Biol. Cybern. 65, 253-265 (1991)

Cybernetics
© Springer-Verlag 1991

Optimising synaptic learning rules in linear associative memories

P. Dayan and D. J. Willshaw

Centre for Cognitive Science, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, Scotland, United Kingdom

Received April 2, 1991/Accepted May 7, 1991

Abstract. Associative matrix memories with real-valued synapses have been studied in many incarnations. We consider how the signal/noise ratio for associations depends on the form of the learning rule, and we show that a covariance rule is optimal. Two other rules, which have been suggested in the neurobiology literature, are asymptotically optimal in the limit of sparse coding. The results appear to contradict a line of reasoning particularly prevalent in the physics community. It turns out that the apparent conflict is due to the adoption of different underlying models. Ironically, they perform identically at their co-incident optima. We give details of the mathematical results, and discuss some other possible derivations and definitions of the signal/noise ratio.

1 Introduction

The immense body of work on the neurophysiology of synaptic plasticity severely tantalises theoreticians. Long term potentiation, LTP (Bliss and Lømo 1973) in the hippocampus and neocortex is generally thought to support the Hebb hypothesis (Hebb 1949) about the facilitation of synapses due to coincident pre- and post-synaptic activity. However, even on theoretical grounds, it is clear that there also has to be some mechanism for reducing their efficacies, and the more recent discovery of long term depression, LTD (Stanton and Sejnowski 1989) points to this. There are various hypotheses about how LTD might work, and this paper presents one way to analyse them.

An associative network that is mathematically tractable has binary inputs and outputs but real-valued synapses (Willshaw 1971; Kohonen 1972). The most straight-forward learning rule for these synapses is then a linear one, in which the contributions from each association are just summed.

For non-linear matrix memories, such as the Associative Net (Willshaw et al. 1969; Willshaw 1971) which has two-valued synapses, it is hard to justify any learning rule other than the Hebb-like one which modifies them on the conjunction of pre-synaptic and post-synaptic activity. In the linear case, though, there is no such intuition. Ignoring the rôle of time, there are four possible conjunctions of activity or quiescence on the input and output fibres, and, in principle, the efficacy of the synapse linking them could change by a different amount for each of these. These four numbers define a learning rule (Palm 1988a, b). The obvious questions are which rule is optimal, and how far from the optimum are other interesting possibilities.

However, determining the optimal learning rule requires some way of judging the quality of the unit. One such metric is the signal/noise ratio (S/N), which has its roots in engineering and has proved useful in a large number of applications. Consider a single unit that is to discriminate between two classes of outputs, the 'lows' and the 'highs', based on a scalar 'return' the *dendritic sum*. For real valued synapses, the distributions of dendritic sums for the two classes are both approximately Gaussian, $\mathcal{G}(\mu_l, \sigma_l^2)$ and $\mathcal{G}(\mu_h, \sigma_h^2)$, say, then it will be easy to separate the two classes if the signal, $\mu_h - \mu_l$, is large (informally, if the peaks are far apart) and/or if the two contributions to the noise, σ_l^2 and σ_h^2, are small (informally, if the peaks are very narrow). Figure 1 shows the two distributions. The S/N is defined as:

$$\varrho \equiv \frac{(\mu_h - \mu_l)^2}{\frac{1}{2}(\sigma_l^2 + \sigma_h^2)}, \qquad (1)$$

and so incorporates both these effects. Maximising the S/N should enhance separability.

Note that the S/N is entirely independent of any threshold θ the unit might actually set to make the discrimination. This is desirable, since it factors out an issue which typically arises that one of the classes will occur more frequently than the other. Such an imbalance might happen, for instance, if the output patterns are sparsely coded, having many more lows than highs. Then, it may be more important to set θ either to preserve the few of the latter, or to make fewer errors by getting the bulk of the former correct. If high and

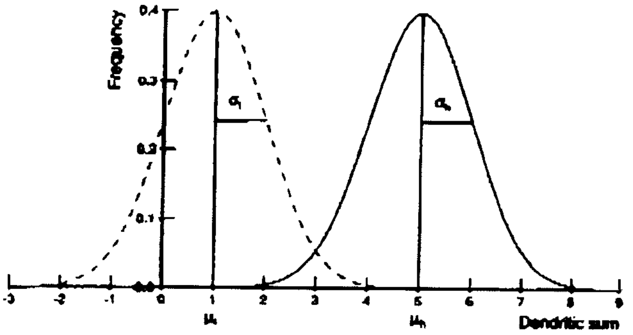


Fig. 1. Distributions of 'Low' and 'High' dendritic sums. - - - Lows; — Highs

low patterns occur with equal frequency, then it is likely to be wise to set $\theta = (\mu_l + \mu_h)/2$. For very large systems, the limit studied in the physics community, the classes will either be perfectly discriminated or perfectly confused, and so the threshold is essentially irrelevant.

In general, the S/N will be some function of both the learning rule and the input and output patterns. Given a strong statistical assumption about these patterns, it is possible to work out a theoretical value for the S/N, and to optimise it with respect to the learning rule. It turns out that care is necessary over exactly how the S/N is defined. At least one incorrect and two different correct values for it are quoted in the literature.

The next section describes the model in the formalism due to Palm (1988a, b), Sect. 3 demonstrates how each of the three possible expressions for the S/N and associated optimal rules arises, and Sect. 4 discusses their properties. Section 5 considers how thresholds might be set, and Sect. 6 compares the results with those current in the physics community. We have previously presented the learning rules (Willshaw and Dayan 1990), but not the mathematics underlying them. The appendix gives the mathematical details.

2 The model

The underlying formalism is based on that of Palm (1988a, b). A matrix memory, of the form shown in Fig. 2, is intended to store Ω associations, indexed by ω. Each component of the associants can take one of two values:

$$a_i(\omega) \in \{c, 1\}\,, \qquad i = 1 \ldots m\,, \qquad c \in \Re\,,$$

and

$$b_j(\omega) \in \{l, h\}\,, \qquad j = 1 \ldots n\,. \tag{1}$$

This is called the $\{c, 1\}$ model for the low and high values of the input respectively. All the patterns are statistically independent and within each set are distributed identically, with probabilities:

$$p = \mathscr{P}[a_i = 1]\,, \qquad 1 - p = \mathscr{P}[a_i = c]\,.$$

$$r = \mathscr{P}[b_j = h]\,, \qquad 1 - r = \mathscr{P}[b_j = l]\,.$$

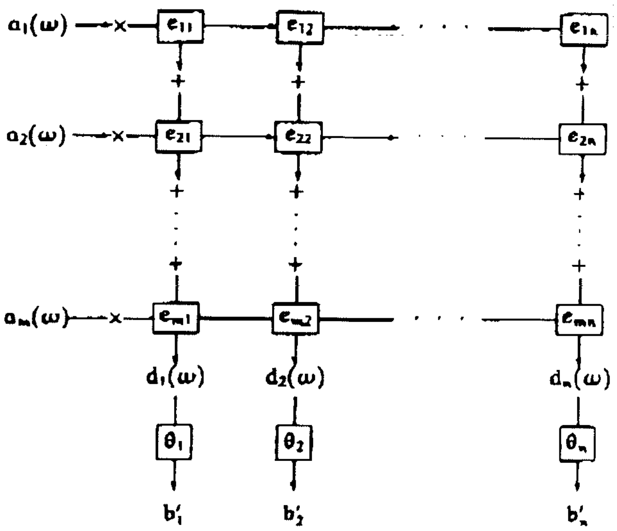


Fig. 2. The matrix shows the steps taken in the retrieval of the pattern $\mathbf{b}(\omega)$ that was previously stored in association with $\mathbf{a}(\omega)$. For good recall, the calculated output $\mathbf{b}'$, the result of thresholding the dendritic sum output by θ, should closely resemble the desired output $\mathbf{b}(\omega)$

Patterns for which $b_j = l(h)$ will be called low (high). For pattern ω^*, define

$\#_c(\omega^*)$ as the number of $i \in [1, m]$ for which $a_i(\omega^*) = c$,

and

$\#_1(\omega^*)$ as the number of $i \in [1, m]$ for which $a_i(\omega^*) = 1$.

The jth unit has synaptic weights, or efficacies, $e_{ij} \in \Re$, and consequent dendritic sum output in response to pattern ω^* input:

$$d_j(\omega^*) = \sum_{i=1}^{m} e_{ij} a_i(\omega^*)\,. \tag{2}$$

The synaptic efficacies are set by the learning rule as:

$$e_{ij} = \sum_{\omega=1}^{\Omega} \Delta_{ij}(\omega)\,,$$

where $\Delta_{ij}(\omega)$ is given in Table 1. The linear dependency on the associations learnt is clear. Any more interesting case, for instance where the synaptic elements saturate, is more difficult to analyse because the effect of a particular association can depend on when it is learnt. Hence, from (2),

$$d_j(\omega^*) = \sum_{i=l}^{m} a_i(\omega^*)\left[\sum_{\omega=l}^{\Omega} \Delta_{ij}(\omega)\right]. \tag{3}$$

Table 1. Local synaptic learning rule

$\Delta_{ij}(\omega)$		Output $b_j(\omega)$ low	 high
Input	c	α	β
$a_i(\omega)$	1	γ	δ

Since the learning rule is *local*, each unit learns separately. The following discussion concerns only one such unit, and the subscript j will be dropped.

For the given pattern ω^*, (3) can be separated into two parts:

$$d(\omega^*) = \sum_{i=1}^{m} a_i(\omega^*)\,\Delta_i(\omega^*)$$

$$+\sum_{i=1}^{m} a_i(\omega^*)\left[\sum_{\omega=1,\,\omega\neq\omega^*}^{\Omega} \Delta_i(\omega)\right]. \qquad (4)$$

The first of these terms, $S(\omega^*)$, determines the *signal* for pattern ω^*, and the second, $N(\omega^*)$, determines the *noise*.

The central limit theorem implies that the dendritic sums $d(\omega^*)$ for both classes of patterns (those to which the unit should respond low and high) will be approximately Gaussian. Figure 1 above gives a possible frequency graph showing the distribution of the dendritic sums. The two peaks, corresponding to the two classes, are clearly evident, as is the fact that there is no threshold θ that would not result in either errors of commission or errors of omission, or both. To see this last point, observe that no vertical line could be drawn that entirely separates the two peaks. As discussed in the introduction, the signal/noise ratio (S/N), defined in (1), is a measure of the average potential fidelity of recall for the unit.

Note also that c, the value contributed by a 'low' input, is a parameter of the system. This is to allow an evaluation of certain claims being made about how dramatically the $\{0, 1\}$ model (i.e. $c = 0$) outperforms the $\{-1, 1\}$ model (i.e. $c = -1$) for sparse patterns. A priori, this seems unlikely, since there is a formal equivalence between these two models. To see this, consider $a_i(\omega^*) = \lambda\hat{a}_i(\omega^*) + \mu$, where $\hat{a}_i(\omega^*) \in \{0, 1\}$ are the 'canonical' inputs for a pattern. By varying λ and μ it is possible to generate any of these models - e.g. $\mu = c$, $\lambda = 1 - c$, gives the $\{c, 1\}$ model. Then

$$d(\omega^*) = \sum_{i=1}^{m} e_i a_i(\omega^*)$$

$$= \lambda\left\{\sum_{i=1}^{m} e_i\hat{a}_i(\omega^*)\right\} + \mu\left\{\sum_{i=1}^{m} e_i\right\}. \qquad (5)$$

where e_i are the unit's weights. The μ term is purely additive, and so cannot affect the S/N. The λ term is multiplicative, but expanding the size of the Gaussian curves for both classes also fails to change the S/N. To see this, note that although the distance between the means goes up by the multiplicative factor, so does the breadth of each of the curves. Operationally, for any given value of θ for, say, the $\{0, 1\}$ model, there is some other threshold for, say, the $\{-1, 1\}$ model, which allows the unit to make *identical* errors. Changing c is essentially a formal step. The apparent contradiction between these results and those enjoying currency in the physics community will be explored in Sect. 6.

3 Establishing the signal/noise ratio

The usual method of calculating ϱ, the S/N, is fairly straightforward. Having separated the dendritic sum for both low and high patterns into *Signal* + *Noise*, as in (4), the numerator for ϱ (sec 1) would be the expected difference between the signals for the two classes:

$$\mu_h - \mu_l = \mathscr{E}_h[S(\omega^*)] - \mathscr{E}_l[S(\omega^*)],$$

and the denominator would be the average variance of their noises

$$\tfrac{1}{2}(\sigma_l^2 + \sigma_h^2) = \tfrac{1}{2}(\mathscr{V}_l[N(\omega^*)] + \mathscr{V}_h[N(\omega^*)]),$$

where $\mathscr{E}_h$ implies that the expectation is taken over those patterns for which the output $b(\omega^*) = h$, and similarly for $\mathscr{V}_h$.

This amounts to making two assumptions:

Expectation of the noise: that $\mathscr{E}_h[N(\omega^*)] = \mathscr{E}_l[N(\omega^*)]$;

Variance of the noise: that it is this quantity rather than some other measure of the spread of the dendritic sums that determines the ability of the unit to perform its discrimination accurately.

Neither assumption is true, although it is possible to reconstruct results in the literature on calculating S/N using either or both of them. The following three subsections demonstrate the effects of accepting and rejecting them.

3.1 The effect of accepting both assumptions

Consider a high pattern, ω_h. The signal $S(\omega_h)$ is the contribution due to terms $\Delta_i(\omega_h)$ for all the input lines and so:

$$S(\omega_h) = \delta \,\#_1(\omega_h) + c\beta \,\#_c(\omega_h).$$

The expectation value of the signal is therefore

$$\mathscr{E}_h[S(\omega_h)] = m[p\delta + (1-p)c\beta].$$

Similarly for a signal ω_l for which the value of the unit should be l,

$$\mathscr{E}_l[S(\omega_l)] = m[p\gamma + (1-p)c\alpha].$$

Assuming that the expectation of the noise is the same for high and low cases,

$$\mu_h - \mu_l = m[p(\delta-\gamma) + (1-p)c(\beta-\alpha)]. \qquad (6)$$

For calculating the noise, there is a lemma that if:

$$\Gamma = \begin{cases}\Phi & \text{with probability } a,\\ \Psi & \text{with probability } 1-a,\end{cases}$$

where Φ and Ψ are random variables, then the variance $\mathscr{V}$ of Γ is

$$\mathscr{V}[\Gamma] = a\mathscr{V}[\Phi] + (1-a)\mathscr{V}[\Psi]$$

$$+ a(1-a)(\mathscr{E}[\Phi] - \mathscr{E}[\Psi])^2. \qquad (7)$$

Now consider in (4), the inner sum in the noise term:

$$\sum_{\omega=1,\,\omega\neq\omega^*}^{\Omega} \Delta_i(\omega).$$

This is made up from contributions from each of $\Omega - 1$ patterns, where, for each pattern,

$$\Delta_i(\omega) = \begin{cases} \begin{cases} \delta & \text{with probability } r, \\ \gamma & \text{with probability } 1-r, \end{cases} & \text{with probability } p, \\ \begin{cases} \beta & \text{with probability } r, \\ \alpha & \text{with probability } 1-r, \end{cases} & \text{with probability } 1-p \end{cases}$$

Applying (7) twice, the variance of $\Delta_i(\omega)$ is:

$$\mathscr{V}[\Delta_i(\omega)] = p[r(1-r)(\delta-\gamma)^2] + (1-p)[r(1-r)(\beta-\alpha)^2] + p(1-p)[r\delta + (1-r)\gamma - r\beta - (1-r)\alpha]^2 . \qquad (8)$$

Equation (4) involves the sum of $\#_c(\omega^*)$ copies weighted by c and $\#_1(\omega^*)$ copies weighted by 1. Under the apparently plausible assumption of independence between $a_i(\omega^*)$ and $\Sigma^{\Omega}_{\omega=1,\omega\neq\omega^*}\Delta_i(\omega)$, over all the patterns,

$$\mathscr{V}[N(\omega^*)] = (\Omega - 1)(c^2\#_c(\omega^*) + \#_1(\omega^*))\mathscr{V}[\Delta_i(\omega)] . \qquad (9)$$

and making the assumption that the variance for each pattern can be averaged over all patterns ω^* would produce

$$\begin{aligned} \sigma_h^2 &= \sigma_l^2 \\ &= m(\Omega-1)[p + c^2(1-p)]\mathscr{V}[\Delta_i(\omega)] \\ &= (\Omega-1)[p + c^2(1-p)]r(1-r) \\ &\quad \times \left\{ p(\delta-\gamma)^2 + (1-p)(\beta-\alpha)^2 \right. \\ &\quad \left. + \frac{p(1-p)}{r(1-r)}[r(\delta-\gamma) - r(\beta-\alpha) + (\gamma-\alpha)]^2 \right\} . \end{aligned} \qquad (10)$$

The S/N, ϱ, can now be calculated from expressions 6 and 10. Maximising it with respect to α, β, γ and δ determines the conditions for an optimum.

Table 2 sets out the consequent rule, where δ has been arbitrarily set to f and γ to $f-g$. The optimal S/N is:

$$\hat{\varrho}_1 = \frac{m}{\Omega-1}\frac{1}{r}\frac{1}{1-r}$$

which, oddly enough, is correct in the general case, as shown later.

Table 2. Optimal $\hat{\varrho}_l$ local synaptic learning rule

$\Delta_i(\omega)$		Output $b(\omega)$ low	high
Input	c	$f - g(1 - r + rc)$	$f - g(1-r)(1-c)$
$a_i(\omega)$	1	$f - g$	f

For $p = r$, one of the special cases of the rule is the one quoted by Palm (1988b).

$$\alpha = cp \qquad \beta = -c(1-p)$$
$$\gamma = -p \qquad \delta = 1 - p .$$

Palm (1988b) also gives the S/Ns for two rules which are not, in general, instances of the optima. They are:

The Hopfield rule:

$$\varrho_1^{\text{Hopfield}} = \frac{m}{\Omega-1}\frac{1}{2p(1-p)}\frac{1}{1-2p(1-p)}$$

$$\begin{matrix} \alpha = 1 & \beta = -1 \\ \gamma = -1 & \delta = 1 \end{matrix} \quad p = r, \quad c = -1$$

(This is optimal for $p = r = 1/2$)

The Hebb rule: $\varrho_1^{\text{Hebb}} = \frac{m}{\Omega-1}\frac{1}{p}\frac{1}{1-p^2}$

$$\begin{matrix} \alpha = 0 & \beta = 0 \\ \gamma = 0 & \delta = 1 \end{matrix} \quad p = r, \quad c = 0$$

Although these results are identical to those by Palm (1988b), it remains unclear to what extent this derivation, and the general expression for the S/N, mirrors his own.

That something is amiss may be appreciated by considering the behaviour of $\varrho_1^{\text{Hopfield}}$ as $p = r \to 0$. One might expect that the S/N should decrease under these circumstances, since the learning rule is incorrectly symmetrical in $a(\omega)$. However, $\varrho_1^{\text{Hopfield}}$ actually increases. Simulations confirm this point; Table 3 shows theoretical and empirical values of $\varrho_1^{\text{Hopfield}}$ and ϱ_1^{Hebb} for various values of p[1]. It is apparent both that the simulations diverge substantially from the theoretical expression, ϱ_1, and that the Hopfield rule does indeed get worse for smaller p. The Hebb rule is not optimal for any values of p or r, but it is asymptotically optimal for sparse patterns, as $p = r \to 0$.

In his treatment, Palm makes the assumption from the very outset that the S/N will be unaffected if all of α, β, γ, and δ are multiplied by the same non-zero number, or if the same number is added to them all. A priori, and, as indeed is borne out by simulations, the last invariance is most unlikely to hold. If a large enough quantity is added to each element in the rule such that all the weight values are large and positive, then the signal which determines the classification of a particular pattern as low or high is likely to be entirely swamped by the noise due to the uncertainty in the number of the inputs that are c or 1. Palm uses this assumption to reduce the number of free variables on which the learning rule depends.

[1] For this and the other simulations in this paper, the $n = 20$ units have $m = 512$ input lines and are fed $\Omega = 200$ patterns. Figures are averages over 50 runs. The observed values of the variances are based on a unit-by-unit calculation

Table 3. Theoretical and empirical values of the S/N for the Hebb and Hopfield rules. After Willshaw and Dayan (1990)

Hebb rule

		Predicted S/N			Actual	
p, r	c	ρ_1	ρ_2	ρ_3	S/N	$\pm\sigma$
0.5	0	6.9	1.7	0.050	0.10	±0.11
0.4	0	7.7	2.8	0.12	0.11	±0.090
0.3	0	9.4	4.6	0.32	0.34	±0.15
0.2	0	13	8.6	1.1	1.2	±0.47
0.1	0	26	21	7.7	7.1	±1.0
0.05	0	52	47	32	28	±18

Hopfield rule

		Predicted S/N			Actual	
p, r	c	ρ_1	ρ_2	ρ_3	S/N	$\pm\sigma$
0.5	0.5	1.0	1.0	10	11	±1.3
0.5	0	5.1	5.1	10	11	±1.3
0.5	−0.5	9.3	9.3	10	11	±1.3
0.5	−1	10	10	10	11	±1.3
0.4	−1	10	9.5	7.5	8.3	±1.5
0.3	−1	11	7.5	1.4	1.3	±0.40
0.2	−1	12	4.8	0.25	0.32	±0.22

3.2 *Correcting for the expectation of the noise*

The first assumption given above was that the expected values of the noise obscuring high and low patterns are the same. This is not true, and so the difference between the expected value of $d(\omega^*)$ for high and low patterns cannot be taken to be equal to the difference between the expected value of the signal $S(\omega^*)$ in the two cases. In (4) the noise term

$$N(\omega^*) = \sum_{i=1}^{m} a_i(\omega^*)\left[\sum_{\omega=1,\,\omega\neq\omega^*}^{\Omega} \Delta_i(\omega)\right]$$

excludes pattern ω^*, and there is a difference between excluding a pattern for which $b(\omega^*) = h$ and one for which $b(\omega^*) = l$. If $\mathcal{N}_h$ patterns have $b(\omega^*) = h$ and $\mathcal{N}_l$ have $b(\omega^*) = l$, so $\mathscr{E}[\mathcal{N}_h] = \Omega r$ and $\mathscr{E}[\mathcal{N}_l] = \Omega(1-r)$, then:

$$\mathscr{E}_h[N(\omega^*)] = m\mathscr{E}_h[a_i(\omega^*)]\mathscr{E}[(\mathcal{N}_h - 1)(p\delta + (1-p)\beta) + \mathcal{N}_l(p\gamma + (1-p)\alpha)],$$

$$\mathscr{E}_l[N(\omega^*)] = m\mathscr{E}_l[a_i(\omega^*)]\mathscr{E}[\mathcal{N}_h(p\delta + (1-p)\beta) + (\mathcal{N}_l - 1)(p\gamma + (1-p)\alpha)].$$

and therefore:

$$\mathscr{E}_h[N(\omega^*)] - \mathscr{E}_l[N(\omega^*)] = -m[p + c(1-p)] \times \underline{[p\delta + (1-p)\beta - p\gamma - (1-p)\alpha]}.$$

Using this contribution to amend the expression for $\mu_h - \mu_l$ in (6) yields

$$\mu_h - \mu_l = m[p(\delta-\gamma) + c(1-p)(\beta-\alpha) - (p + c(1-p)) \times (p\delta + (1-p)\beta - p\gamma - (1-p)\alpha)]$$
$$= mp(1-p)(1-c)[(\delta-\gamma) - (\beta-\alpha)]. \qquad (11)$$

Table 4. Optimal ρ_2 local synaptic learning rule

$\Delta_i(\omega)$		Output $b(\omega)$ low	high
Input	c	$h - g\frac{1-p-r}{1-p}$	$h - g\frac{1-r}{1-p}$
$a_i(\omega)$	1	$h - g$	h

Using (11) and the old expression (10) for the noise gives:

$$\varrho_2 = \zeta\,\frac{p^2(1-p)^2(1-c)^2[(\delta-\gamma)-(\beta-\alpha)]^2}{p(\delta-\gamma)^2 + (1-p)(\beta-\alpha)^2 + \frac{p(1-p)}{r(1-r)} \times [r(\delta-\gamma) - r(\beta-\alpha) + (\gamma-\alpha)]^2}, \qquad (12)$$

where

$$\zeta = \frac{m}{\Omega - 1}\,\frac{1}{(p + c^2(1-p))r(1-r)}.$$

Maximising this with respect to α, β, γ and δ gives the optimal rule shown in Table 4, where, for comparison, $\delta = h$ and $\gamma = h - g$. The optimal S/N is now:

$$\hat{\varrho}_2 = \frac{m}{\Omega-1}\,\frac{1}{r(1-r)}\,\frac{p(1-p)(1-c)^2}{p + c^2(1-p)}$$

$$= \hat{\varrho}_1\,\frac{p(1-p)(1-c)^2}{p + c^2(1-p)}.$$

This derivation has removed the dependence on c of the learning rule, but leaves us free to maximise the S/N with respect to c. The maximum occurs at $\hat{c} = -p/(1-p)$, where the average value of each input is zero. Then, $\hat{\varrho}_2 = \hat{\varrho}_1$.

Not only is this rule somewhat inelegant, but it also violates two empirical principles outlined earlier: the S/N should actually be independent of c, the numerical value of a low input, and the rule should not be additively invariant, i.e. it should not be the case that any number can be added to the rule without affecting its S/N. Table 3 also compares the theoretical ϱ_2 and actual S/Ns for the Hebb and Hopfield rules for various values of $p = r$. It is apparent that ϱ_2 is indeed fallacious. Note again that the Hopfield rule is a special case of the optimum for $p = r = 1/2$ and $h = 1$, $g = 2$.

3.3 *A resolution*

The first pointer to a solution of these problems came from the simulations. There are two possible ways of calculating the mean low and high dendritic sums: either over the whole set of output units, or on an individual, output-unit by output-unit basis. The estimated sample variance will obviously depend on which of these is adopted, and should be lower for the first method than for the second. However, under the second assumption, that it is the variance of the noise that

Cov rule: $\alpha = pr$, $\beta = -p(1-r)$, $\gamma = -(1-p)r$, $\delta = (1-p)(1-r)$
the term $p\delta + (1-p)\beta - p\gamma - (1-p)\alpha = 0$
Hence E(hi noise) = E(lo noise)

determines the theoretical discriminability, they would not differ in the limit of large numbers of inputs. Simulations confirmed that this was not the case.

It was then obvious that it is not enough to calculate the variances of the dendritic sums – the correlations between two dendritic sums are important too. The analysis based solely on the variance ignores the fact that the efficacies e_i are quenched, i.e. although they are determined during learning by the statistics of the patterns, they are fixed by the time of recall. The units can take advantage of this by setting their thresholds independently, each according to its own quenched weights. The correlations in the dendritic sums come about because the synaptic efficacies are determined by the *actual* numbers of low and high patterns the units have learnt rather than just the *mean* numbers.

For instance, using the Hebb rule with $\{0, 1\}$ patterns, a unit that happens to have learnt a large number of high patterns will tend to have dendritic sums that are greater than those for a unit that happens to have learnt only a few. The variance analysis for ϱ_2 just balances these cases out, whereas it is clear that the threshold for a unit of the first type will optimally be larger than the threshold for one of the second.

The following simple didactic example of the effects of correlation between noise terms demonstrates the class of phenomenon that occurs. Imagine that signals $\phi(t) \in \{-1, 1\}$ are corrupted by additive noise $\psi(t)$. There are two possible processes generating $\psi(t)$:

$$\psi_1 \sim \mathscr{G}((1-\pi), \sigma^2),$$

and

$$\psi_2 \sim \mathscr{G}(-\pi, \sigma^2)$$

where each collection is independent and identically distributed. It is not known before the experiment which process will generate the noise; all that is known is that

$$\pi = \mathscr{P}[\psi \text{ is given by } \psi_1],$$

and

$$1-\pi = \mathscr{P}[\psi \text{ is given by } \psi_2].$$

Figure 3 demonstrates the two possibilities. Rather similarly to the effect of changing μ in the analysis of the rôle played by c (see 5), the only difference between the two cases is that the frequency graphs are shifted with respect to each other. The S/Ns are identical, and indeed an appropriate choice of threshold would result in no more and no fewer errors being made.

However, performing the formal analysis as for ϱ_2 gives that

$$\mathscr{E}[\psi(t)] = 0,$$

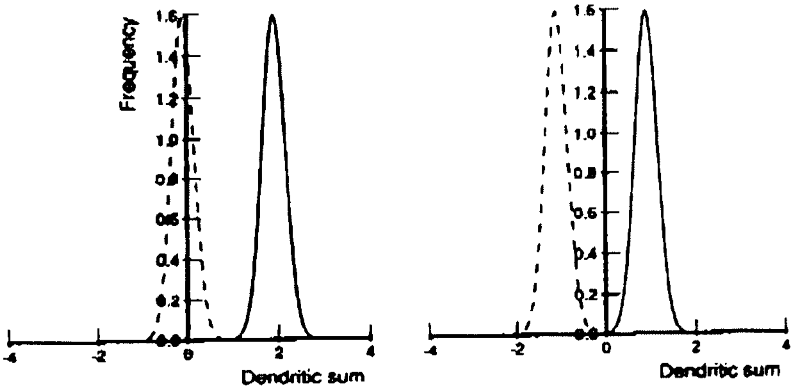
and

$$\begin{aligned}\mathscr{V}[\psi(t)] &= \pi(\sigma^2 + (1-\pi)^2) + (1-\pi)(\sigma^2 + \pi^2) \\ &= \sigma^2 + \pi(1-\pi).\end{aligned}$$

But this is clearly an overestimate of the 'operative variance', which is here defined as the expected dispersion of the corrupted signals about their *actual* means, rather than their *expected* means. So long as the unit can set its own threshold according to which of ψ_1 and ψ_2 occurred, this is the appropriate quantity to calculate, being the factor that disposes it to err. Its value is obviously σ^2, the individual variance of both ψ_1 and ψ_2.

In the simple example, the noise terms are correlated, because one choice (based on the probability π) determines the distributions for them all. Ignoring this, by calculating the true variance rather than the dispersion of the corrupted signals, leads to an incorrect measure of how well the unit will be able to do its job of discriminating between the two possible classes, $\phi(t) = -1$ and $\phi(t) = 1$.

In the case of the associative memory, this issue is slightly more complicated. Here, the distribution of the noise terms is also determined in advance of the operation of the unit as a discriminator, in this case by the quenched weight values that emerge from the particular set of input/output associations it learns. However, the effects of the noise are mediated through the actual $\{c, 1\}$ input values for the patterns. If $c = -p/(1-p)$ then the expected value of any input is zero. This nullifies any effect from the differences between the actual efficacies of the synapses and their expected



Distribution under ψ_1

Distribution under ψ_2

Fig 3. Distributions under ψ_1 and ψ_2 for $\pi = 0.1$, $\sigma = 0.25$ – dotted lines indicate the low signals, solid lines the high ones. Translation is the only difference

values, which are normally the cause of the whole problem. If c does not take this value, then there will be an effect due to the quenching, that will make the variance of the dendritic sums diverge from the dispersion. To re-iterate, it is the dispersion rather than the variance that determines the unit's ability to discriminate, and so it is the dispersion that is the appropriate measure for the S/N.

The mean dispersion is defined as:

$$s_h^2 = \mathscr{E}\left[\frac{1}{\mathscr{N}_h}\sum_{\{\omega|b(\omega)=h\}}[\mathrm{d}(\omega)]^2 - \left(\frac{1}{\mathscr{N}_h}\sum_{\{\omega|b(\omega)=h\}}\mathrm{d}(\omega)\right)^2\right], \tag{13}$$

$\mathscr{N}_h$(and $\mathscr{N}_l$) being the number of ω for which $b(\omega) = h$(or l). s_l^2 is defined similarly as the expected dispersion for low patterns. Symbolically,

Dispersion = Variance − Correlation,

and it is the interaction of the quenched weights with c that introduces the correlations.

The appendix gives details of the calculation of the expected value of the dispersion, done explicitly through writing out the squares of the sums in (3). This gives:

$$\begin{aligned}s_h^2 \simeq\ & mp(1-p)[(1-c)^2(1-2p)^2(\delta-\beta)^2\\ & -2p(1-p)(1-c)^2(\delta-\beta)^2]\\ & + mp(1-p)\mathscr{E}[p(1-p)(1-c)^2(\mathscr{N}_h(\delta-\beta)^2\\ & + \mathscr{N}_l(\gamma-\alpha)^2)] + mp(1-p)\mathscr{E}[2(1-c)^2(1-2p)\\ & \times(\delta-\beta)(\mathscr{N}_h\phi+\mathscr{N}_l\psi)] + mp(1-p)\mathscr{E}[(1-c)^2\\ & \times(\mathscr{N}_h\phi+\mathscr{N}_l\psi)^2].\end{aligned} \tag{14}$$

where $\phi = p\delta + (1-p)\beta$ is the average contribution to the synaptic efficacy from a high pattern, and $\psi = p\gamma + (1-p)\alpha$ is the average contribution from a low one.

For large Ω, the last of these terms

$$\begin{aligned}\mathscr{E}[(\mathscr{N}_h\phi+\mathscr{N}_l\psi)^2] = \ & \Omega r(1-r)(\phi-\psi)^2\\ & + \Omega^2(r\phi+(1-r)\psi)^2\end{aligned} \tag{15}$$

will dominate the noise and swamp the signal unless $r\phi + (1-r)\psi = 0$. In practice this removes the additive degree of freedom in the rules for ϱ_1 and ϱ_2, ensuring that the average value of the efficacy of a synapse must be 0. The component that remains arises from the uncertainty in the values $\mathscr{N}_h$ and $\mathscr{N}_l$.

Ignoring the first terms in s_h^2, which are dominated by the terms which are in Ω and Ω^2, gives

$$\begin{aligned}s_h^2 & \sim s_l^2\\ & \simeq m\Omega p(1-p)(1-c)^2[p(1-p)(r(\delta-\beta)^2\\ & \quad + (1-r)(\gamma-\alpha)^2) + r(1-r)(\phi-\psi)^2\\ & \quad + \Omega(r\phi+(1-r)\psi)^2]\\ & = m\Omega p(1-p)(1-c)^2[r(1-r)(p(\delta-\gamma)^2\\ & \quad + (1-p)(\beta-\alpha)^2) + p(1-p)(r\delta+(1-r)\gamma\\ & \quad - r\beta-(1-r)\alpha)^2 + \Omega(r\phi+(1-r)\psi)^2]\end{aligned}$$

and so:

$$\varrho_3 = \frac{m}{\Omega}\frac{p(1-p)[\delta-\gamma-\beta+\alpha]^2}{p(1-p)[r(\delta-\beta)^2+(1-r)(\gamma-\alpha)^2] + r(1-r)[\phi-\psi]^2 + \Omega(r\phi+(1-r)\psi)^2}$$

Comparing the form of ϱ_3 with that of ϱ_2, it turns out that, excluding the term in Ω^2, they have the same dependence on α, β, γ and δ, but that the dependence on c has finally been excised.

Maximising with respect to α, β, γ and δ, the optimal rule is just as for ϱ_2 apart from the important constraint that $r\phi + (1-r)\psi = 0$. This gives one true optimum:

The Covariance rule **R1** (see Table 5):

$$\begin{array}{ll}\alpha = pr & \beta = -p(1-r)\\ \gamma = -(1-p)r & \delta = (1-p)(1-r)\end{array} \quad \rho_3^{\text{Covariance}} = \frac{m}{\Omega}\frac{1}{r}\frac{1}{1-r}$$

Two other sub-optimal rules have previously been proposed (see the next section for a discussion). As Alessandro Treves has pointed out, our original classification in Willshaw and Dayan (1990) of them as being locally optimal was incorrect. In fact, they are not even optimal under the additional condition that $\alpha = 0$. They are:

The Heterosynaptic rule **R2**:

$$\begin{array}{ll}\alpha = 0 & \beta = -p\\ \gamma = 0 & \delta = 1-p\end{array} \quad \rho_3^{\text{Hetero}} = \frac{m}{\Omega}\frac{1}{r}$$

The Homosynaptic rule **R3**:

$$\begin{array}{ll}\alpha = 0 & \beta = 0\\ \gamma = -r & \delta = 1-r\end{array} \quad \rho_3^{\text{Homo}} = \frac{m}{\Omega}\frac{1}{r}\frac{1-p}{1-r}$$

Table 3 shows the close agreement between the theoretical prediction, ρ_3, of the S/N and the empirical result for the Hebb and Hopfield rules. Table 6 shows the

Table 5. Optimal ϱ_3 local synaptic learning rule

$\Delta_j(\omega)$		Output $b(\omega)$ low	 high
Input	c	pr	$p(1-r)$
$a_j(\omega)$	1	$-(1-p)r$	$(1-p)(1-r)$

Table 6. Theoretical ϱ_3 predictions of the S/N for the optimal (**R1**), sub-optimal (**R2** and **R3**), Hebb, and Hopfield rules for various values of $p = r$. Note that **R2** and **R3** are very close to **R1** as the sparsity increases, but the Hebb rule is significantly worse. After Willshaw and Dayan (1990)

	Signal/Noise Ratios for			
p,r	**R1**	**R2, R3**	Hebb	Hopfield
0.5	10	5.1	0.050	10
0.4	11	6.4	0.12	7.5
0.3	12	8.5	0.32	1.4
0.2	16	13	1.1	0 25
0.1	28	26	7.7	0.045
0.05	54	51	32	0.015

theoretical S/N for the optimal, sub-optimal, and the Hebb and Hopfield rules for various values of $p = r$, based on ϱ_3. The Hopfield rule is optimal for $p = r = 1/2$, but rapidly tails off as the patterns get more sparse. Even though the Hebb rule is asymptotically optimal as p gets small, it is significantly worse than rules **R1** to **R3** optima even for quite tiny but finite values.

4 The optimal and sub-optimal rules

The optimal and two sub-optimal rules in the previous section can be identified with ones suggested in various places in the literature. The covariance rule was originally proposed by Sejnowski (1977a, b), and has since been widely used in connectionist systems. For instance, the Hopfield rule (Willshaw 1971; Hopfield 1982) is a special case of it when $p = r = 1/2$. In fact, in the physics models (Tsodyks and Feigel'man 1988; Buhmann et al. 1989; Perez-Vincente and Amit 1989) discussed in Sect. 6, it is taken as read. The motivation behind it is even clearer from the equivalent form

$$\Delta_i(\omega) \propto (a_i(\omega) - \bar{a})(b(\omega) - \bar{b}) .$$

where $\bar{a}$ is the average value of an input $p + (1 - p)c$ and $\bar{b}$ is the average value of an output.

None of the ϱ_3 rules described is biologically plausible, for reasons discussed below, but $\varrho_3^{\mathrm{Covariance}}$ is particularly difficult to justify because $\alpha > 0$. α is the change in efficacy of a synapse in the absence of either pre- or post-synaptic activity. One could imagine some form of decay process, which would tend to eliminate unused synapses, but for the efficacy actually to rise is counterintuitive. α's 'rôle' is to keep the expected value of a synapse zero, which is the non-additivity condition that Palm ignored.

Various parts of the brain show synaptic plasticity, including the visual system (during development), the cerebellum and the hippocampus. Different underlying mechanisms are believed to be responsible – for instance the analogue of long term potentiation (LTP) in the hippocampus seems to be long term depression (LTD) in the cerebellum (Ito et al. 1982) – and the extent to which the plasticity is merely an artefact of the procedure is also in doubt. Our hetero- and homo-synaptic rules are so called because of their similarities with the eponymous biological rules for LTD. Heterosynaptic LTD has been known about for some time in various parts of the brain, and a theoretical rule like this has been suggested by Stent (1973); Singer (1985); and others. The evidence for homosynaptic LTD in the hippocampus is rather more recent (Stanton and Sejnowski 1989), and disputes remain about its reality and properties. Bienenstock et al. (1982) made an early proposal along the lines of the homosynaptic rule, for plasticity in the visual system[2].

Both the hetero- and homo-synaptic rules perform worse than the covariance rule; $\varrho_3^{\mathrm{Hetero}}$ by a factor $1 - r$, and $\varrho_3^{\mathrm{Homo}}$ by a factor $1 - p$. However, since the regime in which any of the rules work well is where the patterns are sparse (i.e. p and r are small), these factors are relatively small. The nervous system is known to employ sparse coding. For $p = r$, $\varrho_3^{\mathrm{Hetero}}$ and $\varrho_3^{\mathrm{Homo}}$ are equal. The homosynaptic rule has also been used for connectionist systems, such as Kanerva's sparse distributed memory (SDM) (Kanerva 1988). The original version of SDM only considers patterns with $r = 1/2$, and for it to be used optimally with different activity ratios, the analysis here would suggest that the equivalents of γ and δ ought to be suitably juggled.

One notable feature of all the rules is that for sparse patterns, the absolute value of the increment δ is an order of magnitude larger than the decrements β or γ. If this were also true of the real rules, it would make LTD significantly more difficult to detect than LTP. This would require careful experimental design, to ensure the frequency of non-stimulation of input and output fibres was sufficiently high.

All the rules involve both increases and decreases in synaptic efficacy. Unfortunately for their biological relevance, they also require the synapse to take both positive and negative values. The whole scheme works by ensuring that the expected value of every change to a synapse is zero – otherwise the Ω^2 factor lurking in (15) will swamp the signal entirely. Dale's law, that almost no synapse can change its spots from being excitatory to inhibitory, or vice-versa, has the status almost of a theoretical *pons asinorum* – one that these rules cannot cross. The obvious solution to this, which, for instance Hancock et al. (1991) adopt in their related model, is to regard each unit as a composite of two mutually inhibiting units; one which sums up the excitatory inputs, and the other which sums up the inhibitory ones. For this to work in practice, there would have to be a high degree of anatomical specificity in connections and connection types, for which there is no evidence.

A further problem with these rules is that they ignore the crucial rôle of time in the learning, and they rely too heavily on the convenient availability of the **b** patterns with which inputs are associated. It is ironic that the hippocampus is one of the main regions in which the 'static' phenomena of LTP and LTD are studied, since it is known to be important for a variety of temporal tasks such as delayed matching or non-matching to sample (Gaffan 1974). Any model of learning, such as these, that allows no temporal influence, is unlikely to be very accurate. However, even given these constraints, and the added fact of the highly complex time-course of real LTP (Racine et al. 1983), the model does provide a theoretical maximum discriminability for any associative memory built along these lines.

5 Threshold setting

As seen above, the S/N is a threshold-independent measure of the quality of the unit. The unit is susceptible

[2] Note that the optimal rule under the additional condition that $\alpha = 0$ specifies decreases in efficacy under both hetero- and homosynaptic conditions. The latter are an order magnitude greater than the former for sparse patterns

to the two types of error (commission and omission) and the threshold can be set optimally according to how each of these is weighed. Essentially the problem reduces to the standard statistical one of class discrimination (Duda and Hart 1973) with so called 'Type 1' and 'Type 2' errors. As an example, consider the problem of minimising the probability of the unit erring, given that the two distributions are distributed as Gaussians with a common variance, $\mathscr{G}(\mu_l, \sigma^2)$ and $\mathscr{G}(\mu_h, \sigma^2)$ respectively, and with the relative frequencies of low and high patterns being $(1 - r) : r$. This is a fair approximation, as discussed above. Then, for a threshold θ, the overall probability of a mis-classification is

$$\mathscr{P}_M = \frac{1-r}{\sqrt{2\pi\sigma^2}} \int_\theta^\infty e^{-(x-\mu_l)^2/2\sigma^2}\,dx + \frac{r}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\theta} e^{-(x-\mu_h)^2/2\sigma^2}\,dx$$

where the first term is the probability of getting a low pattern wrong and the second is the probability of misclassifying a high one.

Differentiating $\mathscr{P}_M$ with respect to θ gives

$$\frac{d}{d\theta}\mathscr{P}_M = \frac{1}{\sqrt{2\pi\sigma^2}}\left\{re^{-(\theta-\mu_l)^2/2\sigma^2} - (1-r)e^{-(\theta-\mu_h)^2/2\sigma^2}\right\}$$

which is zero at

$$\hat{\theta} = \frac{\mu_h + \mu_l}{2} - \frac{\sigma}{\sqrt{\rho}} \ln\left[\frac{r}{1-r}\right],$$

where ϱ is the S/N. This makes intuitive sense since $\lim_{r \to 0} \hat{\theta} = +\infty$, i.e. if virtually every pattern is low, the threshold will be large and positive, and so almost every pattern will be classed as a low. Conversely, $\lim_{r \to 1} \hat{\theta} = -\infty$, which arranges for the opposite effect. Note also that the larger the S/N, the smaller the effect of any difference between the two frequencies.

Table 7 shows the result of using the Hopfield rule in conjunction with this threshold for various values of $p = r$, demonstrating the close agreement between theory and simulation. Recall that this rule is only optimal (as an example of **R1**) for $p = r = 1/2$. The normal criterion adopted for the Associative Net (Willshaw et al. 1969) is that the expected number of errors across all the outputs should be 1. Since the expected number of errors rises with the number of output units, achieving this criterion for a network of units requires a higher S/N than for a single output unit; either there will need to be more input lines, or sparser patterns must be used, or else fewer patterns can be stored to the same accuracy.

Table 7. Using threshold $\hat{\theta}$, the expected and actual numbers of errors across $n = 20$ output lines. After Willshaw & Dayan (1990)

Hopfield rule

		Expect	Actual		Expect	Actual
p, r	c	S/N	S/N	$\pm\sigma$	Errors	Errors
0.5	−1	10	11	±1.3	1.1	1.1
0.4	−1	7.5	8.3	±1.5	1.7	1.6
0.3	−1	1.4	1.3	±0.40	4.6	4.5
0.2	−1	0.25	0.32	±0.22	4.0	4.2

One interesting feature of Table 7 is that the expected and actual numbers of errors both decrease between $p = 0.3$ and $p = 0.2$, despite the fact that the S/N also decreases, and so the unit might be expected to behave less well. This is because the expected error rate using the above threshold is

$$\mathscr{P}_M = (1-r)\Phi\left(-\frac{\sqrt{\rho}}{2} + \frac{1}{\sqrt{\rho}} \ln\frac{r}{1-r}\right) + r\Phi\left(-\frac{\sqrt{\rho}}{2} - \frac{1}{\sqrt{\rho}} \ln\frac{r}{1-r}\right),$$

where $\Phi(x)$ is the area up to x under a standard Gaussian curve. Figure 4 plots this as a function of r and ρ, and it is apparent that for small S/N, the unit will be expected to make more errors for values of r away from 0 and 1, even at the same S/N. An intuitive feel for this is given by the observation that the maximum error rate is bounded by $\min\{r, 1-r\}$, as the threshold could be set at ∞ or $-\infty$. Note the waxing and waning bimodality of this function.

All this analysis is based on the assumption that each unit can set its own threshold. Section 3 showed that this is only necessary when $c \neq -p/(1-p)$, as otherwise the average value of any input is zero, and so the quenching of the weights cannot affect the overall positioning of the two distributions. This is true for the standard Hopfield rule ($p = r = 0.5, c = -1$), but not for any $\{0, 1\}$ version of the Hebb rule. Also, the effect of the different values of c is not confined to this particular model of associative memory. Buckingham's (1991) work on sparsely connected Associative Nets,

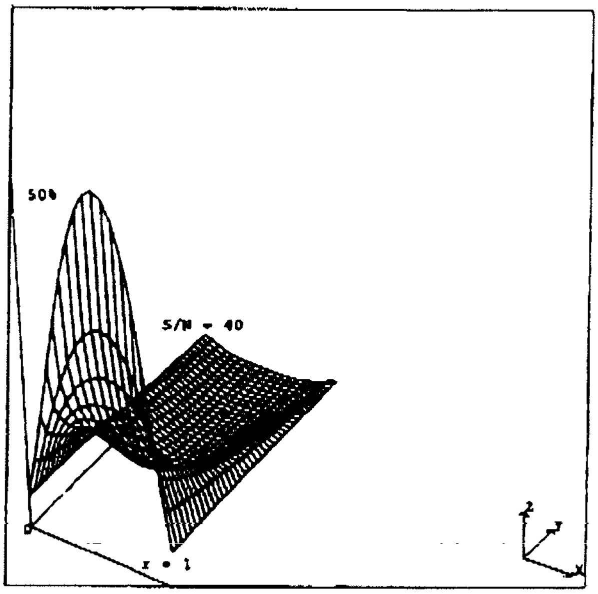


Fig. 4. Expected error rates using the optimal threshold as a function of r and the S/N

following Marr's model of the hippocampus (Marr 1971), shows exactly the same phenomenon, and indeed its extinction when this particular value of c is used. Marr was apparently not aware of this effect.

6 Physics models

The physics connectionist community became interested in exactly the sort of issues aired in this paper, at about the same time as Palm was developing his analysis. They were essentially responding to the poor performance of the Hopfield learning rule ($\alpha = \delta = 1, \beta = \gamma = -1$) for values of p and r (which are simply related to a quantity called the magnetisation) other than 1/2. The first papers were by Tsodyks and Feigel'man (1988) and Buhmann et al. (1989), who studied the case of asymptotic sparsity $p = r \rightarrow 0$, followed by Perez-Vincente and Amit (1989), who published on the case of general $p = r$. Another important contribution came from Gardner (1988), who showed how many patterns **any** such network can store, and how this depends on the magnetisation.

Having assumed a covariance rule, Tsodyks and Feigel'man claimed that:

> 'It should be borne in mind that the "old" theories of associative memory were formulated in terms of the [{0, 1}] model, which seems to be most natural. Then, however, it was replaced by the [{−1, 1}] model without careful analysis of their equivalence. The results of our paper give rise to an amazing conclusion that in some cases such "obvious" simplification may drastically affect the performance of the neural networks.' (Tsodyks and Feigel'man 1988) p. 105.

However, Sect. 3 showed that changing the value of c from 0 to −1, which is equivalent to changing from the {0, 1} to the {−1, 1} model, makes no difference in this model to the ability of the unit to discriminate between low and high patterns. There is thus something rather uncertain about any such remarkable performance in terms of the number of patterns the unit can store. Also, note that the mere number of patterns is not necessarily the appropriate way to judge a learning rule. Sparse patterns inherently contain less information than dense ones (there is less uncertainly as any element is more likely to be a 0), so storing more of them may not increase the informational efficiency. Gardner (1988) showed just this – although as $p = r \rightarrow 0$, the theoretical maximum increases without bound, of the number of potentially retrievable patterns that can be stored, the total information stored by the network actually decreases *whatever* the learning rule.

Perez-Vincente and Amit also assume a covariance rule, and conclude that, in the notation of this paper, the variance of the noise is (Amit, personal communication)

$$\sigma^2 = \{p(\delta - \gamma)^2 + (1 - p)(\beta - \alpha)^2 + \frac{p(1 - p)}{r(1 - r)} \times [r(\delta - \gamma) - r(\beta - \alpha) + (\gamma - \alpha)]^2\} ,$$

that the S/N is essentially ϱ_{γ}, and that it is essential to take $c = -p/(1 - p)$. This also seems to contradict the findings above.

Three possibilities for explaining this divergence spring to mind: the difference between heteroassociative and autoassociative nets or between S/N and mean field analyses, or the non-isomorphism of the underlying models. Although the model considered here is hetero-associative, and the physics models autoassociative, it turns out that this should make no difference so long as the 'identity' synapses, i.e. the diagonal terms in the connection matrix, are absent. In the S/N analysis, they introduce correlations that swamp out all the contributions from the other synapses, whereas for autoassociation their absence is also required for there to be an energy function describing the trajectory of the system as it stabilises to a memory.

S/N studies are generally used as a preliminary to the more exact mean field analyses, and also to confirm their results. The mean field analysis itself is only true in the limit of very many inputs, whereas the S/N can be calculated for finite systems. An example of where this might be important is the threshold; in the limit, the two distributions shown in Fig. 1 are either infinitely far apart or totally indistinguishable, and so the threshold is essentially irrelevant. This turns out to be the case for the mean field analysis too, but it is obviously not true for any finite system.

As became evident in a series of discussions with Daniel Amit, the results really differ because the models do too. The three physics papers mentioned above, apart from Gardner's, all consider the effects of inputting one pattern into a whole set of output units, each of which has learnt its own associations independently of the others, but is entrained to have the same threshold as all the others. The linear associative model considers the effects of inputting many patterns into a single output unit. The rationale behind this is that there is no necessary connection between one unit and the next, and so no a priori reason to tie the threshold for one unit to that of another. The S/N measures the theoretical capability of a single unit to discriminate between its output, **not** the capability of some 'average' unit, which, in principle, cannot exist.

This difference between the models explains the divergence of the results. Lumping together a whole set of output units forces one to measure the variance rather than the dispersion of the dendritic sums, and so to ignore the helpful correlations between them which would be evident for any single unit in isolation. Setting $c = -p/(1 - p)$, the 'optimum' identified by Perez-Vincente and Amit, eliminates the helpful correlations, and so makes the two types of model perform equally well. Likewise, a common threshold can be set across all the units, determined only by the statistics of the associations, rather than their actual values. The 'amazing conclusion' has been reduced to something more mundane. Equivalently, it is well known that the {0, 1} model can apparently store roughly half as many patterns as the {−1, 1} model in the standard Hopfield case (where $p = r = 1/2$), but again this is almost an artefact.

Furthermore, the Associative Net (Willshaw 1971) uses the same threshold for all the units, namely the number of on bits in the input. This is again due to its rather anomalous form.

Tsodyks and Feigel'man analyse the case in which $p \to 0$. For this case, the results here would predict the optimal value of c to be $-p/(1-p)$, which also tends to 0. As seen in the quotation, they actually use $\{0, 1\}$ patterns, which are only asymptotically optimal, but find that this is adequate given the particular manner in which the limit is approached.

Interestingly, Gardner used the linear associative model for her analysis. She therefore also treats the threshold in a different manner from Perez-Vincente and Amit, not needing to introduce it in the first instance. This allowed her rather more elegant results.

7 Conclusions

Adopting the criterion of maximising the signal/noise ratio (S/N) for a class of very simple associative matrix memories leads to one optimal, and two sub-optimal, learning rules. Each of these, a covariance, a heterosynaptic, and a homosynaptic rule, has previously been proposed, but they have not previously been analysed in a common fashion. The covariance rule performs better than the other two, but only negligibly so in the limit of sparse coding. Unlike the other two rules, it also requires synapses to increase in efficacy even if their pre- and post-synaptic units are silent. All the rules have the automatic consequence that the average value of a synapse should be zero, to suppress noise, and so require synapses to take both positive and negative values. The threshold may be set according to an additional criterion, such as minimising the probability of an error, but certain of these criteria may not be monotonic in the S/N.

The rules here, and the lack of dependency of the S/N on the input values for patterns, differ from previous analyses. Some of these analyses are incorrect, ignoring vital correlations in the noise terms. Other analyses are correct, but are based on a different model. The key characteristic discussed here is that each unit is evaluated independently, and so can set its own threshold to allow for its particular quenched weights. Other analyses have lumped collections of output units together, and awarded them the same thresholds. This reduces the apparent quality of the memory quite markedly, unless one particular relationship holds between the high and low values of the input patterns. In that case they perform identically.

Acknowledgements. We are indebted to Alessandro Treves for reading the manuscript and for pointing out an inexactitude in the argument. This work was supported by an MRC Programme Grant to DJ Willshaw and RGM Morris and by an SERC studentship to P Dayan, on a chapter of whose thesis this paper is based

Appendix: calculating the dispersion

Having chosen the expected dispersion (13), repeated below for reference,

$$s_h^2 = \mathcal{E}\left[\frac{1}{\mathcal{N}_h} \sum_{\{\omega|\theta(\omega) = h\}} [\mathrm{d}(\omega)]^2 - \left(\frac{1}{\mathcal{N}_h} \sum_{\{\omega|\theta(\omega) = h\}} \mathrm{d}(\omega)\right)^2\right],$$

as the appropriate measure to calculate, it remained to calculate it.

First, it is necessary to reduce the expression above to a more manageable form. Let $\omega_{h_1}, \ldots, \omega_{h_{\mathcal{N}_h}}$ be the patterns that should be classified as high, and $\omega_{l_1}, \ldots, \omega_{l_{\mathcal{N}_l}}$ be those that should be classified as low. Expanding out the square, we get:

$$\mathcal{E}\left[\frac{1}{\mathcal{N}_h}[\mathrm{d}^2(\omega_{h_1}) + \cdots + \mathrm{d}^2(\omega_{h_{\mathcal{N}_h}}) - \frac{1}{\mathcal{N}_h}\{\mathrm{d}^2(\omega_{h_1}) + \cdots + \mathrm{d}^2(\omega_{h_{\mathcal{N}_h}})\right.$$

$$\left. - (\mathrm{d}(\omega_{h_1})\mathrm{d}(\omega_{h_2}) + \cdots + \mathrm{d}(\omega_{h_{\mathcal{N}_h - 1}})\mathrm{d}(\omega_{h_{\mathcal{N}_h}}))\}]\right]$$

$$= \mathcal{E}\left[\frac{\mathcal{N}_h - 1}{\mathcal{N}_h}(\mathrm{d}^2(\omega_{h_1}) - \mathrm{d}(\omega_{h_1})\mathrm{d}(\omega_{h_2}))\right]$$

since $\mathcal{E}[\mathrm{d}^2(\omega_{h_1})] = \mathcal{E}[\mathrm{d}^2(\omega_{h_2})]$, $\mathcal{E}[\mathrm{d}(\omega_{h_1})\mathrm{d}(\omega_{h_2})] = \mathcal{E}[\mathrm{d}(\omega_{h_3})\mathrm{d}(\omega_{h_4})]$, etc, for a given value of $\mathcal{N}_h$.

It is far more difficult to calculate this quantity than to calculate

$$\mathcal{E}[\mathrm{d}^2(\omega_{h_1}) - \mathrm{d}(\omega_{h_1})\mathrm{d}(\omega_{h_2})]$$

which, by comparison with traditional statistics, might more naturally be regarded as the sample dispersion. In any case, the difference between the two expressions is negligible, so we shall proceed with the latter.

To calculate this, it is necessary to adopt a new notation, which will only be used here. Define:

$$(a_1(\omega_{h_1}), a_2(\omega_{h_1}), \ldots, a_m(\omega_{h_1})) \equiv (a_1, a_2, \ldots, a_m),$$

$$(a_1(\omega_{h_2}), a_2(\omega_{h_2}), \ldots, a_m(\omega_{h_2})) \equiv (b_1, b_2, \ldots, b_m),$$

$$(\Delta_1(\omega_{h_1}), \Delta_2(\omega_{h_1}), \ldots, \Delta_m(\omega_{h_1})) \equiv (A_1, A_2, \ldots, A_m),$$

$$(\Delta_1(\omega_{h_2}), \Delta_2(\omega_{h_2}), \ldots, \Delta_m(\omega_{h_2})) \equiv (B_1, B_2, \ldots, B_m),$$

$$(\Delta_1(\omega_{h_j}), \Delta_2(\omega_{h_j}), \ldots, \Delta_m(\omega_{h_j})) \equiv (C_1, C_2, \ldots, C_m),$$

$$(\Delta_1(\omega_{l_{\mathcal{N}_l - 1}}), \Delta_2(\omega_{l_{\mathcal{N}_l - 1}}), \ldots, \Delta_m(\omega_{l_{\mathcal{N}_l - 1}})) \equiv (Y_1, Y_2, \ldots, Y_m),$$

$$(\Delta_1(\omega_{l_{\mathcal{N}_l}}), \Delta_2(\omega_{l_{\mathcal{N}_l}}), \ldots, \Delta_m(\omega_{l_{\mathcal{N}_l}})) \equiv (Z_1, Z_2, \ldots, Z_m).$$

The point of these is that **a** and **b** represent the input values of two high patterns for $\mathrm{d}(\omega_{h_1})$ and $\mathrm{d}(\omega_{h_2})$, that **A** and **B** represent their associated contributions to the synaptic efficacy, that **C** represents the paradigmatic contribution to the synaptic weight of a high pattern (other than ω_{h_1} or ω_{h_2}), and that **Y** and **Z** represent the paradigmatic contribution to the weight from two different low patterns.

We then have,

$$\mathrm{d}^2(\omega_{h_1}) = [(a_1 A_1 + \cdots + a_m A_m) + (a_1 B_1 + \cdots + a_m B_m)$$

$$+ \cdots + (a_1 Z_1 + \cdots + a_m Z_m)]^2$$

Taking the expectation of this quantity is then merely a tedious process. Using the definitions:

$$\phi = p\delta + (1-p)\beta$$

$$\psi = p\gamma + (1-p)\alpha$$

$$\sigma = p + c(1-p) \qquad (= \mathcal{E}[a_1])$$

$$\kappa = p + c^2(1-p) \qquad (= \mathcal{E}[a_1^2])$$

we derive the quantities in Table 8.

Table 8. Contributions to $\mathcal{E}[d^2(\omega_{k,1})]$

Number of terms	m		$m(m-1)$	
	Paradigm	$\mathcal{E}$	Paradigm	$\mathcal{E}$
1	$a_1^2A_1^2$	$p\delta^2 + c^2(1-p)\beta^2$	$a_1A_1a_mA_m$	$(p\delta + c(1-p)\beta)^2$
$\mathcal{N}_k - 1$	$a_1^2B_1^2$	$\pi(p\delta^2 + (1-p)\beta^2)$	$a_1B_1a_mB_m$	$\sigma^2\phi^2$
$\mathcal{N}_l$	$a_1^2Z_1^2$	$\pi(p\gamma^2 + (1-p)\alpha^2)$	$a_1Z_1a_mZ_m$	$\sigma^2\psi^2$
$2(\mathcal{N}_k - 1)$	$a_1^2A_1B_1$	$(p\delta + c^2(1-p)\beta)\phi$	$a_1A_1a_mB_m$	$(p\delta + c(1-p)\beta)\sigma\phi$
$2\mathcal{N}_l$	$a_1^2A_1Z_1$	$(p\delta + c^2(1-p)\beta)\psi$	$a_1A_1a_mZ_m$	$(p\delta + c(1-p)\beta)\sigma\psi$
$(\mathcal{N}_k - 1)(\mathcal{N}_k - 2)$	$a_1^2B_1C_1$	$\pi\phi^2$	$a_1B_1a_mC_m$	$\sigma^2\phi^2$
$2\mathcal{N}_l(\mathcal{N}_k - 1)$	$a_1^2B_1Z_1$	$\pi\phi\psi$	$a_1B_1a_mZ_m$	$\sigma^2\phi\psi$
$\mathcal{N}_l(\mathcal{N}_l - 1)$	$a_1^2Y_1Z_1$	$\pi\psi^2$	$a_1Y_1a_mZ_m$	$\sigma^2\psi^2$

Table 9. Contributions to $\mathcal{E}[d(\omega_{k,1})d(\omega_{k,2})]$

Number of terms	m		$m(m-1)$	
	Paradigm	$\mathcal{E}$	Paradigm	$\mathcal{E}$
2	$a_1A_1^2b_1$	$(p\delta^2 + c(1-p)\beta^2)\sigma$	$a_1A_1b_mA_m$	$(p\delta + c(1-p)\beta)\sigma\phi$
1	$a_1A_1b_1B_1$	$(p\delta + c(1-p)\beta)^2$	$a_1A_1b_mB_m$	$(p\delta + c(1-p)\beta)^2$
1	$a_1B_1b_1A_1$	$(p\delta + c(1-p)\beta)^2$	$a_1B_1b_mA_m$	$\sigma^2\phi^2$
$2(\mathcal{N}_k - 2)$	$a_1A_1b_1C_1$	$(p\delta + c(1-p)\beta)\sigma\phi$	$a_1A_1b_mC_m$	$(p\delta + c(1-p)\beta)\sigma\phi$
$2\mathcal{N}_l$	$a_1A_1b_1Z_1$	$(p\delta + c(1-p)\beta)\sigma\psi$	$a_1A_1b_mZ_m$	$(p\delta + c(1-p)\beta)\sigma\psi$
$2(\mathcal{N}_k - 2)$	$a_1B_1b_1C_1$	$(p\delta + c(1-p)\beta)\sigma\phi$	$a_1B_1b_mC_m$	$\sigma^2\phi^2$
$2\mathcal{N}_l$	$a_1B_1b_1Z_1$	$(p\delta + c(1-p)\beta)\sigma\psi$	$a_1B_1b_mZ_m$	$\sigma^2\phi\psi$
$(\mathcal{N}_k - 2)$	$a_1C_1^2b_1$	$\sigma^2(p\delta^2 + (1-p)\beta^2)$	$a_1C_1b_mC_m$	$\sigma^2\phi^2$
$\mathcal{N}_l$	$a_1Z_1^2b_1$	$\sigma^2(p\gamma^2 + (1-p)\alpha^2)$	$a_1Z_1b_mZ_m$	$\sigma^2\psi^2$
$(\mathcal{N}_k - 2)(\mathcal{N}_k - 3)$	$a_1C_1b_1D_1$	$\sigma^2\phi^2$	$a_1C_1b_mD_m$	$\sigma^2\phi^2$
$2(\mathcal{N}_k - 2)\mathcal{N}_l$	$a_1C_1b_1Z_1$	$\sigma^2\phi\psi$	$a_1C_1b_mZ_m$	$\sigma^2\phi\psi$
$\mathcal{N}_l(\mathcal{N}_l - 1)$	$a_1Y_1b_1Z_1$	$\sigma^2\psi^2$	$a_1Y_1b_mZ_m$	$\sigma^2\psi^2$

Summing up the terms, we get:

$$\begin{aligned}
\mathcal{E}[d^2(\omega_{k,1})] = m\mathcal{E}[\pi[\mathcal{N}_k\phi + \mathcal{N}_l\psi]^2 &+ \pi p(1-p)[\mathcal{N}_k(\delta-\beta)^2 \\
&+ \mathcal{N}_l(\gamma-\alpha)^2] + 2p(1-p)(1-c^2)(\delta-\beta)[\mathcal{N}_k\phi + \mathcal{N}_l\psi] \\
&+ p(1-p)(1-c^2)(1-2p)(\delta-\beta)^2] \\
&+ m(m-1)\mathcal{E}[\sigma^2[\mathcal{N}_k\phi + \mathcal{N}_l\psi]^2 \\
&+ 2p(1-p)(1-c)(\delta-\beta)\sigma[\mathcal{N}_k\phi + \mathcal{N}_l\psi] \\
&+ p^2(l-p)^2(l-c)^2(\delta-\beta)^2]
\end{aligned}$$

and similarly:

$$\begin{aligned}
\mathcal{E}[d^2(\omega_{l,1})] = m\mathcal{E}[\pi[\mathcal{N}_k\phi + \mathcal{N}_l\psi]^2 &+ \pi p(1-p)[\mathcal{N}_k(\delta-\beta)^2 \\
&+ \mathcal{N}_l(\gamma-\alpha)^2] + 2p(1-p)(1-c^2)(\gamma-\alpha)[\mathcal{N}_k\phi + \mathcal{N}_l\psi] \\
&+ p(1-p)(1-c^2)(1-2p)(\gamma-\alpha)^2] \\
&+ m(m-1)\mathcal{E}[\sigma^2[\mathcal{N}_k\phi + \mathcal{N}_l\psi]^2 \\
&+ 2p(1-p)(1-c)(\gamma-\alpha)\sigma[\mathcal{N}_k\phi + \mathcal{N}_l\psi] \\
&+ p^2(1-p)^2(1-c)^2(\gamma-\alpha)^2]
\end{aligned}$$

Now we have.

$$\begin{aligned}
d(\omega_{k,1})d(\omega_{k,2}) = {}&[(a_1A_1 + \cdots + a_mA_m) + \cdots + (a_1Z_1 + \cdots + a_mZ_m)] \\
&\times [(b_1A_1 + \cdots + b_mA_m) + \cdots + (b_1Z_1 + \cdots + b_mZ_m)].
\end{aligned}$$

Summing these terms as in Table 9. we derive:

$$\begin{aligned}
\mathcal{E}[d(\omega_{k,1})d(\omega_{k,2})] = m\mathcal{E}[\sigma^2[\mathcal{N}_k\phi + \mathcal{N}_l\psi]^2& \\
&+ 4p(1-p)(1-c)(\delta-\beta)\sigma(\mathcal{N}_k\phi + \mathcal{N}_l\psi) \\
&+ p(1-p)\sigma^2[\mathcal{N}_k(\delta-\beta)^2 + \mathcal{N}_l(\gamma-\alpha)^2] \\
&+ 2p^2(1-p)^2(1-c)^2(\delta-\beta)^2 \\
&+ 2\sigma p(1-p)(1-c)(1-2p)(\delta-\beta)^2] \\
&+ m(m-1)\mathcal{E}[\sigma^2[\mathcal{N}_k\phi + \mathcal{N}_l\psi]^2 \\
&+ 2p(1-p)(1-c)(\delta-\beta)\sigma[\mathcal{N}_k\phi + \mathcal{N}_l\psi] \\
&+ p^2(1-p)^2(1-c)^2(\delta-\beta)^2]
\end{aligned}$$

and:

$$\begin{aligned}
\mathcal{E}[d(\omega_{l,1})d(\omega_{l,2})] = m\mathcal{E}[\sigma^2[\mathcal{N}_k\phi + \mathcal{N}_l\psi]^2& \\
&+ 4p(1-p)(1-c)(\gamma-\alpha)\sigma(\mathcal{N}_k\phi + \mathcal{N}_l\psi) \\
&+ p(1-p)\sigma^2[\mathcal{N}_k(\delta-\beta)^2 + \mathcal{N}_l(\gamma-\alpha)^2] \\
&+ 2p^2(1-p)^2(1-c)^2(\gamma-\alpha)^2 \\
&+ 2\sigma p(1-p)(1-c)(1-2p)(\gamma-\alpha)^2] \\
&+ m(m-1)\mathcal{E}[\sigma^2[\mathcal{N}_k\phi + \mathcal{N}_l\psi]^2 \\
&+ 2p(1-p)(1-c)(\gamma-\alpha)\sigma[\mathcal{N}_k\phi + \mathcal{N}_l\psi] \\
&+ p^2(1-p)^2(1-c)^2(\gamma-\alpha)^2].
\end{aligned}$$

So, as in (14):

$$\begin{aligned}
\mathcal{E}[d^2(\omega_{k,1}) - d(\omega_{k,1})d(\omega_{k,2})] = mp(1-p)(1-c)^2& \\
&\times \mathcal{E}[p(1-p)[\mathcal{N}_k(\delta-\beta)^2 \\
&+ \mathcal{N}_l(\gamma-\alpha)^2] + [\mathcal{N}_k\phi + \mathcal{N}_l\psi]^2 \\
&+ 2(1-2p)(\delta-\beta)[\mathcal{N}_k\phi + \mathcal{N}_l\psi] \\
&+ [(1-2p)^2 - 2p(1-p)](\delta-\beta)^2]
\end{aligned}$$

and similarly:

$$\begin{aligned}
\mathcal{E}[d^2(\omega_{l,1}) - d(\omega_{l,1})d(\omega_{l,2}) = mp(1 - p(1-c)^2& \\
&\times \mathcal{E}[p(1-p)[\mathcal{N}_k(\delta-\beta)^2 \\
&+ \mathcal{N}_l(\gamma-\alpha)^2] + [\mathcal{N}_k\phi + \mathcal{N}_l\psi]^2 \\
&+ 2(1-2p)(\gamma-\alpha)[\mathcal{N}_k\phi + \mathcal{N}_l\psi] \\
&+ [(1-2p)^2 - 2p(1-p)](\gamma-\alpha)^2]
\end{aligned}$$

The correctness of the summation performed using the terms of Tables 8 and 9 was verified using the computer program REDUCE.

References

Bienenstock E, Cooper LN, Munro P (1982) Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. J Neurosci 2:32–48

Bliss TVP, Lømo T (1973) Long-lasting potentiation of synaptic transmission in the dentate area of the anæsthetized rabbit following stimulation of the perforant path. J Physiol (London) 232:331–356

Buckingham JT (1991) Computation in the archicortex. PhD Thesis, Department of Artificial Intelligence, University of Edinburgh

Buhman J, Divko R, Schulten K (1989) On sparsely coded associative memories. In: Personnaz L, Dreyfus G (eds) Neural networks: from models to applications. nEURO88, Paris

Duda RO, Hart PE (1973) Pattern classification and scene analysis. Wiley, New York

Gaffan D (1974) Recognition impaired and association intact in the memory of monkeys after transection of the fornix. J Comp Physiol Psychol 86:1100–1109

Gardner E (1988) The space of interactions in neural network models. J Phys A: Math Gen 21:257

Hancock PJB, Smith LS, Phillips WA (1991) A biologically supported error-correcting learning rule. Neural Computation (in press)

Hebb DO (1949) The organization of behavior: a neuropsychological theory. Wiley, New York

Hopfield JJ (1982) Neural networks and physical systems with emergent computational abilities. Proc Natl Acad Sci 79:2554–2558

Ito M, Sakurai M, Tongroach P (1982) Climbing fibre induced depression of both mossy fibre responsiveness and glutamate sensitivity of cerebellar Purkinje cells. J Physiol 324:113–134

Kanerva P (1988) Sparse distributed memory. MIT Press, Cambridge

Kohonen T (1972) Correlation matrix memories. IEEE Trans Comput C-21:353–359

Marr D (1971) Simple memory: a theory for archicortex. Philos Trans Soc London B 262:23–81

Palm G (1988a) On the asymptotic information storage capacity of neural networks. In: Eckmiller R, von der Malsburg C (eds) Neural computers. NATO ASI Series F41. Springer, Berlin, Heidelberg New York, pp 271–280

Palm G (1988b) Local synaptic rules with maximal information storage capacity. In: Haken H (ed) Neural and synergetic computers Springer Series in Synergetics, vol 42. Springer, Berlin, Heidelberg New York, pp 100–110

Perez-Vincente CJ, Amit DJ (1989). Optimised network for sparsely coded patterns. J Phys A: Math Gen 22:559–569

Racine RJ, Milgram NW, Hafner S (1983) Long-term potentiation phenomena in the rat limbic forebrain. Brain Res 260:217–231

Sejnowski TJ (1977a) Storing covariance with nonlinearly interacting neurons. J Math Biol 4:303–321

Sejnowski TJ (1977b) Statistical constraints on synaptic plasticity. J Theor Biol 69:385–389

Singer W (1985) Activity-dependent self-organization of synaptic connections as a substrate of learning. In: Changeux JP, Konishi M (eds) The neural and molecular bases of learning. Wiley, New York, pp 301–335

Stanton P, Sejnowski TJ (1989) Associative long-term depression in the hippocampus: Induction of synaptic plasticity by Hebbian covariance. Nature 339:215–18

Stent GS (1973) A physiological mechanism for Hebb's postulate of learning. Proc Natl Acad Sci 70:997–1001

Tsodyks MV, Fiegel'man MV (1988) The enhanced storage capacity in neural networks with low activity level. Europhys Lett 6:101–105

Willshaw DJ (1971) Models of distributed associative memory. PhD Thesis, University of Edinburgh

Willshaw DJ, Dayan P (1990) Optimal plasticity in matrix memories: what goes up MUST come down. Neural Comput 2:85–93

Willshaw DJ, Buneman OP, Longuet-Higgins HC (1969) Nonholographic associative memory. Nature 222:960–962

Dr. David Willshaw
Centre for Cognitive Science
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW
United Kingdom