POSSIBLE STRATEGIES FOR USING SLEEP TO IMPROVE EPISODIC MEMORY IN THE FACE OF OVERLAP

A.R. Gardner-Medwin. Department of Physiology, University College London, Gower Street, London WC1E 6BT, UK & The Physiological Laboratory, Cambridge CB2 3EG, UK

In Theory and Applications of Neural Networks (1992), Ed. J.G. Taylor & C.L.T. Mannion. Springer-Verlag. London, pp. 129-138

It is not known what biological benefits may derive from information handling during sleep. Several facts about sleep suggest that information handling does take place, in different fashions in the two principal phases of sleep ('Slow Wave' and 'Paradoxical' or 'Rapid Eye Movement' (REM) Sleep). This paper seeks to identify benefits that could arise from such 'off-line' information processing, in relation to one of the simpler, but neurally potentially important, forms of memory: auto-association. One of the interesting outcomes is that the constraints of the theory lead naturally to an algorithm that requires two stages for its implementation, resembling in some respects the two phases of sleep [1].

In both phases of sleep the nervous system is cut off from its sensory and motor systems. Bizarre forms of recall and lines of thought take place, often driven substantially by association. Memory is poor. Activity and internal experiences during sleep can be remembered in some detail, especially on immediate rehearsal after awakening from REM sleep. Nevertheless, memory performance is in no way comparable to what it would be like for similarly unusual, vivid and often emotionally charged experiences in waking life. Physiological evidence indicates synchronous fluctuations of threshold in cortical neurons during Slow Wave Sleep, and activity in the visual pathway during Paradoxical Sleep that arises from the brainstem rather than from the eyes, so-called 'PGO' waves.

Theoretical approaches to the handling of information by neural networks may be able to prompt testable suggestions about benefits from information processing during sleep. Several suggestions already exist in the literature, either arising within specific theoretical models [e.g. 2] or within a more general theoretical framework [3,4]. The approach adopted here is to examine one of the simpler types of memory (auto-association) as rigorously and quantitatively as possible, and to see how algorithms can be applied to relax the constraints that normally would limit performance. Such processing might or might not require the isolation and the poor memory registration that are characteristic of sleep. As it turns out, not only are both these features necessary for some of the major benefits, but also a separation of the algorithm into two interdependent phases seems to be required, with possibly significant similarities to the two phases of sleep. This is consistent with the thesis of Giuditta [5], who argues that hypotheses about sleep function should take account of the interdependence of two sleep phases.

Auto-Association

Auto-association is the development of strong excitatory interactions within a population of nodes, between nodes that have been active together. It permits previously active patterns to be re-elicited on presentation of a subset of the active nodes or a set resembling the previously active set. It is not obvious that it should always be desirable to 'complete' previously experienced patterns in a neural system in this way. Early in a sensory pathway it is probably not desirable to do so. At higher perceptual levels completion seems to be a part of gestalt perception, in relation to which auto-association was probably first proposed as a neural mechanism [6]. Triggered evocation of episodic memory is commonplace in human memory, where it is usually beneficial but may sometimes cause trouble (as for example with victims of horrific experiences). Performance in this area of 'content-addressable memory' is one of the skills at which the human brain seems to excel.

Auto-association can be implemented with a Hebbian modification rule [7, 8], rules involving decreases as well as increases of weights [9, 10] or rules involving both positive and negative activity parameters and weights [11]. Only the first and simplest rule is considered in this article. The same issues in relation to overlap and confusion will arise, at least qualitatively, with any strategy for episodic recall. Therefore it is at least plausible that the principles that arise in the present analysis may have application in more complex settings with other primary algorithms.

In some respects auto-association can be treated analytically as a special case of 'cross-association', in which connections are strengthened between active cells in two separate populations [7]. Unlike cross-association however, auto-association can be carried out iteratively to improve performance by growing full patterns from a seed [8, 12]. This has been termed 'progressive' recall, requiring careful management of neural thresholds by recurrent inhibition [8]. Auto-associative storage can also be used to provide a measure of the familiarity of a pattern: a form of 'recognition' memory [1].

Auto-association permits storage and retrieval of the content of a pattern, after what may be simply a one-trial learning situation. Episodic memory requires such an algorithm: it is the features that are specific to a particular, possibly unique, experience that are important. The identification of features that may commonly be grouped together within patterns (adaptive recoding), or that correlate with external signals (classification learning), are different forms of memory requiring, by definition, presentations of many related patterns (often many times over). These are equally important forms of learning, but they are not considered here. It is envisaged that auto-associative learning of pattern content may contribute to the plasticity of the nervous system at each of many levels of representation (Fig. 1). The projections for recoding and classifying patterns, for seeding the content of patterns at higher levels, and for 'top-down' influences (hatched arrows in Fig. 1) will be subject to their own plasticity according to separate algorithms.

Auto-associative memory employing the Hebb conjunctive rule [6] (i.e. strengthening that occurs when a synapse is active in a situation in which its activity contributes to the firing of a postsynaptic neuron) is subject to two broad constraints. If M patterns are to be learned on N cells, each comprising activity in a fraction α of the neurons, then to avoid serious saturation effects the relation:

$$\alpha < \approx 1/\sqrt{M} \qquad (1)$$

must hold [7]. It is not necessary for each cell to be connected to every other cell, and in general if the number of connections per cell is R, the relation

$$\alpha R \approx 30 \qquad (2)$$

must hold for a good compromise of performance and efficiency. The quantity αR is the mean number of inputs received by individual cells in an active pattern from the other active cells in the pattern. The critical value of order 30 arises because the mean number of inputs must be large enough to ensure that on the basis of a Poisson distribution the actual number is not only reliably non-zero, but reliably greater than the number of inputs onto a typical spurious cell, which is a Poisson variable typically around $0.5\alpha R$ at saturation. If αR greatly exceeds 30, then there is unnecessary redundancy in the connections and the efficiency (i.e. the information capable of being stored and recovered per synapse) is less than maximum [8].

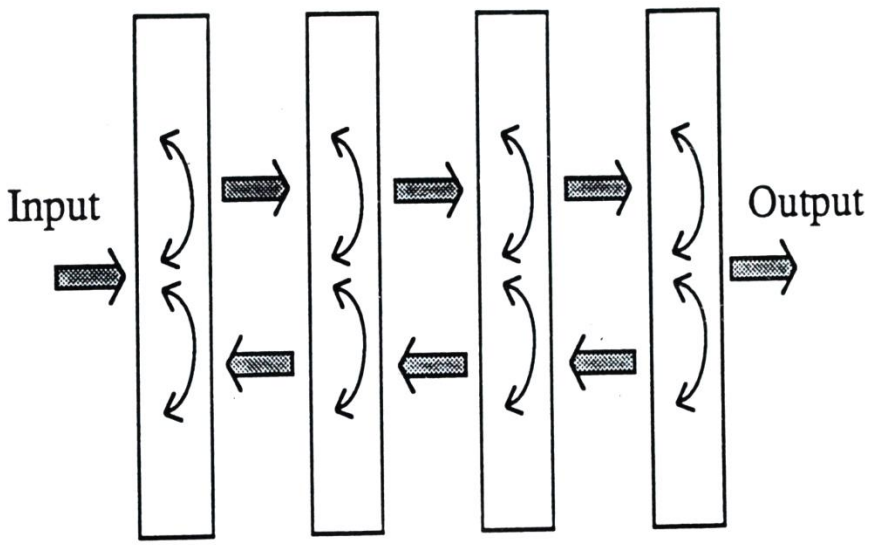

Fig. 1. Schematic representation of the nervous system. Shaded arrows are forward and backward projections reponsible for recoding, pattern classification (feature detection) and for the seeding of learned output patterns. Line arrows are auto-associative connections responsible for completion and correction of the content of patterns experienced at each level of representation.

It follows from Equation (1) that storage of a large number of patterns in these models requires that the patterns be coded into a sparse representation, with a low activity ratio α. Examples of statistics of patterns with different activity ratios and essentially the same information content are given in Table I. In principle it is possible to recode information reversibly from one such form to another, and indeed at early stages in visual processing there are known mechanisms that have the effect of reducing activity ratios (e.g. lateral inhibition, feature detection).

It follows from Equation (2) that in large networks storing patterns with a

substantial information content (>>30 active cells per pattern), the optimal number of connections of any one cell can be substantially less than the number of cells (R<<N). Thus an efficient large auto-associative network in general requires both sparse coding (α<<1) and sparse connectivity (R/N <<1). Both conditions are plausibly realistic in relation to what is known about at least some parts of the cerebral cortex [8].

Table 1. Representation of information at different activity ratios (α). The same amount of information (100 ±5 bits) is required to specify each pattern of W active cells from a total of N. The information content is calculated on the basis of two slightly different assumptions: I_1 is for uniform independent probabilities α, giving rise to W as an expectation value: $I_1=-N(\alpha\log_2(\alpha)-(1-\alpha)\log_2(1-\alpha))$. I_2 is calculated for fixed (integer) values of W (= αN) giving rise to probabilities of activation (α) that are not strictly independent: $I_2=\log_2(N!/(W!(N-W)!))$. I_1 and I_2 differ by at most 4% over the indicated range.

No. Cells	No. Active	Activity Ratio	Information Content	
N	W	α	I_1	I_2
100	50	0.5	100	96
200	22	0.11	100	97
400	17	0.0425	101	98
1000	13	0.013	100	97
2000	11	0.0055	98	95
4000	10	0.0025	101	98
10000	9	0.0009	104	101

The overlap problem

Auto-association provides an algorithm for one-trial storage of patterns. The statistical constraints and handling techniques for sparsely connected nets have been analysed [8] and simulated [1] for simple situations. One of the fundamental limitations on performance arises through overlap between stored patterns (i.e the existence of active elements that are common to two or more patterns). This is illustrated in Fig. 2. The result of overlap is that recall of a pattern P_1 readily leads to activation of elements that do not belong to P_1, but are part of an overlapping pattern P_2. These are called 'intrusion' errors. The recalled pattern can readily become some sort of hybrid or intermediate between P_1 and P_2, even though the disparate elements of P_1 and P_2 may never have been experienced together.

In other contexts, the behaviour of neural networks in response to overlapping patterns can be an advantage. In classification tasks, a closely related phenomenon is that of generalisation, whereby learned responses can transfer to patterns that have never been experienced, but are similar to (i.e. overlap with) ones experienced in the training set.

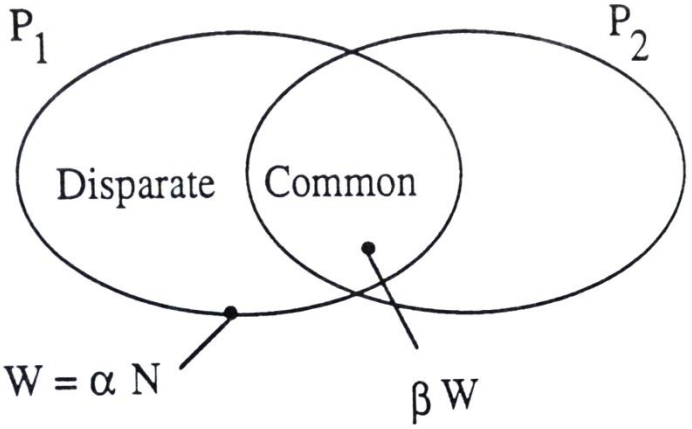

Fig.2. Patterns P_1 and P_2 each consist of W active cells with an overlap fraction ß. The cells active in P_1 and P_2 fall into categories called common and disparate, with the latter comprising cells that are specific to P_1 and specific to P_2.

When the task is to recall and re-evoke an experienced pattern, the consequences of overlap are often undesirable. Note that this is an observation about the way in which recall memory is used, not about the fundamental statistical issues underlying what one may variously call intrusion errors or generalisation. For example, if one visits the Georgian city of Bath, and then tries to recall details of a particular building in Bath, one may generate a sketch that is useless as a representation of the specific prompted building because of numerous intrusion errors from other buildings. It might nevertheless encapsulate the spirit of Georgian architecture even better than any single building might have done. The problem of overlap arises specifically where recall memory is used in an episodic fashion, to identify details of a specific episode or pattern. To pursue the example, it is useless, and indeed positively confusing, to have picked up a general appreciation of the preferred symmetry of Georgian architecture if one is trying to direct someone how to enter a building that happens to have its entrance on the left.

Given that intrusion errors from overlap are a problem in episodic memory, how can one reduce these errors? There are at least three distinct strategies:

1. It may be possible to use classification and pattern recognition techniques at a higher level of representation to identify probable or improbable combinations of elements in the recall of P_1 or P_2. For example, in the illustration, one might use knowledge of the general constraints on the positioning of building entrances.

2. It may be possible to reduce overlap in the representation. In episodic memory, after initial learning, it is too late to reduce overlap in the primary engram. Two things can happen, however:

i) Current experience may lead to alterations in the representation of future inputs so as to reduce overlap for these inputs. For example, once one has learned to classify the stereotypes of Georgian architecture one may on future occasions be able to represent new buildings in terms of combinations of these stereotypes and particular

departures from them. This is a form of 'feature detection', through which inputs may be represented with minimum use of information channels (minimum entropy coding [13]).

ii) It may be possible to generate recoded versions of the current patterns P_1 and P_2 with less overlap. For example, by surmising (either correctly or incorrectly) that minor differences in building style relate to different periods or different personalities, it may be possible to build up coherent networks of associations containing P_1 and P_2 with less overlap. Such a strategy has both benefits and risks: it may reduce intrusion errors due to overlap between P_1 and P_2 while generating errors due to the specific form of recoding.

3. It may be possible to adjust the relative weights of associations between elements of P_1 and P_2 to reduce the effects of the overlap.

Strategies 2ii and 3 both require that episodic memories be held in a temporary robust form that allows algorithms for reducing effects of overlap to operate. The potential for performing such operations is just one of the benefits that can arise from having temporary as well as long term (LT) memory stores with flexible consolidation processes for transfer from one to the other [1]. Strategy 3 is developed here because it is prima facie the simplest, and need relate only to the current representation and to a single, potentially homogeneous, set of cells.

Strategies for adjusting relative weights to reduce the overlap problem

Within a pair of overlapping patterns there are two distinct categories of cells described as 'common' (c) and 'disparate' (d). Between these cells there are 4 categories of associative connection that contribute to recall performance on P_1 and/or P_2 ($c \rightarrow c$, $c \rightarrow d$, $d \rightarrow c$, $d \rightarrow d$). Each has its own significance in relation to the overlap problem. Before discussing them individually, it is helpful to consider briefly the dynamics of recall in the face of overlap.

If recall is prompted by activation of a seed of active cells specific to P_1, then iterative recall may lead to recruitment of other cells specific to P_1, common cells, incorrect cells specific to P_2 (intrusion errors), and possibly spurious cells present in neither P_1 nor P_2. There are actually two inter-related problems arising from overlap:

1. If we suppose that P_1 has been successfully recalled with near total accuracy, then the mean excitation onto specific P_2 cells (from the common cells) may be nearly as great, with substantial overlap, as the mean excitation onto P_1 cells. The statistical separation of the excitation onto the two categories is poor because of the overlap, and intrusion errors are likely. It is the relative strength of $d \rightarrow d$ and $c \rightarrow d$ connections that is relevant in this situation. Improvements can be made if the $d \rightarrow d$ connections can be strengthened relative to $c \rightarrow d$.

2. The second problem arises earlier in the progressive recall process. If the common cells are relatively numerous within P_1, then both by simple probability and by virtue of the strong interactive support that they provide for each other once activated, common cells will tend to be recruited in early iterations of the recall process. This contributes to the excitation of P_2 cells as well as P_1 and diminishes the statistical weighting in favour of recruitment of P_1 rather than P_2 that existed at the start by virtue

of the specific seed from P_1. To reduce the preference for early recall of common cells it is necessary to increase the strength of d→d connections relative to d→c and c→c connections.

The identified changes of relative connection strengths can be achieved for problem (1) either by increasing d→d weights or by reducing c→d. For problem (2) it is necessary to increase d→d weights or reduce d→c and/or c→c. The adjustment that helps with both aspects of the overlap problem is to increase the strength of d→d connections [1]. An alternative strategy proposed for helping with an analogous problem, that of too ready elicitation of spurious 'parasitic' states due to densely interconnected nodes in a network, is to decrease the strength of c→c connections [3]. In the present context this strategy diminishes the early recruitment of common cells, but fails to help with the problem of eliminating intrusion errors once recall is nearly complete.

To change the weights of a particular category of connections it is necessary somehow to identify these connections within the network. It is possible to identify the c→c connections by activating solely the c cells (which tend to be the most readily activated cells [3]). There is no such simple way of identifying directly the d→d connections in order to strengthen them. A two stage algorithm has been proposed, however, which achieves this automatically after storage of pairs of overlapping patterns [1].

In simulations the algorithm for d→d strengthening has been shown to have both of the desired effects: increase of stability of correctly recalled patterns with fewer intrusion errors, and early recall of the specifically correct cells in preference to the common cells. The average quality of recall elicited from a seed was increased substantially (Fig.3).

Implementation of this algorithm in a neural structure would require two stages in sequence, with a carry over of some form of temporary 'fatigue' within cells that are strongly activated in Stage 1 (which would be largely the 'common' cells) to Stage 2. Such a carry over of fatigue between the two phases of sleep is not known, but has probably never been sought experimentally.

Some of the other conditions necessary for implementation of the algorithm bear quite strong resemblances to known facts about sleep. For example, the first stage requires large threshold swings to activate firstly the hybrid union of cells that are within P_1 or P_2, then the common cells that recieve greatest excitation from this hybrid pattern. It is essential that memory traces should not be laid down for the hybrid patterns experienced at this time, since this would lead to associations between the disparate cells of P_1 and P_2 that would simply compound the problem of intrusion errors. Stage 2 (cf. Paradoxical Sleep) must always follow Stage 1 and requires recall from random seeds subject to the tight threshold control that is characteristic of normal recall (as presumably employed during waking). This ensures that coherent sets of disparate cells are activated together to permit enhanced strengthening of the d→d weights. Though there are striking parallels in these requirements to known facts about sleep, such parallels can be no more than suggestive of a true relation between the proposed optimisation procedure and sleep.

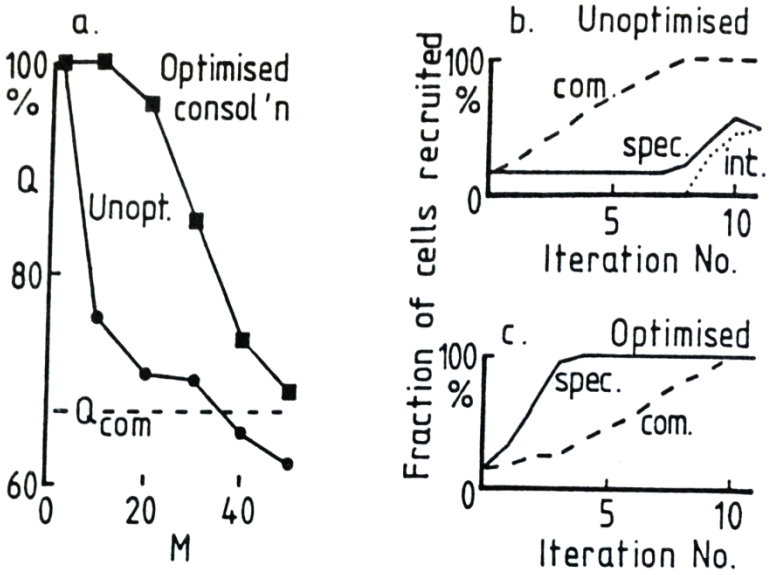

Fig.3. The effect on recall of overlapping patterns, of an algorithm for increasing d→d weights. Patterns were learned in pairs with overlap fraction ß=0.74 (Fig.2). (a) Quality of recall quality with and without the optimising algorithm. Q_{com}=quality corresponding to correct recall of common cells, with chance levels of discrimination between specifically correct cells and intrusion errors. M=number of learned patterns. (b,c) Recruitment sequence for common cells (com), specifically correct cells (spec) and intrusion errors (int) during iterative recall without (b) and with (c) the optimising algorithm. Simulation data from [1].

The advantages derived from specific strengthening of d→d connections can to some extent be analysed analytically rather than by simulation. An expression can be derived for the signal to noise ratio 't' for discrimination between excitation onto specific P_1 and specific P_2 cells once P_1 is correctly recalled. 't' is defined as the ratio of the difference between the mean excitation onto the two types of cells, to the sum of the standard deviations for the excitation onto the two types. A value t=2, for example, permits a threshold to be set between the two levels so as to give approximately 2.5% false positive and false negative rates. For a sparsely connected net with a fraction f of its synapses having been modified, an overlap fraction ß between a pair of patterns, and enhancement of d→d connections by a factor θ, the signal to noise ratio is given by:

$$t = \sqrt{(\alpha R)}\,(1-ß)\,(\theta-f) \,/\, \{\, \sqrt{(ß+\theta^2(1-ß))} + \sqrt{(ß+f(1-ß))}\,\} \qquad (3)$$

For ß=f=0.5, this gives $t=0.13\sqrt{(\alpha R)}$ for θ=1, $t=0.31\sqrt{(\alpha R)}$ for θ=2 and $t=0.71\sqrt{(\alpha R)}$ for $\theta\to\infty$. This analysis assumes that enhancements of d→d connections for an overlapping pair of patterns are made on a background of uniform connection strengths for all the other synapses in the network at which modification conditions have been met. Much of the maximum benefit derived from arbitrarily large increases of d→d weights is obtained with θ=2.

There is a cost associated with increasing θ too far. The heavily weighted synapses affect adversely the performance in other recall situations, where their weight is inappropriate. Fig. 4 shows the results of analysis in which such effects are seen in the signal-noise ratio for different aspects of recall of a paired pattern. The entire experience of the network in this case has consisted of pairs of patterns with overlap fraction ß=0.5. All connections that have at some time in the net's experience been d→d connections (in this case a fraction 0.36 of those that have been modified generally) are taken to have a strength θ relative to the other modified synapses.

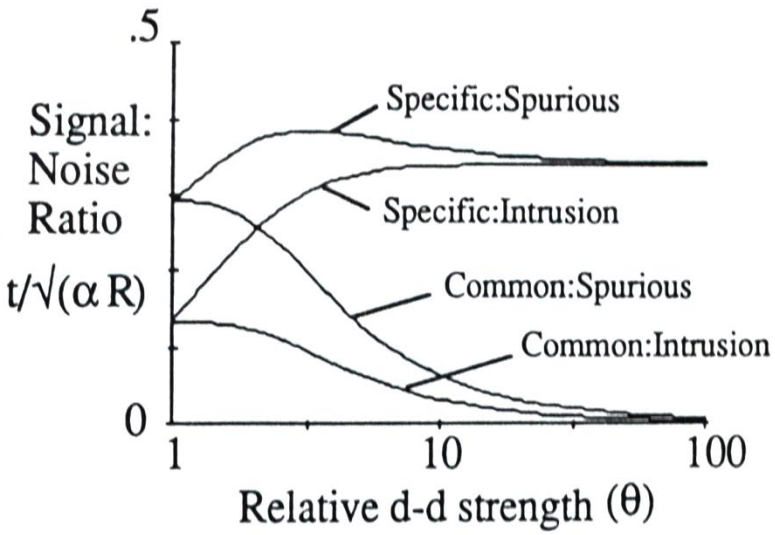

Fig. 4. Effect of enhancement of connections between cells in the disparate parts of overlapping patterns. The ordinate is the signal-noise ratio (in units of $\sqrt{\alpha R}$) for discrimination between each of the two categories of correct cells (disparate and common) and each of the two categories of incorrect cells (intrusion errors from P_2 and spurious cells not part of P_1 or P_2). The abscissa is the factor θ by which d→d connections have been enhanced. Conditions are f=ß=0.5 as described in the text, and it is assumed that the entire experience of the net consists of pairs of patterns with equal overlap ß between pairs.

It can be seen in Fig.4 that the signal-noise ratio for the common cells falls with increasing θ throughout the range. For θ>≈3 it falls also for excitation of the disparate P_1 cells compared to spurious recruits. The signal-noise ratio for recruitment of the specific P_1 cells against intrusion errors rises continuously with θ. A reasonable compromise under these conditions would be achieved with θ=2-3: there are then substantial gains in the rejection of intrusion errors and relatively little loss in the signal-noise ratio for common cells. Precise optimisation would have to depend on the relative costs of errors associated with the different categories of cells.

Conclusion

It is possible to identify strategies for improving performance in auto-association through application of algorithms operating after the initial learning. In this way it may be possible to reduce confusions and intrusion errors resulting from overlap between similar patterns stored in episodic memory. Some of the potential improvements for a specific algorithm have been quantified by both analysis and simulation. The conditions that would be necessary for implementing such an algorithm neuronally are quite constrained and raise unresolved issues in relation to the experimental study of sleep.

Acknowledgement

I thank Horace Barlow, Graeme Mitchison and Peter Foldiak for comments on the ms.

References

1. Gardner-Medwin AR. Doubly modifiable synapses: a model of short and long term auto-associative memory. Proc Roy Soc Lond B 1989; 238:137-154

2. Marr D. A theory for cerebral neocortex. Proc Roy Soc Lond B 1970; 176: 161-234

3. Crick F, Mitchison G. The function of dream sleep. Nature, Lond 1983; 304: 111-114

4. Gardner-Medwin AR. Modifiable synapses necessary for learning. Nature, Lond 1969; 223: 916-918

5. Giuditta A. A sequential hypothesis for the function of sleep. In: Koella WP, Ruther E, Schulz H (ed.) Sleep '84. Gustav Fischer, Stuttgart, 1985

6. Hebb DO. The organization of behaviour. Wiley, New York, 1949

7. Willshaw DJ, Buneman OP, Longuet-Higgins HC. Non-holographic associative memory. Nature, Lond 1969; 222: 960-962

8. Gardner-Medwin AR. The recall of events through the learning of associations between their parts. Proc Roy Soc Lond B 1976; 194:375-402

9. Palm G. Local synaptic rules with maximal information storage capacity. In: Haken H. (ed) Neural and synergetic computers, Springer, Berlin, 1988

10. Willshaw D, Dayan P. Optimal plasticity from matrix memories: what goes up must come down. 1990; Neural Computation In Press

11. Hopfield JJ. Neural networks and physical systems with emergent computational abilities. Proc Natl Acad Sci USA 1982; 79:2554-2558

12. Lansner A, Ekeberg O. Reliability and speed of recall in an associative network. IEEE Trans. PAMI 1985; 7: 490-498

13. Barlow HB, Kaushal TP, Mitchison GJ. Finding minimum entropy codes. Neural Computation 1989; 1: 412-423