Communicated by Terrence J. Sejnowski

Optimal Plasticity from Matrix Memories: What Goes Up Must Come Down

David Willshaw
Peter Dayan
Centre for Cognitive Science and Department of Physics, University of Edinburgh, Edinburgh, Scotland

A recent article (Stanton and Sejnowski 1989) on long-term synaptic depression in the hippocampus has reopened the issue of the computational efficiency of particular synaptic learning rules (Hebb 1949; Palm 1988a; Morris and Willshaw 1989) — homosynaptic versus heterosynaptic and monotonic versus nonmonotonic changes in synaptic efficacy. We have addressed these questions by calculating and maximizing the signal-to-noise ratio, a measure of the potential fidelity of recall, in a class of associative matrix memories. Up to a multiplicative constant, there are three optimal rules, each providing for synaptic depression such that positive and negative changes in synaptic efficacy balance out. For one rule, which is found to be the Stent–Singer rule (Stent 1973; Rauschecker and Singer 1979), the depression is purely heterosynaptic; for another (Stanton and Sejnowski 1989), the depression is purely homosynaptic; for the third, which is a generalization of the first two, and has a higher signal-to-noise ratio, it is both heterosynaptic and homosynaptic. The third rule takes the form of a covariance rule (Sejnowski 1977a,b) and includes, as a special case, the prescription due to Hopfield (1982) and others (Willshaw 1971; Kohonen 1972).

In principle, the association between the synchronous activities in two neurons could be registered by a mechanism that increases the efficacy of the synapses between them, in the manner first proposed by Hebb (1949); the generalization of this idea to the storage of the associations between activity in two sets of neurons is in terms of a matrix of modifiable synapses (Anderson 1968; Willshaw *et al.* 1969; Kohonen 1972). This type of architecture is seen in the cerebellum (Eccles *et al.* 1968) and in the hippocampus (Marr 1971) where associative storage of the Hebbian type (Bliss and Lømo 1973) has been ascribed to the NMDA receptor (Collingridge *et al.* 1983; Morris *et al.* 1986). A number of questions concerning the computational power of certain synaptic

modification rules in matrix memories have direct biological relevance. For example, is it necessary, or merely desirable, to have a rule for decreasing synaptic efficacy under conditions of asynchronous firing, to complement the increases prescribed by the pure Hebbian rule (Hebb 1949)? The need for a mechanism for decreasing efficacy is pointed to by general considerations, such as those concerned with keeping individual synaptic efficacies within bounds (Sejnowski 1977b); and more specific considerations, such as providing an explanation for ocular dominance reversal and other phenomena of plasticity in the visual cortex (Bienenstock *et al.* 1982; Singer 1985). There are two types of asynchrony between the presynaptic and the postsynaptic neurons that could be used to signal a decrease in synaptic efficacy (Sejnowski 1977b; Sejnowski *et al.* 1988): the presynaptic neuron might be active while the postsynaptic neuron is inactive (homosynaptic depression), or vice versa (heterosynaptic depression).

We have explored the theoretical consequences of such issues. We consider the storage of a number Ω of pattern pairs [represented as the binary vectors $\mathbf{A}(\omega)$ and $\mathbf{B}(\omega)$ of length m and n, respectively] in a matrix associative memory. The matrix memory has m input lines and n output lines, carrying information about the **A**-patterns and the **B**-patterns, respectively, each output line being driven by a linear threshold unit (LTU) with m variable weights (Fig. 1). Pattern components are generated independently and at random. Each component of an **A**-pattern takes the value 1 (representing the active state) with probability p and the value c (inactive state) with probability $1 - p$. Likewise, the probabilities for the two possible states 1 and c for a component of a **B**-pattern are r, $1 - r$. In the storage of the association of the ωth pair, the amount Δ_{ij} by which the weight W_{ij} is changed depends on the values of the pair of numbers $[A_i(\omega), B_j(\omega)]$.

Once the entire set of patterns has been learned, retrieval of a previously stored **B**-pattern is effected by the presentation of the corresponding **A**-pattern. The jth LTU calculates the weighted sum of its inputs, $d_j[\mathbf{A}(\omega)]$,

$$d_j[\mathbf{A}(\omega)] = \sum_{i=1}^{m} A_i(\omega) W_{ij} = \sum_{i=1}^{m} A_i(\omega) \sum_{\omega'=1}^{\Omega} \Delta_{ij}(\omega')$$

The state of output line j is then set to c or 1, according to whether $d_j[\mathbf{A}(\omega)]$ is less than or greater than the threshold θ_j.

The signal-to-noise ratio ρ is a measure of the ability of an LTU to act as a discriminator between those $\mathbf{A}(\omega)$ that are to elicit the output c and those that are to elicit the output 1. It is a function of the parameters of the system, and is calculated by regarding $d_j[\mathbf{A}(\omega)]$ as the sum of two components: the signal, $s_j(\omega)$, which stems from that portion of the

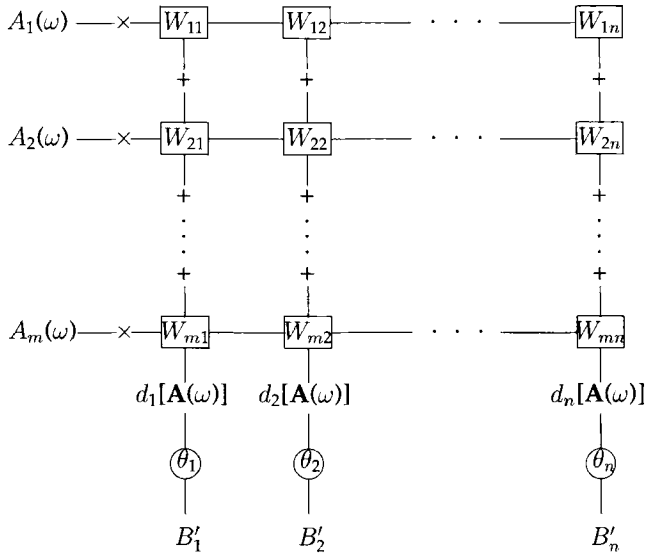

Figure 1: The matrix memory, which associates **A**-patterns with **B**-patterns. Each weight W_{ij} is a linear combination over the patterns:

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

where Δ is given in the table below.

$\Delta_{ij}(\omega)$		Output $B_j(\omega)$	
		c	1
Input $A_i(\omega)$	c	α	β
	1	γ	δ

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 $\mathbf{B}'$ must resemble the desired output $\mathbf{B}(\omega)$.

weights arising from the storage of pattern ω, and the noise, $n_j(\omega)$, which is due to the contribution from all the other patterns to the weights.

$$d_j[\mathbf{A}(\omega)] = \sum_{i=1}^{m} A_i(\omega)\Delta_{ij}(\omega) + \qquad \{s_j(\omega)\}$$
$$\sum_{i=1}^{m} \sum_{\omega'=1,\omega'\neq\omega}^{\Omega} A_i(\omega)\Delta_{ij}(\omega') \qquad \{n_j(\omega)\} \tag{1.1}$$

In most applications of signal-to-noise (S/N) analysis, the noise terms have the same mean and are uncorrelated between different patterns. When these assumptions are applied to the current model, maximizing the signal-to-noise ratio with respect to the learning rule parameters α, β, γ, and δ, leaves them dependent on the parameter c (Palm 1988b). However, the mean of the noise $n_j(\omega)$ in equation 1 is biased by the exclusion of the contribution $\Delta_{ij}(\omega)$, whose value depends on the target output for pattern ω; and the noise terms for two different patterns ω_1 and ω_2 are in general correlated through the $\Omega - 2$ contributions to the value of $\Delta_{ij}(\omega)$, which occur in both terms. Our analysis (Fig. 2) takes account of these factors, and its validity is confirmed by the results of computer simulation (Table 1). Maximizing the expression we obtain for the signal-to-noise ratio in terms of the learning parameters leads to the three c-independent rules, **R1**, **R2**, and **R3**. To within a multiplicative constant they are

$$\begin{aligned}
\textbf{R1:}\ (\alpha,\beta,\gamma,\delta) &= [pr, -p(1-r), -(1-p)r, (1-p)(1-r)],\\
\rho &= \frac{m}{\Omega}\frac{1}{r(1-r)}\\
\textbf{R2:}\ (\alpha,\beta,\gamma,\delta) &= (0, -p, 0, 1-p),\\
\rho &= \frac{m}{\Omega}\frac{1}{r}\\
\textbf{R3:}\ (\alpha,\beta,\gamma,\delta) &= (0, 0, -r, 1-r)\\
\rho &= \frac{m}{\Omega}\frac{1}{r}\frac{1-p}{1-r}
\end{aligned}$$

Rule **R1** is a generalization of the Hebb rule, called the covariance rule (Sejnowski 1977a; Sejnowski *et al.* 1988; Linsker 1986). In this formulation, the synaptic efficacy between any two cells is changed according to the product of the deviation of each cell's activity from the mean. When pattern components are equally likely to be in the active and the inactive states ($p = r = 1/2$), **R1** takes the form of the "Hopfield" rule (Hopfield 1982), and has the lowest signal-to-noise ratio of all such rules. Rule **R1** prescribes changes in efficacy for all of the four possible states of activity seen at an individual synapse, and thus utilizes both heterosynaptic and homosynaptic asynchrony. It also has the biologically undesirable property that changes can occur when neither pre- nor postsynaptic neuron is

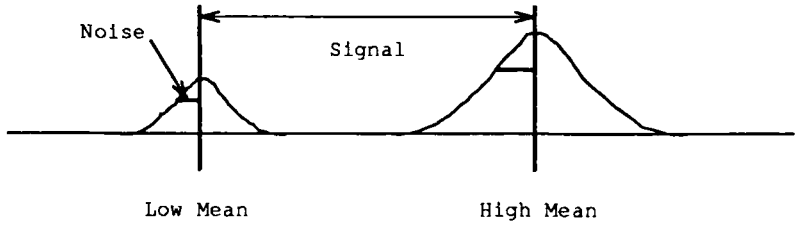

Figure 2: Signal-to-noise ratios. The frequency graph of its linear combinations $d(\omega)$ for a given LTU. The two classes to be distinguished appear as approximately Gaussian distributions, with high mean μ_h, low mean μ_l, and variances σ_h^2, σ_l^2, where $\sigma_h^2 \simeq \sigma_l^2$. For good discrimination the two distributions should be narrow and widely separated.

In our calculation of the signal-to-noise ratio, the mean of the noise $n(\omega)$ (equation 1) differs for high and low patterns, and so the expressions for the expected values of μ_h and μ_l were calculated separately. Second, the correlations between the noise terms obscuring different patterns add an extra quantity to the variance of the total noise. The entire graph of the frequency distributions for high and low patterns is displaced from the expected location, by a different amount for each unit. This overall displacement does not affect the power of the unit to discriminate between patterns. In calculating the signal-to-noise ratio, it is therefore appropriate to calculate the expected dispersion of the noise about the mean for each unit, rather than using the variance, which would imply measuring deviations from the expected mean. The expected dispersion for high patterns is defined as

$$s_h^2 = \varepsilon \left[\frac{1}{H} \sum_{\{\omega | B(\omega)=1\}} [d(\omega)]^2 - \left(\frac{1}{H} \sum_{\{\omega | B(\omega)=1\}} d(\omega) \right)^2 \right]$$

H being the number of ω for which $B(\omega) = 1$, and s_l^2 is defined similarly as the expected dispersion for low patterns.

The signal-to-noise ratio for a discriminator is therefore defined as

$$\rho = \frac{(\varepsilon[\mu_h - \mu_l])^2}{\frac{1}{2}(s_h^2 + s_l^2)}$$

It depends on all the parameters of the system, and may be maximized with respect to those that define the learning rule, α, β, γ, and δ. The maxima are found at the rules **R1**, **R2**, and **R3** described in the text.

The effect of changing c is to shift and compress or expand the distributions. For a given LTU, it is always possible to move the threshold with c in such a way that exactly the same errors are made (Table 1a). The choice of c partly determines the variability of μ_h and μ_l across the set of units, and this variability is minimized at $c = -p/(1-p)$. With this value of c, and in the limit of large m, n, and Ω, its effect becomes negligible, and hence the thresholds for all the units may be set equal.

active ($\alpha \neq 0$). However, the change to be applied in the absence of any activity can be regarded as a constant background term of magnitude pr. In rule **R2**, the so-called Stent–Singer rule (Stent 1973; Rauschecker and Singer 1979), depression is purely heterosynaptic. For a given number of stored associations, the signal-to-noise ratio for **R2** is less than that for **R1** by a factor of $1/(1-r)$. In rule **R3**, which Stanton and Sejnowski (1989) proposed for the mossy fibers in the hippocampal CA3 region, and which is also used in theoretical schemes (Kanerva 1988), depression is purely

homosynaptic. **R3** has a signal-to-noise ratio less than **R1** by a factor of $1/(1-p)$. If $p = r$, **R2** and **R3** have the same signal-to-noise ratio.

All the rules have the automatic property that the expected value of each weight is 0; that is, what goes up does indeed come down. One way of implementing this property that avoids the necessity of synapses switching between excitatory and inhibitory states is to assign each synapse a constant positive amount of synaptic efficacy initially. Our results do not apply exactly to this case, but an informal argument suggests that initial synaptic values should be chosen so as to keep the total synaptic efficacy as small as possible, without any value going negative. Given that it is likely that the level of activity in the nervous system is relatively low ($< 10\%$), it is predicted that the amount of (homosynaptic) long-term potentiation (Bliss and Lømo 1973) per nerve cell will be an order of magnitude greater than the amount of either homosynaptic or heterosynaptic depression. Further, under **R1**, any experimental technique for investigating long-term depression that relies on the aggregate effect on one postsynaptic cell of such sparse activity will find a larger heterosynaptic than homosynaptic effect.

As for the Hopfield case (Willshaw 1971; Kohonen 1972; Hopfield 1982), for a given criterion of error (as specified by the signal-to-noise ratio) the number of associations that may be stored is proportional to the size, m, of the network. It is often noted (Willshaw *et al.* 1969; Amit *et al.* 1987; Gardner 1987; Tsodyks and Feigel'man 1988) that the sparser the coding of information (i.e., the lower the probability of a unit being active) the more efficient is the storage and retrieval of information. This is also true for rules **R1**, **R2**, and **R3**, but the information efficiency of the matrix memory, measured as the ratio of the number of bits stored as associations to the number of bits required to represent the weights, is always less than in similar memories incorporating clipped synapses (Willshaw *et al.* 1969), that is, ones having limited dynamic range.

The signal-to-noise ratio measures only the potential of an LTU to recall correctly the associations it has learned. By contrast, the threshold θ_j determines the actual frequency of occurrence of the two possible types of misclassification. The threshold may be set according to some further optimality criterion, such as minimizing the expected number of recall errors for a pattern. For a given LTU, the optimal value of θ will depend directly on the actual associations it has learned rather than just on the parameters generating the patterns, which means that each LTU should have a different threshold. It can be shown that, as m, n, and Ω grow large, setting c at the value $-p/(1-p)$ enables the thresholds of all the LTUs to be set equal (and dependent only on the parameters, not the actual patterns) without introducing additional error.

Although natural processing is by no means constrained to follow an optimal path, it is important to understand the computational consequences of suggested synaptic mechanisms. The signal-to-noise ratio

1a

p,r	c	Expect S/N	Actual S/N $\pm\sigma$	Previous S/N	Expect errors	Actual errors
0.5	-1	10	11 ± 1.3	10	1.1	1.1
0.4	-1	7.5	8.3 ± 1.5	10	1.7	1.6
0.3	-1	1.4	1.3 ± 0.40	11	4.6	4.5
0.2	-1	0.25	0.32 ± 0.22	12	4.0	4.2
0.5	-1	10	11 ± 1.3		1.1	1.1
0.5	-0.5	10	11 ± 1.3		1.1	1.1
0.5	0	10	11 ± 1.3		1.1	1.1
0.5	0.5	10	11 ± 1.3		1.1	1.1

1b

p,r	c	Expect S/N	Actual S/N $\pm\sigma$	Previous S/N	Expect errors	Actual errors
0.5	0	0.05	0.10 ± 0.11	6.8	9.1	8.7
0.4	0	0.11	0.11 ± 0.09	7.6	7.8	7.6
0.3	0	0.31	0.34 ± 0.15	9.4	5.8	5.9
0.2	0	1.1	1.2 ± 0.47	13	3.6	3.4
0.1	0	5.9	5.3 ± 1.8	26	0.92	1.2
0.05	0	16	28 ± 18	51	0.16	0.15

1c

p, r	**R1**	**R2, R3**	Hebb	Hopfield
0.5	10	5.1	0.050	10
0.4	11	6.4	0.11	7.5
0.3	12	8.5	0.31	1.4
0.2	16	13	1.1	0.25
0.1	28	26	5.9	0.045
0.05	54	51	16	0.015

Table 1: Simulations. The object of the simulations was to check the formulae developed in our analysis and compare them with a previous derivation (Palm 1988b). The matrix memory has m = 512 input lines and n = 20 output lines. To ensure noticeable error rates, the number of pattern pairs was set at Ω = 200. In all cases $p = r$.

1a: The Hopfield (1982) rule $(\alpha, \beta, \gamma, \delta) = (1, -1, -1, 1)$. Columns 3 and 4 compare the S/N ratio expected from our analysis and that measured in the simulation, the latter also showing the standard error measured over the output units; column 5 gives the S/N ratio calculated on the basis of previous analysis (Palm 1988b). Columns 6 and 7 compare the expected and measured numbers of errors per pattern, the threshold being set so that the two possible types of error occurred with equal frequency. For good recall (< 0.03 errors per unit) the S/N ratio must be at least 16. The lack of dependence on the value of c is demonstrated in rows 5–8. The same patterns were used in each case.

1b: Similar results for the nonoptimal Hebb (1949) rule $(\alpha, \beta, \gamma, \delta) = (0, 0, 0, 1)$.

1c: Values of the signal-to-noise ratio for the rules **R1**, **R2**, and **R3** and the Hebb and the Hopfield rules. **R1** has higher signal-to-noise ratio than **R2** and **R3**, but for the latter two it is the same since $p = r$ here. The Hebb rule approaches optimality in the limit of sparse coding; conversely, the Hopfield rule is optimal at $p = r = 1/2$.

indicates how good a linear threshold unit may be at its discrimination task, and consequently how much information can be stored by a network of a number of such units. Synaptic depression is important for computational reasons, independent of any role it might play in preventing saturation of synaptic strengths. Up to a multiplicative constant, only three learning rules maximize the signal-to-noise ratio. Each rule involves both decreases and increases in the values of the weights. One rule involves heterosynaptic depression, another involves homosynaptic depression, and in the third rule there is both homosynaptic and heterosynaptic depression. All rules work most efficiently when the patterns of neural activity are sparsely coded.

Acknowledgments

We thank colleagues, particularly R. Morris, T. Bliss, P. Hancock, A. Gardner-Medwin, and M. Evans, for their helpful comments and criticisms on an earlier draft. This research was supported by grants from the MRC and the SERC.

References

Amit, D. J., Gutfreund, H., and Sompolinsky, H. 1987. Information storage in neural networks with low levels of activity. *Phys. Rev. A.* **35**, 2293–2303.

Anderson, J. A. 1968. A memory storage model utilizing spatial correlation functions. *Kybernetik* **5**, 113–119.

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

Bliss, T. V. P. and Lømo, T. 1973. Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *J. Physiol. (London)* **232**, 331–356.

Collingridge, G. L., Kehl, S. J., and McLennan, H. J. 1983. Excitatory amino acids in synaptic transmission in the Schaffer collateral-commissural pathway of the rat hippocampus. *J. Physiol. (London)* **334**, 33–46.

Eccles, J. T., Ito, M., and Szentágothai, J. 1968. *The Cerebellum as a Neuronal Machine.* Springer Verlag, Berlin.

Gardner, E. 1987. Maximum storage capacity of neural networks. *Europhys. Lett.* **4**, 481–485.

Hebb, D. O. 1949. *The Organization of Behavior.* Wiley, New York.

Hopfield, J. J. 1982. Neural networks and physical systems with emergent computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* **79**, 2554–2558.

Kanerva, P. 1988. *Sparse Distributed Memory.* MIT Press/Bradford Books: Cambridge, MA.

Kohonen, T. 1972. Correlation matrix memories. *IEEE Trans. Comput.* **C-21**, 353–359.

Linsker, R. 1986. From basic network principles to neural architecture: Emergence of spatial opponent cells. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 7508–7512.

Marr, D. 1971. Simple memory: A theory for archicortex. *Phil. Trans. R. Soc. London B* **262**, 23–81.

Morris, R. G. M., Anderson, E., Baudry, M., and Lynch, G. S. 1986. Selective impairment of learning and blockade of long term potentiation *in vivo* by AP5, an NMDA antagonist. *Nature (London)* **319**, 774–776.

Morris, R. G. M. and Willshaw, D. J. 1989. Must what goes up come down? *Nature (London)* **339**, 175–176.

Palm, G. 1988a. On the asymptotic information storage capacity of neural networks. In *Neural Computers*, R. Eckmiller, and C. von der Malsburg, eds. NATO ASI Series **F41**, pp. 271–280. Springer Verlag, Berlin.

Palm, G. 1988b. Local synaptic rules with maximal information storage capcity. In *Neural and Synergetic Computers, Springer Series in Synergetics*, H. Haken, ed., Vol. 42, pp. 100–110. Springer-Verlag, Berlin.

Rauschecker, J. P. and Singer, W. 1979. Changes in the circuitry of the kitten's visual cortex are gated by postsynaptic activity. *Nature (London)* **280**, 58–60.

Sejnowski, T. J. 1977a. Storing covariance with nonlinearly interacting neurons. *J. Math. Biol.* **4**, 303–321.

Sejnowski, T. J. 1977b. Statistical constraints on synaptic plasticity. *J. Theor. Biol.* **69**, 385–389.

Sejnowski, T. J., Chattarji, S., and Stanton, P. 1988. Induction of synaptic plasticity by Hebbian covariance. In *The Computing Neuron*, R. Durbin, C. Miall, and G. Mitchison, eds., pp. 105–124. Addison-Wesley, Wokingham, England.

Singer, W. 1985. Activity-dependent self-organization of synaptic connections as a substrate of learning. In *The Neural and Molecular Bases of Learning*, J. P. Changeux and M. Konishi, eds., pp. 301–335. Wiley, New York.

Stanton, P. and Sejnowski, T. J. 1989. Associative long-term depression in the hippocampus: Induction of synaptic plasticity by Hebbian covariance. *Nature (London)* **339**, 215–218.

Stent, G. S. 1973. A physiological mechanism for Hebb's postulate of learning. *Proc. Natl. Acad. Sci. U.S.A.* **70**, 997–1001.

Tsodyks, M. V. and Feigel'man, M. V. 1988. The enhanced storage capacity in neural networks with low activity level. *Europhys. Lett.* **6**, 101–105.

Willshaw, D. J. 1971. Models of Distributed Associative Memory. Ph.D. Thesis, University of Edinburgh.

Willshaw, D. J., Buneman, O. P., and Longuet-Higgins, H. C. 1969. Non-holographic associative memory. *Nature (London)* **222**, 960–962.

Received 28 November 1989; accepted 19 December 1989.