Network 1 (1990) 105–122. Printed in the UK

REVIEW

Learning in neural network memories

L F Abbott
Physics Department, Brandeis University, Waltham, MA 02254, USA

Received 19 September 1989

Abstract. Various algorithms for constructing a synaptic coupling matrix which can associatively map input patterns onto nearby stored memory patterns are reviewed. Issues discussed include performance, capacity, speed, efficiency and biological plausibility.

1. Introduction

The term 'learning' is applied to a wide range of activities associated with the construction of neural networks ranging from single-layer binary classifiers [1] to multilayered systems performing relatively sophisticated tasks [2]. Any reviewer hoping to cover this field in a reasonable amount of time and space must do so with a severely restricted viewpoint. Here, I will concentrate on a fairly simple task, associative memory, accomplished by a single-layered iterative network of binary elements [3–5]. This area is considered because there are now available a large number of precise analytic results and a wealth of ideas and approaches have appeared and been analysed in detail.

Most neural network modelling relies crucially on the assumption that synaptic plasticity [6] is a (or perhaps *the*) key component in the remarkable adaptive behaviour of biological networks. The various unrealistic simplifications made in the construction of mathematical models are more palatable if viewed in this light. In fact, we might say that the fundamental goal of neural network research is to test the importance and probe the limitations of neural plasticity as a primary learning mechanism. As a result all the attention in these models is focused on the synaptic strengths. The wide variety of behaviours exhibited by individual neurons are almost completely ignored, not because they are uninteresting or even inessential, but rather because the synaptic plasticity hypothesis is thus tested in its most extreme form. In mathematical networks, synaptic plasticity is the only non-trivial element available to produce interesting behaviour. If model networks can achieve anything approaching the behaviour of their biological counterparts then it will be clear that synaptic plasticity is remarkably powerful and likely to be of crucial importance. On the other hand, if the mathematical models cannot approach biological complexity then other elements such as more accurate descriptions of individual cell behaviour will have to be included in the models until we learn what minimum set of behaviours is needed to mimic biological systems. Of course the advantage of starting with the simplest models (those having synaptic plasticity as their only non-trivial element) is that computations can be performed which might be impossible in a more complete model. The results of such computations are the subject of this review.

0954-898X/90/010105+18$03.50

The model we will concentrate on here [3] takes the synaptic plasticity hypothesis to its extreme and models individual neurons trivially. Each neuron is characterised by a variable S which takes the value $+1$ if the neuron is firing and -1 if the neuron is not firing. Thus, the actual value of the membrane potential, the firing rate and, as a result, such features as firing rate adaptation and postburst hyperpolarisation are ignored. In the model, time is measured in discrete intervals which may be taken to be the refractory period and will be the basic unit of time in our discussion. At time $t+1$ the neuron labelled by the index i, where $i = 1, 2, 3, \ldots, N$ for a system of N cells, fires or does not fire based on whether the total signal it is receiving from other cells to which it is synaptically connected is positive or negative. Thus, the basic dynamic rule is

$$S_i(t+1) = \operatorname{sgn}\left(\sum_{j=1}^{N} J_{ij} S_j(t)\right) \tag{1.1}$$

where J_{ij} represents the strength of the synapse connecting cell j to cell i. The dynamic updating (1.1) may be parallel, sequential or in a random, asynchronous sequence. For simplicity we do not include any offset or threshold factors in the dynamic rule so all self-couplings are set to zero, $J_{ii} = 0$. Note that in addition to having an extremely simple description of the cell, $S_i = \pm 1$, the model imposes an extremely simple dynamics on the cell and such features as postinhibitory rebound, delayed excitation and plateau or bursting behaviour are not implemented. In addition, the synaptic strength is characterised by a single number J_{ij} which means that numerous features of real biological synapses are ignored. There is no analogue of a reversal potential in the model or more precisely the model assumes that the magnitude of the reversal potential is much larger than the magnitude of the cell potential. In addition, synaptic delay and accommodation are not modelled.

Having given up so much one might well ask whether anything interesting can come out of the dynamics of this model? One possibility is that the dynamics (1.1) can map an initial state of firing and non-firing neurons, $S_i(0)$, to a fixed pattern, ξ_i, which remains invariant under the transformation (1.1). This is the basis of a network associative memory. Various memory patterns ξ_i^μ for $\mu = 1, 2, 3, \ldots, P$ which do not change under the transformation (1.1) act as fixed-point attractors and initial inputs $S_i(0)$ are mapped to an associated memory pattern ξ_i^μ if the overlap $\sum \xi_i^\mu S_i(0)/N$ is close enough to one. How close this overlap must be to one, or equivalently how well the initial pattern must match the memory pattern in order to be mapped to it and thus associated with it, is determined by the radius of the domain of attraction of the fixed point.

The issue of domains of attraction associated with a fixed point has never been completely resolved. The sum of all synaptic inputs at site i,

$$h_i^\mu = \sum_{j=1}^{N} J_{ij} \xi_j^\mu \tag{1.2}$$

known as the local field, is the signal which tells cell i whether or not to fire when $S_j = \xi_j^\mu$ for all $j \neq i$. In order for a memory pattern to be a stable fixed point of the dynamics (1.1) the local field must have the same sign as ξ_i^μ, or equivalently

$$h_i^\mu \xi_i^\mu > 0. \tag{1.3}$$

We will call the quantities $h_i^\mu \xi_i^\mu$ the aligned local fields. It seems reasonable to assume that the larger the aligned local fields are for a given μ value the stronger the attraction of the corresponding fixed point ξ_i^μ and so the larger its domain of attraction. This reasoning is almost right, but it leaves out an important feature of the dynamics (1.1). Multiplying J_{ij} by any set of constants A_i has absolutely no effect on (1.1) since the dynamics depends only on the sign and not on the magnitude of the quantity $\sum J_{ij} S_j$. Since the quantities $h_i^\mu \xi_i^\mu$ change under this multiplication they alone cannot determine the size of the basin of attraction. Instead, several investigations [7, 8] have found that quantities known as stability parameters and given by

$$\gamma_i^\mu = \frac{h_i^\mu \xi_i^\mu}{|J|_i} \tag{1.4}$$

where we define

$$|J|_i = \left(\sum_{j=1}^{N} J_{ij}^2 \right)^{1/2} \tag{1.5}$$

provide an important indicator of the size of the basin of attraction associated with the fixed point ξ_i^μ. Roughly speaking the larger the values of the γ_i^μ the larger the domain of attraction of the associated memory pattern. The presence of the normalising term $|J|_i$ will be an important feature in our discussion of learning algorithms. This is because many algorithms are based on increasing the values of the quantities $h_i^\mu \xi_i^\mu$ to provide stronger local fields attracting inputs to the memory pattern ξ_i^μ. However, the relevant quantity is not $h_i^\mu \xi_i^\mu$ but γ_i^μ and in studying learning we must explore how *this* quantity is affected by the algorithm.

In order to construct an associative memory we must find a matrix of synaptic strengths J_{ij} which satisfies the condition of stability of the memory fixed points (1.3) and has a specified distribution of values for the γ_i^μ giving the domain of attraction which is desired. Although associative memory is a fairly simple task a great advantage of considering this example is now apparent: the problem is now well posed and amenable to mathematical analysis.

2. Capacities and gamma distributions

The job of a learning algorithm is to find a coupling matrix J_{ij} which will achieve an assigned goal which has been specified in terms of the number of memory patterns and the sizes of the domains of attraction required. If the specified task is impossible, initiating the learning process would be pointless so it is important to know whether any matrix satisfying the preassigned criteria actually exists. Using an approach pioneered by Elizabeth Gardner [9] a great deal is known about this matter. The Gardner approach searches the space of all coupling matrices for any matrices which achieve the learning goals. It does not find these matrices, that being the task of the learning algorithm, but rather indicates whether or not they exist by giving the fractional volume in the space of all couplings occupied by matrices satisfying the learning criteria.

To assign a learning task we must first specify what type of distribution of γ_i^μ values is desired. We will consider here three classes of models characterised by different such distributions. It may seem extremely restrictive to consider only three classes of models

but if we are willing to concentrate on associative memories near their saturation point (that is, storing almost the maximum number of memory patterns possible) this is not the case. It has been shown [11] that network models of associative memory fall into universality classes which may have markedly different behaviour away from saturation but which have the same behaviour as they approach the saturation limit. Although in biological systems we may not always be interested in the saturation limit, in cases where this limit does apply the universality provides a tremendous benefit.

Universality is a concept which arose in the study of critical phenomena. When , as in the case of critical phenomena and here in the case of networks near saturation, there are classes of behaviour shared by many models, it is not essential that the model being studied be a very accurate representation of the real system being modelled. Instead, we must merely require that the model being computed lies in the same universality class as the real system. Then, since all models in the class have the same limiting behaviour, a calculation done on one of the simpler members of the class containing the real system is guaranteed to give the correct answers even if it seems a gross simplification of the real system. The realisation that network behaviour is universal near saturation provides the hope that the shortcomings of unrealistic models may not be such a severe limitation if models in the appropriate universality class can be found. Also, because of universality, it will suffice to find algorithms which construct one member of each class if we are interested in studying behaviour near the saturation limit.

We will write the number of memory patterns being stored as

$$P = \alpha N \tag{2.1}$$

and the maximum storage capacity of a model with a given γ distribution as

$$P_{\max} = \alpha_{\max} N. \tag{2.2}$$

Let $\rho(\gamma)\mathrm{d}\gamma$ be the fraction of γ_i^μ values lying between γ and $\gamma + \mathrm{d}\gamma$. The three classes we will discuss are based on three forms for the distribution of γ values. The first has a γ distribution given by a Gaussian

$$\rho(\gamma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\gamma - \bar{\gamma})^2}{2\sigma^2}\right) \tag{2.3}$$

and a maximum capacity

$$\alpha_{\max} = \frac{1}{\bar{\gamma}^2 + (1-\sigma)^2}. \tag{2.4}$$

Models of this type will be termed of the Hopfield class because the well known Hopfield model [4]

$$J_{ij} = (1 - \delta_{ij}) \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu \tag{2.5}$$

corresponds to the above formulae with $\sigma = 1$ and $\bar{\gamma} = 1/\sqrt{\alpha}$ provided that $\alpha < 0.14$ [12]. The value 0.14 is known as α_c and gives the maximum storage capacity for a coupling matrix of the form (2.5). This is different from $\alpha_{\max}$ which gives the maximum

storage capacity for any model having a specified Gaussian γ distribution (2.3). Note that models in this class make errors; that is, the memory patterns ξ_i^μ are not exactly fixed points. This is because the Gaussian γ distribution has support for negative γ so some of the elements of ξ_i^μ are unstable. The fraction of unstable sites is given by

$$F(\gamma < 0) = \int_{\bar{\gamma}/\sigma}^{\infty} \mathrm{D}z \tag{2.6}$$

where we use the notation

$$\mathrm{D}z = \mathrm{d}z \frac{\exp(-\frac{1}{2}z^2)}{\sqrt{2\pi}}. \tag{2.7}$$

For example, the Hopfield model at saturation, when $\alpha = 0.14$, has an error rate of about 1.5% [12]. The above analysis shows [11], however, that a matrix should exist with a narrower Gaussian γ distribution ($\sigma = 0.12$ is optimal) which makes no more errors than the Hopfield model at saturation but which has $\alpha = 1.14$. It would be interesting to have a construction for such a matrix.

The second class of models assumes that all the γ_i^μ are set to a specific value γ_0 so that

$$\rho(\gamma) = \delta(\gamma - \gamma_0). \tag{2.8}$$

For models of this type the maximum storage capacity is given by [11,13]

$$\alpha_{\max} = \frac{1}{1 + \gamma_0^2}. \tag{2.9}$$

I will refer to the class of models with this limiting behaviour as the pseudo-inverse class since this is the best known example. For the pseudo-inverse model [13]

$$J_{ij} = (1 - \delta_{ij}) \sum_{\mu,\nu}^{P} C_{\mu\nu}^{-1} \xi_i^\mu \xi_j^\nu \tag{2.10}$$

where

$$C_{\mu\nu} = \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu \xi_j^\nu \tag{2.11}$$

the γ distribution is given by a δ function with

$$\gamma_0 = [(1 - \alpha)/\alpha]^{1/2}. \tag{2.12}$$

A critical capacity $P_c = \alpha_c N$ is also defined for the pseudo-inverse model. It is the value of $\alpha_{\max}$ when $\gamma_0 = 0$ and thus $\alpha_c = 1$.

The final class of models to be consider has a clipped γ distribution

$$\gamma_i^\mu \geq \kappa \tag{2.13}$$

for all i and μ. By choosing the value of κ the size of the basins of attraction associated with the memory patterns can be controlled. For such models near the saturation point [9]

$$\alpha_{\max} = \left(\int_{-\kappa}^{\infty} \mathrm{D}z(\kappa + z)^2 \right)^{-1} \tag{2.14}$$

which satisfies $\alpha_{\max} < 2$ [10], the γ distribution is given by [8]

$$\rho(\gamma) = \frac{\exp(-\frac{1}{2}\gamma^2)}{\sqrt{2\pi}} \theta(\gamma - \kappa) + \delta(\gamma - \kappa) \int_{-\kappa}^{\infty} \mathrm{D}z. \tag{2.15}$$

Alternatively, we can invert the relation for $\alpha_{\max}$ in terms of κ to get a maximum value $\kappa_{\max}$ corresponding to a given value of α

$$\int_{-\kappa_{\max}}^{\infty} \mathrm{D}z(\kappa_{\max} + z)^2 = \frac{1}{\alpha}. \tag{2.16}$$

This will be useful in what follows. For $\alpha = 2$, $\kappa_{\max} = 0$ and $\kappa_{\max}$ increases monotonically with decreasing α going through $\kappa_{\max} \approx 0.5$ at $\alpha = 1$, $\kappa_{\max} = 1$ at $\alpha \approx 0.5$ and $\kappa_{\max} = 2$ at $\alpha \approx 0.2$. This class of models will be called the Gardner class. It is important to realise that within all these classes there are many models with very different behaviours away from saturation but all members of a given class converge to the above results near saturation.

3. Learning algorithms

The above results for the three classes of models determine whether or not a specific learning task can be achieved. From now on we assume that the specified learning task is possible (for example, $\alpha < \alpha_{\max}$ for given κ or $\kappa < \kappa_{\max}$ for given α) so at least one matrix J_{ij} capable of doing the job exists. The learning task is to find this matrix or one equally good at accomplishing the learning goal. A typical task might be to learn a set of P memory patterns and assure large values of γ_i^μ giving large basins of attraction. All of the learning algorithms discussed here are based on a learning mode of operation known as supervised learning in which the network is presented with the patterns to be learned and synapses are adjusted in a way which depends on the firing patterns of the pre- and postsynaptic cells and perhaps on the local field at the postsynaptic cell h_i^μ, the stability parameter γ_i^μ, the normalisation of the synapses terminating at cell i, $|J|_i$ and/or the synaptic strength itself J_{ij}. For our discussion of learning algorithms it is important to keep track of the relevant quantities used for the modification of the synaptic strength J_{ij}, namely: the state of cell i when the pattern to be learned is presented, ξ_i^μ, similarly the state of cell j, ξ_j^μ, the aligned local field $h_i^\mu \xi_i^\mu = \sum J_{ij} \xi_i^\mu \xi_j^\mu$, the stability parameter $\gamma_i^\mu = h_i^\mu \xi_i^\mu / |J|_i$ and the normalisation factor $|J|_i$ where $|J|_i^2 = \sum J_{ij}^2$.

The learning process begins with a random matrix of couplings or more frequently with zero coupling $J_{ij} = 0$ and repeatedly modifies the synaptic strengths in a specified way which hopefully improves the situation until a successful matrix of couplings is found. The learning process proceeds from site to site (each of which learns independently) and from pattern to pattern either sequentially or in a random order.

Besides biological plausibility, the only real figure of merit for a learning algorithm is the time it takes to find a suitable set of couplings. First, we must be assured that the algorithm converges if any matrices satisfying the established criteria exist. All of the algorithms discussed below have been shown to converge if the required matrix exists. Most of the convergence proofs are variants of the original perceptron learning proof [1] and will not be given here. We will concentrate instead on results. All derivations and proofs can be found in the literature cited. Once we know that an algorithm converges we are interested in how long it takes to achieve the desired goal. We will define the learning time T to be the number of times that the learning rule changes the coupling matrix at a given site before an acceptable matrix is found. Since this may vary from site to site we will consider mean values and/or distributions of values for T.

A very general modification of the synaptic strength J_{ij} in learning the memory pattern ξ_i^μ takes the form

$$\Delta J_{ij} = (f/N)(\xi_i^\mu \xi_j^\mu + a\xi_i^\mu + b\xi_j^\mu + c)(1 - \delta_{ij}) \tag{3.1}$$

where f, a, b and c may in general be functions of h_i^μ (or more often $h_i^\mu \xi_i^\mu$), J_{ij}, γ_i^μ and $|J|_i$.

Although neural plasticity has been demonstrated in biological systems the form it takes is not well known. Synaptic strengthening when both pre- and postsynaptic cells are firing has been seen [14] and the original Hebb rule [6], stating that the strength of excitatory synapses increases, in this case corresponds to $a = b = c = 1$ in (3.1). Synaptic weakening when either pre- or postsynaptic cells fire but the opposite partner does not fire, known as the anti-Hebb rule, has also been discussed for both excitatory [15] and inhibitory synapses [16]. A rule which incorporates both the Hebb and anti-Hebb rules in a simple way is, for example, $a = b = \frac{1}{2}$ and $c = 0$.

The effect of the learning change (3.1) on the aligned local field $h_i^\mu \xi_i^\mu$ is (for large N)

$$\Delta h_i^\mu \xi_i^\mu = (1 + am_\mu + (b + cm_\mu)\xi_i^\mu) f \tag{3.2}$$

where

$$m_\mu = \frac{1}{N} \sum_{j=1}^{N} \xi_j^\mu. \tag{3.3}$$

The whole point of the learning process is to increase the value of the aligned local field $h_i^\mu \xi_i^\mu$. For unbiased patterns $m_\mu = O(1/\sqrt{N})$ so it appears that non-zero a and c are not so bad. However, a value of b with magnitude greater than one would be disastrous since sometimes $h_i^\mu \xi_i^\mu$ would decrease instead of increasing. Peretto has studied the effects of non-zero a, b and c in more detail [17]. Here we will follow convention and assume that a and c are small enough to be irrelevant and so set them to zero, and assume b is small enough ($\frac{1}{2}$?) to be ignored as well. Thus we consider learning algorithms which are of the form

$$\Delta J_{ij} = (f/N)\xi_i^\mu \xi_j^\mu (1 - \delta_{ij}). \tag{3.4}$$

The function f then determines the size of the correction made to the coupling matrix while the pre- and postsynaptic firing patterns determine its sign. Although this form

for the learning rule is almost universally used it has the distinctively unrealistic feature that couplings increase when neither the pre- nor the postsynaptic cell is firing ($\xi_i^\mu = -1$ and $\xi_j^\mu = -1$).

Learning algorithms will be classified by the form of the function f. For example we will term learning conditional or unconditional depending on whether or not f vanishes identically for any finite range of its arguments. Models in the Hopfield or pseudo-inverse classes can be constructed using unconditional algorithms but to get a clipped distribution characterising the Gardner class it is necessary to have a conditional algorithm. Algorithms are further distinguished by the variables on which the function f depends. It is not unreasonable to assume that the values of the aligned local field $h_i^\mu \xi_i^\mu$ are available at the synapse since h_i^μ is just the total postsynaptic signal coming into the cell i. Thus we will consider algorithms for which $f = f(h_i^\mu \xi_i^\mu)$. The disadvantage of such algorithms is that they contain no direct information about the quantities relevant for adjusting the basins of attraction, the stability parameters γ_i^μ.

In order to contain a dependence on γ_i^μ the function f must depend on the normalisation factor $|J|_i$ as well as on $h_i^\mu \xi_i^\mu$. This information is not directly available when the pattern ξ_i^μ is imposed on the system during learning and so it might be considered less plausible in a biological system that $f = f(h_i^\mu \xi_i^\mu, |J|_i)$. However, we can imagine a way in which such information could be transmitted to the cell [18]. Suppose there is noise in the network so that at any given time, when the pattern to be learned is imposed on the system the firing pattern S_i does not equal ξ_i^μ exactly but rather

$$S_i = \xi_i^\mu + \delta S_i \tag{3.5}$$

where δS_i is a random variable which when averaged over time satisfies

$$\langle \delta S_i \rangle = 0 \tag{3.6}$$

$$\langle \delta S_i \delta S_j \rangle = \epsilon \delta_{ij}. \tag{3.7}$$

Here ϵ is a measure of the noise in the system. If the learning process takes place on a fairly slow time scale then the presence of this noise will have no appreciable effect on learning because its time average is zero. However, the expectation value of the square of the total synaptic signal coming into cell i is given by

$$\left\langle \sum_{j=1}^{N} J_{ij} S_j \sum_{k=1}^{N} J_{ik} S_k \right\rangle = (h_i^\mu)^2 + \epsilon \sum_{j=1}^{N} J_{ij}^2 = (h_i^\mu)^2 + \epsilon |J|_i^2 \tag{3.8}$$

providing a direct measure of the quantity $|J|_i$. Thus, it is perhaps not so unreasonable to suppose that some dependence of f on $|J|_i$ is possible in a biological system.

Finally, we can include in the learning rule (3.4) a dependence on the synaptic strength J_{ij} itself. Such a dependence is quite reasonable especially because the strength of a given synapse is certainly bounded and such a dependence can assure that the bound is not violated. In addition, synaptic plasticity probably does not extend to the value of the sign of the synapse and a dependence on J_{ij} can assure that sign flips are not allowed.

4. Unconditional learning algorithms

The simplest learning algorithm is just the case $f = 1$ which constructs the Hopfield matrix (2.5) after a single pass through all the sites i and patterns μ if we start from

the null matrix $J_{ij} = 0$. As mentioned in the introduction this constructs a model with a Gaussian γ distribution with $\sigma = 1$ provided that $\alpha < \alpha_c = 0.14$ [12]. However, the Hopfield matrix has a fairly limited capacity, makes errors and has limited basins of attraction. By introducing some dependence of f on the aligned local field we can construct the pseudo-inverse model. This is done [19, 20] by choosing

$$f = 1 - h_i^\mu \xi_i^\mu. \tag{4.1}$$

Starting from a null coupling matrix the application of this learning rule is equivalent to the Gauss–Seidel construction (see, for example, [21]) of the pseudo-inverse coupling matrix (2.10). The behaviour of the method for linearly dependent patterns is also very good [22]. In learning αN patterns the rule converges (in infinite time) to a δ function γ distribution with all $\gamma_i^\mu = \alpha/(1+\alpha)$ and all $h_i^\mu \xi_i^\mu = 1$ provided that $\alpha < 1$.

Unlike most of the algorithms discussed here, the algorithm given by (4.1) takes an infinite number of learning steps to actually produce the pseudo-inverse matrix. Therefore it is essential to analyse the time dependence of the approach to this goal so that we can determine what happens in a finite period of training. Since ultimately $h_i^\mu \xi_i^\mu \to 1$ we can define an error function at time t during the learning process as

$$E(t) = \frac{1}{NP} \sum_{i,\mu} (1 - h_i^\mu \xi_i^\mu)^2. \tag{4.2}$$

The behaviour of E as a function of time for a slight generalisation of (4.1)

$$f = \eta(1 - h_i^\mu \xi_i^\mu) \tag{4.3}$$

has been computed [23, 24] and is given by

$$E(t) = \frac{\alpha - 1}{\alpha}\theta(\alpha - 1) + \frac{1}{2\pi\alpha}\int_{\lambda_-}^{\lambda_+} \frac{d\lambda}{\lambda}(1 - \eta\lambda)^{2t}\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)} \tag{4.4}$$

where

$$\lambda_\pm = (1 \pm \sqrt{\alpha})^2. \tag{4.5}$$

This shows immediately that the algorithm will not converge even in infinite time if $\alpha > 1$. However, even for $\alpha > 1$ only a fraction of the bits are unstable for each pattern, so if errors are allowed the algorithm is still very useful. For $\alpha < 1$ a value of $\eta < (1 + \sqrt{\alpha})^2/2$ can always be chosen so that $E(t)$ decays exponentially to zero for $t \to \infty$. From this exponential decay of the error function we can define a learning lifetime τ. As $\alpha \to 1$ the learning lifetime diverges. If we demand that most of the patterns be learned then, as $\alpha \to 1$,

$$\tau \sim \frac{1}{1 - \alpha} \tag{4.6}$$

while if we demand that all the patterns be learned to the desired level of accuracy then for the optimal value

$$\eta = \frac{1}{1 + \alpha} \tag{4.7}$$

we find [24]

$$\tau = \left[\ln\left(\frac{1+\alpha}{2\sqrt{\alpha}}\right)\right]^{-1} \tag{4.8}$$

which as $\alpha \to 1$ diverges like

$$\tau \sim \frac{1}{(1-\alpha)^2}. \tag{4.9}$$

The algorithm (4.1) is an extremely successful unconditional learning rule since it converges even for linearly related patterns and the dynamics of the learning process is well known. At finite learning times it produces a matrix which is quite acceptable and which ultimately approaches the pseudo-inverse coupling matrix as the learning process continues.

5. Conditional algorithms

If we want to construct models of the Gardner class we must put a strict bound on the γ distribution $\gamma_i^\mu > \kappa$. This is most easily done by including a term $\theta(\kappa - \gamma_i^\mu)$ in f. This will be done in the next section but for now we will restrict ourselves to rules which do not involve the coupling normalisation factor $|J|_i$ and so which do not involve γ_i^μ directly. Instead we let

$$f = \theta(c - h_i^\mu \xi_i^\mu) g(h_i^\mu \xi_i^\mu) \tag{5.1}$$

and thus require that the learning algorithm be applied if the aligned local field is smaller than some value c. A behaviour involving some threshold in the total synaptic signal does not seem unreasonable for a biological system although it would of course be more realistic to use a smoother function than the θ function with the same general behaviour. Smooth functions in place of the θ function have been considered by Peretto [25].

Various forms for g have been considered [26] and shown to converge to a matrix satisfying

$$h_i^\mu \xi_i^\mu > c \tag{5.2}$$

provided such a matrix exists. (Since we have not specified anything about the normalisation all that is actually needed is that a matrix satisfying $h_i^\mu \xi_i^\mu > 0$ exists because by multiplying this matrix by a suitable constant A_i we can achieve (5.2).) An interesting case is

$$g = \sqrt{(h_i^\mu \xi_i^\mu)^2 + B} - h_i^\mu \xi_i^\mu \tag{5.3}$$

for arbitrary positive B. This algorithm converges and has the interesting property that it increases the normalisation $|J|_i^2$ by a fixed amount B/N upon each application. A non-trivial function g is of course useful because we can adjust the learning step size to maximise convergence time [26]. For example, the step size given by (5.3) is larger for aligned local fields which are far from the goal than for those that are near to it.

We postpone discussion of an optimal choice for g until we discuss algorithms which depend on the normalisation $|J|_i$ because in this case the advantage of a variable step size can be exploited most fully.

The case $g = 1$ has been studied most thoroughly [20, 27, 28]. The learning time for $g = 1$ is bounded by [27]

$$T \leq \frac{2c+1}{\kappa_{\max}^2} N \tag{5.4}$$

where $\kappa_{\max}$ is given by equation (2.16). In addition, the normalisation factors of the resulting matrix although not specified in the learning rule satisfy

$$|J|_i \leq \frac{2c+1}{\kappa_{\max}} \tag{5.5}$$

so that the γ distribution is of the Gardner class where $\gamma_i^\mu > \kappa$ with [27]

$$\kappa \geq \frac{c}{2c+1} \kappa_{\max}. \tag{5.6}$$

It is important that we know that $|J|_i$ is bounded for this learning rule since unlimited growth of the synaptic strengths would be highly unrealistic. In addition the limit on the κ value achieved is important since it at least puts a bound on the radius of the domain of attraction even if this is not known exactly. Note that for $c \to \infty$, $\kappa > \kappa_{\max}/2$. In fact, numerical simulation shows that for sufficiently large c, values of κ close to $\kappa_{\max}$ can be obtained. Faster convergence can be achieved by using non-trivial g functions.

Krauth and Mézard [27] have given an interesting variant of the $g = 1$ algorithm which yields a definite value of κ which is in fact $\kappa_{\max}$. The learning rule itself is unchanged, but it is applied at any given time only using the pattern ξ_i^μ with the minimum value of $h_i^\mu \xi_i^\mu$ at site i. This procedure has been shown to provide a model of the Gardner class with optimal stability $\gamma_i^\mu > \kappa = \kappa_{\max}$ in the limit $c \to \infty$ provided that such a matrix exists. The dynamics of the Krauth and Mézard learning process has been analysed by Opper [24, 29]. These results are also approximately valid for the general algorithm with sufficiently large values of c. Opper shows that the fraction of memory patterns which require a learning time between cx and $c(x + \mathrm{d}x)$ is given by $w(x)\,\mathrm{d}x$ where

$$w(x) = P_0 \delta(x) + \frac{\theta(x)}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) \tag{5.7}$$

where

$$m = \frac{\kappa_{\max}^2}{\lambda} \qquad \sigma = \frac{\kappa_{\max}}{\lambda} \qquad P_0 = \int_{-\infty}^{-\kappa_{\max}} \mathrm{D}z \tag{5.8}$$

with

$$\lambda = \alpha \kappa_{\max}^3 \int_{-\kappa_{\max}}^{\infty} \mathrm{D}z\ (z + \kappa_{\max}). \tag{5.9}$$

The average learning time required to learn all the given patterns is

$$\langle T \rangle = \frac{cN}{\kappa_{\max}^2}. \tag{5.10}$$

As α goes to its maximum value of 2 this diverges like $(2 - \alpha)^{-2}$.

6. Algorithms involving the magnitude of the coupling matrix

Much of the complication in the last section on conditional algorithms came about because the algorithm makes reference only to the local aligned fields $h_i^\mu \xi_i^\mu$ while the condition desired for Gardner-type models refers to the stability parameters γ_i^μ, namely $\gamma_i^\mu > \kappa$. These complications can be avoided and considerably efficiency gained by considering rules which allow f to be a function of $|J|_i$ as well as h_i^μ. The first such algorithm, considered by Gardner [9], was

$$f = \theta(\kappa - \gamma_i^\mu) \tag{6.1}$$

which was shown to converge in T sweeps where

$$\frac{T}{\ln T} \leq \frac{N}{2\delta(\kappa + \delta)} \tag{6.2}$$

provided that a matrix exists satisfying the condition

$$\sum_{j=1}^{N} J_{ij}^* \xi_i^\mu \xi_j^\mu > (\kappa + \delta)|J^*|_i \tag{6.3}$$

for all μ and i and arbitrary positive δ. The learning rule of course produces a Gardner-type matrix with $\gamma_i^\mu > \kappa$.

Since the Gardner learning algorithm involves the quantities γ_i^μ a dependence on the coupling normalisation $|J|_i$ has entered f. Much more efficient algorithms can be constructed [30] if we take full advantage of dependence on $|J|_i$ to resolve a problem which we have not yet discussed or faced. All the algorithms considered thus far involve either a fixed learning step size, or one that depends on the value of the local aligned field $h_i^\mu \xi_i^\mu$. Since the step size does not depend on the normalisation of the coupling matrix in these previous algorithms the same step size will be taken whether the elements of the coupling matrix are small or gigantic. This is clearly inefficient. In addition step size has in previous algorithms depended at most on $h_i^\mu \xi_i^\mu$ not on γ_i^μ which is the relevant quantity. Both problems can be solved by considering learning rules of the form [30]

$$f = \theta(\kappa - \gamma_i^\mu) g(\gamma_i^\mu)|J|_i \tag{6.4}$$

which have been shown to converge for any $g(\gamma)$ satisfying

$$0 < g(\gamma) < 2(\kappa + \delta - \gamma) \tag{6.5}$$

where δ is again given by the condition (6.3). These algorithms converge in a time bounded by

$$T \leq 2N/\delta^2. \tag{6.6}$$

The most efficient algorithm can be found by maximising the step size without destroying the convergence rate bound that applies in general to these algorithms. The optimal choice seems to be

$$g(\gamma) = \kappa + \delta - \gamma + \sqrt{(\kappa + \delta - \gamma)^2 - \delta^2} \tag{6.7}$$

for small values of δ. This essentially saturates the upper limit of the bound for convergent g functions and it provides a remarkably fast algorithm for constructing a matrix of the Gardner type. For example, in a network with $N = 100$ the algorithm converged at least 10 times faster that the $g = 1$ algorithm over a range $0.2 < \alpha < 1.5$ and $1.5 > \kappa > 0.04$.

7. Algorithms with restricted synaptic strengths and signs

Learning rules which involve a dependence of the change in the coupling matrix, ΔJ_{ij}, on J_{ij} itself, are introduced to ensure that the magnitude of individual synaptic strengths remains bounded or that synaptic strengths cannot change from excitatory to inhibitory or vice versa. In the algorithms we have discussed thus far, nothing prevents an individual element J_{ij} from growing or shrinking without bound, a highly unrealistic situation. Various modifications in the learning rule which assure that the magnitude of any given synaptic strength is bounded have been proposed [31], the simplest being

$$\Delta J_{ij} = -\lambda J_{ij} + (1/N)\xi_i^\mu \xi_j^\mu. \tag{7.1}$$

Such a decrease in the amplitude of the coupling strength over time could also be the result of some aging process [32] rather than of the learning rule itself. Bounding the synaptic strengths has the interesting consequence of introducing learning with forgetting. In the usual Hopfield model, constructed by the above algorithm with $\lambda = 0$, memories can be added to a network until at the critical capacity $\alpha = 0.14$ there is a transition to a state where all memories are lost [12]. With non-zero λ, adding new memories past the critical limit has a much less drastic effect. As new memories are added old ones are lost so that asymptotically the network always stores the latest set of patterns which it has learned.

Obviously the idea of bounding the synaptic strengths can be included in any of the algorithms we have discussed. Krauth and Mézard [27] have pointed out that the problem of maximising the aligned local fields while keeping the J_{ij} within specified bounds is a standard problem in linear programming which may be solved, for example by the simplex algorithm [33].

In addition to restricting the magnitude of synaptic strengths, the more severe constraint of binary synapses has been studied. This is perhaps of more interest for electronic circuit applications than biological modelling. The binary Hopfield model

$$J_{ij} = \operatorname{sgn}\left(\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu\right) \tag{7.2}$$

has a capacity about three quarters of that of the unconstrained model (see, for example, [34]) while in general [35] the capacity of any model with synaptic strengths restricted to $J_{ij} = \pm 1$ seems to be $\alpha < 0.83$.

Another shortcoming of the algorithms considered thus far is that they allow a given synapse to change from excitatory to inhibitory or from inhibitory to excitatory. Biological synapses are not only believed to be prohibited from making such sign changes but, in the cortex at least, they also seem to obey Dale's rule [36] stating that synapses emanating from a given neuron are all either excitatory or inhibitory. This constraint can be imposed by introducing the quantity g_i which is $+1$ if neuron i has excitatory synapses so that $J_{ij} \geq 0$ for all j and -1 if they are inhibitory so that $J_{ij} \leq 0$ for all j. In other words we constrain the synaptic matrix so that

$$J_{ij} g_i \geq 0. \tag{7.3}$$

A simple way of imposing the sign constraint on synaptic weights is to eliminate any synapses which after application of one of the unconstrained learning rules have

the wrong sign. If this is done for the Hopfield model the maximum storage capacity is $\alpha_c = 0.09$ [34] down from $\alpha_c = 0.14$ for the unconstrained model. Work on such diluted models has continued as part of a general program to study diluted models with reduced firing rates [37]. The sign constrained model has, in addition to the learned patterns, a uniform fixed point which can act as an attractor for unrecognised patterns [38].

Recently, the Gardner calculation of the storage capacity and stability parameter bound for arbitrary coupling matrices has been repeated for sign-constrained synaptic weights [39]. The results of this calculation are surprisingly simple. The maximum storage capacity of a sign-constrained network at fixed κ value is independent of the particular set of g_i being used and is exactly half of the maximum capacity of the unconstrained network given by equation (2.14). A learning algorithm capable of finding such matrices if they exist has also been formulated [40]. We start with an initial matrix satisfying (7.3) and then apply a standard algorithm with the additional condition that no change be applied if it would result in a new coupling matrix violating this constraint. For example, the algorithm with

$$f = \theta(-\gamma_i^\mu)\theta(J_{ij}(J_{ij} + \xi_i^\mu \xi_j^\mu)) \tag{7.4}$$

has been shown to converge [40]. It would be interesting to explore the convergence and dynamics of all the learning algorithms we have discussed with this extra constraint imposed.

An old model which incorporates many of the features discussed in this section and which has received recent attention [41] is the Willshaw model [42]. This stores patterns in a purely excitatory synaptic matrix constrained to take on the values $J_{ij} = 0, 1$. The Willshaw learning rule is extremely simple, J_{ij} is set equal to one if neuron i and neuron j are both active in any of the memory patterns, and to zero otherwise:

$$J_{ij} = \theta\left(\sum_{\mu=1}^{P}(\xi_i^\mu + 1)(\xi_j^\mu + 1)\right). \tag{7.5}$$

The model only works well if the memory patterns are highly biased towards non-firing cells, i.e. most $\xi_i^\mu = -1$, but in this case the model can form the basis of an associative memory with low overall and local firing rates which improves agreement with firing data taken from the cortex [43].

8. Conclusions

There is no doubt that the results reviewed here, and the many interesting developments which could not be covered, represent a significant achievement and a dramatic advance in our understanding of mathematical network models. What is much less clear is whether we have learned anything of biological relevance from all this work. Synaptic plasticity has been shown to be an enormously powerful adaptive force in network behaviour and both the extent and the limits of its capabilities have been explored. However, application to biological systems has been hampered by several unanswered questions.

How big a role do dynamic properties of individual neurons play in network behaviour? The idealised binary neurons we have discussed are clearly unrealistic.

More sophisticated neuronal behaviour can be modelled [44] and so it should be possible to address this question theoretically and of course experimentally. Of special interest is the role of oscillating or burster neurons in network behaviour.

What is the correct form for neuronal plasticity? Perhaps the biggest roadblock to making the mathematical models more realistic is our lack of knowledge about the real form that neuronal plasticity takes. This may be completely different for excitatory and inhibitory synapses. Clearly more experimental results are needed here, but in addition attempts at more realistic learning rules can be explored theoretically.

How does learning take place as a dynamic process? We have considered learning only in a controlled, supervised mode of operation. In an isolated biological network learning is part of the dynamic process by which the network operates. Work on dynamic, unsupervised learning has begun [45] but much remains to be learned.

It may be that a further difficulty concerns the approach taken by researchers to learning problems. Typically, in both computations and simulations networks are pushed to their limits, saturating their capacities and making the basins of attraction as deep as possible. Likewise, researchers are tempted to devise clever algorithms which work with maximum efficiency and speed. It is only natural to rise to such intellectual challenges. However, biological systems probably work far from the limits of their capacities and learning in real biological systems is unlikely to be maximally efficient by our measures of efficiency and for simple tasks we might devise as tests. Perhaps we must learn to appreciate the inherently convoluted and redundant nature of biological design, for despite their apparent lack of optimisation, biological networks are capable of achieving behaviours which modellers have yet to touch.

Acknowledgments

This work was supported byDepartment of Energy Contract DE-AC0276-ER03230. I thank Tom Kepler and Charlie Marcus for their help.

References

[1] Rosenblatt F 1961 *Principles of Neurodynamics* (Washington, DC: Spartan)
Minsky M and Papert S 1969 *Perceptrons* (Cambridge, MA: MIT Press)

[2] Rumelhart D E and McClelland J L (eds) 1986 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* vols I and II (Cambridge, MA: MIT Press)

[3] McCulloch W S and Pitts W 1943 A logical calculus of the ideas immanent in nervous activity *Bull. Math. Biophys.* **5** 115–33
Little W A 1975 The existence of persistent states in the brain *Math. Biosci.* **19** 101–20

[4] Hopfield J J 1982 Neural networks and physical systems with emergent selective computational abilities *Proc. Natl Acad. Sci. USA* **79** 2554–258

[5] For reviews and other viewpoints see, for example:
Amari S and Maginu K 1988 Statistical neurodynamics of associative memory *Neural Networks* **1** 63–73
Grossberg S 1988 Nonlinear neural networks: principles, mechanisms and architecures *Neural Networks* **1** 17–61
Kohonen T 1988 An introduction to neural computing *Neural Networks* **1** 3–16; 1984 *Self Organisation and Associative Memory* (Berlin: Springer)
Amit D 1989 *Modelling Brain Function: The World of Attractor Neural Networks* (Cambridge: Cambridge University Press)

[6] Hebb D O 1949 *The Organisation of Behaviour: A Neuropsychological Theory* (New York: Wiley)

[7] Forrest B M 1988 Content-addressability and learning in neural networks *J. Phys. A: Math. Gen.* **21** 245–55

Kohring G A 1989 Dynamical interference between the attractors in a neural network *Europhys. Lett.* (to be published)

Opper M, Kleinz J and Kinzel W 1989 Basins of attraction near the critical capacity for neural networks with constant stabilities *J. Phys. A: Math Gen.* **22** L407–11

Krätzschnar J and Kohring G A 1989 Retrieval dyanamics of neural networks constructed from local and nonlocal learning rules *Preprint* Bonn

[8] Kepler T B and Abbott L F 1988 Domains of attraction in neural networks *J. Physique* **49** 1657–62

Krauth W, Nadal J-P and Mézard M 1988 The role of stability and symmetry in the dynamics of neural networks *J. Phys. A: Math Gen.* **21** 2995–3011

[9] Gardner E 1987 Maximum storage capacity in neural networks *Europhys. Lett.* **4** 481–5; 1988 The space of interactions in neural network models *J. Phys. A: Math Gen.* **21** 257–70

Gardner E and Derrida B 1988 Optimal storage properties of neural network models *J. Phys. A: Math Gen.* **21** 271–84

[10] Cover T M 1965 Geometrical and statistical properties of systems of linear inequalities with application in pattern recognition *IEEE Trans. Electromagnet. Compat.* **EC-14** 326–34

Venkatesh S 1986 Epsilon capacity of a neural network *Neural Networks for Computing* (Proc. Conf., Snowbird, Utah, 1986) (*AIP Conf. Proc.* **151**) ed J S Denker (New York: American Institute of Physics) pp 440–5

Baldi P and Venkatesh S 1987 The number of stable points for spin-glass and neural networks of higher order *Phys. Rev. Lett.* **58** 913–6

[11] Abbott L F and Kepler T B 1989 Universality in the space of interactions for network models *J. Phys. A: Math Gen.* **22** 2031–8

[12] Amit D J, Gutfreund H and Sompolinsky H 1985 Storing infinite numbers of patterns in a spin glass model of neural networks *Phys. Rev. Lett.* **55** 1530–3; 1985 Spin glass models of neural networks *Phys. Rev.* A **32** 1007–18; 1987 Information storage in neural networks with low levels of activity *Phys. Rev.* A **35** 2293–303; 1987 Statistical mechanics of neural networks near saturation *Ann. Phys., NY* **173** 30–67

[13] Kohonen T, Reuhkala E, Mäkisara K and Vainio L 1976 Associative recall of images *Biol. Cyber.* **22** 159–68

Personnaz L, Guyon I and Dreyfus G 1985 Information storage and retrieval in spin-glass like neural networks *J. Physique* **46** L359–65

Kanter I and Sompolinsky H 1986 Associative recall of memories without errors *Phys. Rev.* A **35** 380–92

[14] Kelso S R, Ganong A H and Brown T H 1986 Hebbian synapses in the hippocampus *Proc. Natl Acad. Sci. USA* **83** 5326–30

diPrisco G V 1984 Hebb synaptic plasticity *Prog. Neurobiol.* **22** 89–102

Levy W B 1985 Associative changes at the synapse: LTP in the hippocampus *Synaptic Modification, Neuron Selectivity and Nervous System Organisation* ed W B Levy, J A Anderson and S Lehmkuhle (Hillsdale, NY: L Erlbaum Associates) pp 5–33

Brown T H *et al* 1988 Long-term synaptic potentiation *Science* **242** 724–8

[15] Rauschecker J P and Singer J P 1981 The effects of early visual experience on the cat's visual cortex and their possible explanation by Hebb synapses *J. Physiol.* **310** 215–39

Singer W 1985 Hebbian modification of synaptic transmission as a common mechanism in experience-dependent maturation of cortical functions *Synaptic Modification, Neuron Selectivity and Nervous System Organisation* ed W B Levy, J A Anderson and S Lehmkuhle (Hillsdale, NY: L Erlbaum Associates) pp 35–64

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

Reiter H O and Stryker M P 1988 Neural plasticity without postsynaptic action potentials: less-active inputs become dominant when kitten visual cortex cells are pharmacologically inhibited *Proc. Natl Acad. Sci. USA* **85** 3623–7

[17] Peretto P 1988 On learning and memory storage abilities of asymmetrical neural networks *J. Physique.* **49** 711–26

[18] This was suggested to me by T Kepler (private communication)

[19] Widrow B and Hoff M E 1960 Adaptive switching circuits *WESCON Convention Report IV* pp 96–104

[20] Diederich S and Opper M 1988 Learning of correlated patterns in spin-glass networks by local learning rules *Phys. Rev. Lett.* **58** 949–52

[21] Carnahan B, Luther H A and Wilkes J O 1969 *Applied Numerical Methods* (New York: Wiley)

[22] Berryman K W, Inchiosa M E, Jaffe A M and Janowsky S A 1989 Convergence of an iterative neural network learning algorithm for linearly dependent patterns *Preprint* HUTMP 89/B237, Havard University

[23] Hertz J A, Thorbergson G I and Krogh A 1989 Dynamics of learning in simple perceptrons *Phys. Scr.* **T25** 149–51

[24] Kinzel W and Opper M 1989 Dynamics of learning *Physics of Neural Networks* ed J L van Hemmen, E Domany and K Schulten (Berlin: Springer) (to appear)

[25] Peretto P 1988 On the dynamics of memorisation processes *Neural Networks* **1** 309–22

[26] Agmon S 1954 The relaxation method for linear inequalities *Can. J. Math.* **6** 382–92

Jacobs R 1988 Increased rates of convergence through learning rate adaptation *Neural Networks* **1** 295–307

[27] Krauth W and Mézard M 1987 Learning algorithms with optimal stability for neural networks *J. Phys. A: Math Gen.* **20** L745–52

[28] Gardner E, Stroud N and Wallace D J 1988 Training with noise and the storage of correlated patterns in neural network models *Neural Computers* ed R Eckmiller and C vander Malsburg (Berlin: Springer) pp 251–60

Pöppel G and Krey U 1987 Dynamical learning process for recognition of correlated patterns in symmetric spin glass models *Europhys. Lett.* **4** 979–85

[29] Opper M 1988 Learning times of neural networks: exact solution for a perceptron algorithm *Phys. Rev.* A **38** 3824–6

[30] Abbott L F and Kepler T B 1989 Optimal learning in neural network memories *J. Phys. A: Math Gen.* **22** L711–7

[31] Parisi G 1986 A memory which forgets *J. Phys. A: Math Gen.* **19** L616–20

van Hemmen J L, Keller G and Kühn R 1988 Forgetful memories *Europhys. Lett.* **5** 663–8

Derrida B and Nadal J-P 1987 Learning and forgetting on asymmetric, diluted neural networks *J. Stat. Phys.* **49** 993–1009

Toulouse G, Dehaene S and Changeux J-P 1986 Spin-glass model of learning by selection *Proc. Natl Acad. Sci USA* **83** 1695–8

Geszti T and Pazmandi F 1987 Learning within bounds and dream sleep *J. Phys. A: Math Gen.* **20** L1299–303

Gordon M B 1987 Memory capacity of neural networks learning within bounds *J. Physique* **48** 2053–8

Nadal J-P, Toulouse G, Changeux J-P and Dehaene S 1986 Networks of formal neurons and memory palimpsests *Europhys. Lett.* **1** 535–42

[32] Mézard M, Nadal J-P and Toulouse G 1986 Solvable models of working memories *J. Physique* **47** 1457–62

[33] Papadimitriou D and Steiglitz K 1982 *Combinatorical Optimisation: Algorithms and Complexity* (Englewood Cliffs NJ: Prentice Hall)

[34] Sompolinsky H 1986 Neural networks with nonlinear synapses and a static noise *Phys. Rev.* A **34** 2571–4

van Hemmen J L 1987 Nonlinear neural networks near saturation *Phys. Rev.* A **36** 1959–62

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

[35] Krauth W and Opper M 1989 Critical storage capacity of the $J = \pm 1$ neural network *J. Phys. A: Math Gen.* **22** L519–23

Krauth W and Mézard M 1989 Storage capacity of memory networks with binary couplings *Preprint*

[36] Dale H H 1935 Pharmacology and nerve-endings *Proc. R. Soc. Med.* **28** 319–32

Eccles J 1964 *Physiology of Synapses* (Berlin: Springer)

Breitenberg V 1986 *Brain Theory* ed G Palm and A Aersten (Berlin: Springer)

Peters A and Jones E G 1984 *The Cerebral Cortex* ed A Peters and E G Jones (New York: Plenum)

[37] Treves A and Amit D J 1988 Metastable states in asymmetrical diluted Hopfield nets *J. Phys. A: Math Gen.* **21** 3155–69; 1989 Low firing rates: an effective Hamiltonian for excitatory neurons *J. Phys. A: Math Gen.* **22** 2205–26

Campbell C 1989 Neural network models with sign constrained weights *Neural Computing* ed J G Taylor and C Mannion (Bristol: Adam Hilger)

Tsodyks M V and Feigel'man M V 1988 The enhanced storage capacity of neural networks with low activity level *Europhys. Lett.* **6** 101–5

Tsodyks M V 1988 Associative memory in asymmetric diluted network with low level of activity *Europhys. Lett.* **7** 203–8

Buhmann J, Divko V and Schulten K 1988 Associative memories with high information content *Preprint*

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

[38] Shinomoto S 1987 A cognitive and associative memory *Biol. Cybern.* **57** 197–206

Kohring G A 1989 Coexistence of global and local attractors in neural networks *Preprint* Bonn

[39] Amit D J, Campbell C and Wong K Y M 1989 The interaction space of neural networks with sign-constrained synapses *Preprint*

[40] Amit D J, Wong K Y M and Campbell C 1989 Perceptron learning with sign-constrained weights *J. Phys. A: Math Gen.* **22** 2039–46

[41] Nadal J-P and Toulouse G 1989 Information storage in sparsely-coded memory nets *Network* **1** 61–74

Rubin N and Sompolinsky H 1989 Neural networks with low local firing rates *Europhys. Lett.* (to appear)

Golomb D, Rubin N and Sompolinsky H 1989 The Willshaw model: associative memory with sparse coding and low firing rates *Preprint*

[42] Willshaw D J, Buneman O P and Longuet-Higgins H C 1969 Non-holographic associative memory *Nature* **222** 960–2

[43 Abeles M 1982 *Local Cortical Circuits* (Berlin: Springer)

Miyashita Y and Chang H S 1988 Neuronal correlate of pictorial short-term memory in the primate temporal cortex *Nature* **331** 68–70

Fuster J M and Jervey J P 1982 Neuronal firing in the inferotemporal cortex of the monkey in a visual memory task *J. Neurosci.* **2** 361–75

[44] Buhmann J and Shulten K 1987 Noise driven temporal association in neural networks *Europhys. Lett.* **4** 1205–9

Coolen A C C and Gielen C C A M 1988 Delays in neural networks *Europhys. Lett.* **7** 281–5

Abbott L F 1989 A network of oscillators *Preprint* Brandeis University

[45] Shinomoto S 1987 Memory maintenence in neural networks *J. Phys. A: Math Gen.* **20** L1305–9

Hertz J A, Thorbergson G I and Krough A 1989 Phase transitions in simple learnings *J. Phys. A: Math Gen.* **22** 2133–50

Meir R and Domany E 1988 Iterated learning in a layered feed-forward neural network *Phys. Rev.* A **37** 2660–8

Domany E, Meir R and Kinzel W 1986 Storing and retrieving information in a layered spin system *Europhys. Lett.* **2** 175–85

Linsker R 1986 From basic network principles to neural architecture: emergence of spatial-opponent cells *Proc. Natl Acad. Sci. USA 83* 7508–12; 1986 From basic network principles to neural architecture: emergence of orientation-selection cells *Proc. Natl Acad. Sci. USA 83* 8390–4; 1986 From basic network principles to neural architecture: emergence of orientation columns *Proc. Natl Acad. Sci. USA 83* 8779–83