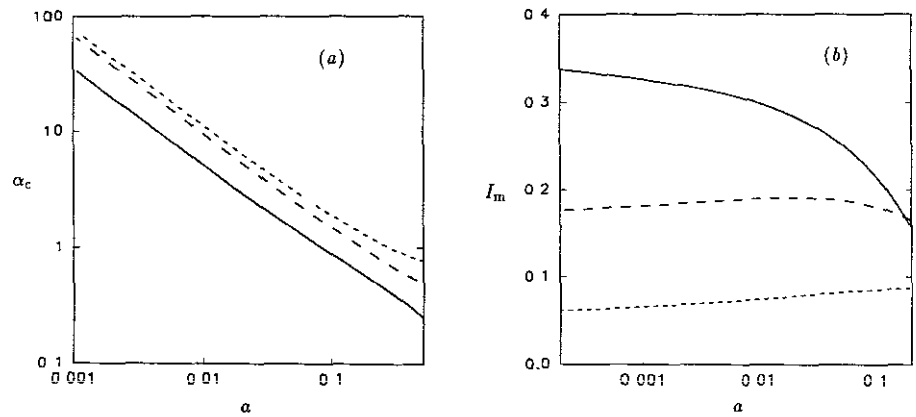


Figure 7. The effect of a nonlinearity in the learning rule (*a*) $\alpha_c(a)$ and (*b*) $I_m(a)$, for the same cases as in figure 6 but with the model 'NMDA' rule of equation 19 rather than the linear covariance rule

for non-binary codes), the information content of each stored pattern (which clearly is higher for non-binary codes), and the fraction of this information which is lost during retrieval. Apparently this last factor more than compensates for the other two, and the poorer retrieval quality brings down the total information retrievable using non-binary codes. Note that in the case of continuous distributions, such as the exponential one of the example, the actual stored information per pattern is infinite, but such a vanishing fraction of this infinity can be extracted from memory, due to the natural resolution limit induced by extensive memory loading, that the final result

is essentially the same as for the ternary distribution, and not very different from the binary one.

The use during storage of the nonlinear, 'NMDA'-like learning rule of equation (19) results in the capacities of figure 7. The graph for $\alpha_c(a)$ looks remarkably similar to that produced by the covariance rule, in figure 6(*a*). For ternary patterns, however, and more markedly for exponential ones, the number of patterns that can be stored is higher with the NMDA rule, for essentially all values of a. For binary patterns the nonlinearity in itself does not produce any difference (subsection 3.3), and the only deviation from the covariance rule behaviour occurs when the coding is not sparse, $a \simeq 0.5$†.

The effect of the nonlinearity on the information content is, once more, reversed: while for binary patterns $I_m(a)$ is again essentially unaffected, for ternary patterns there is a decrease in the retrievable information, which becomes very marked for the continuous, exponential pattern distribution. This result is easy to understand: the nonlinearity in the way firing patterns are stored in the memory produces a distortion, which results in retrieved representations which do not faithfully reproduce the information present in the original representations [42]. The distortion is quite naturally particularly serious in the case of continuous distributions (and it would be all the more serious the stronger the nonlinearity), and irrelevant with binary distributions.

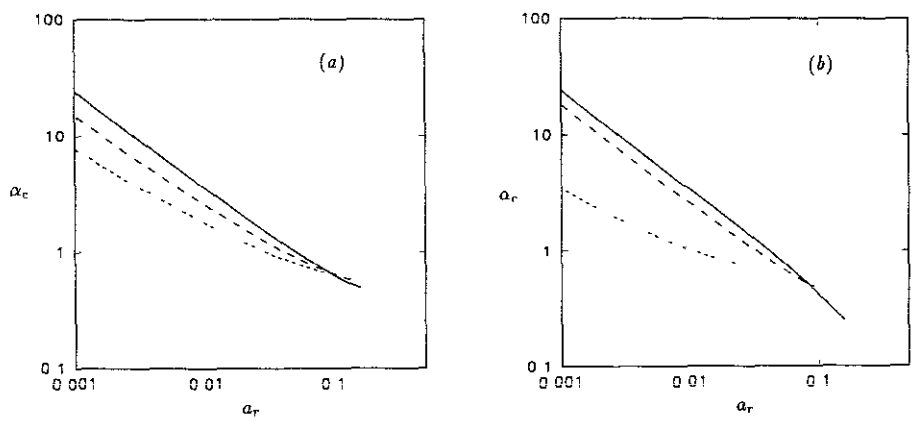


Figure 8. α_c plotted against a_r, for the same coding structures as in figures 6–7, and for (*a*) the linear covariance rule, (*b*) the nonlinear NMDA rule (and model $\mathcal{M}3$).

Also the fact that more non-binary than binary patterns can be retrieved for equal values of the parameter a can be understood, as well as the fact that this effect is enhanced by the use of the NMDA-like learning rule of equation (19). They result from the *retrieved* patterns being sparser, in all these cases, than the ones stored during learning, i.e. $a_r < a$ (cf equation (10)). α_c grows with sparse patterns, and

† In that case the capacity is sensitive to whether $a_F = 0$ (as in the covariance rule) or not (as in the NMDA rule), whereas for sparse coding c_F gives the dominant contribution, e.g. to λ_2, and a_F is irrelevant.

it grows further if the retrieved patterns are even sparser. To make that clear, it is useful to plot α_c as a function of a_r rather than a. Figure 8 shows that the relative advantages of non-binary representations disappear when compared in this fashion, all the more when a nonlinear rule, which accentuates the sparsity of retrieved continuous representations, is used.

5. Discussion

These results show that autoassociative memory models based on cooperative parallel processing, can function, and function efficiently, also when freed of some of the unrealistic features introduced in earlier versions of the models. Biologically plausible graded response nets can store and retrieve non-binary patterns, even patterns with exponentially scarce fast-firing neurons. The appearance, in earlier schemes, of bimodal distributions of firing rates is to be regarded as a simplification (due to the representation of neuronal current-to-frequency transduction as binary gating), and *not* as associated with the convergence towards dynamical attractors, or with the role of feedback in this process.

On a more quantitative level, the use of threshold-linear units allows in particular the analytical study of the storage capacity of different models, in a way that helps address biologically relevant issues. α_c and I_m have been shown to depend rather mildly on the detailed structure of the coding distribution and on the detailed form of the learning rule. It should be remembered that the learning rules considered both belonged to a general 'Hebbian' class representing strictly associative mechanisms of synaptic plasticity, with the further restriction imposed in equation (14). The only case in which there is a rather markedly reduced performance—in fact, reduced retrievable information I_m—is when a *combination* of a nonlinear learning rule with a very structured, graded or continuous, pattern distribution results in severe information distortion. This is interesting, in view of the possibility that precisely this combined condition describes the situation in certain real systems. Such a reduction in capacity would clearly not affect the efficiency of a system that was intended to operate merely as a stimulus classifier rather than to actually retrieve detailed information from memory; in that case the number of stored patterns would be the relevant measure of efficiency, and that is not reduced. But it seems unlikely that any autoassociative system in the brain would afford to waste extensive storage space just to produce afferent stimulus classification, which could be achieved with other means.

These results are derived for threshold-linear units, but would apply to any similar type of graded response units, such as generically sigmoidal, provided it is inhibition that controls the overall activity level, rather than single-unit saturation (in which case sigmoidal units approach binary ones). In this respect, it could be objected that in layers IV and V of neocortex, and even more in the hippocampus, especially in the CA3 region, many pyramidal cells tend to display bursting behaviour, evidenced *in vitro* as a characteristic response to orthodromic excitation in which several spikes ride on top of a single depolarizing wave. To the extent that bursting were an all-or-none phenomenon, it would suggest an essentially binary processing. Bursting activity, however, appears to be graded, and may also be a less prominent in primates than in certain subprimate species [9]; and Θ-rythmicity, which in rodents is thought to shut down hippocampal processing every 100 ms, has hardly been observed in primates [41]. It is interesting to note, in any case, that a representation of the firing behaviour

of neurons engaged in memory retrieval, which is the present objective, which does not take bursting into account, could still be compatible with bursts having a specific role during learning.

The analysis of different patterns of connectivity provides useful insight into the role of feedback in autoassociative systems, or at least in the way it affects their capacity. Consider, for example, the multilayered feedforward autoassociator $\mathcal{M}2$. It can be thought of [18] as a version of the fully connected feedback net $\mathcal{M}1$, which is unfolded in time over different layers. In the unfolding, the positive effect of feedback, in iterating the cooperative motion towards retrieval attractors, is preserved, but negative effects are washed out, in that noisy correlations are reset at each layer and not let to reverberate. This 'purification' of the signal results in a higher capacity—cf inequalities (32),(36)—but of course costs in neural components, as for a given output size the number of neurons and synapses has to be multiplied by the number of layers. The highly diluted net $\mathcal{M}3$ represents the limit in which the signal is further purified by removing the correlations that stem from sharing inputs among different units, and of course an even higher capacity is achieved in this limit. A discussion in these terms had been suggested in [18]. Now, the interesting result here is that the relative advantages of purifying signals in this fashion disappear as the coding becomes sparse (figure 5). What remains is the obvious advantage of the compact feedback net in terms of hardware components†. What are the implications for real systems, where coding does seem to be relatively sparse? In CA3 for example, where the connectivity is of the order of 2% (in the rat [3]) and shows no marked topographical organization, the suggestion would be that having many more cells ($\simeq$ 600 000 in the rat) than axon collateral synapses per cell (12 000, [3]) is *not* in order for the net to be able to store more patterns thanks to its diluted connections. The reasons for having many more cells than just $C = 12\,000$ are likely to be of a very different kind. For example, a certain minimal size of the output representation might be required to carry the information funnelled through the hippocampal formation.

Finally, it might be useful to take advantage once more of the data concerning the rat hippocampus in order to provide an illustrative numerical example of the estimated capacity of a real system. Thus, if the various factors discussed in this paper (in particular, the sparseness of coding) are such as to result in $\alpha_c \approx 3$ and $I_m \approx 0.2$—the middle ranges in the figures included in the text—then the CA3 system in the rat could store in the order of 36 000 memories, or if maximizing I_m instead, 1 400 Mbits of retrievable information. In primates, and especially in man, these numbers would grow considerably due to the larger number of modifiable inputs to each pyramidal cell, to the coding being probably sparser (which would mainly affect the number of memories) and to the larger number of cells in CA3 itself (which affects the amount of information, but not the number of memories). Clearly, the most crucial experimental support for the hypothesis that CA3 is an autoassociative memory with extensive capacity would come from the evidence that the synapses on its collaterals are associatively modifiable, and display both LTD and LTP, as in equation (14). But the models indicate which other factors it is important to investigate experimentally: the number of synapses per cell, the sparseness of the firing activity, and how the sparseness varies in different modes of operation of the system.

† And also, on a different level, the ease with which new patterns can be stored, in contrast with the difficulties arising with a feedforward network.

Acknowledgments

This work was supported in part by the Medical Research Council (PG 8513790) and by the EEC BRAIN Initiative (grant 88300446/JU1).

Appendix

In this appendix the capacity equation is derived in the sample case of the multilayered network A12.

The starting point is the 'microscopic' equation describing the input to a given unit of layer $l+1$ in terms of the outputs of the units in layer l:

$$h_i^{l+1} = \sum_{j=1}^{C} J_{ij}^{c} V_j^l + b\left(\sum_{j=1}^{C} V_j^l / C\right) \tag{37}$$

with

$$J_{ij}^{c} = \frac{1}{C}\sum_{\mu=1}^{p} F(\eta_i^{l+1,\mu})\left(\frac{\eta_j^{l,\mu}}{a} - 1\right). \tag{38}$$

It is convenient to define the average activity of the layer

$$X^l = \frac{1}{C}\sum_{i=1}^{C} V_i^l \tag{39}$$

and the subtracted overlaps

$$\hat{X}^{l,\mu} = \frac{1}{C}\sum_{i=1}^{C}\left(\frac{\eta_i^{l,\mu}}{a} - 1\right) V_i^l. \tag{40}$$

For the non-stochastic feedforward process considered here, these quantities are fully determined by the activities in the first layer, $\{V_i^1\}$, and by the sets $\{\eta_i^{k,\mu}\}$ for $k = 1,\ldots,l$ and $\mu = 1,\ldots,p$. From them, by averaging over initial conditions and over the random assignment of patterns, one obtains quantities of the type introduced in subsection 3.2,

$$x^l = \langle X^l \rangle \qquad \hat{x}^{l,\mu} = \langle \hat{X}^{l,\mu} \rangle. \tag{41}$$

To derive equations describing the retrieval of a pattern (e.g. $\mu = 1$), the average $\langle\,\rangle$ is over a set of initial conditions such that the subtracted overlaps with all other patterns on average vanish, $\hat{x}^{1,\mu\neq 1} = 0$. In that case the inputs to layer $l+1$

$$h_i^{l+1} = F(\eta_i^{l+1,1})\hat{X}^{l,1} + \sum_{\mu\neq 1} F(\eta_i^{l+1,\mu})\hat{X}^{l,\mu} + b(X^l) \tag{42}$$

can be written [18] in terms of a signal plus a Gaussian noise term

$$h_i^{l+1} = F(\eta_i^{l+1,1})\hat{x}^{l,1} - T_0\varrho^{l+1}z + b(x^l) \tag{43}$$

where z has a Gaussian distribution with unit variance, and the variance of the noise is

$$(T_0\varrho^{l+1})^2 = \sum_{\mu\neq 1}\langle F^2(\eta_i^{l+1,\mu})(\hat{X}^{l,\mu})^2\rangle = \sum_{\mu\neq 1}(c_F + a_F^2)\langle(\hat{X}^{l,\mu})^2\rangle. \tag{44}$$

The fully connected layered architecture determines the average propagation of noisy fluctuations

$$\begin{aligned}\langle(\hat{X}^{l,\mu})^2\rangle &= \frac{1}{C^2}\sum_i\left\langle\left(\frac{\eta_i^{l,\mu}}{a}-1\right)^2(V_i^l)^2\right\rangle + \frac{1}{C^2}\sum_{i\neq j}\left\langle\left(\frac{\eta_i^{l,\mu}}{a}-1\right)\left(\frac{\eta_j^{l,\mu}}{a}-1\right)V_i^lV_j^l\right\rangle \\ &= \frac{1}{C}T_0y^l + \frac{1}{C^2}\sum_{i\neq j}\left\langle\left(\frac{\eta_i^{l,\mu}}{a}-1\right)\right. \\ &\quad\left.\times\left(\frac{\eta_j^{l,\mu}}{a}-1\right)\frac{dV_i^l}{dh_i^l}\frac{dV_j^l}{dh_j^l}F(\eta_i^{l,\mu})F(\eta_j^{l,\mu})(\hat{X}^{l-1,\mu})^2\right\rangle \\ &= \frac{1}{C}T_0y^l + (d_F-a_F)^2g^2\left\langle\phi\left(\frac{h-T_{hr}}{T_0\varrho^l}\right)\right\rangle_{\eta^1}^2\langle(\hat{X}^{l-1,\mu})^2\rangle\end{aligned} \tag{45}$$

with

$$y^l = \frac{1}{C}\sum_{i=1}^{C}\langle V_i^2\rangle \tag{46}$$

$$\phi(x) = \int_{-\infty}^{x}\frac{dy}{\sqrt{2\pi}}e^{-y^2/2} \tag{47}$$

and the other symbols have been defined in the main text. Therefore

$$(\varrho^{l+1})^2 = \alpha\lambda_2y^l + (gT_0)^2\lambda_1\left\langle\phi\left(\frac{h-T_{hr}}{T_0\varrho^l}\right)\right\rangle_{\eta^1}^2(\varrho^l)^2. \tag{48}$$

At the fixed point, macroscopic quantities do not vary from layer to layer. The asymptotic relationships are then simply obtained by suppressing the layer index. In addition to the above equation for ϱ, one has for example

$$\begin{aligned}\hat{x}^1 &= g\left\langle\left(\frac{\eta^\mu}{a}-1\right)\int_{h>T_{hr}}\mathrm{Dz}(\mathrm{h}-\mathrm{T_{hr}})\right\rangle_{\eta^1} \\ y &= g^2\left\langle\int_{h>T_{hr}}\mathrm{Dz}(\mathrm{h}-\mathrm{T_{hr}})^2\right\rangle_{\eta^1}.\end{aligned} \tag{49}$$

Introducing the two signal-to-noise ratios:

$$\begin{aligned}w &= (b(x)-\hat{x}^1-T_{hr})/(T_0\varrho) \qquad \text{(uniform)} \\ v &= (\hat{x}^1)/(T_0\varrho) \qquad \text{(specific)}\end{aligned} \tag{50}$$

and the A-averages defined in the main text leads to the system

$$\begin{aligned} \varrho^2 &= \alpha\lambda_2 y + \lambda_1 (T_0 g)^2 (A_2 - A_1)^2 \varrho^2 \\ \hat{x}^1 &= (T_0 g) A_2 \hat{x}^1 \\ y &= (T_0 g)^2 A_3 \varrho^2. \end{aligned} \tag{51}$$

The system has a solution corresponding to retrieval, $\hat{x}^1 > 0$, for some value of g if the equation

$$A_2^2 - \lambda_1 (A_2 - A_1)^2 - \alpha\lambda_2 A_3 = 0 \tag{52}$$

is satisfied somewhere in the w, v plane. This requirement determines both α_c and the I_m.

References

[1] Abeles M, Vaadia E and Bergman H 1990 Firing patterns of single units in the prefrontal cortex and neural network models *Network* **1** 13–25

[2] Amaral D G, Insausti R and Cowan W M 1984 The commissural connection of the monkey hippocampal formation *J Comput. Neurol.* **224** 307–36

[3] Amaral D G, Ishizuka N and Claiborne B 1990 Neurons, numbers and the hippocampal network *Prog. Brain Res* **83** 1–11

[4] Amit D J 1989 *Modelling Brain Function* (Cambridge: Cambridge University Press)

[5] Amit D J, Gutfreund H and Sompolinsky H 1985 Spin-glass models of neural networks *Phys Rev* A **32** 1007–18

[6] Amit D J, Gutfreund H and Sompolinsky H 1987 Statistical mechanics of neural networks near saturation *Ann. Phys. NY* **173** 30–67

[7] Amit D J, Gutfreund H and Sompolinsky H 1987 Information storage in neural networks with low levels of activity *Phys. Rev.* A **35** 2293–303

[8] Amit D J, Parisi G and Nicholis S 1990 Neural potentials as stimuli for attractor neural networks *Network* **1** 75–88

[9] Avoli M and Olivier A 1989 Electrophysiological properties and synaptic responses in the deep layers of the human epileptogenic neocortex in vitro *J. Neurophysiol.* **61** 589–606

[10] Bear M F, Cooper L N and Ebner F F 1987 A physiological basis for a theory of synapse modification *Science* **237** 42–8

[11] Bienenstock E L, Cooper L N and Munro P W 1982 Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex *J. Neurosci.* **2** 32–48

[12] Brown T H, Kairiss E W and Keenan C L 1990 Hebbian synapses: biophysical mechanisms and algorithms *Ann. Rev. Neurosci.* **13** 475–511

[13] Buhmann J, Divko R and Schulten K 1989 Associative memory with high information content *Phys. Rev.* A **39** 2689–92

[14] Cohen M A and Grossberg S 1983 Absolute stability of global pattern formation and parallel memory storage by competitive neural networks *IEEE Trans. Systems* **SMC-13** 815–26

[15] Collingridge G L and Singer W 1990 Excitatory amino acid receptors and synaptic plasticity *Trends Pharm. Sci.* **11** 290–6

[16] Cotman C W, Monaghan D T and Ganong A H 1988 Excitatory amino acid neurotransmission: NMDA receptors and Hebb-type synaptic plasticity *Ann. Rev. Neurosci.* **11** 61–80

[17] Derrida B, Gardner E and Zippelius A 1987 An exactly solvable asymmetric neural network model *Europhys Lett.* **4** 167–73

[18] Domany E, Kinzel W and Meir R 1989 Layered neural networks *J. Phys. A: Math. Gen.* **22** 2081–102

[19] Evans M R 1989 Random dilution in a neural network for biased patterns *J. Phys. A: Math. Gen.* **22** 2103–18

[20] Fregnac Y, Smith D and Friedlander M J F 1990 Postsynaptic membrane potential regulates synaptic potentiation and depression in visual cortical neurons *Soc. Neurosci. Abs.* **16** 798

[21] Gardner E 1988 The space of interactions in neural network models *J. Phys. A: Math. Gen.* **21** 257–70

[22] Gardner-Medwin A R 1976 The recall of events through the learning of associations between their parts *Proc. R Soc.* B **194** 375–402
[23] Gray C M and Singer W 1989 Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex *Proc. Natl. Acad. Sci. USA* **86** 1698–702
[24] Gustafsson B and Wigstrom H 1981 Shape of frequency-current curves in CA1 pyramidal cells in the hippocampus *Brain Res.* **223** 417–21
[25] Hopfield J J 1982 Neural networks and physical systems with emergent collective computational abilities *Proc. Natl. Acad. Sci. USA* **79** 2554–8
[26] Ishizuka N, Weber J and Amaral D G 1990 Organization of intrahippocampal projections originating from CA3 pyramidal cells in the rat *J. Comp Neurol.* **295** 580–623
[27] Kohonen T 1977 *Associative Memory* (Berlin: Springer)
[28] Kree R and Zippelius A 1990 Asymmetrically diluted neural networks *Preprint* (Göttingen)
[29] Lanthorn T, Storm J and Andersen P 1984 Current-to-frequency transduction in CA1 hippocampal pyramidal cells slow prepotentials dominate the primary range firing *Exp. Brain Res.* **53** 431–43
[30] Little W A 1974 The existence of persistent states in the brain *Math. Biosci.* **19** 101–20
[31] Marr D 1970 A theory for cerebral neocortex *Proc. R Soc.* B **176** 161–234
[32] Marr D 1971 Simple memory: a theory for archicortex *Phil. Trans. R. Soc.* B **262** 24–81
[33] Mason A and Larkman A 1990 Correlations between morphology and electrophysiology of pyramidal neurones in slices of rat visual cortex. II Electrophysiology *J. Neurosci.* **10** 1415–28
[34] McCullough W S and Pitts W 1943 A logical calculus of the ideas immanent in nervous activity *Bull. Math. Biophys.* **5** 115–37
[35] McNaughton B L and Nadel L 1990 Hebb–Marr networks and the neurobiological representation of action in space *Neuroscience and Connectionist Theory* ed M A Gluck and D E Rumelhart (Hillsdale, NJ Erlbaum)
[36] McNaughton B L, Barnes C A and Anderson P 1981 Synaptic efficacy and EPSP summation in granule cells of rat fascia dentata studied in vitro *J Neurophysiol.* **46** 952–66
[37] Peters A and Jones E G (ed) 1984 *Cerebral Cortex* (New York: Plenum)
[38] Rolls E T 1989 Functions of neuronal networks in the hippocampus and neocortex in memory *Neural Models of Plasticity* ed J H Byrne and W O Berry (San Diego: Academic) pp 240–65
[39] Rolls E T 1989 The representation and storage of information in neuronal networks in the primate cerebral cortex and hippocampus *The Computing Neuron* ed R Durbin, C Miall and G Mitchison (Wokingham: Addison-Wesley) 125–59
[40] Rolls E T 1990 Functions of the primate hippocampus in spatial processing and memory *Neurobiology of Comparative Cognition* ed D S Olton and R P Kesner (Hillsdale, NJ· Erlbaum) 339–62
[41] Rolls E T, Miyashita Y, Cahusac P M B, Kesner R P, Niki H, Feigenbaum J D and Bach L 1989 Hippocampal neurons in the monkey with activity related to the place in which a stimulus is shown *J Neurosci.* **9** 1835–45
[42] Rolls E T and Treves A 1990 The relative advantages of sparse versus distributed encoding for associative neuronal networks in the brain *Network* **1** 407–21
[43] Sejnowski T 1977 Storing covariance with nonlinearly interacting neurons *J. Math. Biol.* **4** 303–21
[44] Shannon C E and Weaver W 1949 *The Mathematical Theory of Communication* (Urbana, IL. University of Illinois Press)
[45] Squire L R, Shinamura A P and Amaral D 1989 Memory and the hippocampus *Neural Models of Plasticity* ed J H Byrne and W O Berry (San Diego. Academic) 208–39
[46] Tsodyks M V and Feigel'man M V 1988 The enhanced storage capacity in neural networks with low activity level *Europhys. Lett.* **6** 101–05
[47] Treves A 1990 Graded-response neurons and information encodings in autoassociative memories *Phys. Rev* A **42** 2418–30
[48] Treves A 1991 Dilution and sparse coding in threshold-linear nets *J. Phys. A: Math. Gen.* **24** 327–35
[49] Treves A and Rolls E T 1992 Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network *Hippocampus* **2** in press
[50] Trotter Y, Thorpe S J, Celebrini S, Pouget A and Imbert M 1989 Processing of orientation in V1 of the awake monkey *Soc. Neurosci. Abs.* **15** 1056
[51] Willshaw D J, Buneman O P and Longuet-Higgins H C 1969 Non-holographic associative memory *Nature* **222** 960–2
[52] Willshaw D J and Dayan P 1990 Optimal plasticity from matrix memories: what goes up must come down *Neural Computation* **2** 85–93