Dilution and sparse coding in threshold-linear nets

1991 J. Phys. A: Math. Gen. 24 327

(http://iopscience.iop.org/0305-4470/24/1/038)

J. Phys. A: Math. Gen. 24 (1991) 327–335. Printed in the UK

Dilution and sparse coding in threshold-linear nets

Alessandro Treves

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

Received 3 July 1990

Abstract. The storage capacity of an autoassociative memory with extremely diluted connectivity and with threshold-linear elementary units is studied in its dependence on the graded structure and on the sparseness of the coding scheme, and on the form of the learning rule used. As the coding becomes sparse, more patterns can be stored, and the difference in capacity (measured for a given number of modifiable synapses per unit) between fully connected and highly diluted systems vanishes. Graded (non-binary) codings, especially when used with learning rules nonlinear in their post-synaptic factor, further increase the number of patterns that can be stored by making their retrieved representation even sparser.

The ability of neurons to produce graded responses in the form of continuously variable firing rates might be exploited in the processing of information in the brain. In cases, for example, in which a complex signal is coded in the form of short-time averaged firing rates of a bundle of axonal fibres, quite clearly the amount of information that can be transferred increases as the rates are allowed to take more values. Studying the implications of this potentially advantageous feature requires, however, an understanding of the factors limiting the resolution available in the relevant systems. This is beyond current knowledge. A specific framework in which the question may be tractable is that of associative memories, when conditions are such that the dominant factor affecting the resolution is the effective noise induced by extensive memory loading. In particular it is possible to study how the storage capacity of an autoassociative network of graded-response neurons varies with the code used to store information.

Threshold-linear formal neurons have been proposed for this purpose [1], and the storage capacity of a fully connected model network with a covariance learning rule has been studied [2]. The results are extended here to the case of a network with extremely diluted connectivity, and with a more general learning rule.

Consider an autoassociative net of N neurons, in which each cell receives on average inputs from the axon collaterals of C others, and suppose C is very large (in fact, the limit $C \to \infty$ shall be assumed) but fixed, for example determined by constraints on the physical size of a neuron. Different biological (or artificial) systems might be modelled by giving different values for N, e.g. perhaps $N \simeq 20C$ for the CA3 region of the Hippocampus [3]. What are the effects of varying the size of the net? Does a larger net perform better, in any sense? Or, on the contrary, is a more compact connectivity necessary to harness beneficial feedback effects? A partial answer may come from a comparison of two limit cases. If $N = C + 1$, as in the case previously studied [2], feedback loops dominate the dynamical evolution. If $N \simeq \exp(C)$, i.e. the highly diluted case [4], the net is effectively a feedforward system. Formal models predict somewhat different behaviours in the two situations, in particular as concerns capacity measures. The remarkable fact has been noted [5], however, that systems of binary units, endowed with a covariance learning rule, in the limit of very sparse coding can be described by the same equations both in the fully connected and in the highly diluted cases. If this result applies also in a range of moderately sparse coding, and carries over to other types of elementary units and more general coding schemes, it would suggest that the presence or absence of feedback may not be a relevant factor in the performance of certain biologically interesting networks, at least as measured by their capacity. The role of feedback in determining other features, such as the basins of attraction of retrieval states, which appear to have a more delicate dependence on the details of the dynamics, may be less meaningfully investigated with simplified formal models.

The model and the notation used are the same as in [2]. Positive variables V_i, $i = 1, \ldots, N$ denote the short-time averaged firing rate of N formal units representing cortical pyramidal cells. The synaptic connections between these units are taken to encode information about p patterns of firing activity. The activity of unit i while pattern $\mu = 1, \ldots, p$ is being learned is denoted as η_i^μ. For the purpose of evaluating storage capacities, the η are supposed to be drawn at random, independently for each μ and i, from a common probability distribution P_η. The average value of η over P_η is written a.

Inputs affecting the cells are integrated in the 'membrane potentials' h_i, which are written [2]

$$h_i = \sum_{j(\neq i)} J_{ij}^c V_j + \sum_\mu s^\mu \frac{\eta_i^\mu}{a} + b\left(\sum_j \frac{V_j}{N}\right) . \tag{1}$$

J_{ij}^c stands for the change in the efficacy of the synaptic connection between cells i and j after encoding the patterns, and the first term in the right-hand side of the above equation represents the components of post-synaptic potentials from other units due to these modifications in the efficacies. The second term in h_i represents external inputs tending to elicit some of the encoded patterns, each with relative strength s^μ. A variety of other inputs is lumped together in the last term, including uniform (not pattern-specific) external stimulation, interactions mediated by inhibitory interneurons, and the components of direct exchanges due to baseline synaptic efficacies, as they were before encoding the patterns. The approximation is made of considering this term as dependent only on the average activity of the N units (and on external conditions). An important feature of the model [2] is that although the b term may have a complicated form, it does not include any information about the encoded patterns. As a result, it contributes in setting the overall scale of the network response, but it does not affect its storage capacity.

In the version of the model considered here, the connectivity through modifiable synapses is sparse and asymmetric. Each efficacy change J_{ij}^c is written

$$J_{ij}^c = c_{ij} \frac{1}{C} \sum_{\mu=1}^{p} F(\eta_i^\mu)(\eta_j^\mu/a - 1) \tag{2}$$

where the factor c_{ij}, which specifies whether the modifiable synapse is present or not, is drawn at random from the distribution [4]

$$P_c(c) = \frac{C}{N-1}\delta(c-1) + [1 - C/(N-1)]\delta(c). \tag{3}$$

The modification due to each pattern is written as the product of a pre-synaptic factor, proportional to the deviation of the activity from its average, times a post-synaptic factor which is just a generic function $F(\eta)$ of the post-synaptic activity. In particular, to compare with the fully connected symmetric network of [2], the case considered first will be $F(\eta) = \eta/a - 1$, so that the synaptic modification takes the form of a covariance rule [6], as studied in most symmetric models [7–9].

The dynamics is modelled by assuming that each V_i is updated at random intervals, with a fixed probability per unit time. The updated state is determined by the potential h_i according to a threshold-linear transfer function:

$$V(t) = \begin{cases} g(h(t) - T_{\rm hr}) & h(t) > T_{\rm hr} \\ 0 & h(t) < T_{\rm hr} \end{cases} \tag{4}$$

where g is a gain parameter and $T_{\rm hr}$ a threshold.

The microscopic evolution of the activities depends on the details of the initial conditions, on the updating order and on the quenched assignments $\{c_{ij}\}$ and $\{\eta_i^\mu\}$. Averaging over these factors and over single units one can describe the state of the system with macroscopic quantities such as the correlations with the encoded patterns

$$\hat{x}^\mu(t) = \frac{1}{N}\sum_{i=1}^{N}(\eta_i^\mu/a - 1)\langle V_i(t)\rangle \tag{5}$$

the overall mean activity

$$x(t) = \frac{1}{N}\sum_{i=1}^{N}\langle V_i(t)\rangle \tag{6}$$

and the mean square activity

$$y(t) = \frac{1}{N}\sum_{i=1}^{N}\langle V_i^2(t)\rangle. \tag{7}$$

While the equations describing the evolution of the above quantities are in general quite complicated, only the simple situation is considered here, in which (i) the thermodynamic limit with high dilution is assumed, i.e. $C \to \infty$, $C/N \to 0$ [13]; (ii) memory loading is extensive, $p \to \infty$ with $\alpha \equiv (p-1)/C$ finite; (iii) a single pattern, say $\mu = 1$, has non-zero correlation (the system is stable with respect to the growth of correlations with other patterns if $s^\mu = 0$ for $\mu > 1$); and (iv) the evolution of $\hat{x}^\mu(t)$, $x(t)$ and $y(t)$ has reached a fixed point (which does not necessarily imply a fixed point for the $\{V_i(t)\}$). Then the average potential can be split, as with binary neurons [5], into a signal term, dependent on the encoded activity value η^1

and with negligible variance, and a noise term due to uncorrelated fluctuations in the correlations with all other patterns. The variance of the noise term is

$$(T_0\rho)^2 = \alpha(c_F + a_F^2)T_0 y \tag{8}$$

where T_0 denotes the variance of the pre-synaptic factor in (2) under P_η [2] (i.e. $T_0 \equiv \int P_\eta(\eta)(\eta/a-1)^2 \mathrm{d}\eta$), and a_F, c_F are, respectively, the average and variance of $F(\eta)$ under P_η (so that $c_F + a_F^2 \equiv \int P_\eta(\eta)F^2(\eta)\mathrm{d}\eta$).

The fixed-point equations are

$$\begin{aligned} x &= g\left\langle\!\!\left\langle \int_{h>T_{\mathrm{hr}}} \mathrm{D}z(h - T_{\mathrm{hr}}) \right\rangle\!\!\right\rangle \\ \hat{x}^1 &= g\left\langle\!\!\left\langle (\eta^1/a - 1)\int_{h>T_{\mathrm{hr}}} \mathrm{D}z(h - T_{\mathrm{hr}}) \right\rangle\!\!\right\rangle \\ y &= g^2\left\langle\!\!\left\langle \int_{h>T_{\mathrm{hr}}} \mathrm{D}z(h - T_{hr})^2 \right\rangle\!\!\right\rangle \end{aligned} \tag{9}$$

with

$$\begin{aligned} h &= F(\eta^1)\hat{x}^1 + (\eta^1/a)s^1 + b(x) - T_0\rho z \\ \mathrm{D}z &= \frac{\mathrm{d}z}{\sqrt{2\pi}}e^{-z^2/2} \end{aligned}$$

and $\langle\langle\ldots\rangle\rangle$ denotes averages over the distribution $P_\eta(\eta^1)$.

One can compare these equations with the saddle-point equations obtained for the fully connected model in the corresponding limit [2]. Two features do not appear in the highly diluted case: the renormalization of the gain parameter and the dependence of the variance of the noise on the 'degree of freezing' i.e. on the correlations between the various configurations concurring in the attractor state of the fully connected network. The second simplification occurs also in highly diluted networks of binary-threshold units [10], while the first is present, of course, only with analogue units (cf [11,12]).

The range of capacity values α for which a fixed point exists that corresponds to the retrieval of pattern 1 (i.e., with $\hat{x}^1 > 0$) is independent of the form of $b(x)$. This is a feature of the threshold-linear neuron representation, as becomes clear following the analysis used for the fully connected model [2]. Let the ratio of the pattern-specific external stimulus s^1 to the correlation measuring the pattern-specific collective response be denoted as

$$\delta = s^1/\hat{x}^1 \tag{10}$$

and introduce the two signal-to-noise ratios:

$$\begin{aligned} w &= (b(x) - \hat{x}^1 - T_{\mathrm{hr}})/T_0\rho \qquad && \text{uniform} \\ v &= (\hat{x}^1 + s^1)/T_0\rho && \text{specific.} \end{aligned} \tag{11}$$

For any given distribution P_η and function $F(\eta)$, one can compute the averages over z and P_η:

$$\begin{aligned} A_2(w,v) &= \frac{1}{vT_0}\left\langle\!\!\left\langle (\eta/a-1)\int^{+} \mathrm{D}z\left(w + \frac{v}{1+\delta}[1 + (\eta/a)\delta + F(\eta)] - z\right)\right\rangle\!\!\right\rangle \\ A_3(w,v) &= \left\langle\!\!\left\langle \int^{+} \mathrm{D}z\left(w + \frac{v}{1+\delta}[1 + (\eta/a)\delta + F(\eta)] - z\right)^2\right\rangle\!\!\right\rangle \\ A_4(w,v) &= \frac{1}{v}\left\langle\!\!\left\langle \int^{+} \mathrm{D}z\left(w + \frac{v}{1+\delta}[1 + (\eta/a)\delta + F(\eta)] - z\right)\right\rangle\!\!\right\rangle \end{aligned} \tag{12}$$

where the superscript + indicates that the z-average has to be carried out only in the range where $w + [v/(1+\delta)][1 + \delta(\eta/a) + F(\eta)] - z > 0$.

In terms of the A, the fixed-point equations reduce to

$$\begin{aligned} x/gT_0 &= A_4 v\rho \\ 1/gT_0 &= A_2(1+\delta) \\ (1/gT_0)^2 &= \alpha A_3(c_F + a_F^2)/T_0. \end{aligned} \tag{13}$$

The last two equations are the important ones: they determine the pair (w, v) of the retrieval solution. The first of equations (13), together with the definitions of w and v, just sets the absolute scale for the response, measured in $x, \hat{x}^1, y$ and ρ. For a given value of δ and choice of $F(\eta)$ and P_η, eliminating g yields an equation that gives the maximum capacity α_c: the solutions of

$$E_1(w, v) \equiv (1+\delta)^2 A_2^2 - \alpha A_3 \frac{(c_F + a_F^2)}{T_0} = 0 \tag{14}$$

are on a closed line on the w, v plane, which shrinks as α grows and disappears at α_c. For $\alpha < \alpha_c$ a solution to the last two of equation (13) exists if the gain g falls in a certain intermediate range. The discontinuous disappearance of the solution implies, *inter alia*, that a second-order transition with α [4] is not a general feature of networks with very diluted connectivity, but rather is peculiar to highly symmetric 'spin' models.

The way specific choices of $F(\eta)$ and P_η affect the maximum capacity α_c is now studied. The dependence on δ was found, for the fully connected model [2], to reduce to a rather uninteresting increase of α_c with increasing (small) δ, decoupled from the dependence on other factors. Therefore, in the following, only the case $\delta = 0$ will be considered, corresponding to retrieval elicited only by the initial (as opposed to persistent) stimulation.

While one can choose any arbitrary form for the statistical distribution of the encoded patterns, possibly inspired by some real neurobiological data, the focus here will be on studying the effects on the capacity of the network that can arise when the use of graded response units allows one to retrieve graded patterns. These effects can be explored, at the most basic level, by comparing a binary with a ternary distribution. The storage capacity, however, turns out to depend most strongly on the sparseness of the coding scheme, rather than on the structure of its statistical distribution. To quantify that dependence by introducing a sparse coding parameter, the traditional, if somewhat confusing, notation will be adopted here [9]. Setting the average of η over P_η to a fixes the scale of the positive variable η, which only affects (just as the term $b(x)$) the overall scale of the response. One can turn a into a sparse coding parameter by also requiring that

$$\langle\langle \eta^2 \rangle\rangle = a \tag{15}$$

which implies $T_0 = (1-a)/a$. The binary distribution is then

$$P_\eta(\eta) = (1-a)\delta(\eta) + a\delta(\eta - 1) \tag{16}$$

with $\delta(x)$ the Dirac δ function. A specific ternary distribution is

$$P_\eta(\eta) = (1 - \tfrac{4}{3}a)\delta(\eta) + a\delta(\eta - \tfrac{1}{2}) + \tfrac{1}{3}a\delta(\eta - \tfrac{3}{2}). \tag{17}$$

This particular choice examplifies well the effects seen with any ternary distribution. This has been shown [2], for the fully connected network, by exploring the space of all possible ternary distributions that satisfy (15) (and thus allow a quantitative comparison with the binary code of equal sparseness, equation (16)).

Turning to the choice of the post-synaptic factor $F(\eta)$, for a straightforward comparison of the storage capacity of the highly diluted model with that of the fully connected model it is useful to consider first the 'covariance' rule

$$F(\eta) = \frac{\eta - a}{a} \tag{18}$$

in which case $a_F = 0$ and $c_F = T_0$. In this case asymmetry results only from the diluted connectivity, and not from a different form of the pre- and post-synaptic factors. In figure 1 the resulting α_c is plotted as a function of a for the two P_η considered, together with the results for the fully connected model. The number of patterns that can be stored (for a given number C of modifiable synapses per neuron) is higher with the ternary distribution than with the binary distribution, and in both cases it increases roughly as $[a \ln(1/a)]^{-1}$ as the coding becomes sparser. The interesting feature of figure 1 is that as the coding becomes sparser the differences between the fully connected and the highly diluted models disappear, as with binary models [5]. The decrease in the capacity due to the noisy correlations characterizing fully connected feedback networks is marked only when the coding is not sparse at all.

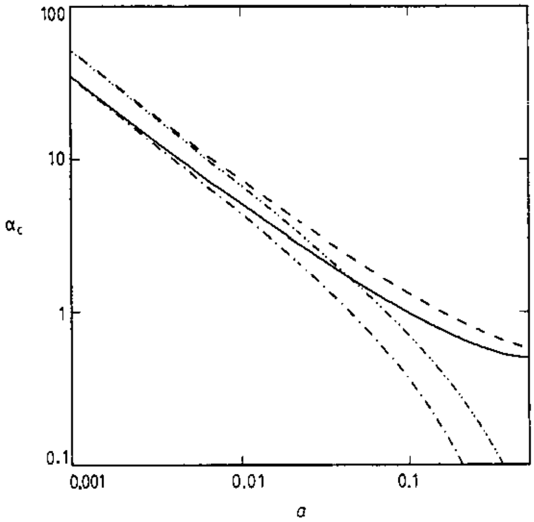

Figure 1. Storage capacity, as measured by α_c, against the sparse coding parameter a for binary and ternary pattern distributions (16), (17) and for a covariance learning rule, (18). Highly diluted model: full curve (binary distribution) and broken curve (ternary); fully connected model: chain curve (binary) and double-dot chain curve (ternary).

The same feature appears when calculating the amount of information that can be stored and retrieved by the network [2]. This is lower for ternary distributions (see figure 2), as a poorer retrieval quality more than compensates for the higher number of patterns that can be stored, and for the larger amount of information each one of

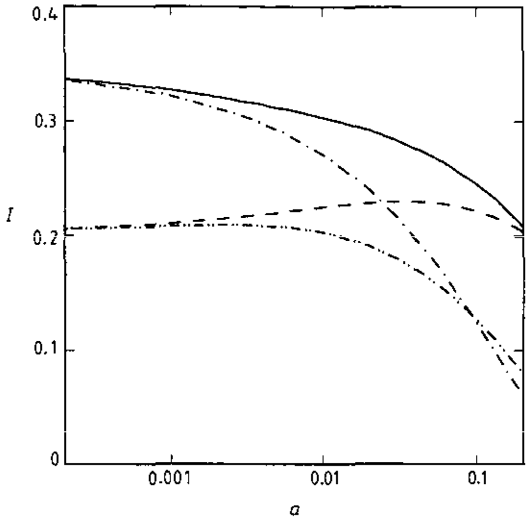

Figure 2. Information capacity I against the sparse coding parameter a, for the same cases and with the same line codes as in figure 1.

them carries [2]. Again, differences between the fully connected and highly diluted models are important only when the coding is not sparse.

The fact that more ternary than binary patterns can be retrieved for equal values of the parameter a measuring the sparseness of the encoded representation can be traced to the fact that the *retrieved* patterns are sparser in the ternary case. To make that clear, it is useful to introduce a second parameter, measuring the sparseness of the representation of the retrieved information. It can be defined in analogy with a as

$$a_r = \frac{\langle V\rangle^2}{\langle V^2\rangle} \tag{19}$$

where the averages are over the pattern distribution and over the noise. Figure 3 shows that when the capacity is plotted against a_r, it is lower in the ternary case than in the binary case for essentially all values of a_r (in the fully connected case, for all values of a_r).

Are the above features peculiar to the choice of the 'covariance' rule, equation (18)? The limit of high dilution allows one to explore alternative choices for $F(\eta)$ [13]. It is interesting to use this freedom to try to model some current neurobiological hypotheses on the mechanisms of synaptic plasticity. In particular, mechanisms of long-term potentiation (LTP) based on the activation of NMDA-receptors seem to occur only when the post-synaptic membrane is very depolarized [14]. This feature might be modelled by setting a threshold for synaptic modification (in learning a pattern, for example) which is higher than the threshold above which the post-synaptic 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 post-synaptic rate, and the form of $F(\eta)$ between the two values is irrelevant. New effects could be seen, instead, with a ternary pattern. To be specific, having in mind the distributions of (16) and (17) let

$$F(\eta) = \frac{2\eta^2 - \eta}{a} \tag{20}$$

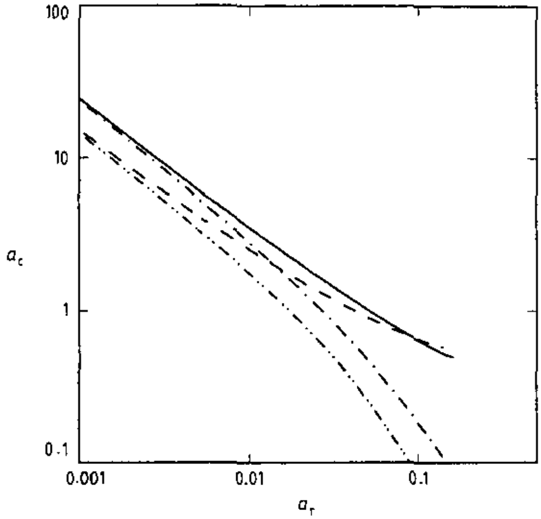

Figure 3. Storage capacity α_c against the parameter a_r measuring the sparseness of the retrieved representation, for the same cases and with the same line codes as in figure 1.

which implies that the modification occurs, both in the binary and in the ternary case, only for the highest post-synaptic firing rate. In the ternary case of (17) synapses onto neurons that fire at the intermediate rate while encoding the pattern are not affected. Note that

$$c_F + a_F^2 = \begin{cases} 1/a & \text{binary} \\ 3/a & \text{ternary.} \end{cases} \tag{21}$$

Figure 4 compares the number of patterns that can be stored with this model 'NMDA' rule to that of the covariance rule, for binary and ternary patterns. It is seen that for binary patterns the only differences occur when the coding is not sparse, in which case the capacity is sensitive to whether $a_F = 0$ or not, whereas for sparse coding c_F dominates a_F anyway. For ternary patterns α_c is higher in the case of the NMDA rule, for almost any a. This was expected, as this rule effectively enhances the *sparseness of the retrieved representation.*

To summarize, the capacity of an autoassociative network of threshold-linear units, as measured by the number of patterns that can be retrieved (for a given number of modifiable synapses per unit), depends very strongly on the sparseness of the coding scheme. Moreover (i) with a covariance learning rule, the difference between fully connected and highly diluted networks vanishes in the biologically meaningful region of sparse codings; (ii) more ternary than binary patterns can be stored for equal values of the sparse coding parameter, as ternary patterns enhance the sparseness of the retrieved representation; (iii) learning rules nonlinear in the post-synaptic factor may also increase the storage capacity in conjunction with non-binary pattern distributions, as they also increase the sparseness of the retrieved representation.

Another measure of performance, the amount of information that can be stored and retrieved, has a different dependence on the sparseness and on the structure of the coding (in particular, it decreases as ternary rather than binary patterns are used), but the difference with fully connected networks again vanishes for sparse codings. Finally, it should be noted that an essential ingredient of the analysis yielding the above results

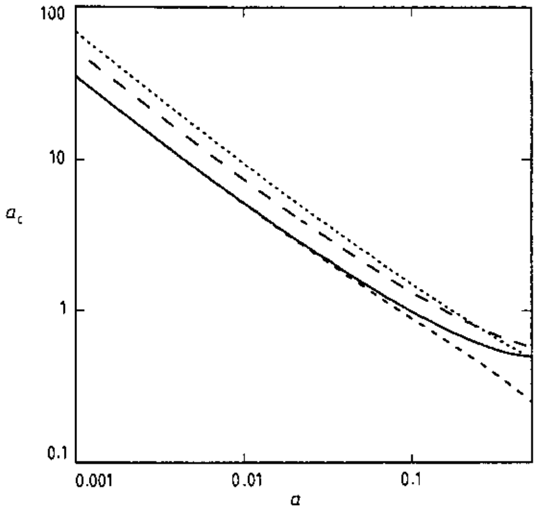

Figure 4. Storage capacity α_c against the sparse coding parameter a, in the highly diluted case, for the covariance and NMDA learning rules (18), (20). Covariance rule, same line codes as in figure 1: full curve (binary distribution) and long-dash broken curve (ternary); NMDA rule: short-dash broken curve (binary) and dotted curve (ternary).

is the pre-synaptic factor used in the learning rule. The form chosen implies, from a neurobiological perspective, a careful balance between effects of long-term potentiation and long-term depression, tuned to the average activity of the firing patterns to be stored. The implications of this and other aspects of biological significance for the study of neuronal networks in the brain will be discussed elsewhere.

Acknowledgments

The author wishes to thank E T Rolls and S Mertens for stimulating discussions, and acknowledges support from the MRC and from a BRAIN grant.

References

[1] Treves A 1990 *J. Phys. A: Math. Gen.* **23** 2631
[2] Treves A 1990 *Phys. Rev.* A **42** 2418
[3] Treves A and Rolls E T 1990 *Neural Networks (Proc. XI Sitges Conf.)* (Berlin: Springer) in press
[4] Derrida B, Gardner E and Zippelius A 1967 *Europhys. Lett.* **4** 167
[5] Evans M R 1989 *J. Phys. A: Math. Gen.* **22** 2103
[6] Sejnowski T 1977 *J. Math. Biol.* **4** 303
[7] Hopfield J J 1982 *Proc. Natl Acad. Sci.* **79** 2554
[8] Amit D J 1989 *Modelling Brain Function* (New York: Cambridge University Press)
[9] Buhmann J, Divko R and Schulten K 1989 *Phys. Rev.* A **39** 2689
[10] Domany E, Kinzel R and Meir R 1989 *J. Phys. A: Math. Gen.* **22** 2081
[11] Kuhn R 1990 *Neural Networks (Proc XI Sitges Conf.)* (Berlin: Springer) in press
[12] Reiger J 1990 *Neural Networks (Proc XI Sitges Conf.)* (Berlin: Springer) in press
[13] Gardner E, Mertens S and Zippelius A 1989 *J. Phys. A: Math. Gen.* **22** 2009
[14] Cotman C W, Monaghan D T and Ganong A H 1988 *Ann. Rev. Neurosci.* **11** 61