Statistical mechanics of unsupervised Hebbian learning

1993 J. Phys. A: Math. Gen. 26 2343

(http://iopscience.iop.org/0305-4470/26/10/009)

J. Phys. A: Math. Gen. 26 (1993) 2343–2369. Printed in the UK

Statistical mechanics of unsupervised Hebbian learning

Adam Prügel-Bennett and Jonathan L Shapiro
Department of Computer Science, University of Manchester, Oxford Road, Manchester, UK

Received 23 January 1992, in final form 7 October 1992

Abstract. A model describing the dynamics of the synaptic weights of a single neuron performing Hebbian learning is described. The neuron is repeatedly excited by a set of input patterns. Its response is modelled as a continuous, nonlinear function of its excitation. We study how the model forms a self-organized representation of the set of input patterns. The dynamical equations are solved directly in a few simple cases. The model is studied for random patterns by a signal-to-noise analysis, and by introducing a partition function and applying the replica approach. As the number of patterns is increased a first-order phase transition occurs where the neuron becomes unable to remember one pattern but learns instead a mixture of very many patterns. The critical number of patterns for this transition scales as N^b, where N is the number of synapses and b is the degree of nonlinearity. The leading order finite-size corrections are calculated and compared with numerical simulations. It is shown how the representation of the input patterns learned by the neuron depends upon the nonlinearity in the neuron's response. Two types of behaviour can be identified depending on the degree of nonlinearity: either the neuron learns to discriminate one pattern from all the others, or it will learn to discriminate a complex mixture of many of the patterns.

1. Introduction

The mechanism first outlined by Hebb [1] is widely accepted as an important form of learning in neural systems. He proposed that when the activation of a neuron coincides with that of an input neuron, the synaptic weight from the input neuron increases. In this paper we will study a simple mathematical model of a single neuron which takes the mechanism of Hebb very literally. Our aim is to model the behaviour of neurons with many synapses (10^4–10^5), so that statistical mechanics can be used. In the model a neuron learns in an unsupervised fashion—the neuron responds to a stimuli and the synaptic weight is changed proportionally to the stimuli and the response. A regularizing mechanism is incorporated to prevent the weights from growing too large. The response of the neuron to its inputs is described by a simple nonlinear function.

A neuron learning in this way, when repeatedly presented with a set of input stimuli, develops a self-organized representation of those stimuli. The precise nature of this representation is important in determining what processing functions can be computed by networks made of these neurons. We find that the representation which the neuron learns is strongly dependent upon the shape of the activation function—that is the function which determines the neuron's response for a given excitation. If the activation function is linear, the neuron learns a statistical property of the ensemble of patterns. When the activation function is sufficiently nonlinear, the neuron becomes a discriminator, learning to distinguish one pattern from the others.

Over the past ten years, a number of authors have considered unsupervised Hebbian learning in neurons with linear activation functions, see, for example, [2–5]. In a model

0305-4470/93/102343+27$07.50

very similar to the one considered here, but with a linear activation function, Oja [2] found that the neuron learns to discriminate the maximal eigenvector of the correlation matrix of the patterns (this is described in section 2.1). Using those neurons as building blocks, Oja and others have developed neural network architectures which perform principal component analysis [6–8], a well known method of dimensional reduction. What is new in our work is that we study the model with nonlinear response functions and also that we apply methods of statistical mechanics to the model. Our results show that neurons with nonlinear activation functions learn very differently from linear neurons. Instead of learning to find the principal component of all of the patterns, they learn to discriminate one pattern from the others.

Experimental measurements of the shape of the activation function are still an open and active area of research. It is known from measuring the spike frequency against the injected currents (so called F–I plots) that the activation functions differ in different cell types [9]. However, these measurements are still too crude to be able deduce the form of the activation function. In this paper we have therefore modelled the activation function as simply as possible, that is as a power law with a threshold. This can only be regarded as a first approximation as it does not incorporate any saturation of the firing for large post-synaptic potentials. However, with random patterns the vast majority of patterns will produce very small excitations, where the activation function may be well described by a power law. Much of the behaviour of the neuron will depend on the response of the neurons to these small excitations. In particular, if the activation function is initially concave then these small excitations will act like a 'noise', the magnitude depending on the power of the nonlinearity. While if the activation function is convex then the neuron will learn to discriminate a mixture of these patterns. Thus the form of the activation function might provide an important clue as to the nature of the neuron function. More complicated activation functions will give rise to more complicated behaviour. For example, neurons with sigmoid activation functions might learn to discriminate between a mixture of a few patterns. In the conclusion we will discuss how some of the calculations presented here can be extended to sigmoid activation functions.

The model presented here also has an interesting statistical interpretation. As already mentioned the Oja model learns the principal component (technically this is only true for patterns drawn from a distribution with zero mean). That is the synaptic weights become the vector which maximizes the second moment of the pattern distribution. In our model the weights learn to maximize higher moments of the pattern distribution. This is discussed in more detail in section 2.2. Although, in this paper, we will discuss this model mainly as a neuron model, with very minor modifications the results can equally well be applied to this statistical interpretation. An important difference emerges between the principal component analysis performed by the Oja model and the higher moment analysis. In the Oja model all eigenvectors of the correlation matrix are stationary solutions, but they are all unstable in the direction of the principal eigenvector. Thus the Oja model will, with probability one, find the maximum eigenvalue. In the case of higher moments there exist many local maxima so that a nonlinear generalization of Oja's model is no longer guaranteed to find the global maximum.

The model is solved using two approaches: the dynamical equation is studied directly, and a partition function is introduced which has the same stationary states as the dynamical equation. For two patterns and for patterns with a single correlation between them, the dynamical equations can be solved exactly. From this one finds that the ability to learn to discriminate one pattern from many others is determined by the degree of nonlinearity of the activation function and the correlation between patterns. A type of signal-to-noise analysis is used to study the dynamical equations when many random patterns are learned.

We identify two transitions as the number of patterns is increased. When only a few patterns are shown to the neuron, it will learn to discriminate one of the patterns from all the others. As more patterns are shown the neuron increasingly feels the influence of the other patterns and becomes less well 'tuned' to the pattern it has learned. The first transition occurs when the neuron experiences so many patterns that it is no longer able to learn a new pattern, although if it has learned a pattern it will retain a 'memory' of it. As the number of patterns is increased there is a second transition where the neuron loses any memory of a pattern.

The case of many random patterns is also studied using a replica analysis of the partition function. This analysis confirms the results obtained from the signal-to-noise analysis. It also allows us to study the effects of introducing additive stochastic noise. These effects are similar to those caused by introducing many random patterns. Again the noise produces a deterioration in the ability to learn a pattern and two transitions analogous to those already described can be identified. The replica calculation is unusual as the energy function is not quadratic. As a consequence the critical number of patterns required for the neuron to be unable to remember a single pattern scales as N^b where N is the number of synapses and b measures the degree of nonlinearity in the activation function. This is the first time, to the authors' knowledge, that a replica calculation has given rise to this type of scaling.

The leading finite-size effects, for $b < 4$, in the case of many random patterns have been calculated using signal-to-noise analysis. For small nonlinearities, the size of these finite-size effects are very large and will be noticeable even in large neurons. The theoretical calculations are compared with numerical simulations and found to agree, although the simulations are only possible on relatively small systems as the number of patterns that can be learned rapidly becomes prohibitively large.

The organization of the paper is as follows. In the next section the model is defined and its biological motivations are discussed. In section 3 the dynamical equations are solved, first in the case when the neuron is presented with orthogonal patterns (in section 3.1), and then for two correlated patterns and for many patterns each with the same correlation (section 3.2). In section 4.1, the signal-to-noise analysis for many random patterns is developed. In the second part of this section, the solution to the signal-to-noise analysis is discussed. The next section describes the mean-field solution for the partition function. Section 5.1 contains the solution when only a few patterns are present. Section 5.2 contains the solution for many random patterns where the replica approach is used. These results confirm those obtained from the signal-to-noise analysis. To test for replica symmetry breaking we examined the solution with one step of replica breaking, this is described in the appendix. No evidence for replica symmetry breaking was found, suggesting these results are exact. In section 6 the finite-size corrections and simulations of the model are briefly described. The final section discusses some of the biological implications of the results obtained in the preceding sections.

2. The model

In this section we present the model of the neuron. First, we give the the dynamical equation describing Hebbian learning. The biological motivation is briefly discussed, although most features of the model are standard in neural modelling. The only novel aspect is our form for the activation function. In the second part of this section, the partition function for this model is described, and the relationship between it and the dynamics is discussed.

2.1. The dynamical equations

We consider a neuron which receives stimuli through N synapses with coupling weights denoted by w_i, where i labels the N different synapses. The stimuli experienced by a real neuron comes in the form of a train of spikes. We use the standard assumption that the the inputs vary sufficiently slowly in time that we can replace the spike train by a single continuous valued quantity, the 'pre-synaptic activity', related to the frequency of the spike train. We will treat the pre-synaptic activities across all the axons at any one time as a pattern vector. The pre-synaptic activity experienced by axon i when pattern μ is present is denoted by ξ_i^μ. Each pattern produces a post-synaptic potential, V^μ, given by the weighted sum of the input activities

$$V^\mu = \sum_{i=1}^{N} \xi_i^\mu w_i = \boldsymbol{\xi}^\mu \cdot \boldsymbol{w}. \tag{2.1}$$

The post-synaptic potential will excite the neuron which then fires according to its activation function, $A(V^\mu)$. We assume that there is some threshold ϕ below which the neuron does not fire and above which the neuron fires according to a simple power law, that is

$$A(V^\mu) = (V^\mu - \phi)^b \Theta(V^\mu - \phi) = \begin{cases} (V^\mu - \phi)^b & V^\mu > \phi \\ 0 & V^\mu \leqslant \phi \end{cases} \tag{2.2}$$

where b measures the degree of nonlinearity of the response.

The synapses are assumed to modify according to the learning rule

$$w_i \to w_i' = w_i + rA(V^\mu)(\xi_i^\mu - V^\mu w_i). \tag{2.3}$$

When the learning rate, r, is very small, so that w_i hardly changes during a single presentation of all the patterns, we can make the (adiabatic) approximation, that after all the patterns have been presented once, the synaptic weights change by an amount

$$\delta w_i = r \sum_{\mu=1}^{P} A(V^\mu)(\xi_i^\mu - V^\mu w_i). \tag{2.4}$$

The first term is a 'linear Hebbian' term—the change in weight is proportional to the input activity and the cell activation. The second term prevents the synapses growing unboundedly by tending to normalize the weights. In order to see this, we observe that

$$\boldsymbol{w} \cdot \delta \boldsymbol{w} = \tfrac{1}{2}\delta |\boldsymbol{w}|^2 = r(1 - |\boldsymbol{w}|^2) \sum_\mu A(V^\mu) V^\mu. \tag{2.5}$$

When the neuron has finished learning ($\delta \boldsymbol{w} = 0$), the weight factor will be normalized. Furthermore as $r \sum_\mu A(V^\mu) V^\mu$ is always positive (strictly only for $\phi \geqslant 0$) $|\boldsymbol{w}|^2$ will always move towards one and thus the steady-state solutions will always be stable to perturbations in the direction of $\boldsymbol{w}$.

Learning is unsupervised in the sense that there is no teacher. The system starts with some (non-zero) initial weights, then equation (2.4) is applied until the system converges. The neuron forms a representation of the pattern set. This depends on the initial weights, the form of the activation function and the properties of the patterns.

The dynamical equation (2.4) with $A(V^\mu) = V^\mu$ has been solved by Oja [2]. He showed that the weight vector learns to recognize the eigenvector with the largest eigenvalue (i.e. the principal component) of the correlation matrix $M_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu$. This is a property of the ensemble of patterns. When the principal component is unique, the weight vector after learning will be independent of the initial weights. A lot of work has still be done on this model, for a review see [10, pp 204–7]. The model in this paper can be viewed as an extension of Oja's model to the nonlinear activation function (2.2). As we shall see this nonlinearity can radically alter what the neuron learns.

2.2. *The partition function*

We have also studied a partition function which has the same stationary states as the dynamical equation (2.4). This allows us to average over random patterns to find the probability distribution of weights after learning. We solve this model in the framework of mean-field theory. A further bonus of this approach is that it allows us to study the influence of additive noise in the dynamical equations. We consider the partition function

$$Z = \prod_{i=1}^{N} \int \mathrm{d}w_i \, \delta\Big(\sum_{i=1}^{N} w_i^2 - N \Big) \mathrm{e}^{-\beta E(w_i)} \tag{2.6}$$

where β is the inverse temperature and

$$E(w_i) = -\frac{N}{b+1} \sum_\mu (V^\mu - \phi)^{b+1} \Theta(V^\mu - \phi) \tag{2.7}$$

where it is now more convenient to define $V^\mu = \xi^\mu \cdot w/N$. Using this definition of V^μ rather than (2.1) one finds that the dynamics imply that at the fixed point $|w^*|^2 = N$. The dynamics is not affected by this redefinition.

It might not be obvious that the partition function and the dynamical equation (2.4) are related. Indeed, the energy defined in (2.7) is not the integral of the right-hand side of equation (2.4). The difference is that in the partition function formulation, the weights must lie on the unit sphere, whereas in the dynamical equation the weights are merely attracted to the unit sphere. However, these equations (2.6) and (2.7) do indeed have the same stationary states as the dynamical equation (2.4) with (2.2). To see this, we note that the corresponding Langevin equation is given by

$$\delta w_i = r \sum_j \hat{P}_{ij} \frac{\partial E}{\partial w_j} + \eta_i = r \sum_\mu A(V^\mu)(\xi_i^\mu - V^\mu w_i) + \eta_i \tag{2.8}$$

where $\hat{P}_{ij}$ is the projection operator $\delta_{ij} - w_i w_j/N$ which imposes the constraint $|w|^2 = N$. $A(V^\mu)$ is the activation function given in equation (2.2) and η_i is a random Gaussian noise (orthogonal to w) with variance σ^2 related to β by the usual Einstein relationship, $2\beta\sigma^2 = r$. Although equations (2.4) and (2.8) are identical for $\eta_i = 0$, the partition function describes the equilibrium behaviour for (2.8) only in the limit $r \to 0$ and when $|w|^2 = N$. However, the latter condition will be satisfied for stationary solutions and the stability is found to be independent of r so the partition function will correctly give the stability of the stationary solutions of (2.4).

The statistical interpretation of our model is made clear by considering the energy function. Principal component analysis can be viewed as finding the unit vector w which maximizes the second moment

$$\sum_\mu (\xi^\mu \cdot w)^2. \tag{2.9}$$

This is equivalent to minimizing the energy function

$$E \propto -\sum_{\mu}(\xi^{\mu} \cdot w)^2 \tag{2.10}$$

subject to the constraint $|w| = 1$. The corresponding Langevin equation is the Oja model. The most natural generalization to higher moments is to find the vector w which maximizes

$$\sum_{\mu}|\xi^{\mu} \cdot w|^a \tag{2.11}$$

or minimizes the energy

$$E \propto -\sum_{\mu}|\xi^{\mu} \cdot w|^a. \tag{2.12}$$

The Langevin equation for this model is the familiar dynamical equation (2.8) but with

$$A(V^{\mu}) = \mathrm{sgn}(V^{\mu})\,|V^{\mu}|^b \tag{2.13}$$

where $b = a - 1$. For positively correlated patterns this model is identical to the neuron model discussed throughout this paper. With very small modifications all the calculations presented in here can be applied to this higher moment model.

3. The stationary solutions of the dynamical equations

Here we examine the stationary solutions of the dynamical equations. Clearly this only makes sense in terms of the adiabatic approximation (2.4) rather than equation (2.3) where the stationary solutions correspond to small cycles. The stationary solutions w^* satisfy the equation

$$\delta w = r\sum_{\mu}A(V^{\mu})(\xi^{\mu} - V^{\mu}w^*) = 0 \tag{3.1}$$

and are thus independent of the learning rate, r. We will examine two simple cases where the model can be solved exactly. In the first part of this section we will consider orthogonal patterns. In the second part we will examine two correlated patterns and, in a very special case, many correlated patterns.

3.1. Orthogonal patterns

For orthogonal patterns there are stationary solutions at $w = \xi^{\mu}/|\xi^{\mu}|$ for each pattern ξ^{μ}. There are also stationary solutions corresponding to mixtures of patterns but these turn out to be unstable when the activation function is nonlinear (more precisely when $b > 1$). To see this, it is useful to examine the dynamical equations in the directions of each of the patterns. Resolving in the direction ξ^1 equation (2.4) becomes

$$\delta V^{\mu} = \xi^{\mu} \cdot \delta w = r\sum_{\nu}A(V^{\nu})(\xi^{\mu} \cdot \xi^{\nu} - V^{\mu}V^{\nu}). \tag{3.2}$$

For simplicity we consider normalized patterns so that $\xi^{\mu} \cdot \xi^{\nu} = \delta^{\mu,\nu}$, and we set the threshold, ϕ, to zero. The mixed solutions are of the form $w^* = \sum_{\mu\in S_n}\xi^{\mu}/\sqrt{n}$ where

$\mathcal{S}_n$ is any set of n of the patterns. We first examine the case of a two-patterns mixed state $w^* = (\xi^1 + \xi^2)/\sqrt{2}$. We consider perturbation in the direction $\delta w = \epsilon^1 \xi^1 + \epsilon^2 \xi^2$. Substituting $w = w^* + \delta w$ into the equations for δV^μ and only keeping terms up to first order in ϵ^1 and ϵ^2

$$\begin{pmatrix} \delta V^1 \\ \delta V^2 \end{pmatrix} = 2^{-(b+1)/2} r \begin{pmatrix} b-3 & -(b+1) \\ -(b+1) & b-3 \end{pmatrix} \begin{pmatrix} \epsilon^1 \\ \epsilon^2 \end{pmatrix} \tag{3.3}$$

which has an eigenvalue of $r(b-1)2^{(1-b)/2}$ in the direction $\xi^1 - \xi^2$ and an eigenvalue $-r2^{(3-b)/2}$ in the direction $\xi^1 + \xi^2$. If we consider perturbations in the direction of an unlearned pattern, ξ^ν say, then to leading order $\delta V^\nu = r(\epsilon^\nu)^b - r2^{(1-b)/2}\epsilon^\nu$. Thus, for $b > 1$, the leading order term is $\delta V^\nu = -r2^{(1-b)/2}\epsilon^\nu$ and the mixed fixed point is stable in the direction of the unlearned patterns and in the direction bisecting the two patterns but unstable in the direction towards either of the patterns. For $b = 1$, $\delta V^\nu = 0$ as is the eigenvalue in the direction $\xi^1 - \xi^2$ so that all mixtures of patterns are marginally stable. For $b < 1$, $\delta V^\nu = r(\epsilon^\nu)^b$ so the two-pattern mixed state is unstable in the direction of any unlearned pattern. The generalization to the n-pattern mixed solution is quite straightforward. The matrix for all P of the δV^μs has three distinct eigenvalues. A non-degenerate eigenvalue in the direction bisecting the n learned patterns

$$\lambda_1 = -2rn^{(1-b)/2} \tag{3.4}$$

a $(P-n)$ degenerate eigenvalue in the direction of each of the unlearned patterns (for $b > 1$)

$$\lambda_2 = -rn^{(1-b)/2} \tag{3.5}$$

and an $(n-1)$ degenerate eigenvalue

$$\lambda_3 = r(b-1)n^{(1-b)/2} \tag{3.6}$$

describing mixing between the learned pattern. For $b > 1$, this last eigenvalue is positive indicating that this mixed fixed point is unstable in the direction of any of the patterns. The only stable states are the single-pattern states. As in the two-pattern case, the perturbation in the direction of an unlearned pattern, ξ^ν, is, to leading order, $\delta V^\nu = r(\epsilon^\nu)^b - rn^{(1-b)/2}\epsilon^\nu$, so as b becomes less than one the n-pattern mixed state becomes unstable in the directions of the unlearned patterns. The only stable state for $b < 1$ is the therefore the P-pattern mixed state.

As discussed in section 2, when $A(V^\mu) = V^\mu$ equation (2.4) corresponds to Oja's rule. In this case the only stable solutions are those in the direction of the principal component of the matrix $M_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu$. For orthonormal patterns, all the patterns and all mixtures of patterns have the same maximum eigenvalue of one. Thus, the principal component is degenerate and any mixture is also a principal component. For sufficiently many random patterns the principal component will be unique.

3.2. Two correlated patterns

Although the analysis for orthogonal patterns provides some insight into how neurons learn, in general patterns will be correlated with each other. This situation is much more complicated and can only be solved completely for two patterns. We therefore consider

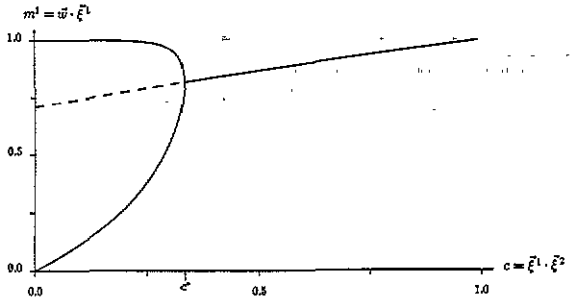

Figure 1. The curves show the post-synaptic potential, $V^1 = w \cdot \xi^1$, against the correlation $c = \xi^1 \cdot \xi^2$ between two patterns of equal length for $b = 2$. The continuous (broken) line shows the stable (unstable) stationary states. As long as the correlation is not too great ($c < c^*$), the neuron learns near one of the patterns and can discriminate between them. For $c > c^*$, the neuron learns the symmetric mixture and cannot distinguish the patterns.

two correlated patterns ξ^1 and ξ^2. For simplicity we will assume that both patterns have the same length, one, although the generalization to patterns of different lengths is quite straightforward. We look for solutions of the form

$$w = u\xi^1 + v\xi^2 \tag{3.7}$$

and we will assume $|w|^2 = 1$. (As demonstrated by equation (2.5) this will hold at the fixed point which will be stable to perturbation parallel to w.) Then $u^2 + v^2 + 2uvc = 1$, where $c = \xi^1 \cdot \xi^2$, and

$$\begin{aligned} \delta V^1 &= rv^2(1-c^2)(A(V^1) - (u/v)A(V^2)) \\ \delta V^2 &= -ruv(1-c^2)(A(V^1) - (u/v)A(V^2)). \end{aligned} \tag{3.8}$$

Thus the condition for w to be a stationary solution is

$$\frac{u}{v} = \frac{A(V^1)}{A(V^2)} = \left(\frac{u + vc - \phi}{v + uc - \phi}\right)^b \tag{3.9}$$

or writing $\zeta = u/v$ then the stationary value, ζ^*, satisfies $f(\zeta^*) = 0$ where

$$f(\zeta) = \zeta - \left(\frac{\zeta + c - \phi\sqrt{\zeta^2 + 2c\zeta + 1}}{1 + c\zeta - \phi\sqrt{\zeta^2 + 2c\zeta + 1}}\right)^b. \tag{3.10}$$

The stationary solutions for $b = 2$ and $\phi = 0$ against c are shown in figure 2. The condition for the stationary solution to be stable is $f'(\zeta^*) < 0$. One notices that $\zeta = 1$ is always a solution and corresponds to the symmetric mixed state $w \propto \xi^1 + \xi^2$. This mixed state becomes unstable when $b > b^*$, where

$$b^* = \frac{1 + c - \phi\sqrt{2(1+c)}}{1 - c}. \tag{3.11}$$

For $b > b^*$ two new stable solutions emerge one close to each of the patterns. Equation (3.11) can be solved for c^*, the correlation above which the mixture of the two patterns is stable. This is given by

$$c^* = \frac{b-1}{b+1} + \frac{\phi^2 - \phi\sqrt{\phi^2 + 4b(b+1)}}{(b+1)^2}. \tag{3.12}$$

This analysis can be extended to a greater number of patterns but the number of stationary solutions rapidly increases and they cannot, in general, be examined analytically. Nevertheless in the special case when all the patterns have the same mutual correlation, the stability of the completely mixed state can be studied. For P patterns with mutual correlation c one finds that the mixed state $w \propto \sum_{\mu=1}^{P} \xi^{\mu}$ is stable to perturbations in the direction of one of the patterns provided $c > (b-1)/(b+P-1)$ (where $\phi = 0$). From this analysis we can understand what would happen in the case of simple clusters. Assuming the patterns were generated by making a random perturbation away form a set of orthogonal (or randomly generated) prototypes so that within each cluster the correlation between patterns would be c, while the correlation between patterns in different clusters are zero (or $1/\sqrt{N}$). Then the neuron could learn to recognize the centre of one of the clusters if $c > (b-1)/(b+p-1)$, where p is the number of patterns in that cluster, or it would learn to recognize an individual pattern if $c < (b-1)/(b+p-1)$.

4. Signal-to-noise analysis for many random patterns

In the first part of this section we derived the steady-state equation for a neuron which has a macroscopic overlap with one of the patterns and a microscopic overlap with all the the other patterns. The second part of this section discusses the solutions of this steady-state equation.

4.1. Signal-to-noise calculation

We consider input activities, ξ_i^{μ} which are independently chosen from a distribution with zero mean and variance $1/\sqrt{N}$. The patterns will therefore have an average length of one and will have correlations of order $1/\sqrt{N}$. These correlations will prevent any pattern from being learned perfectly. As more and more patterns are shown the neuron will eventually be unable to remember any single pattern but will learn some mixture of very many patterns. We will assume that there is one pattern with a macroscopic correlation (of order one) with the weight vector. This pattern will be treated as the 'signal'. The other $P-1$ patterns are assumed to have microscopic correlations (of order $1/\sqrt{N}$) with the neuron. These patterns will be treated as 'noise'. Note, however, that this noise is not real stochastic noise—the neuron behaves deterministically—but comes from the large number of random patterns. In the derivation given in the following we do not attempt to give a full justification for every step. The results derived here will also be derived using mean-field theory where the approximations are more easily controlled.

We consider the case of a neuron that has a macroscopic overlap with pattern ξ^1 and microscopic overlaps with all the other patterns. The weight vector can be resolved into a component in the direction of pattern ξ^1 and a component in an orthogonal direction x which will depend on all the other patterns

$$w = u\xi^1 + vx \tag{4.1}$$

where ξ^1 and x are unit vectors and where $\xi^1 \cdot x = 0$. We assume that $|w|^2 = 1$ so that $u^2 + v^2 = 1$. Setting $\delta w = 0$ in equation (3.1) it is easily verified that

$$x = \frac{1}{C} \sum_{\mu} A(V^{\mu})(\xi^{\mu} - \Delta_1^{\mu}\xi^1) \tag{4.2}$$

where $\Delta_1^\mu = \xi^1 \cdot \xi^\mu$ are independent Gaussianly distributed random variables with variance $1/\sqrt{N}$ and where C is the normalization factor

$$C = \left| \sum_\mu A(V^\mu)(\xi^\mu - \Delta_1^\mu \xi^1) \right| = \sqrt{\sum_\mu A^2(V^\mu)(1 + \mathcal{O}(1/N))}. \tag{4.3}$$

The direction x has a small dependence on each of the patterns ξ^μ. Separating out the part that depends on ξ^μ,

$$\xi^\mu \cdot x = \Delta_x^\mu + \frac{A(V^\mu)\left(1 - (\Delta_1^\mu)^2\right)}{C} \tag{4.4}$$

where

$$\Delta_x^\mu = \xi^\mu \cdot \frac{1}{C} \sum_{\nu \neq \mu} A(V^\nu)(\xi^\nu - \Delta_1^\nu \xi^\nu) \tag{4.5}$$

is essentially the overlap between two independently chosen, random unit length N-vectors. Thus the Δ_x^μs are independent Gaussian variables with variance $1/\sqrt{N}$. The term $(\Delta_1^\mu)^2$ is of order $1/N$ and will be neglected.

The dynamical equation in the direction ξ^1 is

$$\delta u = \xi^1 \cdot \delta w = \sum_\mu A(V^\mu)(\xi^\mu \cdot \xi^1 - u V^\mu). \tag{4.6}$$

For $\mu \neq 1$,

$$V^\mu = \xi^\mu \cdot (u\xi^1 + v x) = u\Delta_1^\mu + v\Delta_x^\mu + vA(V^\mu)/C \tag{4.7}$$

where we have used (4.4). The last term depends on V^μ and can be expanded out *ad infinitum*. However, the magnitude of this term is of order $N^{b/2}$ and thus it will be smaller than the other terms provided $b > 1$. Thus we need to keep only the first few terms to find the large-N behaviour. Substituting equation (4.7) into equation (4.6) we find

$$\delta u = A(u)(1 - u^2) + \sum_\mu A\left(u\Delta_1^\mu + v\Delta_x^\mu + \frac{vA(V^\mu)}{C}\right)$$
$$\left((1 - u^2)\Delta_1^\mu - uv\Delta_x^\mu - uv\frac{A(V^\mu)}{C}\right). \tag{4.8}$$

Using $1 - u^2 = v^2$, and expanding to first order in $A(v^\mu)/C$, the last term becomes

$$v \sum_\mu \left(A(u\Delta_1^\mu + v\Delta_x^\mu) + \frac{vA(V^\mu)}{C} A'(u\Delta_1^\mu + v\Delta_x^\mu)\right)\left(v\Delta_1^\mu - u\Delta_x^\mu - u\frac{A(V^\mu)}{C}\right). \tag{4.9}$$

To find the stationary equation for u in the large N limit we replace the sum over μ by its average value. This is straightforward as Δ_1^μ and Δ_x^μ are independent Gaussian variables. It is simpler to make the change of variables $z_1 = \sqrt{N}(u\Delta_1 + v\Delta_x)$ and $z_2 = \sqrt{N}(v\Delta_1 - u\Delta_x)$.

Both z_1 and z_2 are Gaussianly distributed random variables. After averaging the leading order term equation (4.9) becomes

$$vuPk/N^bC \tag{4.10}$$

where

$$k = \int \mathrm{D}z_1 \, \bar{A}^2(z_1) \qquad \mathrm{D}z = \frac{\mathrm{d}z}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} \tag{4.11}$$

and where

$$\bar{A}(z) = (z-\kappa)^b \Theta(z-\kappa) \tag{4.12}$$

and $\kappa = \phi/\sqrt{N}$ is the rescaled threshold. When $\kappa = 0$, $k = 2^{b-1}\Gamma(b+1/2)/\sqrt{\pi}$ (for integer b, $k = (2b-1)!!/2$). Similarly we find the average of the constant C, defined in equation (4.3), which, to leading order, is equal to $\sqrt{kP/N^b}$. Thus in the limit $N \to \infty$ equation (4.8) becomes

$$\delta u = v\left(A(u)v - u\sqrt{Pk/N^b}\right). \tag{4.13}$$

In deriving this equation we used the value of x at the fixed point so this equation is strictly only valid at the fixed point. The fixed point equation is found by putting $\delta u = 0$. Squaring we find

$$(1-u^2)A^2(u)/u^2 = \alpha k \tag{4.14}$$

where $\alpha = P/N^b$. We will discuss the solution to this equation in detail in the next part of this section. We can also extract the finite-size corrections from this formalism. This will be discussed in section 6.

4.2. Solutions of the signal-to-noise equation

In the remainder of this section we discuss the solutions to equation (4.14). Provided α is less than some critical value, equation (4.14) has two solutions: a stable solution u^* and an unstable solution $\bar{u}$ ($u^* \geqslant \bar{u}$). The stable solution corresponds to the point where the neuron has learned the pattern ξ^1. There is a small shift away from the pattern caused by the other patterns. The unstable solution corresponds to the point where the attraction towards the pattern stored at $u = u^*$ is exactly balanced by the attraction towards the totally mixed solution at $u = 0$. In other words, $\bar{u}$ is where the strength of the signal equals that of the noise. Thus $\bar{u}$ measures the size of the basin of attraction to the fixed point u^*.

As more patterns are added and α increases towards α_c, u^* decreases and $\bar{u}$ increases until they coalesce at α_c. Above α_c only the $u = 0$ solution exists and the neuron fails to recognize any pattern. This transition point is given by

$$\alpha_c = \frac{1}{k(b-1)}\left(\frac{b-1}{b}\right)^b \tag{4.15}$$

at which point $u^* = \bar{u} = \sqrt{(b-1)/b}$. These fixed points are shown in figure 2 plotted against $\alpha = P/N^b$ for $b = 2$ and $\kappa = 0$. Increasing the threshold κ decreases the noise and

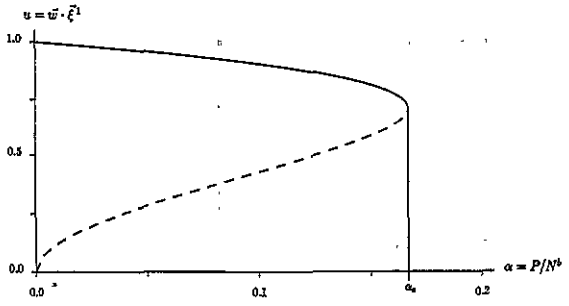

Figure 2. The fixed points of the signal-to-noise equation, equation (4.14) against $\alpha = P/N^b$. The stable solution, u^*, is shown in the full curve; the unstable solution, $\bar{u}$, is shown as the broken curve. Here, $b = 2$, $\kappa = 0$ and the phase transition occurs at $\alpha = 1/6$.

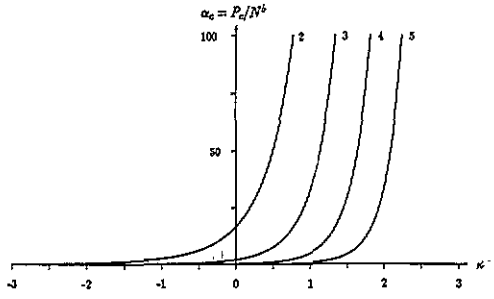

Figure 3. The critical capacity α_c for a neuron to lose all memory of a pattern against the rescaled threshold $\kappa = \phi/\sqrt{N}$ for $b = 2$, 3, 4 and 5. Note that the number of patterns scale as N^b so that the curves for different b should not be compared.

thus increases the capacity. Figure 3 shows α_c against κ for different values of b. (Note that $\alpha = P/N^b$ so the curves at different values of b cannot be directly compared—for any reasonable size of neuron the number of patterns that can be shown to the neuron before it becomes overloaded increases substantially with b.)

The transitions just described represent 'forgetting'—for $\alpha > \alpha_c$ no memory of any pattern can be retained. This is true even if the weights were set to be equal to the pattern before learning. However, a neuron will only learn a pattern in the first place if the weights at the start of learning were within the basin of attraction of the stored pattern state u^*. A glance at figure 2 shows that this could be unlikely, since $\bar{u}$ determines the size of the basin of attraction. For example, for the parameters in that figure, if α is 0.1, then the initial weights would have to have an overlap with a pattern of almost 0.5 in order to learn that pattern. Assuming no prior knowledge of the patterns to be learned, this will almost never happen. Thus, we address the question: how small must α be in order for the neuron to learn a pattern from random starting weights.

The neuron will learn a pattern only if the initial overlap of the weights with that pattern is larger than $\bar{u}$. Thus, there is another transition which occurs when $\bar{u}$ is greater than the initial overlap of the weights with all patterns. This could be deemed the 'learning' transition because it separates the system which can learn a pattern, from one which will not forget a learned pattern but cannot learn. To compute where this transition occurs, consider the overlap between the initial weights and the pattern with the largest overlap

with those weights. Call this quantity $V_0^{\max}$. The magnitude of $V_0^{\max}$ will depend on the initial distributions of the input activities and the synaptic weights. Furthermore it will be sample dependent. Assuming the same Gaussian distribution for the initial weights as for the patterns, the average value, $\bar{V}_0^{\max}$, will asymptotically go like $\sqrt{2\ln(P)/N}$. The corresponding typical number of patterns $\tilde{P}_c$ that can be shown to a neuron before it is unable to learn a new pattern is given by $\bar{V}_0^{\max}(\tilde{P}_c) = \bar{u}(\tilde{P}_c)$. This implies an asymptotic result $\tilde{P}_c \approx kN(\ln kN)^{b-1}$. This asymptotic result is not terribly useful in finite systems, however, because convergence to it is only logarithmic in N. Figure 4 shows $\tilde{P}_c/N$ against N computed numerically for $b = 2$, 2.5, 3 and 3.5 and with $\kappa = 0$. The learning transition occurs for $\tilde{P}_c$ much larger than N. It increases with b because the nonlinearity damps out noise.

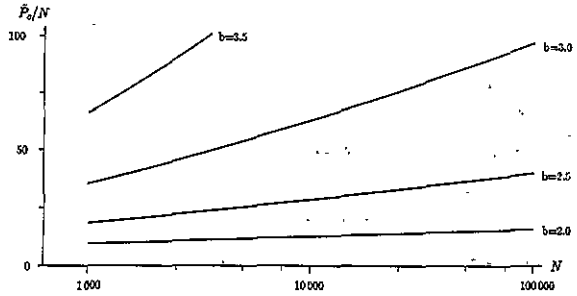

Figure 4. The neuron can learn to recognize a pattern from random starting weights only if the number of patterns is less than $\tilde{P}_c$. This plot shows $\tilde{P}_c/N$ against the number of synapses N for $\kappa = 0$ and for $b = 2$, 2.5, 3 and 3.5. The initial correlations between the pattern vectors and weight vector are Gaussian distributed.

The effect of κ on $\tilde{P}_c/N$ is shown in figure 5, for fixed N. Again increasing b reduces the amount of noise and thus increases the ability of the neuron to learn a new pattern. Note that the neuron will only learn if $V_0^{\max} > \kappa$ so for $\kappa > 0$ there will typically be a minimum number of patterns that must be presented in order for the neuron to learn.

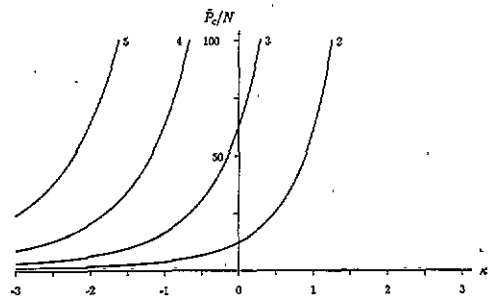

Figure 5. $\tilde{P}_c/N$ against the rescaled threshold $\kappa = \phi/\sqrt{N}$ for $N = 10\,000$ and $b = 2$, 3, 4 and 5.

In summary, the neuron either learns to distinguish one pattern from another, or it learns to distinguish some mixed state. As the number of random patterns P increases there are

two transitions. The first occurs at $\tilde{P}_c \approx kN(\ln kN)^{b-1}$. Below this point, the neuron can learn one of the patterns from random initial starting weights. Above this point but below the second transition point $P_c = \alpha_c N^b$, the neuron will not forget a pattern which it has learned, but is unlikely to learn from a random start. Above the second transition, the neuron learns a mixture of very many patterns. We have not examined this mixed state as for random patterns there is nothing in the distribution of the patterns for the neuron to learn. For more complicated distributions the nature of the mixed state will be more important.

We have seen that increasing b and κ increases both α_c and $\tilde{P}_c$. However increasing b and κ will slow the rate of learning. To see this, we consider learning a single pattern from a small initial overlap V_0. The time required for the overlap to grow to macroscopic size is proportional to $rb(V_0 - \kappa/\sqrt{N})^{1-b}$. For random initial weights $V_0^{\max}$ is of order $1/\sqrt{N}$. Thus, for random patterns, the time taken will scale as $N^{(b-1)/2}$. The effect of having many random patterns will be to introduce a 'friction'-like term which will slow down the learning. For a neuron to learn quickly it is advantageous to choose b and κ to be as small as possible. Of course they must be chosen sufficiently large so that the neuron is able to learn new patterns and this will depend on how many random patterns the neuron experiences and on how noisily the neuron behaves.

5. Mean-field theory for Hebbian learning

In section 2.2 we introduced a partition function which describes the distribution of stationary solutions of the dynamical equations. Here we will study this partition function using mean-field theory. This is done in two stages. In the first part of this section, we examine the case when P/N^b is negligible and the partition function self-averages. In the second part, we consider the case when $P = \alpha N^b$ for arbitrary α where the replica approach is required. The calculations presented here resemble those of Amit *et al* for the Hopfield model [11–13] and Gardner's calculation for the capacity of the spherical perceptron [14–16]. We therefore give only a brief outline of the calculation. Throughout this section we assume $b > 1$. The limit $b \to 1$ cannot be taken because terms of order N have been ignored when compared with terms of order N^b. When $A(V^\mu) = V^\mu$, the partition function is just that of the spherical Hopfield model. This is also the partition function which corresponds to Oja's rule. We will not discuss the mean-field theory for this case, because the only stable solution for this model is the principal component of the correlation matrix of patterns.

The significance of the calculation presented here is that it validates the signal-to-noise analysis. The resulting equation, (5.54), and the signal-to-noise equation, (4.14), are the same. In addition, it provides a more general formalism which allows that result to be extended to the consideration of more than one macroscopic overlap and to additive noise. The replica calculation is of interest as the energy function is not quadratic and the position of the first-order phase transition scales as N^b in contrast to the Hopfield and perceptron calculation where the energy is quadratic and the position of the first-order phase transition scales linearly with N.

5.1. A few random patterns

The partition function in equation (2.6) can re-written as

$$Z = \prod_i \int \mathrm{d}w_i \, \delta\Big(\sum_i (w_i)^2 - N\Big) \prod_\mu \int \mathrm{d}V^\mu \, \delta\Big(V^\mu - \frac{1}{N}\sum_i w_i \xi_i^\mu\Big) \exp\Big\{\beta N \sum_\mu u(V^\mu)\Big\} \tag{5.1}$$

with

$$u(V^\mu) = \frac{1}{b+1}(V^\mu - \phi)^{b+1}\,\Theta(V^\mu - \phi) \tag{5.2}$$

so that $u'(V^\mu) = A(V^\mu)$, the activation function. Introducing the integral representation for the delta functions

$$\delta\Big(\sum_i (w_i)^2 - N\Big) = \int_{-i\infty}^{i\infty} \frac{d\lambda}{4\pi} \exp\Big\{-\frac{\lambda}{2}\Big(\sum_i (w_i)^2 - N\Big)\Big\} \tag{5.3}$$

and

$$\prod_\mu \delta\Big(V^\mu - \frac{1}{N}\sum_i w_i \xi_i^\mu\Big) = \prod_\mu \int_{-i\infty}^{i\infty} \frac{dt^\mu}{2\pi/N} \exp\Big\{-N\sum_\mu t^\mu\Big(V^\mu - \frac{1}{N}\sum_i w_i \xi_i^\mu\Big)\Big\} \tag{5.4}$$

then the sums over the different synapses i all decouple and we can replace w_i and ξ_i^μ by a representative weight and pattern w and ξ^μ, respectively. The average of the partition function can then be written as

$$\langle Z\rangle = C\prod_\mu \int dV^\mu\, dt^\mu \int d\lambda\, e^{NG(\lambda, V^\mu, t^\mu)} \tag{5.5}$$

where C is a constant, $\langle\cdots\rangle$ denotes averaging over the patterns, and

$$G(\lambda, V^\mu, t^\mu) = \frac{\lambda}{2} - \sum_\mu t^\mu V^\mu + \beta\sum_\mu u(V^\mu) + \left\langle \log\Big[\int \frac{dw}{\sqrt{2\pi}} \exp\Big\{-\frac{\lambda}{2}w^2 + w\sum_\mu \xi^\mu t^\mu\Big\}\Big]\right\rangle. \tag{5.6}$$

Performing the integral over w

$$\left\langle \log\Big[\int \frac{dw}{\sqrt{2\pi}} \exp\Big\{-\frac{\lambda}{2}w^2 + w\sum_\mu \xi^\mu t^\mu\Big\}\Big]\right\rangle = -\frac{1}{2}\log[\lambda] + \frac{1}{2\lambda}\left\langle\Big(\sum_\mu t^\mu \xi^\mu\Big)^2\right\rangle. \tag{5.7}$$

Since the patterns are independent $\langle \xi^\mu \xi^\nu\rangle = \delta^{\mu,\nu}$ and

$$\left\langle\Big(\sum_\mu t^\mu \xi^\mu\Big)^2\right\rangle = \sum_\mu (t^\mu)^2 \tag{5.8}$$

thus

$$G(\lambda, V^\mu, t^\mu) = \frac{\lambda}{2} - \frac{1}{2}\log[\lambda] - \sum_\mu t^\mu V^\mu + \beta\sum_\mu u(V^\mu) + \frac{1}{2\lambda}\sum_\mu (t^\mu)^2. \tag{5.9}$$

In the large N limit the integral in the partition function (5.5) is dominated by its saddle-point values. To find the saddle point we minimize with respect to λ, t^μ and V^μ, that is we seek solutions to the saddle-point equations

$$\left(\frac{\partial G}{\partial \lambda}\right)_{V^\mu, t^\mu} = \left(\frac{\partial G}{\partial t^\mu}\right)_{\lambda, V^\mu} = \left(\frac{\partial G}{\partial V^\mu}\right)_{\lambda, t^\mu} = 0. \tag{5.10}$$

These equations are satisfied when

$$t^\mu = \lambda V^\mu \qquad V^\mu = \frac{\beta}{\lambda} A(V^\mu) \qquad \lambda = \frac{1}{1 - \sum_\mu (V^\mu)^2} \tag{5.11}$$

where we have used $u'(V^\mu) = A(V^\mu)$. Substituting back into equation (5.9) and assuming the partition function self-averages, we find that the free energy is given by

$$f = -\frac{1}{2\beta} - \frac{1}{2\beta} \log \left[1 - \sum_\mu (V^\mu)^2 \right] - \sum_\mu u(V^\mu) \tag{5.12}$$

where V^μ satisfies the equation

$$V^\mu = \beta \left(1 - \sum_\nu (V^\nu)^2 \right) A(V^\mu). \tag{5.13}$$

The only solutions to this equation are permutations of the solutions

$$\boldsymbol{V} = (\underbrace{V_n, V_n, \ldots, V_n}_{n}, 0, 0, \ldots). \tag{5.14}$$

Substituting (5.14) into equation (5.13) and taking $\phi = 0$ we find V_n satisfies

$$(1 - n V_n^2) V_r^{b-1} = T. \tag{5.15}$$

Calculating the free energies for these solutions one finds the single-pattern solutions $V^\mu = V_1 \delta^{\mu,\nu}$ are the ground states. To examine the stability of these solutions we look at the Hessian matrix

$$\frac{\partial^2 f}{\partial V^\mu \partial V^\nu} = \left(\frac{1}{\beta(1 - \sum_\lambda (V^\lambda)^2)} - A'(V^\mu) \right) \delta^{\mu,\nu} + \frac{2 V^\mu V^\nu}{\beta (1 - \sum_\lambda (V^\lambda)^2)^2}. \tag{5.16}$$

The Hessian matrix has three distinct eigenvalues at the n-pattern mixed solutions (for $n > 1$): a non-degenerate eigenvalue

$$\lambda_1 = (1 - b) V_n^{b-1} + 2\beta n V_n^{2b} \tag{5.17}$$

an eigenvalue with degeneracy $P - n$,

$$\lambda_2 = V_n^{b-1} \tag{5.18}$$

and an eigenvalue with degeneracy $n - 1$,

$$\lambda_3 = (1 - b) V_n^{b-1}. \tag{5.19}$$

For $b > 1$ this last eigenvalue, which is associated with fluctuations of anisotropy in the space of the n learned patterns, is always negative. Thus the mixed solutions are all unstable to fluctuation in the direction of any of the patterns. The only stable solutions are the single-pattern solutions.

When $T = 0$ and $\phi = 0$ we retrieve the symmetric solutions of section 3.1 with $V_n = 1/\sqrt{n}$. Note that for $T = 0$ the eigenvalues λ_2 and λ_3 in this section are identical

to those in section 3.1 up to a constant factor of $-r$. Whereas, from (5.17), $\lambda_1 \to \infty$ as $T \to 0$ reflecting the strong constraint $|\boldsymbol{w}|^2 = 1$ in contrast to section 3.1 where λ_1 is finite reflecting the fact that the condition $|\boldsymbol{w}^*|^2 = 1$ is a consequence of the dynamics.

As T increases, V_n decreases until it reaches a critical transition temperature, T_c, at which point the only solution is $V^\mu = 0$ for all μ. This behaviour is reminiscent of the behaviour described in the previous section when the neuron was overloaded with patterns. For the single-pattern solution the critical temperature is given by

$$T_c = \frac{2}{b+1}\left(\frac{b-1}{b+1}\right)^{(b-1)/2} \tag{5.20}$$

at which point V_1 decreases to $(b-1)/(b+1)$ and λ_1 goes to zero. The fraction of the $\boldsymbol{w}$-sphere explored by the synaptic weight is $\sqrt{1-\sum_\mu (V^\mu)^2}$: when $T \to 0$ this fractional volume goes to zero as $\sqrt{T}$.

5.2. *Many random patterns:* $P = \alpha N^b$

When the number of patterns is of order αN^b the partition function will no longer self-average so we use the replica trick. The n-replica partition function is

$$Z^n = \prod_{i=1}^{N}\prod_{a=1}^{n}\int \mathrm{d}w_i^a \prod_{a=1}^{n}\delta\Big(\sum_{i=1}^{N}(w_i^a)^2 - N\Big)\exp\left\{\beta N \sum_{a=1}^{n}\sum_{\mu=1}^{P} u\Big(\frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu w_i^a\Big)\right\}. \tag{5.21}$$

We divide up the patterns into two sets: those with macroscopic overlaps (of order one) with the synaptic weight vector which we label $\mu = 1, 2, \ldots, s$ and those with microscopic overlaps (of order $1/\sqrt{N}$) which we label $\bar{\mu} = s+1, \ldots, P$. The partition function can then be written as

$$\begin{aligned} Z^n = \prod_{i,a}\int \mathrm{d}w_i^a \prod_a \delta\Big(\sum_i (w_i^a)^2 - N\Big) \prod_{a,\mu}\int \mathrm{d}V^{\mu,a} \prod_{a,\mu}\delta\Big(V^{\mu,a} - \frac{1}{N}\sum_i \xi_i^\mu w_i^a\Big) \\ \times \prod_{a,\bar{\mu}}\int \mathrm{d}\bar{V}^{\bar{\mu},a} \prod_{a,\bar{\mu}}\delta\Big(\bar{V}^{\bar{\mu},a} - \frac{1}{\sqrt{N}}\sum_i \xi_i^{\bar{\mu}} w_i^a\Big) \\ \times \exp\left\{\beta N \sum_{a,\mu} u(V^{\mu,a}) + \beta N^{(1-b)/2}\sum_{a,\bar{\mu}} \bar{u}(\bar{V}^{\bar{\mu},a})\right\} \end{aligned} \tag{5.22}$$

where we have used

$$\begin{aligned} u\Big(\frac{\bar{V}^{\bar{\mu},a}}{\sqrt{N}}\Big) &= \frac{1}{b+1}\Big(\frac{\bar{V}^{\bar{\mu},a}}{\sqrt{N}} - \phi\Big)^{b+1}\Theta\Big(\frac{\bar{V}^{\bar{\mu},a}}{\sqrt{N}} - \phi\Big) \\ &= N^{-(b+1)/2}\frac{1}{b+1}(\bar{V}^{\bar{\mu},a} - \kappa)^{b+1}\Theta(\bar{V}^{\bar{\mu},a} - \kappa) = N^{-(b+1)/2}\bar{u}(\bar{V}^{\bar{\mu},a}) \end{aligned} \tag{5.23}$$

and where again we have introduced a rescaled threshold $\kappa = \phi/\sqrt{N}$. To integrate out the microscopic overlaps we introduce the integral representation for the delta function

$$\prod_{a,\bar{\mu}}\delta\Big(\bar{V}^{\bar{\mu},a} - \frac{1}{\sqrt{N}}\sum_i \xi_i^{\bar{\mu}} w_i^a\Big) = \prod_{a,\bar{\mu}}\int \frac{\mathrm{d}\bar{t}^{\bar{\mu},a}}{2\pi}\exp\left\{\mathrm{i}\sum_{\bar{\mu},a}\bar{t}^{\bar{\mu},a}\Big(\bar{V}^{\bar{\mu},a} - \frac{1}{\sqrt{N}}\sum_i \xi_i^{\bar{\mu}} w_i^a\Big)\right\}. \tag{5.24}$$

To perform the average over $\xi_i^{\bar{\mu}}$ for $\bar{\mu} = s+1, \ldots, P$ we expand the exponential, and use $\langle \xi_i^{\mu} \rangle = 0$ and $\langle (\xi_i^{\mu})^2 \rangle = 1$, then

$$\left\langle \prod_{i,\bar{\mu}} \exp\left\{ \frac{\mathrm{i}}{\sqrt{N}} \xi_i^{\bar{\mu}} \sum_a \bar{t}^{\bar{\mu},a} w_i^a \right\} \right\rangle_{\xi_i^{\bar{\mu}}} = \prod_{i,\bar{\mu}} \left\{ 1 - \frac{1}{2N} \left(\sum_a \bar{t}^{\bar{\mu},a} w_i^a \right)^2 + \mathcal{O}\left(\frac{1}{N^{3/2}} \right) \right\}$$

$$= \exp\left\{ -\frac{1}{2} \sum_{\bar{\mu}} \sum_a (\bar{t}^{\bar{\mu},a})^2 - \sum_{\bar{\mu}} \sum_{a<b} q^{ab} \bar{t}^{\bar{\mu},a} \bar{t}^{\bar{\mu},b} \right\} \tag{5.25}$$

where, in the last step, we used $\sum_i (w_i^a)^2 = N$ and

$$q^{ab} = \frac{1}{N} \sum_i w_i^a w_i^b. \tag{5.26}$$

We impose (5.26) using the integral representation of the delta function

$$\prod_{a<b} \delta\left(q^{ab} - \frac{1}{N} \sum_i w_i^a w_i^b \right) = \prod_{a<b} \int \frac{\mathrm{d}p^{ab}}{2\pi/N} \exp\left\{ \mathrm{i}N \sum_{a<b} p^{ab} \left(q^{ab} - \frac{1}{N} \sum_i w_i^a w_i^b \right) \right\} \tag{5.27}$$

and similarly to impose the constraint on the $V^{\mu,a}$s and the w_i^as

$$\prod_{a,\mu} \delta\left(V^{\mu,a} - \frac{1}{N} \sum_i \xi_i^{\mu} w_i^a \right) = \prod_{a,\mu} \int \frac{\mathrm{d}t^{\mu,a}}{2\pi/N} \exp\left\{ \mathrm{i}N \sum_{\mu,a} t^{\mu,a} \left(V^{\mu,a} - \frac{1}{N} \sum_i \xi_i^{\mu} w_i^a \right) \right\} \tag{5.28}$$

$$\prod_a \delta\left(\sum_i (w_i^a)^2 - N \right) = \prod_a \int \frac{\mathrm{d}\lambda^a}{4\pi} \exp\left\{ \frac{\mathrm{i}}{2} \sum_a \lambda^a \left(\sum_i (w_i^a)^2 - N \right) \right\} \tag{5.29}$$

then the sums on i and the sums on $\bar{\mu}$ decouple.

Thus the n-replica partition function can be written as

$$\langle Z^n \rangle = C \prod_{a<b} \int \mathrm{d}p^{ab}\, \mathrm{d}q^{ab} \prod_a \int \mathrm{d}\lambda^a \prod_{\mu=1}^{s} \prod_a \int \mathrm{d}V^{\mu,a}\, \mathrm{d}t^{\mu,a}$$

$$\times \exp\left\{ N G(p^{ab}, q^{ab}, \lambda^a, V^{\mu,a}, t^{\mu,a}) \right\} \tag{5.30}$$

where C is a constant and

$$G(p^{ab}, q^{ab}, \lambda^a, V^{\mu,a}, t^{\mu,a}) = \mathrm{i} \sum_{a<b} p^{ab} q^{ab} - \frac{\mathrm{i}}{2} \sum_a \lambda^a + i \sum_a \sum_{\mu=1}^{s} t^{\mu,a} V^{\mu,a} + \beta \sum_a \sum_{\mu=1}^{s} u(V^{\mu,a})$$

$$+ G_1(q^{ab}) + G_2(p^{ab}, \lambda^a; t^{\mu,a}) \tag{5.31}$$

with

$$G_1(q^{ab}) = \frac{P}{N} \log\left\{ \prod_a \int \frac{\mathrm{d}\bar{V}^a\, \mathrm{d}\bar{t}^a}{2\pi} \exp\left\{ -\frac{1}{2} \sum_a (\bar{t}^a)^2 - \sum_{a<b} q^{ab} \bar{t}^a \bar{t}^b \right.\right.$$

$$\left.\left. - \mathrm{i} \sum_a \bar{t}^a \bar{V}^a + \beta N^{(1-b)/2} \sum_a \bar{u}(\bar{V}^a) \right\} \right\} \tag{5.32}$$

and

$$G_2(p^{ab}, \lambda^a, t^{\mu,a}) = \left\langle \log\left[\prod_a \int \frac{\mathrm{d}w^a}{\sqrt{2\pi}} \exp\left\{ \frac{\mathrm{i}}{2} \sum_a \lambda^a (w^a)^2 - \mathrm{i} \sum_{a<b} p^{ab} w^a w^b \right.\right.\right.$$
$$\left.\left.\left. - \mathrm{i} \sum_a \sum_{\mu=1}^{s} t^{\mu,a} w^a \xi^\mu \right] \right\} \right\rangle_{\xi^\mu}. \tag{5.33}$$

To find the free energy we make the usual ansatz that the macroscopic order parameters are symmetric in the different replicas

$$p^{ab} = \mathrm{i}p \qquad q^{ab} = q \qquad \forall a < b$$
$$\lambda^a = \mathrm{i}\lambda V^{\mu,a} = V^\mu \qquad t^{\mu,a} = \mathrm{i}t^\mu \qquad \forall a.$$

Then the n-replica partition function becomes

$$\langle Z^n \rangle = C \int \mathrm{d}p\, \mathrm{d}q\, \mathrm{d}\lambda \prod_{\mu=1}^{s} \int \mathrm{d}V^\mu\, \mathrm{d}t^\mu \exp\{NG(p, q, \lambda, V^\mu, t^\mu)\} \tag{5.34}$$

where to first order in n

$$G(p, q, \lambda, V^\mu, t^\mu) = \frac{n}{2} pq + \frac{n}{2}\lambda - n \sum_{\mu=1}^{s} t^\mu V^\mu + n\beta \sum_{\mu=1}^{s} u(V^\mu) + G_1(q) + G_2(p, \lambda, t^\mu) \tag{5.35}$$

with

$$G_1(q) = \frac{P}{N} \log\left[\prod_a \int \frac{\mathrm{d}\bar{V}^a\, \mathrm{d}\bar{t}^a}{2\pi} \exp\left\{ -\frac{1}{2} \sum_a (\bar{t}^a)^2 - q \sum_{a<b} \bar{t}^a \bar{t}^b \right.\right.$$
$$\left.\left. - \mathrm{i} \sum_a \bar{t}^a \bar{V}^a + \beta N^{(1-b)/2} \sum_a \bar{u}(\bar{V}^a) \right\} \right] \tag{5.36}$$

and

$$G_2(p, \lambda, t^\mu) = \left\langle \log\left[\prod_a \int \frac{\mathrm{d}w^a}{\sqrt{2\pi}} \exp\left\{ -\frac{\lambda}{2} \sum_a (w^a)^2 + p \sum_{a<b} w^a w^b + \sum_{\mu=1}^{s} t^\mu \xi^\mu \sum_a w^a \right\} \right] \right\rangle. \tag{5.37}$$

The $\bar{t}^a$s in G_1 can be decoupled and the integral over $\bar{V}^a$ and $\bar{t}^a$ performed. Then

$$G_1(q) = \frac{P}{N} \log\left[\int \mathrm{D}z_1 \left(\int \mathrm{D}z_2 \exp\left\{ \beta N^{(1-b)/2} \bar{u}(\sqrt{q} z_1 + \sqrt{1-q} z_2) \right\} \right)^n \right] \tag{5.38}$$

where $\mathrm{D}z$ is the Gaussian measure defined in equation (4.11). Using the trick of writing

$$\log\left[\int \mathrm{D}z f^n(z) \right] = n \int \mathrm{D}z \log[f(z)] + \mathcal{O}(n^2) \tag{5.39}$$

and expanding in powers of $N^{(1-b)/2}$ we find after some algebra that, to first order in n and neglecting terms of order $PN^{(1-3b)/2}$,

$$G_1(q) = n\beta \frac{P}{N^{(b+1)/2}} \int \mathrm{D}z\bar{u}(z) + n\beta^2 \frac{P}{2N^b} \times \int \mathrm{D}z_1 \left(\bar{u}^2(z_1) - \left(\int \mathrm{D}z_2 \bar{u}(\sqrt{q}z_1 + \sqrt{1-q}z_2) \right)^2 \right). \tag{5.40}$$

Decoupling the weights in G_2 and evaluating the integrals we find

$$G_2(p, \lambda, t^\mu) = -\frac{n}{2}\log[\lambda + p] + \frac{n}{2(\lambda+p)} \sum_{\mu=1}^{s} (t^\mu)^2 + \frac{np}{2(\lambda+p)} + \mathcal{O}(n^2). \tag{5.41}$$

Using the replica trick we find the free energy per synapse is given by

$$f = -\lim_{N\to\infty} \frac{1}{N\beta} \lim_{n\to 0} \frac{\langle Z^n \rangle - 1}{n} = -\frac{1}{\beta} \lim_{n\to 0} \operatorname*{extr}_{p,q,\lambda,V^\mu,t^\mu} \frac{G(p, q, \lambda, V^\mu, t^\mu)}{n} \tag{5.42}$$

where we have exchanged the order of the limits to allow us to evaluate the n-replica partition function using the saddle-point approximation. Thus the free energy is given by the extremum of

$$\beta f = -\frac{pq}{2} - \frac{\lambda}{2} + \sum_{\mu=1}^{s} t^\mu V^\mu - \beta \sum_{\mu=1}^{s} u(V^\mu) + \frac{1}{2}\log[\lambda + p] - \frac{p}{2(\lambda+p)} - \frac{1}{2(\lambda+p)} \sum_{\mu=1}^{s} (t^\mu)^2 - G_1^0(q) \tag{5.43}$$

where

$$G_1^0(q) = \lim_{n\to 0} \frac{G_1(q)}{n}. \tag{5.44}$$

Using the saddle-point equations

$$\frac{\partial f}{\partial t^\mu} = \frac{\partial f}{\partial \lambda} = \frac{\partial f}{\partial p} = 0 \tag{5.45}$$

we find

$$t^\mu = (\lambda + p)V^\mu \qquad p = \frac{q}{(1-q)^2} - \sum_{\mu=1}^{s} (t^\mu)^2 \qquad \lambda = \frac{1-2q}{(1-q)^2} + \sum_{\mu=1}^{s} (t^\mu)^2 \tag{5.46}$$

and thus the free energy (up to an additive constant) becomes

$$f = -\frac{1}{2\beta} + \frac{1}{2\beta(1-q)} \sum_{\mu=1}^{s} (V^\mu)^2 - \sum_{\mu=1}^{s} u(V^\mu) - \frac{q}{2\beta(1-q)} - \frac{1}{2\beta}\log[1-q] - \frac{1}{\beta} G_1^0(q). \tag{5.47}$$

The last two saddle-point equations

$$\frac{\partial f}{\partial V^\mu} = \frac{\partial f}{\partial q} = 0 \tag{5.48}$$

imply

$$V^\mu = (1-q)\beta \sum_{\mu=1}^{s} A(V^\mu) \tag{5.49}$$

$$\sum_{\mu=1}^{s} (V^\mu)^2 = q + 2(1-q)^2 \frac{\mathrm{d}G_1^0}{\mathrm{d}q} \tag{5.50}$$

where again we have use $u'(V^\mu) = A(V^\mu)$. Putting $\alpha = P/N^b$, then

$$\frac{\mathrm{d}G_1^0}{\mathrm{d}q} = \frac{-\alpha\beta^2}{2\sqrt{q(1-q)}} \int \mathrm{D}x\, \mathrm{D}y\, \mathrm{D}z\, y\, \bar{A}(x)\, \bar{u}(qx + \sqrt{q(1-q)}y + \sqrt{1-q}z) \tag{5.51}$$

where we have used the change of variables $x = \sqrt{q}\, z_1 + \sqrt{1-q}\, z_2$ and $y = \sqrt{1-q}\, z_1 - \sqrt{q}\, z_2$ and where $\bar{A}(x) = \bar{u}'(x)$.

For zero noise, $\beta \to \infty$, we find $q \to 1$ while $\beta(1-q)$ remains finite. Expanding to first order in $\sqrt{1-q}$, we find

$$\frac{\mathrm{d}G_1^0}{\mathrm{d}q} = -\frac{\alpha\beta^2}{2} \int \mathrm{D}x\, \bar{A}^2(x) \tag{5.52}$$

this is just $-\alpha\beta^2 k/2$, where k is defined in equation (4.11). Thus the fixed-point equation for q becomes

$$1 - \sum_{\mu=1}^{s} (V^\mu)^2 = (1-q)^2 \beta^2 \alpha k. \tag{5.53}$$

When there is only a single pattern with macroscopic overlap, then equations (5.49) and (5.53) imply

$$(1-V^2)\frac{A^2(V)}{V^2} = \alpha k. \tag{5.54}$$

This is identical to equation (4.14) derived using the signal-to-noise analysis.

There is also a solution $V^\mu = 0$ for all μ. In this case the synaptic weight vector aligns itself with a mixture of many patterns, but does not have a macroscopic overlap with any one of them. For this solution $\beta(1-q) = 1/\sqrt{\alpha k}$. At zero temperature this phase becomes the global ground state at $\alpha_M = (b^2-1)^{b-1}/kb^{2b}$. For $b = 2$ and $\phi = 0$ this state is the ground state for $\alpha > 1/8$. Note, however, even for $\alpha < \alpha_M$, the basin of attraction of the $V^\mu = 0$ state is much larger than that of the single-pattern state, resulting from the nonlinearity of the activation function.

The average energy per synapse, ϵ, for this model is given by

$$\epsilon = -\sum_{\mu} u(V^\mu) - \frac{P}{N^{(b+1)/2}} \int \mathrm{D}z\, \bar{u}(z)$$
$$- \alpha\beta \int \mathrm{D}z_1 \left(\bar{u}^2(z_1) - \left(\int \mathrm{D}z_2\, \bar{u}(\sqrt{q}z_1 + \sqrt{1-q}z_2) \right)^2 \right). \tag{5.55}$$

The entropy per synapse, s is given by

$$s = \tfrac{1}{2}\log[1-q]. \tag{5.56}$$

This has been normalized so that $s = 0$ implies the whole of the w-sphere is equally probable. Again when $T \to 0$ the fractional volume of the N-sphere visited by w for a typical sample goes to zero as $\sqrt{T}$.

We have looked for replica symmetry breaking by examining the mean-field equations assuming one step of replica symmetry breaking. This is presented in appendix A. No evidence for replica symmetry breaking was found. We therefore postulate that the mean-field solution is exact.

6. Finite-size effects and simulations

In the signal-to-noise analysis described in section 4.1 we obtained the fixed-point equation (4.14) by keeping only the leading order terms in N. To find the dominant finite-size corrections we must keep the next largest terms. What these terms are depend on the size of the nonlinearity b. For $b < 3$ the next largest terms in (4.9) are of the form

$$\sum_{\mu} A^2(u\Delta_1^\mu + v\Delta_x^\mu)A'(u\Delta_1^\mu + v\Delta_x^\mu) \tag{6.1}$$

which are of order $N^{(b-1)/2}$ relative to the leading order terms. Keeping only these terms the fixed-point equation (4.14) becomes

$$\frac{A^2(u)}{u^2}(1-u^2) = \alpha\left(k + \frac{2k_1 v}{\sqrt{\alpha k N^{b-1}}}\right) \tag{6.2}$$

where

$$k_1 = \int \mathrm{D}z\, A^2(z)A'(z). \tag{6.3}$$

When $\kappa = 0$, $k_1 = 2^{(3b-3)/2}b\Gamma(3b/2)/\sqrt{\pi}$. Equation (6.2) again gives a first-order phase transition but at a reduced α which depends on N. The position of this phase transition, $\alpha_c(N)$, against the number of synapses, N, is shown in figure 6 for the case $b = 2$ and $\phi = 0$. In the large N limit this transition point $\alpha_c(N)$ converges to the infinite system result (4.15). The next largest terms again depend on the b. For $b = 2$ these terms are of order $1/N$ relative to the leading order term. They include terms similar to those just discussed but also fluctuation terms.

We have performed simulations of this model. Starting from $w = \xi^1$ the dynamical equation (2.4) was iterated until the weights effectively stopped changing. The final overlap between w and ξ^1 against $\alpha = P/N^b$ was computed. Figure 7 shows a typical simulation with $b = 2$, $\kappa = 0$, $r = 0.1$ and $N = 256$. In this case 2000 iterations were required before w effectively stopped changing. We also show the theoretical curve for the large N limit (4.14) as well as the theoretical curve (6.2) for the leading order finite-size corrections. We see that in the finite system the sharp transition has been smoothed. We find that above the transition the final overlap with the original pattern always remained positive and was of

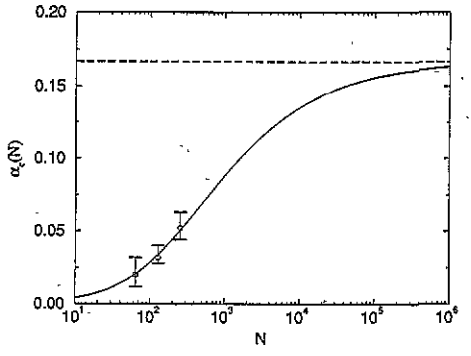

Figure 6. This figure shows the leading order finite-size correction in the critical capacity $\alpha_c(N)$ against the number of synapses N plotted on a logarithmic scale. The curve is plotted for $b = 2$ and $\kappa = 0$. The upper dotted line shows the $N \to \infty$ limit. Also shown are estimates of $\alpha_c(N)$ made from simulations on systems of size $N = 64$, 128 and 256.

order $1/\sqrt{N}$. This suggests the the mixed state is not unique but depends on the starting position.

In figure 6 we also show numerical estimates of $\alpha_c(N)$ for $b = 2$, $\kappa = 0$ and $N = 64$, 128 and 256. The point and the error bars show where the simulation curve cross the point $V^1 = 0.8$, 0.6 and 0.4. We have used these crossing points as the curves round off at lower overlaps due to the residual overlap. The simulations are in reasonable agreement with the theoretical results although a better analysis of the higher order correction would be required to establish the theoretical results more firmly. We note that for realistic sizes of neurons the finite-size effects are very important. However, these finite-size effects depend on b and will be smaller as b is increased.

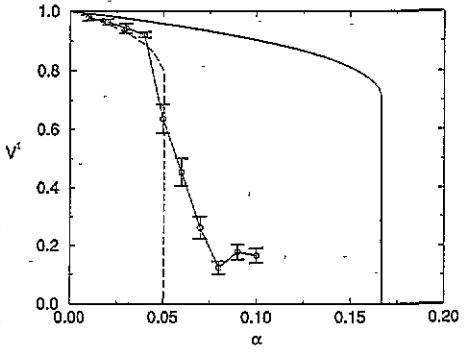

Figure 7. This plot shows a simulation of the overlap, V^1, between the synaptic vector w and pattern ξ^1 against α after iterating the dynamical equations until the synaptic weights effectively stopped changing. Each point represents 50 samples. The error bars show the statistical estimates of the errors in the mean. The simulation was performed with $b = 2$, $\kappa = 0$, $r = 0.1$ and $N = 256$. The full curve shows the theoretical prediction in the infinite volume limit while the broken curve shows the leading order finite-size corrections.

7. Conclusions

We have demonstrated that neurons with nonlinear activation functions can learn to perform very different functions than neurons with linear activation functions. Depending on the size of the nonlinearity b and inter-pattern correlations, a neuron will either learn a single pattern or some mixture of patterns. When the neuron's final state is close to a single pattern, it has learned to recognize that pattern in the presence of the other patterns. When it learns to the mixture, it cannot recognize any individual pattern. The nature of the mixed state was not investigated here, although in the case of two or several patterns with single correlation, the mixed state is the symmetric mid-point of the patterns. From the signal-to-noise analysis we solved the dynamics with many random patterns. Here we found that, for a given nonlinearity b and threshold $\kappa = \phi/\sqrt{N}$, the neuron could learn a random pattern provided the number of random patterns was below some critical number. Increasing the nonlinearity and the threshold increases the critical capacity, however, it also slows down the speed of learning. Finite-size corrections were also obtained from the signal-to-noise calculation. The theoretical predictions were compared with simulations on finite systems and found to be consistent. By introducing a partition function describing the stationary states of the neuron we were able to reproduce and extend the signal-to-noise results, thus giving more confidence in their validity. The mean-field calculation is of interest as the energy function is not quadratic and the position of the first-order phase transition scales as N^b rather than N as found in the Hopfield model [12] or in the perceptron [14].

We have not examined what happens above α_c as there is nothing in the distribution of random patterns for the neuron to learn. For inputs drawn from a more complicated distribution the behaviour above α_c would be more interesting. An indication of what we might expect is provided by the clustered patterns discussed in section 3.2. In this case as the number of patterns in the cluster is increased above a critical number the neuron stops learning to recognize a single pattern but instead learns to recognize the centre of one of the clusters.

This model is a highly idealized version of a real neuron. Some of the neurophysiologically implausible features make very little difference to the result. For example, in some situations it would be more realistic to restrict the weights and patterns to positive values. An appropriate redefinition of the learning rule (2.4) and the threshold ϕ could model this. In addition, we have assumed that the learning rule is linear and the nonlinearity is due to the activation function. However, certain combinations of nonlinear learning rules with the activation functions would have the same form as the learning equation. In this sense, the distinction does not exist; it does not matter whether the nonlinearity is in the activation function or in the learning rule. Many relevant features of real neurons such as the shape of their activation function and the mechanism they use to prevent their synaptic weights from growing unboundedly are still open areas of research. We would hope that this work might provide a useful link between the characteristics of a single cell and its functionality in a larger network.

As mentioned in the introduction, for real neurons there is no reason to believe that the activation function takes the form of a simple power law. Indeed, the firing rate of all neurons must saturate at some point as the excitation is increased, so a sigmoid is probably more realistic. It would be straightforward to extend these results to sigmoids or other functional forms. Any part of the activation function can be approximated by a simple power law and the analysis given can be applied to these parts separately. In the derivation of the critical capacity the part of the activation function associated with the microscopic post-synaptic potentials—denoted by $\tilde{A}(V^\mu)$—was treated separately from the

part associated with the macroscopic post-synaptic potentials. The noise depends only on a very small part—of order $1/\sqrt{N}$—of the activation function around the average post-synaptic potential of the unlearned patterns and will be independent of the shape of the activation function elsewhere. Provided that the small excitation part of the activation function can be described by a power law, then the noise term—αk in equation (4.14)—will remain unchanged. The macroscopic part of the activation function, $A(V^\mu)$, can be substituted into equation (4.14) no matter what form it takes. If the activation function is sigmoid then a mixture of a few random pattern may be stable, although whether such a mixture is learned will depend on the structure of the input world and how it is experienced by the neuron.

In the brain, neurons function as part of a network and will interact. One role of the interactions could be to equipartition the patterns among the neurons. Imagine a collection of non-interacting neurons with random initial weights. These neurons would learn grandmother cell representations of the input patterns. Unfortunately, some patterns would be represented by many neurons, while others would be recognized by none at all. A more favourable representation would have every pattern represented by the same number of neurons. This could be accomplished by inhibitory connections between nearby neurons. This would form a competitive network, where the activation of a neuron suppresses the firing of its neighbours (for a review of competitive networks see [10, ch 9] and [17, pp 63–70]). In such a network each neuron would tend to learn to discriminate a different pattern.

We have seen that for this model the shape of the activation function is important in determining what a neuron computes. Although this model involves many simplifications, it is fairly typical of the models used to describe neurons. With no teacher, this model can learn to discriminate a single pattern from many others. Of course real neurons would operate as part of a complex interacting neural network. The real potential for useful processing would come from this higher architecture. Nevertheless, much can be learned by studying a single neuron in isolation.

Acknowledgments

This work was supported by SERC, whom we would like to thank. We would also like to thank the Manchester Computing Centre for the use of a VP1100 on which the simulations were carried out.

Appendix. Replica symmetry breaking

In the signal-to-noise analysis of section 4.1 the noise was found to be proportional to

$$\sum_{\mu\neq 1} A(V^\mu)\xi^\mu. \tag{A1}$$

Since $A(V^\mu)$ is nonlinear this noise will be very dependent on the position of the synaptic weight vector w. As a consequence we might expect that there would be a number fixed-point solutions close to each pattern and this would allow for the possibility of replica symmetry breaking. To investigate this we consider one step of replica symmetry breaking *à la* Parisi [18]. The matrix q^{ab} is divided up into n/θ blocks of sizes $\theta \times \theta$. The diagonal

elements q^{aa} are zero. The other elements of the blocks on the diagonal are q_1 while the elements of all the other blocks are q_0. Physically, q_1 is a measure of the thermally averaged overlap between replicas in the same state while q_0 is the thermally averaged overlap between replicas in different states; θ measures the probability of two replicas being in different states. We assume that the matrix p^{ab} has a similar structure as q^{ab} while replica symmetry holds among all the other macroscopic order parameters.

Using this ansatz

$$\sum_{a<b} q^{ab}\bar{t}^a\bar{t}^b = \frac{q_0}{2}\left(\sum_a \bar{t}^a\right)^2 + \frac{q_1-q_0}{2}\sum_{\text{block}=1}^{n/\theta}\left(\sum_{a\in\text{block}}\bar{t}^a\right)^2 - \frac{q_1}{2}\sum_a(\bar{t}^a)^2 \tag{A2}$$

and similarly for the other terms in q^{ab} and p^{ab}. After some algebra we find

$$\beta f = \text{extr}\left\{\frac{\theta p_0 q_0}{2} + \frac{(1-\theta)p_1q_1}{2} + \frac{\lambda}{2} - \sum_{\mu=1}^{s} t^\mu V^\mu + \beta\sum_{\mu=1}^{s} u(V^\mu) + \frac{(1-\theta)}{2\theta}\log[\lambda + p_1]\right.$$
$$\left. - \frac{1}{2\theta}\log[\lambda + p_1 + \theta(p_0-p_1)] + \frac{p_0 + \sum_{\mu=1}^{s}(t^\mu)^2}{2(\lambda + p_1 + \theta(p_0-p_1))} + G_1^0(q_0,q_1,\theta)\right\} \tag{A3}$$

where the extremum is taken with respect to all the order parameters and where

$$G_1^0(q_0,q_1,\theta) = \frac{P\beta}{N^{(b-1)/2}}\int \mathrm{D}z\,\bar{u}(z)$$
$$+ \frac{P\beta^2}{2N^b}\int \mathrm{D}z_1\left\{\bar{u}^2(z) - \theta\left(\int \mathrm{D}z_2\,\bar{u}(\sqrt{q_0}z_1 + \sqrt{1-q_0}z_2)\right)^2\right.$$
$$\left. + (\theta-1)\left(\int \mathrm{D}z_2\bar{u}(\sqrt{q_1}z_1 + \sqrt{1-q_1}z_2)\right)^2\right\}. \tag{A4}$$

Performing the extremization with respect to p_0, p_1, λ and t^μ we find

$$f = -\frac{1}{2\beta} + \frac{\sum_{\mu=1}^{s}(V^\mu)^2 - q_0}{2\beta(1-q_1+\theta(q_1-q_0))} - \sum_{\mu=1}^{s} u(V^\mu) - \frac{1}{2\beta}\log[1-q]$$
$$- \frac{1}{2\beta\theta}\log\left[\frac{1-q_1+\theta(q_1-q_0)}{1-q_1}\right] - \frac{G_1^0(q_0,q_1,\theta)}{\beta}. \tag{A5}$$

We have looked for solutions to the saddle-point equations for V^μ, q_0, q_1 and θ with $0 \geqslant q_0 \geqslant q_1 \geqslant 1$ and with V^μ and θ in the interval from 0 to 1. The only solutions which are consistent with all the saddle-point equations are the replica symmetric solutions. This indicates that there are not a sufficient number of local minima to break the replica symmetry.

References

[1] Hebb D O 1949 *The Organization of Behavior* (New York: Wiley) ch 4
[2] Oja E 1982 *J. Math. Biol.* **15** 267–73
[3] Hertz J A 1990 Statistical dynamics of learning *Statistical Mechanics of Neural Networks* ed L Garrido (New York: Springer) pp 137–54
[4] Kohonen T, Oja E and Lehtio P 1981 Storage and processing of information in distributed associative memory systems *Parallel Models of Associative Memory* ed G E Hinton and J A Anderson (New York: Lawrence Erlbaum) pp 129–67
[5] Linsker R 1988 *Computer IEEE* 105–17
[6] Oja E 1989 *Int. J. Neural Sys.* **1** 61–8
[7] Sanger T D 1989 *Neural Networks* **2** 459–73
[8] Mougeot M and Azencott R 1991 Unsupervised learning for the visual cortex (layer iv): model and simulations. *Proc. IJCNN* (New York: IEEE) pp II-613–II-8
[9] Getting P A 1989 *Ann. Rev. Neurosci.* **12** 185–204
[10] Hertz J A, Krogh A S and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (New York: Addison-Wesley) ch 8, 9
[11] Amit D J, Gutfreund H and Sompolinsky H 1985 *Phys. Rev.* A **32** 1007–18
[12] Amit D J, Gutfreund H and Sompolinsky H 1985 *Phys. Rev. Lett.* **55** 1530–3
[13] Amit D J, Gutfreund H and Sompolinsky H 1987 *Ann. Phys* **173** 30–67
[14] Gardner E 1987 *Europhys. Letts.* **4** 481–5
[15] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257–270
[16] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271–84
[17] Hecht-Nielsen R 1989 *Neurocomputing* (New York: Addison-Wesley) section 3.4, pp 64–8
[18] Parisi G 1980 *J. Phys. A: Math. Gen.* **13** 1101–12