Threshold-linear formal neurons in auto-associative nets

1990 J. Phys. A: Math. Gen. 23 2631

(http://iopscience.iop.org/0305-4470/23/12/037)

J. Phys. A: Math. Gen. 23 (1990) 2631–2650. Printed in the UK

Threshold–linear formal neurons in auto-associative nets

Alessandro Treves†

Racah Institute of Physics, Hebrew University, 91904 Jerusalem, Israel

Received 20 November 1989, in final form 8 March 1990

Abstract. Most analytical results concerning the long-time behaviour of associative memory networks have been obtained by using binary elementary units. Here, the use of alternative types of neuron-like processing elements is considered as a way of testing the generality of those results and of approaching biological realism. In particular, threshold–linear units are proposed as appropriate in models designed to reproduce low firing rates, in which long-time stability does not rely on single unit saturation. Such units are simple enough to allow detailed analytical understanding of the properties of the network. This is demonstrated by analysing the attractor states of a network operating at low rates. It is shown that while the interesting retrieval behaviour persists, the roles of the different parameters as well as the nature of the stable states change completely with respect to the binary implementation.

1. Introduction

In most attractor neural networks (ANN) [1, 2] that model auto-associative memory retrieval, the instantaneous activity state representing the output of a formal neuron is a binary variable, which at prescribed times is assigned one of two values according to some integrated input variable reproducing the signals coming from other units. Alternatively, when a continuous representation of the activity state is used, the neuron input–output relationship is typically described by a sigmoid function, so that the output has a lower bound representing a quiescent neuron, and saturates at a given high activity level. These bounds on the activity of single units have an important role in ensuring the global stability of the network, in the long time limit, whenever no alternative mechanism regulates the overall activity level. Conversely, networks endowed with the same type of interactions, but made up of simple linear units [3], fail to produce meaningful long-time behaviour.

When considering network models as providing clues for the analysis of cortical systems such as, for example, the CA3 region of the hippocampus, one should keep in mind that, in those systems, neural activity levels (firing rates) appear to be determined by a complex interplay of afferent and intrinsic excitation and inhibition (e.g. [4]), rather than by the (relatively high) saturation limits of single cell activity. In recording experiments, very rarely do activated cells sustain high spiking frequencies in the range of several hundred spikes s^{-1}; most often the most active ones fire at only tens of spikes s^{-1}. On the other hand, high frequencies of the order of the

† Present address: Department of Experimental Psychology, Oxford University, Oxford OX1 3UD, UK.

0305-4470/90/122631+20$03.50

inverse absolute refractory period would seem to be implied, in the models, by the stabilisation of the network into a given attractor.

In order to address this problem, associative memory network models have been proposed [5–9], in which individual units, still realised as binary variables, display activity levels which are kept low compared with the maximal levels occurring when units persist in their active state. The saturation of single units then ceases to be the factor that determines the overall stability of the network. Instead, this is ensured by a tunable balance of excitatory and inhibitory contributions.

It becomes appropriate in this modified context, in which single unit saturation effects play no important role, to reconsider alternative representations of single neuron behaviour to the usual binary one. Granted that formal neurons, that allow analysis and simulation of very large networks, still need to be extremely simplified processing units, one may ask a number of questions.

(i) In what sense can an alternative (simple) representation claim to come anywhere nearer to realistically reproducing the highly complex behaviour of real neurons?

(ii) To what extent may it enhance the biological plausibility of the whole network as a model of a neural function?

(iii) Does the use of alternative representations meaningfully affect the behaviour of known models?

(iv) Is it still possible to use the more powerful analytical tools developed in studying previous models?

The first two points, of a more general scope, will be addressed elsewhere. This paper will be concerned with the last two questions. In particular, we shall study a network operating at low firing rates, in which a simple threshold–linear, rather than binary, elementary unit is used. First, this will demonstrate how analytical techniques can be easily extended to the new framework. Second, it will exemplify some of the effects of adopting the alternative type of elementary unit. The main conclusion that may be drawn is in a sense familiar. It is the observation that 'high level' features, e.g. retrieval behaviour as an emergent property of the network, are more robust and independent of details than 'low level' ones, such as the dynamical nature of retrieval attractors or the nature of transitions between different types of long-time behaviour. While this provides additional evidence for the importance of theoretical approaches to neural networks, it also underscores the need to test the robustness of results obtained under very specific assumptions.

In sections 2 and 3 we recall the structure of the model considered, and how its long time behaviour can be described. In section 4 the threshold–linear representation is introduced, while the encoding of memories is dealt with in section 5. The analytical treatment set up in section 6 is specialised in the next two sections to describe the attractor states occurring at low memory loading and low noise levels. Section 9 reports the results of numerical simulations, and conclusions are drawn in the last section.

2. A low firing rate attractor neural network

In references [5, 6] an associative memory model was proposed as a modified version of the Hopfield network [1], that displays low individual 'spiking activity' rates. The model distinguishes between excitatory and inhibitory neurons, and only the pattern of firing activities of the former is assumed to bear information related to the memory stored on synaptic strengths. An effective representation of inhibition is used, in which inhibitory effects contribute a term to the input of excitatory neurons, expressed as a nonlinear (quadratic) function of an averaged activity rate of the excitatory neuron themselves.

Synaptic efficacies encode information about p memorised patterns of activity η_i^μ, where $\mu = 1, \ldots, p$, $i = 1, \ldots, N$, and N is the number of excitatory neurons. Each η_i^μ is a positive number related to the degree of firing activity of neuron i in pattern μ, and in the implementation of [5, 6] it was supposed to take the value 0 or 1 with independent probabilities, respectively $1 - a$ and a. The single variable describing the activity of neuron i is V_i, which again was assumed binary (0 or 1). It depends on an input variable, the local field $h_i = h_i^{\mathrm{E}} + h_i^{\mathrm{I}}$. The excitatory part of the local field is written as

$$h_i^{\mathrm{E}} = \sum_{j \neq i}^{N} J_{ij} V_j \tag{1}$$

where the (direct) synaptic couplings J_{ij} assume the form

$$J_{ij} = \frac{1}{Na^2} \sum_{\mu=1}^{p} \eta_i^\mu \eta_j^\mu \,. \tag{2}$$

The inhibitory contribution was taken to be [5]

$$h_i^{\mathrm{I}} = -\frac{1}{\lambda} \left(\sum_{\nu=1}^{p} \frac{\eta_i^\nu}{a} \right) \left(\frac{1}{Np} \sum_{\mu=1}^{p} \sum_{j=1}^{N} \frac{\eta_j^\mu}{a} V_j \right)^2 \tag{3}$$

where the parameter λ regulates the strength of the term relative to h_i^{E}.

3. Dynamical evolution and attractor states

Neuronal states are updated in random order, and the new state is chosen stochastically [10], in the binary case as

$$P(V_i = 1) = [\exp(-\beta h_i) + 1]^{-1} \tag{4}$$

where the 'temperature' $T \equiv \beta^{-1}$ measures the amount of stochastic noise in the process.

Subject to the above interactions, the network evolves dynamically towards one of a set of attractor states. In a given attractor, the network may still wander among a variety of configurations due to the stochastic noise, but it reaches a stationary probability distribution of being in any particular configuration at any instant of time.

The correlation of the attractor state of the network with the stored patterns can be measured by the overlaps

$$x^\mu = \frac{1}{N} \sum_{i=1}^{N} \frac{\eta_i^\mu}{a} \langle V_i \rangle \tag{5}$$

where $\langle\ldots\rangle$ denotes averaging over the probability distribution characterising the attractor.

The overall mean activity of the network (i.e., of the excitatory neurons explicitly represented) can be measured by

$$x = \frac{1}{N}\sum_{i=1}^{N}\langle V_i\rangle . \tag{6}$$

One can further define

$$y_0 = \frac{1}{N}\sum_{i=1}^{N}\langle V_i^2\rangle \tag{7}$$

and

$$y_1 = \frac{1}{N}\sum_{i=1}^{N}\langle V_i\rangle^2 . \tag{8}$$

Clearly, $y_1 - x^2$ is a measure of the variance in the distribution of activity between excitatory neurons, while $y_2 \equiv y_0 - y_1$ will measure the correlation among the various configurations concurring in the attractor state. Thus, $y_1 = x^2$ implies that all neurons have the same average activity, while $y_2 = 0$ implies that the network has frozen into a single configuration, i.e., each neuron has stabilised on a definite activity rate. Note that if V can assume only the values 0 or 1, $x \equiv y_0$, so that for the analysis of [5] one needed just x and y_1 as global order parameters.

In the binary implementation, the scenario appropriate to memory retrieval was identified as the relaxation of the network into an attractor state, in which one of the overlaps would be much larger than the rest, e.g. $x^1 \gg x^{\mu\neq 1}$, but each single unit would have an averaged activity rate considerably below the maximal one, $\langle V_i\rangle \ll 1, \forall i$. Such a scenario was found to be possible for intermediate values of the noise parameter T. In fact, the requirement that single units do not 'freeze' into the saturated activity state $\langle V_i\rangle = 1$ implied a substantial amount of noise, which still had to be low enough in order to break the symmetry between the various patterns and allow the emergence of a single one [5].

4. A threshold–linear neuron representation

In considering alternative representations for the elementary units of the network, we shall restrict, as stated above, to the simplest case in which the abstract neuron is a single variable, V, corresponding to a short-time averaged firing rate. This variable is updated at discrete times, independently of its previous history. The choice of representation reduces to choosing an appropriate input–output function, where the input is also taken to be a single variable, the local field h, summarising the effect of the instantaneous firing rates of the other neurons in the network. Three possible representations are shown diagramatically in figure 1.

One would like to preserve in the input–output relationship some of the most basic general features of the behaviour of real neurons. For example, when the input is below a given threshold the neuron does not fire. Above threshold, there is a region where the firing frequency depends strongly on the input level. At very high input

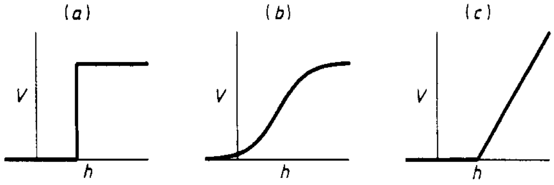

Figure 1. Some of the functions often used the represent the input–output dependence in a simple formal neuron: (*a*) Heaviside; (*b*) sigmoidal; (*c*) threshold–linear.

levels, the firing rate reaches saturation, since the neuron has an absolute refractory period.

When the binary representation is used [11], it is assumed that the neuron spends most of its time either not firing or firing at saturation levels. Accordingly, the region of intermediate firing levels is squeezed to a single point, and the input–output relation acquires the stepwise form

$$V = \theta(h - T_{\mathrm{hr}}) \tag{9}$$

with $\theta(x)$ the Heaviside function, and T_{hr} a threshold.

More general forms for the input–output function, that would allow for the output to assume continuous values, have been considered in the literature in a variety of contexts (see [12,13]). A common choice is a sigmoidal function of the type

$$V = F_{\mathrm{sigm}}(h) = \tfrac{1}{2}\{1 + \tanh[g(h - T_{\mathrm{hr}})]\} \tag{10}$$

that varies rapidly in a region of width g^{-1} between a zero ouput for $h \ll T_{\mathrm{hr}}$ and a saturated one for $h \gg T_{\mathrm{hr}}$. It has been argued [14] that as long as the *gain* g is high enough, the behaviour of an attractor neural network built of such neurons is similar to that of the corresponding network built of binary neurons. In the energy formalism, in those cases in which it can be used, the sums over the two states of binary neurons become weighted integrals over all possible values of the neurons' firing rates. The weight factor [14,15] can be expressed as an additional term $\sum_i \int F_{\mathrm{sigm}}^{-1}(V_i)\,\mathrm{d}V_i$ in the energy of each configuration. With a high-gain sigmoidal function, the weight of the rates close to zero and to saturation dominates, and the neurons are expected to behave in a manner similar to binary ones.

Both with binary neurons and with analogue neurons, it is possible to account for noise effects via the introduction of an effective temperature [16]. In some cases the noise can have a very important role. The network recalled above is a case in point [6]. It operates properly only when the noise is strong enough to make the system wander stochastically over small free-energy barriers. A non-zero noise level is required to keep active neurons jumping frequently between the 0 and the 1 state.

However, as the interactions in the network are designed to model the effect of inhibition, and to keep neurons away from the saturation region of high rates, using binary or indeed sigmoidal functions loses much of its *raison d'être*. In fact, one can renounce representing the saturation effect at all. Let us consider the simple alternative for the input–output function:

$$V = \begin{cases} g(h - T_{\mathrm{hr}}) & h > T_{\mathrm{hr}} \\ 0 & h < T_{\mathrm{hr}} \end{cases} \tag{11}$$

i.e. a function that is linear above threshold and zero below it. Such a function has been widely used in different models (e.g., in [17]). As it neglects to describe the saturation intervening at high firing rates, it is a viable choice, in the long-time limit, only if a cooperative effect prevents the neurons in the network from approaching saturation levels. Close to the threshold, it describes in a simplified way the dual nature of neuronal response: below threshold the neuron is in its resting state (a discrete feature), while above threshold its output varies continuously with the input (here, linearly). Contrasted with the binary representation, the linear regime can be argued to be a closer approximation of the response above threshold, but still away from saturation levels, of a more realistic model neuron. Nevertheless, this threshold-linear representation allows, in contrast with more complex ones, a complete analytical study of network properties, as the following treatment will exemplify.

This description can be easily generalised to include noise effects. One way to do it, while keeping the Glauber dynamics with discrete time and stochastic updatings, is to introduce again an effective temperature. One associates with the firing states of each neuron the weight factor

$$k\delta(V_i) + \exp[-\beta(V_i T_{\mathrm{hr}} + V_i^2/2g)] \,. \tag{12}$$

Beside the gain g, the threshold T_{hr}, and the inverse temperature β, a new parameter k has been introduced. It gives the relative weight of the resting state of the neuron. If h_i is, as before, the local field acting on neuron i, the updating process can then be defined by giving the probability that the updated firing rate of the neuron will fall between V_i and $V_i + \mathrm{d}V$, namely

$$P(V_i, \mathrm{d}V) = D^{-1} \exp[\beta(V_i h_i - V_i T_{\mathrm{hr}} - V_i^2/2g)]\mathrm{d}V \tag{13}$$

and the probability of the new firing rate being zero

$$P(0) = D^{-1} k \tag{14}$$

with

$$D = \int_0^\infty \mathrm{d}V_i \; [k\delta(V_i) + \exp[\beta(V_i h_i - V_i T_{\mathrm{hr}} - V_i^2/2g)]] \;. \tag{15}$$

In the zero-temperature limit, $\beta \to \infty$, it is easy to see that the updating becomes deterministic, and proceeds according to the input–output relation expressed in (11). If $\beta < \infty$, β^{-1} measures the typical deviation of the stochastically updated frequency, V_i, from $g(h_i - T_{\mathrm{hr}})$, provided $h_i > T_{\mathrm{hr}}$. If $h_i < T_{\mathrm{hr}}$, β^{-1} determines the likelihood that the neuron will nevertheless fire at non-zero frequency. Thus, equation (12) allows one to perform 'finite temperature' statistical calculations, in a natural generalisation of the deterministic rule, equation (11). In particular, for a network whose evolution can be described in terms of an energy function, one can study the minima of the free energy [2] in order to investigate the long-time behaviour.

5. General distributions of memories

In the class of associative memory models that we are considering, a set $\{\eta_i^\mu\}$ of stored memories is assigned *a priori*, and determines the strengths of the synaptic connections. There is, of course, an infinite choice of ways in which one can assign the

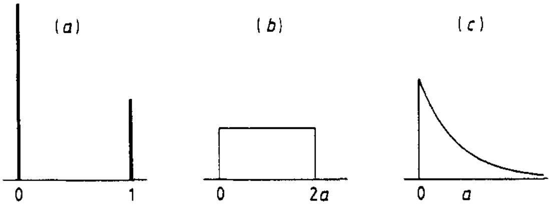

Figure 2. The three probability density distributions P_η considered: (*a*) binary; (*b*) flat; (*c*) exponential.

memories. We restrict ourselves to the case in which each η_i^μ is assigned ***independently*** as a quenched random variable, according to some probability density distribution $P_\eta(\eta)$, which is the same for all $\mu = 1, \ldots, p$ and $i = 1, \ldots, N$. The independence of the η_i^μ is a very useful assumption from a technical point of view and, although arbitrary, it is commonly used to study general properties of networks. The form of P_η, however, can still be chosen freely.

There are at least two aspects to the choice of P_η, when considering its significance in models inspired by neurobiology. First, as long as the synaptic matrix is written in the so-called Hebbian form of (2), the form chosen for P_η determines the interactions occurring in the model, in that it sets the probability distribution of the synaptic strengths. Second, inasmuch as synaptic imprinting is supposed to have occurred via mechanisms of synaptic plasticity during a learning phase, the precise relation between each η_i^μ and the firing activity of neuron i in memory state μ could conceivably be made to model e.g. a certain LTP mechanism. Both aspects are rather overshadowed by the question, currently debated [18], of the biological plausibility of the detailed structure implied in (1)–(3) or in their analogue in other model networks [1, 19].

A common choice is to define P_η as the sum of two delta functions centred, e.g., at $\eta = 0$ and $\eta = 1$. This choice reduces each memory to a binary word of length N. It is a most natural ***ansatz*** in a computer science context, where all variables are decomposed into binary form. It also provides a helpfully intuitive picture, by simply dividing in each memory a group of 'active' neurons from the rest of 'passive' ones. On the other hand, the use of this form for P_η has contributed to generate confusion in the neurobiological community. In fact, when recording from animals performing cognitive tasks, the firing rates of individual neurons hardly appear to fall into two clear-cut classes, characterisable by high and low activity. It has been asked whether this fact would not disprove the neurobiological relevance of attractor neural networks.

To treat more general probability distributions P_η is straightforward. We shall consider a number of specific cases that satisfy the following general requirements:

$$\text{(i)} \qquad P_\eta(\eta) \geq 0 \qquad \int P_\eta(\eta)\,\mathrm{d}\eta = 1$$

where P_η is a probability density distribution;

$$\text{(ii)} \qquad P_\eta(\eta) = 0 \qquad \text{for } \eta < 0$$

i.e. negative firing rates have probability zero;

$$\text{(iii)} \qquad \int P_\eta(\eta)\eta\,\mathrm{d}\eta = a$$

(to set a common normalisation).

In particular we shall consider in detail the following possibilities illustrated in figure 2:

$$P_\eta(\eta) = (1-a)\delta(\eta) + a\delta(\eta - 1) \tag{16}$$

i.e. the binary-word choice;

$$P_\eta(\eta) = 1/2a \qquad (0 \le \eta \le 2a) \qquad P_\eta(\eta) = 0 \qquad (\eta > 2) \tag{17}$$

which assumes that the η are flatly distributed from 0 to $2a$;

$$P_\eta(\eta) = \frac{1}{a}\mathrm{e}^{-\eta/a} \tag{18}$$

where the distribution is exponential. The three forms chosen all lend themselves to explicit analytical treatment. The final goal of the analysis is not to provide specific answers for specific cases, but rather to test to what extent the retrieval behaviour of the model persists irrespective of the form chosen for the distribution of the memories and, consequently, of the synaptic strengths.

6. Mean-field theory calculations in the general case

The analysis of the (thermodynamic) attractor states of the network [2] that was presented in [5] can be repeated along the same lines in the case of a more general probability density distribution P_η governing the assignment of random patterns, and of a more general neuron representation. Again, one expresses the free-energy in terms of a few parameters, and then looks for its local minima. These corresponds to possible attractor states. We note here a few points involved in the free-energy calculation, while formal derivations are relegated to an appendix.

As in reference [20], the average over the quenched pattern distribution is performed in two stages. First, over a possibly infinite number of 'uncondensed' patterns, i.e. those that are only randomly correlated with the thermodynamic state of the system. Later, over a finite number of 'condensed' patterns, i.e. those that contribute explicitly in determining the firing distribution in the thermodynamic state. Accordingly, the form of the probability density P_η enters twice in the calculation. In the first stage, the only relevant parameters are the first moment of P_η, which has been set to be a, see above, and the second moment, which is written

$$\int_0^\infty P_\eta(\eta)\eta^2\,\mathrm{d}\eta \equiv a^2 + a_2\,. \tag{19}$$

In the second stage, instead, the full shape of P_η becomes important. It turns out that the relevant temperature scale, which measures the competition between the fast 'thermal' noise and the cold noise due to random correlations between the patterns, is determined by the ratio a_2/a^2, which will be denoted by the symbol T_0. For the three forms of P_η considered, T_0 takes the values

$$T_0 = \begin{cases} (1-a)/a & \text{binary} \\ 1/3 & \text{flat} \\ 1 & \text{exponential}\,. \end{cases} \tag{20}$$

The use of order parameters in the mean-field theory calculation allows to decouple the different degrees of freedom. The Hamiltonian is expressed as a sum of single-neuron terms, each of the form

$$H^{(1)} = -hV - h_2V^2 \tag{21}$$

h, h_2 being functions of the order parameters introduced. Then, one performs the trace over the single-neuron states, which can be written formally

$$\mathrm{Tr}\,(h, h_2) \equiv \mathrm{Tr}_{\{V\}} \exp[\beta(hV + h_2V^2)]\,. \tag{22}$$

If one uses the threshold-linear representation, it is clear from (12) and (22) that the interaction has been re-expressed as an effective single-neuron input h, plus a modification to its gain parameter, which is now $g' = g/(1 - 2gh_2)$. Explicitly

$$\mathrm{Tr}\,(h, h_2) = k + \sqrt{(\pi g'/2\beta)}\exp[\tfrac{1}{2}\beta g'(h - T_{\mathrm{hr}})^2]\{1 + \mathrm{erf}[\sqrt{(\beta g'/2)}(h - T_{\mathrm{hr}})]\} \tag{23}$$

where $\mathrm{erf}(x)$ denotes the error function.

The free-energy calculation is performed (see [5]) in the limits $N \to \infty, p \to \infty$, with $\alpha \equiv p/N$, and use is made of a replica symmetric *ansatz*. The end result for the free-energy is (see the appendix)

$$\begin{aligned} f = &-\frac{1}{\beta}\left\langle\!\!\left\langle \int Dz \ln \mathrm{Tr}\,(h, h_2)\right\rangle\!\!\right\rangle - \Delta\lambda + \frac{\beta}{2}(\rho_2 y_0 + \rho_1 y_2) + \frac{1}{2}\sum_{\sigma}[(x^\sigma)^2 - \lambda^2] \\ &+ \frac{\alpha}{2\beta}\left[\ln(1 - T_0\beta y_2) - \frac{T_0\beta(y_0 - y_2)}{1 - T_0\beta y_2}\right] \end{aligned} \tag{24}$$

where

$$\begin{aligned} h &= \sum_{\sigma}(x^\sigma - \lambda)\frac{\eta^\sigma}{a} - \Delta - z\sqrt{\rho_1} \\ h_2 &= \tfrac{1}{2}\beta\rho_2 \\ Dz &= \frac{\mathrm{d}z}{\sqrt{2\pi}}\mathrm{e}^{-z^2/2}\,. \end{aligned} \tag{25}$$

Besides the order parameters already defined, the above expression includes ρ_1, ρ_2 which have been introduced to enforce the definition of y_1, y_2, respectively. The saddle-point equations are

$$\begin{aligned} x^\mu &= \frac{1}{\beta}\left\langle\!\!\left\langle \frac{\eta^\mu}{a}\int Dz\frac{\mathrm{d}}{\mathrm{d}h}\ln \mathrm{Tr}\,(h, h_2)\right\rangle\!\!\right\rangle \\ \lambda &= \frac{1}{\beta}\left\langle\!\!\left\langle \int Dz\frac{\mathrm{d}}{\mathrm{d}h}\ln \mathrm{Tr}\,(h, h_2)\right\rangle\!\!\right\rangle \\ y_0 &= \frac{1}{\beta}\left\langle\!\!\left\langle \int Dz\frac{\mathrm{d}}{\mathrm{d}h_2}\ln \mathrm{Tr}\,(h, h_2)\right\rangle\!\!\right\rangle \\ y_2 &= \frac{1}{\beta^2}\left\langle\!\!\left\langle \int Dz\frac{\mathrm{d}^2}{\mathrm{d}h^2}\ln \mathrm{Tr}\,(h, h_2)\right\rangle\!\!\right\rangle \\ \rho_1 &= \frac{\alpha T_0}{\beta}\frac{T_0\beta(y_0 - y_2)}{[1 - T_0\beta y_2]^2} \\ \rho_2 &= \frac{\alpha T_0}{\beta}\frac{1}{[1 - T_0\beta y_2]} \end{aligned} \tag{26}$$

The second equation expresses the fact that λ, the parameter introduced to regulate the strength of the inhibition, turns out to set the overall activity level (cf reference [5]). The emergence of one or more condensed patterns is manifested by their overlap parameters assuming values distinct from λ (the first equation).

7. The limits $\alpha \to 0, \beta \to \infty$

We shall not analyse equations (26) for general α and β. As in reference [5], the limit $\alpha \to 0$ will already yield a non-trivial structure, which can be considered a good approximation in cases in which $p \ll N$. An analysis of the effects of extensive memory loading on a network built of threshold-linear formal neurons will be published elsewhere [21].

The expressions derived so far are valid for arbitrary neuron representations and for any noise level. We shall focus now on the specific threshold-linear neuron representation proposed in section 4, and on the limit $\beta \to \infty$. One should note the differences from the case of spin-like neurons, treated in [5]. There, to consider the zero-temperature limit does not make sense, if one is interested in the low-firing rate retrieval behaviour of the network. At low temperature the system undergoes a spin-glass type of freezing, and single neurons either fire at maximal frequency or remain quiescent. If, however, the neuron representation allows for a region of analogue neuron response at intermediate rates, as in the present case, the thermal noise ceases to have the vital role of keeping the system 'unfrozen'. A system frozen into a single configuration is, in fact, 'acceptable'. The freezing by itself no longer implies that some of the neurons fire at their (implausibly high) saturation rates. An interesting question, then, is whether the non-trivial retrieval behaviour persists in this case. The problem can be addressed analysing the minima of the free-energy (which for $\beta \to \infty$ is just the energy), and the answer will be found to depend essentially on the gain g of the analogue response function.

As in reference [5], the analysis of (26) simplifies considerably when the frozen noise due to random correlations between the patterns is negligible. This is the case provided that, as $\beta \to \infty$

$$1 - T_0 \beta y_2 \to \kappa > 0 \tag{27}$$

Notice that this implies that $y_2 \to 0$ (at least as fast as T), i.e. each V_i has a fixed value in the thermodynamic state, or in other words the system has indeed frozen into a single configuration. Then, the expression to be averaged, $\ln[\mathrm{Tr}\,(h, h_2)]$, acquires, for our choice of neuron representation and in the $\beta \to \infty$ limit, a very simple form

$$\lim_{\beta \to \infty} \frac{1}{\beta} \ln \mathrm{Tr}\,(h, h_2) = \begin{cases} 0 & \text{for } h < T_{\mathrm{hr}} \\ \frac{1}{2} g (h - T_{\mathrm{hr}})^2 & \text{for } h > T_{\mathrm{hr}} . \end{cases} \tag{28}$$

The relevant saddle-point equations reduce to

$$\begin{aligned} x^\mu &= \frac{1}{\beta} \left\langle\!\!\left\langle \frac{\eta^\mu}{a} \frac{\mathrm{d}}{\mathrm{d}h} \ln \mathrm{Tr} \right\rangle\!\!\right\rangle = g \int_{h > T_{\mathrm{hr}}} \Pi_\sigma P_\eta(\eta^\sigma)\, \mathrm{d}\eta^\sigma \frac{\eta^\mu}{a} (h(\{\eta^\nu\}, \{x^\nu\}) - T_{\mathrm{hr}}) \\ \lambda &= \frac{1}{\beta} \left\langle\!\!\left\langle \frac{\mathrm{d}}{\mathrm{d}h} \ln \mathrm{Tr} \right\rangle\!\!\right\rangle = g \int_{h > T_{\mathrm{hr}}} \Pi_\sigma P_\eta(\eta^\sigma)\, \mathrm{d}\eta^\sigma (h(\{\eta^\nu\}, \{x^\nu\}) - T_{\mathrm{hr}}) \end{aligned} \tag{29}$$

where $h(\{\eta^\nu\},\{x^\nu\}) = \sum_\sigma (x^\sigma - \lambda)\eta^\sigma/a - \Delta$. One can extract the meaning of these equations: only neurons whose local field is above threshold contribute to the average activity parameters, and h_2 has become irrelevant. Note, however, that this simple situation does not follow necessarily from the $\alpha \to 0, \beta \to \infty$ limits, as an example below will show.

A remark should be made concerning the number of free parameters left in the model. Adopting the above neuron representation, one has introduced three parameters that were not present in the spin-neuron version, namely the gain g, the threshold T_{hr} and the constant k. In a model like the present one, however, where an order parameter (Δ in our notation) acts as a chemical potential to regulate the overall firing rate, T_{hr} can be reabsorbed into Δ and does not play any special role. Moreover, the fact that the model operates as desired even for $\beta \to \infty$ allows, in practice, to restrict the analysis to the zero-temperature case. Finally, the value of the constant k is irrelevant (provided it is finite) in the $\beta \to \infty$ limit. Therefore one ends up with a single relevant parameter describing the neuron, the gain g. This is a rather natural outcome, as it reflect the simplicity of the linear input–output response function one has started from. The only feature peculiar to the particular ANN model is the fact that even the value of the threshold is immaterial, due to the inhibition fixing the overall firing activity.

The remaining part of the analysis, whenever (27) holds, reduces to performing explicitly the averages over the condensed patterns, with a given form for P_η, solving for the order parameters x^μ and Δ, and considering the possible solutions for the free-energy minima.

8. Attractor states

8.1. The uniform state

The simplest solution of (29) occurs when no pattern condenses, and the system settles in a state characterised by equal overlaps with all the patterns. Then the chemical potential is $\Delta = \lambda/g$ and it sets the overall firing rate to λ. An inspection of (27) yields the condition under which it is consistent, within this solution, to neglect the effect of random fluctuations in the correlations between the uncondensed patterns and the state of the system. The condition reads

$$g < T_0^{-1} \equiv g_0 \tag{30}$$

i.e. the gain has to be below a critical value g_0, which is determined by the ratio between the first and second moments of the distribution P_η. This holds for any choice of P_η. Within this solution all neurons are subject to the same local field (obviously, in the limit $N \to \infty$), and therefore the system chooses a fixed configuration with the activity of each neuron set at λ.

It can be verified that, for any P_η, this *uniform* solution is, in this gain range, the unique true minimum of the free-energy (i.e., the energy), whose value is

$$f = \lambda^2/2g\,. \tag{31}$$

Therefore, when the gain is low the encoded memory structure fails to affect the behaviour of the network, which just settles into the configuration in which all neurons share the same prescribed activity level.

The long-time behaviour for $g > g_0$ depends, instead, on the form adopted for P_η.

3.2. The high-gain behaviour for binary memories

Perhaps the simplest case to analyse is when $P_\eta(\eta) = (1-a)\delta(\eta) + a\delta(\eta - 1)$. One can then derive a general solution for a state representing a symmetric mixture of n different patterns. For $n = 1$ these are just *retrieval states*, and there are p of them. For $n = 2$, there are $p(p-1)/2$ symmetric mixtures [20] corresponding to the possible pairs of stored patterns, and so on.

The result of the analysis is that n-mixture states are the ground states of the system if

$$\frac{g_0}{a^{n-1}} < g < \frac{g_0}{a^n}\,. \tag{32}$$

In other words, as g increases from g_0 to ∞, symmetric mixtures of an increasing number of patterns become, in succession, the global energy minima, and thus the attractors relevant in the long-time limit. Each attractor consists, in the $\beta \to \infty$ limit, of a single configuration. Solving (29), one obtains that, in the interval (32), each of the n patterns of a mixture has an overlap $x = \lambda/a$ with this configuration. This overlap is due only to those neurons that are active in *all* the patterns of the mixture. All other neurons are quiescent in the configuration. The energy of the mixture turns out to be

$$f = \frac{\lambda^2}{2}\left(\frac{1}{ga^n} - \frac{n(1-a)^2}{a^2}\right). \tag{33}$$

For values of the gain g higher than those in the interval (32), n-symmetric solutions violate (27). They are unstable to the growth of small random correlations with other patterns into full macroscopic correlations. In fact, m-mixtures, with a certain $m > n$, are the ground states of the system for such g values.

For values of g smaller than those in the corresponding interval, n-symmetric solutions correspond to unstable saddle-points rather than to true energy minima. They are unstable to the decay of some of the macroscopic correlations, so that eventually the system reaches an l-mixture, with $l < n$.

As far as the retrieval behaviour is concerned, it is clear that the network, close to the limits assumed valid in the above analysis, will operate as a genuine classifier only if g falls in the interval where 1-mixtures, i.e. retrieval states, are the stable solutions. Therefore, for this choice of P_η, the proper long-time behaviour occurs when

$$\frac{a}{1-a} < g < \frac{1}{1-a}\,. \tag{34}$$

3.3. The high-gain behaviour for a flat pattern distribution

For $P_\eta(\eta) = 1/2a \quad (0 \le \eta \le 2a)$ a different set of ground states emerge as the gain is increased. Initially the situation is similar to the previous case, that is, at $g = g_0 = 3$, when the disordered state is destabilised, it is the p retrieval states that become the ground states of the system. The parameters describing the retrieval states can be found by writing down (29) explicitly for the case of a single condensed overlap, e.g. x^1, using the appropriate P_η, and combining them into a single equation for the variable $\phi \equiv (x^1 - \lambda)/(x^1 - \lambda - \Delta/2)$. The equation reads

$$\phi^3 - g(\phi - \tfrac{2}{3}) = 0\,. \tag{35}$$

Its solution varies continuously from $\phi = 1$ at $g = 3$ (in which case $x^1 = \frac{4}{3}\lambda$) to $\phi = 2$ at $g = 6$ (then $x^1 = \frac{5}{3}\lambda$). The energy of these states can then be written as

$$f = \frac{\lambda^2}{2g}(2\phi - \phi^2) . \tag{36}$$

In such retrieval states, active neurons display a variety of activity rates, depending on their η values. When pattern 1 is retrieved, the fastest firing neurons are those with $\eta^1 = 2a$, and their firing rate turns out to be $x_{max} = 2\lambda\phi$.

At $g = 6$ retrieval states are destabilised. In contrast with the previous case, no n-mixture state exists that satisfies (27). In other words, one is not justified, for $g > 6$, in neglecting the fluctuations in the overlaps. In fact, the intervening ground state is of the ***spin-glass*** disordered type [20], and has overlaps with all the p patterns which display strong random deviations from their average values. That causes the order parameter ρ_1 to have a macroscopically non-zero value. To describe analytically this spin glass phase (in the context of the replica-symmetric theory treated here) one cannot use (29), but has to rederive the correct $\alpha \rightarrow 0$ limit of (26). This is achieved by keeping ρ_1 finite, whereas $\rho_2 \rightarrow 0$; after the limit $\beta \rightarrow \infty$ has also been taken the relevant saddle point equations then become

$$\begin{aligned} \lambda &= g \int_{z\sqrt{\rho_1}>\Delta} \frac{\mathrm{d}z}{\sqrt{2\pi}} \exp(-z^2/2)(z\sqrt{\rho_1} - \Delta) \\ g_0 &= g \int_{z\sqrt{\rho_1}>\Delta} \frac{\mathrm{d}z}{\sqrt{2\pi}} \exp(-z^2/2) . \end{aligned} \tag{37}$$

These equations yield values for Δ and ρ_1, which satisfy

$$\psi \equiv \frac{\Delta}{\sqrt{\rho_1}} = \sqrt{2}\,\mathrm{erf}^{-1}\left(1 - \frac{2g_0}{g}\right) \tag{38}$$

and the resulting energy of the spin-glass state is

$$\begin{aligned} f &= -\Delta\lambda + \frac{\rho_1 g_0}{2} - \frac{g}{2} \int_{z\sqrt{\rho_1}>\Delta} \frac{\mathrm{d}z}{\sqrt{2\pi}} \exp(-z^2/2)(z\sqrt{\rho_1} - \Delta)^2 \\ &= \frac{\lambda^2}{2}\left[g_0 - g\frac{\exp(-\psi^2/2)}{\psi\sqrt{2\pi}}\right]^{-1} . \end{aligned} \tag{39}$$

Thus, for $g > 6$, the long-time behaviour of the model cannot be used to retrieve any information encoded in the synaptic strength. The gain range in which retrieval does occur is, for this flat distribution,

$$3 < g < 6 . \tag{40}$$

8.4. The high-gain behaviour for an exponential pattern distribution

The situation for $g > g_0$ is again different if one chooses $P_\eta(\eta) = \exp(-\eta/a)/a$. In that case, it can be shown that the ground states are the p retrieval states, all the way to $g \rightarrow \infty$, and no other states are stable. Equations (29) can now be used to yield x^1 and Δ. One obtains a single equation in terms of the variable $\chi \equiv \Delta/(x^1 - \lambda)$

$$e^\chi = g(1+\chi)\,. \tag{41}$$

In terms of this variable, the ground state energy turns out to be

$$f = \frac{\lambda^2}{2}\left[1 + g\frac{\chi^2}{1-\chi}e^{-\chi}\right]^{-1}. \tag{42}$$

Therefore, apparently the gain range in which the model exhibits retrieval is

$$1 < g < \infty\,. \tag{43}$$

One should note, however, that as the gain increases the neurons that do contribute to the total firing activity of the network are the increasingly small fraction that, by virtue of their high η value in the retrieved pεttern, manage to exceed threshold. Their individual activity rates, being the global one fixed at λ, tend to be higher and higher as g increases. Therefore, one reaches a situation in which those few neurons that are active fire at unacceptably high rates, and this limits in practice the viable range of g. The limit cannot be located precisely, as formally there are, even for low gains, individual neurons firing at infinite rates in the $N \rightarrow \infty$ limit, due to the exponential tail of the P_η distribution which extends to infinity.

8.5. Overview

The states characterising the long-time limit behaviour of the model, and their energies, are shown graphically in figure 3, for the three possible distributions P_η considered above.

For a gain $g < g_0$ the system always settles into a configuration characterised by uniform activity of all the neurons. In contrast, the three different networks, whose synaptic strength are determined by the three distributions P_η considered, exhibit widely different high-gain behaviour. A P_η made up of two δ-functions causes the network to settle into n-mixture state with increasingly high n. For a flat P_η there is a spin-glass transition at high gain, while for an exponentially shaped P_η the retrieval states remain the ground states up to $g \rightarrow \infty$. As far, though, as their retrieval performance is concerned, the three networks share the property that their desired operation requires intermediate gain regimes, in the range immediately above g_0. At higher gains, even the third network, that retains retrieval states as the unique fixed-points of its evolution, is beset by the problem that some neurons exhibit implausibly high rates, violating the condition that made the 'threshold-linear' representation acceptable.

A feature common to all three cases is the absence of metastable states at energies higher than the ground states, for any given gain. The absence of such stable solutions at higher energies, which were, instead, present in the model utilising a binary neuron representation [5], can be traced to the analogue description of the elementary units.

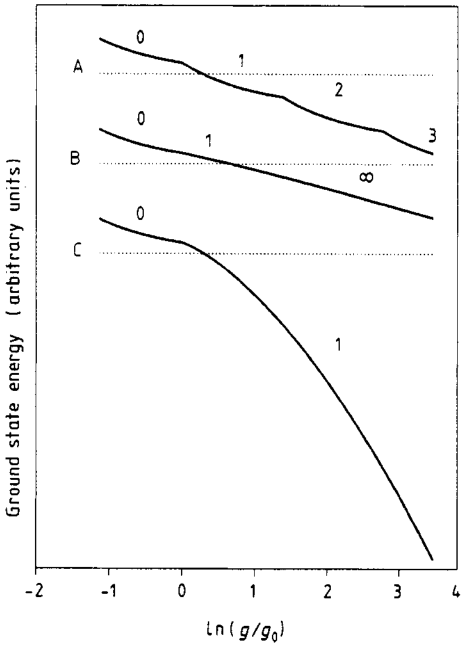

Figure 3. The energies of the different ground states plotted on an arbitrary scale against the logarithm of the gain relative g relative to g_0 defined in the text, for the three distributions P_η considered: curve A, equation (16) ($a = 0.25$); curve B, equation (17); curve C, equation (18). 0 symbolises the uniform ground state, 1 the retrieval states, $2, 3, \ldots$ symmetric n-mixtures, and ∞ the spin-glass phase.

The use of continuous variables turns candidate metastable states into unstable saddle points of the free-energy.

It has to be stressed that only the limit of zero noise has been treated here. It is straightforward, but analytically cumbersome, to extend the above analysis to the case of finite noise. One notes that in this noiseless limit the effect of decreasing the gain is, as could have been expected, in some respect similar to the effect of increasing the noise level in a network of binary units [5].

9. Simulations

Numerical simulations have been performed on networks of size $N = 5000$, with $p = 50$ patterns encoded in the synaptic strengths. The main intent was to test whether the types of behaviour predicted analytically in the limits $p \to \infty$, $\alpha \to 0$ persisted with finite (and rather low) values of N and p. In the simulations the 'threshold-linear' neuron representation was used in the absence of noise. This amounted to randomly selecting each time a neuron to be updated, computing its local field h_i, and updating it to $V_i = g(h_i - T_{\rm hr})$ if $h_i > T_{\rm hr}$, and to $V_i = 0$ otherwise. A 'time cycle' consisted of N updatings. Initial configuration were chosen in a variety of ways depending on the most important features to be tested in any given gain regime. The synaptic strengths were determined according to the three cases for P_η examined above.

A first general observation concerns the relevance of the value chosen for the threshold, $T_{\rm hr}$. Whereas, as shown above, this value is immaterial in the $p \to \infty$ limit, in simulations it does affect the dynamics of the system. In particular, a high positive threshold (how high depends on the gain and the other parameters) does invariably 'turn off' the network, bringing all neurons to the quiescent state. Below such values, varying the threshold resulted mainly in a mild variation of the mean activity level of the network, which could typically deviate as much as $30-40\%$ from the value set for λ. In most of the cases reported in the following, the threshold was set to zero.

The first case to be considered was that of a binary pattern assignment, equation (16). We set, as in the following, $a = 0.25$ and $\lambda = 0.1$. For low gain ($g < 0.2 \sim 0.25$) the system, started in an initial configuration having a much stronger overlap with one of the patterns than with the rest, always evolved rapidly ($<$ 10 time cycles) towards configurations bearing similar overlaps (within 100% fluctuations) with all the patterns. Eventually (over a longer time scale, of tens of time cycles) it settled into a fixed configuration. Note that for $a = 0.25$, $g_0 = 0.33$.

For $0.2 \sim 0.25 < g < 0.7 \sim 1.0$ the system, both when started in a completely random configuration or in one having a strong overlap with two or more pattern, typically evolves toward configurations in which a single overlap dominates the others. When the initial configuration is itself prepared with a single 'high' overlap, in most cases the same overlap stays high (corresponding to proper retrieval) while in others a different pattern is selected at the expense of the former. This is a manifestation of the fact, apparent throughout the simulations, that finite size fluctuations are much more important in determining the 'strongest' patterns than in modifying drastically the nature of the typical fixed points of the dynamics. The strongest patterns are those that have been 'favoured' in the random assignment of the η by having more neurons active in them. Thus, in a small system a strong pattern may easily dominate, after a transient, over the pattern favoured by the initial configuration.

For higher g values, the network tends to evolve towards configurations in which an increasing (with g) number of patterns have a distinctively high overlap with respect to the average ($0.5 \sim 0.9$ cf $0.07 \sim 0.1$). This behaviour reproduces qualitatively the one expected from the above analysis, although such 'mixed' states hardly appear as symmetric n-mixtures, and the number of partecipating patterns, if increasing with g, shows marked fluctuations.

Next, P_η was assumed to have the flat distribution of (17). For $g < 2.5 \sim 3.5$, again any initially high overlap would be washed away in a few time cycles. The network reached quickly an almost fixed configuration with a high degree of uniformity (typical fluctuations of a few percent in the values of the various overlaps). For slightly higher values of g, the network did preserve a single high overlap present in the initial configuration to a higher level than the rest (of some 40%), although when started from a random configuration in most cases no pattern was selected to have a distinctly higher overlap. Finally, for even higher gains, the situation becomes more confused, as the increased spread in the values of all the overlaps (30–40%) dominates over any clear classification into 'high' and 'low' ones.

The last case, of an exponential distribution, (18), is the one that yields clearer results. Around $g \sim 1.1$ there is a transition from a low- to a high-gain regime. At low gain the system always reaches a a fixed point with nearly uniform values for all the overlaps from any initial configuration. At high gain, the initially dominant pattern is enhanced during the evolution to an overlap value which grows with g, while the spread in the other overlaps is limited. In the transition region, sometimes a different pattern is selected to be high, from the one dominant in the initial configuration.

10. Discussion

This paper shows how the same statistical techniques that have been widely used in analysing the attractor neural networks modelling associative memory can be extended easily to the case in which a more general representation is used for the elementary neuron variables.

In particular, the threshold-linear representation is a very simplified scheme, that is possibly more appropriate than the conventional binary representation in studying network models in which individual firing rates are kept well below saturation levels by inhibitory mechanisms. In contrast to the binary network [6], the present one also exhibits interesting retrieval behaviour at zero-noise levels. The constraint of low rates is in fact compatible, in this case, with the freezing of the system into a single configuration. Such freezing does not imply that individual neuron persists in firing at saturation levels, or in the quiescent state. It simply reflects the fact that each neuron settles, in the long-time limit, into a given firing frequency consistent with those of all other neurons in the network.

Whereas in the case of binary neurons [5] the parameter essential in determining the long-time behaviour was the noise level T, with threshold-linear neurons (at $T = 0$) this role is performed by the gain, g. In some loose sense, increasing gain values is analogous to decreasing noise levels. However, in the binary case the structure of stable states, even in the simplifying $\alpha \rightarrow 0$ limit, presented complex features such as a variety of critical noise levels and of possibly metastable states. In the analogue case, instead, there is, at least in the limits used in the above analysis, a very simple pattern of solutions of a single type, for each g and each choice of the distribution P_η determining synaptic strengths. Intermediate g values, in the region above a critical g_0 determined by the moments of P_η, provide the appropriate regime for a network operating according to the requirement, motivated by biological plausibility, of retrieving a memorised pattern without driving individual neurons towards firing at saturation levels.

With respect to those plausibility requirements, the use of threshold-linear neurons causes a different interplay between model parameters and time scales. In the binary case, the explicit representation of firing saturation effects implies [5] that: (i) if the average firing level λ is to be suitably below saturation, one needs $\lambda \ll 1$ and (ii) if the firing rate of the most active neurons is also to be much below saturation, then $\lambda/a \ll 1$. In the threshold-linear case, saturation is not represented, so the average firing activity λ remains a free parameter, that itself sets the time scale in accordance with that of biological systems. The constraint of low individual firing rates is now reflected in the requirement that the most active neurons should not exhibit frequencies higher than the average value by some pre-chosen enhancement factor. If, for example, λ is taken to correspond to an average activity of a few spikes s^{-1}, a typical 'realistic' enhancement could be of an order of magnitude. The implementation of such a constraint in the model is easy, but the g value required depends strongly on the choice of P_η.

One should note that within this model there remains a problem in relating the gain g itself to some experimentally measurable parameter. This is due to the fact that g is related to the normalisation of the terms entering the local field h of each neuron. In the simplified scheme adopted in the model, involving the effective representation of inhibition and the $p \rightarrow \infty$ limit, that normalisation cannot be related to any 'realistic' physiological parameter.

An interesting question that has not been treated here concerns the storage capacity of networks that use threshold–linear neurons. A capacity calculation provides a meaningful test of the generality of results obtained with binary neurons. This is presented in a separate paper [21].

Acknowledgments

The author is grateful to M Abeles for a stimulating remark, to D J Amit for many discussions and for reviewing the manuscript, and to M Griniasty and R Payne for helpful suggestions.

Appendix

We indicate here the main lines of the calculation used to derive equation (24), using the replica method, from the definition of the free-energy

$$f = \lim_{\substack{n\to 0\\ N\to\infty}} \frac{-1}{\beta n N}[\langle\langle Z^n\rangle\rangle - 1]. \tag{A1}$$

Here

$$Z^n = \mathrm{Tr}_{\{V^\gamma\}} \exp\left(-\beta\sum_\gamma H^\gamma\right) \tag{A2}$$

is the partition function for n identical replicas of the systems, labelled with the indices $\gamma, \delta, \ldots$.

One starts by introducing the overlap parameters through δ-functions, so that

$$Z^n = \left(\frac{N\beta}{2\pi}\right)^{pn} \int \mathrm{d}t^{\mu\gamma}\,\mathrm{d}x^{\mu\gamma}\mathrm{Tr}_{\{V^\gamma\}} \exp\beta N \sum_\gamma \Bigg[\mathrm{i}\sum_\mu t^{\mu\gamma}\left(x^{\mu\gamma} - \frac{1}{N}\sum_i \frac{\eta_i^\mu}{a}V_i^\gamma\right) + \frac{1}{2}\sum_\mu (x^{\mu\gamma})^2 - \frac{1}{3\lambda p^2}\left(\sum_\mu x^{\mu\gamma}\right)^3 - \frac{1}{2Na}\sum_\mu x^{\mu\gamma}\Bigg]. \tag{A3}$$

where the $t^{\mu\gamma}$ are Lagrange multipliers which impose the definitions of the order parameters.

Then one assumes that only a finite number s of patterns can condense, i.e. they have an overlap x^μ that deviates from the average λ by a finite amount (as $N \to \infty$), and one takes an average over the η distribution of the remaining $p - s$ patterns. The integrals over the $t^{\mu\gamma}$ for $\mu > s$ reduce to Gaussian integrals, after neglecting terms which vanish as $N \to \infty$. Next one introduces the global order parameters

$$\begin{aligned} y^{\gamma\delta} &= \frac{1}{N}\sum_i V_i^\gamma V_i^\delta \\ x^\gamma &= \frac{1}{N}\sum_i V_i^\gamma \end{aligned} \tag{A4}$$

via δ-functions. Denoting by Y the matrix with elements $y^{\gamma\delta}$, one obtains

$$\begin{aligned}
\langle\!\langle Z^n \rangle\!\rangle = \left(\frac{N\beta}{2\pi}\right)^{\{sn+[(p-s)/2]n+[(n+3)/2]n\}} & \int \mathrm{d}x^{\mu\gamma}\,\mathrm{d}t^{\sigma\gamma}\,\mathrm{d}t^{\gamma}\,\mathrm{d}x^{\gamma}\,\mathrm{d}y^{\gamma\delta}\,\mathrm{d}r^{\gamma\delta} \\
& \left\{\exp \beta N \sum_{\gamma}\left[\frac{1}{2}\sum_{\mu}(x^{\mu\gamma})^2 - \frac{1}{3\lambda p^2}\left(\sum_{\mu} x^{\mu\gamma}\right)^3 - \frac{1}{2Na}\sum_{\mu} x^{\mu\gamma}\right]\right. \\
& \times \exp\left[-\left(\frac{p-s}{2}\mathrm{Tr}_{\{\gamma\}} \ln Y + \frac{Na^2}{2a_2}\sum_{\gamma,\delta,\nu}(x^{\nu\gamma} - x^{\gamma})Y^{-1}_{\gamma\delta}(x^{\nu\delta} - x^{\delta})\right)\right] \\
& \times \mathrm{Tr}_{\{V^\gamma\}} \left\langle\!\!\left\langle \exp \mathrm{i}\beta N \left[\sum_{\sigma,\gamma} t^{\sigma\gamma}\left(x^{\sigma\gamma} - \frac{1}{N}\sum_{i}\frac{\eta_i^{\mu}}{a}V_i^{\gamma}\right)\right.\right.\right. \\
& \left.\left.\left.\left. + \sum_{\gamma} t^{\gamma}\left(x^{\gamma} - \frac{1}{N}\sum_{i} V_i^{\gamma}\right) + \sum_{\gamma,\delta} r^{\gamma\delta}\left(y^{\gamma\delta} - \frac{1}{N}\sum_{i} V_i^{\gamma}V_i^{\delta}\right)\right]\right\rangle\!\!\right\rangle\right\}
\end{aligned} \tag{A5}$$

(the index σ runs over the condensed patterns only, the index ν over the uncondensed ones). The symbol a_2 has been defined in the main text.

To handle the cubic term, one introduces one more order parameter

$$\chi^{\gamma} = \frac{1}{p}\sum_{\mu} x^{\mu\gamma} - x^{\gamma} \tag{A6}$$

to be able to perform the integrals over $x^{\mu\gamma}$ for $\mu > s$. The main effect of these integrals is to produce a factor $\exp\{-(p-s)\mathrm{Tr}_{\gamma} \ln[\frac{1}{2}(Y^{-1} - \beta a_2/a^2)]\}$, which, together with the similar looking factor already present in (A5), can give a finite contribution to the free energy if $p = \alpha N$ with α finite. Considering that the χ and $(x - \lambda)$ are fluctuations of order $\mathrm{O}(1/\sqrt{p})$ and $\mathrm{O}(1/p)$ respectively, one can also integrate over them. Then $\langle\!\langle Z^n \rangle\!\rangle$ can be evaluated at the saddle point as $\exp(-n\beta N f)$ where f is given by the terms which remain finite as $p, N \to \infty$

$$\begin{aligned}
f = -\frac{1}{n}\sum_{\gamma}\sum_{\sigma}\frac{(x^{\sigma\gamma} - \lambda)^2}{2} + \frac{\alpha}{2\beta n}\mathrm{Tr}_{\{\gamma\}} \ln\left(1 - \beta\frac{a_2}{a^2}Y\right) & \\
- \frac{\mathrm{i}}{n}\left(\sum_{\sigma,\gamma} t^{\sigma\gamma}x^{\sigma\gamma} + \sum_{\gamma} t^{\gamma}\lambda + \sum_{(\gamma,\delta)} r^{\gamma\delta}y^{\gamma\delta}\right) & \\
- \frac{1}{\beta n}\left\langle\!\!\left\langle \ln \mathrm{Tr}_{\{V^\gamma\}} \exp\left[-\mathrm{i}\beta\left(\sum_{\sigma,\gamma} t^{\sigma\gamma}\frac{\eta^{\sigma}}{a}V^{\gamma}\right.\right.\right.\right. & \\
\left.\left.\left.\left. + \sum_{\gamma} t^{\gamma}V^{\gamma} + \sum_{(\gamma,\delta)} r^{\gamma\delta}V^{\gamma}V^{\delta}\right)\right]\right\rangle\!\!\right\rangle &
\end{aligned} \tag{A7}$$

One then uses a replica symmetry *ansatz* and takes the limit $n \to 0$. Using the saddle point equations for the x^{σ} to eliminate the t^{σ}, and writing $\rho = -2\mathrm{i}r/\beta$, $\Delta = \mathrm{i}t$ one arrives at equations (24)–(26).

References

[1] Hopfield J J 1982 *Proc. Natl Acad. Sci.* **79** 2554
[2] Amit D J 1989 *Modelling Brain Function* (Cambridge: Cambridge University Press)
[3] Kohonen T 1988 *Self Organization and Associative Memory* (Berlin: Springer)
[4] Buzsáki G 1989 *Neurosci.* **31** 551
[5] Treves A and Amit D J 1989 *J. Phys. A: Math. Gen.* **22** 2205
[6] Amit D J and Treves A 1989 *Proc. Natl Acad. Sci.* **86** 7871
[7] Rubin N and Sompolinsky H 1989 *Europhys. Lett.* **10** 465
[8] Golomb D, Rubin N and Sompolinsky H 1990 *Phys. Rev.* A **41** 1843
[9] Buhmann J 1989 *Phys. Rev.* A **40** 4145
[10] Glauber R J 1963 *J. Math. Phys.* **4** 294
[11] McCullough W S and Pitts W 1943 *Bull. Math. Biophys.* **5** 115
[12] MacGregor R J 1987 *Neural and Brain Modelling* (New York: Academic)
[13] Grossberg S 1988 *Neural Networks* **1** 17
[14] Hopfield J J 1984 *Proc. Natl Acad. Sci.* **81** 3088
[15] Cohen M A and Grossberg S 1983 *IEEE Trans. SMC* **13** 815
[16] Shaw G and Vasudevan R 1974 *Math. Biosci.* **21** 207
[17] Ratliff F 1965 *Mach Bands: Quantitative Studies of Neural Networks in the Retina* (San Francisco: Holden-Day)
[18] Morris R G M and Willshaw D J 1989 *Nature* **339** 175
[19] Willshaw D J, Buneman O P and Longuet-Higgins H C 1969 *Nature* **222** 960
[20] Amit D J, Gutfreund H and Sompolinsky H 1987 *Ann. Phys., NY* **173** 30
[21] Treves A 1990 Graded-response neurons and information encodings in auto-associative memories *Preprint* Oxford Univeristy