Low firing rates: an effective Hamiltonian for excitatory neurons

1989 J. Phys. A: Math. Gen. 22 2205

(http://iopscience.iop.org/0305-4470/22/12/020)

J. Phys. A: Math. Gen. 22 (1989) 2205-2226. Printed in the UK

Low firing rates: an effective Hamiltonian for excitatory neurons

A Treves† and Daniel J Amit‡§

† Racah Institute of Physics, Hebrew University, Jerusalem, Israel
‡ INFN, Dipartimento di Fisica, Università 'La Sapienza', Roma, Italy

Received 19 January 1989

Abstract. We analyse the behaviour of an attractor neural network which exhibits low mean temporal activity levels, despite the fact that the intrinsic neuronal cycle time is very short (2-3 ms). Information and computation are represented on the excitatory neurons only. The influence of inhibitory neurons, which are assumed to react on a shorter timescale than the excitatory ones, is expressed as an effective interaction of the excitatory neurons. This leads to an effective model, which describes the interplay of excitation and inhibition acting on excitatory neurons in terms of the excitatory neural variables alone.

The network operates in the presence of fast noise, which is large relative to the frozen randomness induced by the stored patterns. The overall fraction of active neurons is controlled by a single free parameter, which expresses the relative strength of the effective inhibition. Associative retrieval is identified, as usual, with the breakdown of ergodicity in the dynamics of the network, in particular with the presence of dynamical attractors corresponding to the retrieval of a given pattern. In such an attractor, the activity of neurons corresponding to active sites in the stored patterns increases at the expense of other neurons. Yet only a small fraction of the neurons active in the pattern are in the active state in each elementary time cycle, and they vary from cycle to cycle in an uncorrelated fashion, due to the noise. Hence, the observed mean activity rate of any individual neuron is kept low. This scenario is demonstrated by an analytical study based on the replica method, and the results are tested by numerical simulations.

1. Introduction

The attractor neural networks, which have been proposed in recent years (Hopfield 1982, 1984) to model associative memory in biological systems, combine collective properties emerging from very strong feedback with transparency to analytic investigation (Amit *et al* 1985a, b, 1987a). A number of drastic simplifying assumptions have been called upon to produce this result. Subsequently, the properties of the network have been shown to be robust when many of the simplifications are lifted: full connectivity is not essential, nor are detailed synaptic efficacies (Sompolinsky 1987); the synaptic connections need not be symmetric, and can be extremely diluted (Gutfreund 1987, Derrida *et al* 1987); individual neurons can be assigned a unique excitatory or inhibitory function, according to Dale's Law (Shinomoto 1987); the spatial mean of the network activity need not be $\frac{1}{2}$ (Amit *et al* 1987b, Tsodyks and Feigelman 1988, Buhmann *et al* 1988).

§ On leave from the Racah Institute of Physics, Hebrew University, Jerusalem, Israel.

0305-4470/89/122205+22$02.50

However, other assumptions, as well as some resulting features, remain objectionable from a biological point of view. We address here another important problem of the models proposed so far: the resulting high spike rates of individual neurons upon retrieval. This issue is at the core of the attractor interpretation of memory retrieval. In the attractor dynamics, such as the 'standard' Hopfield scheme in the absence of noise, the network arrives, in its flow in configuration space, to a fixed point. Presence in a fixed point implies that some of the neurons are firing at their maximal rate, which corresponds to a few hundred spikes per second (Amit 1987). But available recordings from cells in associative areas of the cortex do not show such high rates: the most active neurons appear to fire in the order of tens, up to a hundred, spikes per second (Abeles 1982, Anderson and Mountcastle 1983, Sur *et al* 1984, Goldberg and Bruce 1985, Miyashita and Chang 1988).

The root of the problem is that, during retrieval, a fraction of the neurons is *constantly* in the active state. A constant neuronal activity is interpreted in terms of spike rates per second, and the correspondence is fixed by the dynamical characteristics of the neurons being modelled. A consistent representation of biological neurons by two-state variables that are updated at typical time intervals τ, assumes that τ is of the order of the absolute refractory period of the neuron and of its effective integration time constant, which in turn are of the same order of magnitude. It then follows that a model neuron that remains for a certain duration in its active state simulates a real neuron that fires at the high rates mentioned above (Hopfield 1984). Introducing noise does not help. Although some of the neurons may now change their state, the bulk of them will still be essentially frozen either 'on' or 'off'. At most, when the noise level is already so high as to almost restore ergodicity (and destroy retrieval), neurons will have close to equal probabilities to be in either state, thereby reducing the resulting spike rates by only $\frac{1}{2}$. But then also the mean activity of neurons that should be quiescent will be close to $\frac{1}{2}$.

We have recently proposed an attractor neural network (Amit and Treves 1989), which is a modified version of the standard model. This model demonstrates that the high rate problem is not a necessary feature of the attractor paradigm, but only a consequence of realising it as a flow toward fixed points in the configuration space of the network. In the proposed model, ergodicity is broken at suitably low levels of noise, but still the noise level is high enough to keep changing the state of almost all neurons participating in retrieval, at every elementary cycle. The attractors that dominate the dynamics lie close to the configuration with all the N neurons quiescent, but at a finite Hamming distance Na, which is the same for all attractors and is determined by a free parameter of the model. Thus, a small fraction a of the neurons are active in each attractor. This fraction, however, is not made up to Na neurons that freeze in their active state (this is essentially the case in models with low *spatial* mean activity), but rather of a much wider set of neurons that have a definite (and low) probability of being active. In terms of the time evolution of the network, the neurons that are in the active state are chosen at *random* (i.e. by fast noise) at each time step, according to certain probability distributions characterising each attractor. In other words, each attractor represents a stochastic sequence of configurations, and each individual neuron is quiescent in most of the configurations of the sequence. It is via this strong stochasticity of the attractors that we bridge the gap between the neuronal fast cycle time and the low activity rates upon retrieval.

The model is based on the separation between the dynamics of the excitatory neurons and that of the inhibitory neurons. The former are assumed to carry all the

information relevant for retrieval, while the latter have the role of regulating the overall activity of the network of the excitatory neurons. The dynamics of the inhibitory neurons is assumed to proceed on a faster timescale, possibly modelling the more local nature of the inhibition. The inhibition resulting from a given excitatory activity distribution may then be expressed in an effective way as a function of the excitatory activity itself. In this way one can reduce the dynamical variables to the activity states of excitatory neurons only. Within this picture the resulting firing rates are determined, at an arbitrary low level, by a single additional free parameter, which represents the strength of the inhibitory effective coupling relative to the excitatory interactions.

In addition to suggesting a solution to the rates problem, the model is biologically more plausible on several other counts.

1. The distinction between excitatory and inhibitory neurons is not merely formal, but attempts to reproduce and interpret functional differences. In particular, a scheme is suggested that takes into account the more local nature of inhibition, and the non-linear summation of inhibitory inputs.

2. The information content of the network is stored, essentially, only in the excitatory-excitatory synapses, the ones that are expected to be plastic (see, e.g., Eccles 1964).

3. The excitatory synaptic efficacies are closer to the original Hebb rule (Hebb 1949), i.e. they are enhanced only by coincidences in the activity of the pre- and post-synaptic neurons. Recall that in the standard model they were enhanced also when both neurons were quiescent.

4. Only firing neurons, among the excitatory ones, participate in the retrieval process, while in the standard model also quiescent neurons intervened in stabilising the attractors.

Despite these differences, and the fact that the pertinent asymptotic behaviour is of a different nature, the model is still amenable to a detailed analysis, similar to the one developed for the standard model (Amit *et al* 1985a, 1987a). It therefore provides the basic insights necessary for more complex situations that can be only partially explored by analytic means.

The present paper is structured as follows: in § 2 we consider a network comprising both excitatory and inhibitory neurons, and we suggest a description of its relevant behaviour in terms of an effective model consisting of excitatory neurons only. In § 3 we discuss the scenario appropriate to the cognitive function of associative memory retrieval, and in § 4 we formulate a mean-field analysis of the model. In § 5 we illustrate the phase diagram at low memory loading, while the corrections to this simplified picture are treated in § 6. In § 7 we present the results of sample simulations, and in the last section we conclude and discuss possible future developments.

2. An effective Hamiltonian for excitatory neurons

We consider a network composed of N excitatory and $\bar{N}$ inhibitory neurons, represented by two-state (0, 1) variables V_i, $i = 1, \ldots, N$; $\bar{V}_k$, $k = 1, \ldots, \bar{N}$. A '1' represents a neuron that emits a single spike in a given discrete time slice which is of the order of the absolute refractory period; a '0' represents a neuron that is quiescent in that time slice. We assume that only the excitatory neurons are involved in associative retrieval. Accordingly, the network would store p patterns, or activity distributions of the excitatory neurons, defined by p N-bit words $\eta_i^\mu (=0, 1)$, $\mu = 1, \ldots, p$. If the excitatory neuron i has $\eta_i^\mu = 1$, it is expected to have an enhanced activity rate upon retrieval of

pattern μ, while if $\eta_i^\mu = 0$, its rate should be depressed relative to the overall mean activity rate. The η are random and are chosen independently with a probability distribution

$$P(\eta) = a\delta(\eta - 1) + (1 - a)\delta(\eta). \tag{1}$$

Note the following.

1. We require only an *enhanced* activity rate for neurons participating in retrieval of a given pattern, and not the *maximal* rate corresponding to a constant $V_i = 1$.

2. The mean number of such neurons will be aN, and we shall assume $0 < a \ll 1$; this *sparse coding* implies that spatial firing rates will also be low (Amit *et al* 1987b, Tsodyks and Feigelman 1988, Buhmann *et al* 1988).

2.1. A dynamics with two timescales—fast inhibition

The dynamical evolution of the excitatory sub-network is asynchronous, which is captured by assuming that it takes place in a time discretised in units $\Delta t = \tau / N$, where τ is the timescale of the single neuron dynamics (Peretto and Niez 1986). At each time step one of the neurons will update its state according to a Glauber process (Glauber 1963)

$$P(V_i(t + \Delta t) = 1) = [\exp(-\beta h_i(t)) + 1]^{-1} \tag{2}$$

where $\beta^{-1} \equiv T$ measures the amount of stochastic noise in the process, and h_i is the local field (post-synaptic potential) representing the total effect of the network activity at time t on excitatory neuron i.

We assume h_i to be the sum of excitatory and inhibitory terms. The excitatory part is the sum of contributions from all other excitatory neurons through direct synaptic coupling:

$$h_i^E = \sum_{j \neq i}^{N} J_{ij} V_j \tag{3}$$

where the synaptic couplings assume the form

$$J_{ij} = \frac{1}{Na^2} \sum_{\mu=1}^{p} \eta_i^\mu \eta_j^\mu. \tag{4}$$

Note that since $\eta = 0, 1$, a pair of neurons which are inactive in a pattern does not contribute to strengthening the synaptic efficacies†.

The inhibitory part of the local field reflects the activity of the inhibitory neurons. The dynamics of the inhibitory neurons is assumed to be of the same type, but taking place at a faster rate, $\bar{\tau}^{-1}$. This assumption is intended to account for the typically shorter spatial range of the inhibition, which results in a shorter mean time needed by inhibitory neurons to react on excitatory ones, as mentioned at the end of this section. The post-synaptic potential of inhibitory neurons is supposed to include excitatory and inhibitory contributions, but is taken to be roughly *uniform*, i.e. the same for all inhibitory neurons. This is another aspect of the hypothesis that the only structured information carried by the synaptic weights concerns the activity of excitatory neurons. In the same spirit, we assume that the weight of inhibitory synapses affecting excitatory

† This should be compared and contrasted with Willshaw's matrix (Willshaw *et al* 1969) employed recently by Rubin and Sompolinsky (1989) in a model which also addresses the issue of low firing rates.

neurons does not depend on the pre-synaptic neuron. Then, the firing activities of single inhibitory neurons do not have individual significance, but rather it is their overall mean activity.

$$I = \sum_{k=1}^{\tilde{N}} \frac{\bar{V}_k}{\tilde{N}}$$

that is relevant for the behaviour of the excitatory part of the network. If $\bar{\tau}$ is sufficiently shorter than τ, the mean inhibitory activity I relaxes to the value determined by the current excitatory activity distribution, before excitatory neurons have had the time to modify appreciably their configuration. Then the inhibitory contribution to the local field of excitatory neurons can be written

$$h_i^I(t) = h_i^I(I(t)) \simeq h_i^I(I(\{V_i(t)\})) \tag{5}$$

i.e. the inhibition can be expressed as a function of the V_i at the same time t.

2.2. *Effective dynamics of the excitatory neurons*

The above considerations suggest the elimination of the inhibitory neurons as dynamical variables. They regulate the mean activity level of the network by providing an inhibition which is a function of a suitably averaged *excitatory* activity. Such a mechanism is robust if the inhibitory contribution h^I is stronger than a linear function of the excitatory activities, in the sense that the balance between a non-linear h^I and the linear h^E, equation (3), will produce a well defined mean activity rate at equilibrium. An instantaneous activity lower than this mean will grow due to a stronger excitation, while an excess of activity will be damped by overwhelming inhibition. We make a simple choice and restrict ourselves to a quadratic function. The inhibitory contribution to excitatory neuron i is written as

$$h_i^I = -\frac{1}{\nu}\left(\sum_{\lambda=1}^{p} \frac{\eta_i^\lambda}{a}\right)\left(\frac{1}{Np}\sum_{\mu=1}^{p}\sum_{j=1}^{N} \frac{\eta_j^\mu}{a} V_j\right)^2 . \tag{6}$$

Here the first term in brackets is a factor that depends on the post-synaptic excitatory neuron, and enhances the inhibition of those neurons that are active in more patterns. There is a new free parameter, ν, which determines the overall strength of the inhibition compared to the excitation. The reason for taking the average of the excitatory activity weighted with the η, in the term in the second set of brackets, will become clearer in the following.

The form chosen for the inhibition is equivalent to a set of triadic couplings between excitatory neurons. However, we do not assume the existence of such couplings. It is of the nature of effective interactions, mediated by invisible dynamical variables, to take the form of multiparticle interactions. Nor is the limit $\bar{\tau} \ll \tau$ intended to correspond to a real physiological situation. Rather, the aim of these assumptions is to describe in a simplified, effective way a plurality of biological features.

1. Inhibitory neurons are cells with different morphological and physiological properties from excitatory ones. Their action is of a more local nature, and this implies shorter transmission times for inhibitory signals. They may react on the nearby excitatory neurons in a shorter time than the typical time needed by the excitatory neurons, which are on average more distant, to exchange information (Bower 1988). This situation can be modelled by assuming $\bar{\tau} < \tau$, which is an effective representation in a network which is not endowed with geometrical structure.

2. Inhibitory mechanisms are intrinsically different from excitatory ones (Fatt and Katz 1953, Segev and Rall 1987). Rather than provide an hyperpolarising potential, an activated inhibitory synapse will, usually, only increase the membrane conductance of the post-synaptic neuron, thereby tending to shunt the local dendritic branch (Segev and Parnas 1983). In a model network in which the geometrical structure is neglected, this may be described by assuming a strongly non-linear function h^l, as in (6).

The main advantage of the specific interaction, equation (6), is that the dynamics of the excitatory network can be described in terms of an effective Hamiltonian. This Hamiltonian has a single new free parameter, ν. In fact, we can assign an 'energy' to each firing configuration of the excitatory neurons as

$$H\{V\} = -\frac{1}{2Na^2}\sum_{i,j\,i\neq j}\sum_{\mu}\eta_i^{\mu}\eta_j^{\mu}V_iV_j + \frac{1}{3\nu N^2p^2a^3}\sum_{\mu,\lambda,\sigma}\sum_{i,j,k}\eta_i^{\mu}\eta_j^{\lambda}\eta_k^{\sigma}V_iV_jV_k \tag{7}$$

and thus the behaviour of the network can be analysed by means of the same kind of thermodynamical treatment developed for the Hopfield model (Amit *et al* 1985a, b, 1987a).

3. The new retrieval scenario

In a system described by a Hamiltonian and subject to stochastic noise, the non-ergodic behaviour manifests itself in the free-energy landscape (Amit 1989). This is expected to have several local minima, and to classify them we define a set of order parameters. The overlaps with the stored patterns will be

$$x^{\mu} = \frac{1}{N}\sum_{i=1}^{N}\frac{\eta_i^{\mu}}{a}\langle V_i\rangle \tag{8}$$

where the angle bracket stands for a temporal average. The average $\langle V_i\rangle$ gives directly the firing rate of neuron i, relative to the maximal rate τ^{-1}. The parameter x^{μ} is, therefore, the relative mean firing rate of those neurons that should have an enhanced activity in pattern μ (the η_i^{μ} exclude other neurons from the average).

The overall mean activity of the excitatory network will be measured by

$$x = \frac{1}{N}\sum_{i=1}^{N}\langle V_i\rangle \tag{9}$$

and

$$y = \frac{1}{N}\sum_{i=1}^{N}\langle V_i\rangle^2 \tag{10}$$

will measure the correlation in time of consecutive network activity states.

Clearly, $0 \leqslant x^{\mu} \leqslant 1$ and $0 \leqslant y \leqslant x \leqslant 1$. The special case $y = x$ implies that the system is frozen into a single configuration, i.e. that the same neurons fire at each time step, while $y = x^2$ represents a situation in which consecutive network states are completely uncorrelated. In this case $x^{\mu} = x$ for all μ. Intermediate values for y will indicate some amount of correlation which may or may not be related to retrieval.

Retrieval of a stored pattern will be realised when one of the overlaps, say x^1, will be distinctly higher than the mean overall activity, while the other x^{μ} will retain a uniform value. We emphasise again that an enhanced value for x^1 signifies a higher

spike rate of the neurons selected by pattern 1, only. Their mean firing rate will be x^1/τ spikes per second. Since this rate is to be considerably lower than the maximal rate τ^{-1}, x^1 will have to be much smaller than 1. The mean activity x of the network should be considerably lower than x^1, to afford clear recognition of pattern 1, and to comply with experimental evidence on mean firing rates in real systems (Abeles 1982, Anderson and Mountcastle 1983, Sur *et al* 1984, Goldberg and Bruce 1985, Miyashita and Chang 1988).

The mean activity of those neurons that are *not* selected by pattern 1, to be denoted by $\bar{x}^1$, is determined by the identity $ax^1+(1-a)\bar{x}^1=x$. Hence,

$$\bar{x}^1=x-\frac{a}{1-a}(x^1-x) \tag{11}$$

which is even lower than x. In other words, the enhanced activity of the neurons selected by a pattern is at the expense of the rest of the neurons. In the following, we shall refer to the neurons which should have an enhanced activity upon retrieval of pattern μ (those for which $\eta_i^\mu=1$) as *active* in pattern μ, and to the rest of the neurons as *passive* in that pattern. Since the other patterns, $\mu\neq 1$, are uncorrelated, we shall have $x^{\mu\neq 1}=x$.

The simplest scenario would be if the activity of the neurons in the network were to depend only on whether a neuron is to be active or passive in pattern 1. Then the mean activity in each group would be uniform. In such a case, the global correlation parameter y would be

$$y=a(x^1)^2+(1-a)(\bar{x}^1)^2. \tag{12}$$

This represents a maximally disordered phase, within the constraints imposed by the emergence of pattern 1. The probability that any of the aN neurons active in pattern 1 fires, in any given time slice, is uniform and equal to x^1. These neurons, therefore, give a contribution $a(x^1)^2$ to y. The $(1-a)N$ neurons passive in pattern 1 fire with probability $\bar{x}^1$, and hence contribute the second term to y. This uniformity, however, represents an extreme case. Empirically, the activity rates of different neurons are not uniform. A model affecting retrieval can operate with non-uniform activities, as long as the emergence of a single pattern is clearly signalled by the overlap parameter. The only additional requirement should be that high rates of individual neurons should be very rare. In practice, we shall usually encounter situations in which (12) is nearly satisfied. This represents a simple way of controlling N local quantities through the use of a single global parameter.

Finally, to give some concrete numerical content to our retrieval scenario, consider an example with a basic neuronal timescale of $\tau\simeq 2.5$ ms. To reproduce mean overall firing rates of $10\ s^{-1}$ that increase for active neurons to $80\ s^{-1}$ upon retrieval, one would need $x^1\simeq 0.2$, $x\simeq 0.025$. The analytical treatment to be developed will make sense only if within the proposed model this scenario can be realised in the meaningful parameter range.

4. Mean-field theory in the $p\to\infty$ limit

We shall focus on the case in which both N and p are very large numbers, so we shall take the limits N, $p\to\infty$. All the quantities will be functions of a, ν, and $\alpha\equiv p/N$. While $N\to\infty$ is the usual thermodynamic limit, some care must be exercised in taking the limit $p\to\infty$.

4.1. The large-p limit

The energy per neuron of a generic configuration is, in the present normalisation, a quantity of order p. If the configuration is uncorrelated with any of the patterns, i.e. if $x^\mu \simeq x$ for all μ we have

$$E/N \simeq p(-\tfrac{1}{2}x^2 + (1/3\nu)x^3). \tag{13}$$

Moreover, it can be shown that no configuration can have a finite (extensive) correlation with more than a finite number of patterns. This follows from the equality

$$\left\langle\left\langle \sum_\mu (x^\mu - x)^2 \right\rangle\right\rangle = \alpha x \frac{1-a}{a} \tag{14}$$

which holds for any given configuration (see e.g. appendix 1). The double brackets stand for averaging over the η. Consequently, the error one makes by neglecting the differences $x^\mu - x$ does not grow linearly with p, and the energy is given, to leading order, by (13). In addition, the entropy per neuron does not grow with p. Therefore, the minima of the free energy will occur when (13) is minimised, i.e. for

$$x = \nu + \text{terms vanishing as } p \to \infty.$$

In other words, in the $p \to \infty$ limit, the mean activity of the network is constrained by the overall inhibition to be ν. This happens, of course, provided $\nu \leqslant 1$. In the following we shall always assume $0 < \nu < a$. It is then convenient to use a set of properly subtracted order parameters

$$\hat{x}^\mu \equiv x^\mu - \nu \qquad \hat{x} \equiv x - \nu \qquad \hat{y} \equiv y - \nu. \tag{15}$$

Furthermore, a finite amount of noise will cause finite fluctuations in the energy, even when $p \to \infty$. So $\hat{x}$ will be a quantity of order $1/p$ (as indicated by (13)), while the $\hat{x}^\mu$ will be of order $1/\sqrt{p}$ (equation (14)), except for at most a finite number of 'condensed' patterns, for which $\hat{x}^\mu$ may take finite values. The free energy per neuron will also be finite, provided one subtracts the constant $-\nu^2 p/6$, which is the value of the energy (equation (13)), at its minimum.

4.2. Replica-symmetric theory

The free energy per spin is computed using the replica method (Sherrington and Kirkpatrick 1978, Amit *et al* 1987a):

$$g = \lim_{n\to 0} \lim_{N\to\infty} \frac{-1}{\beta n N} [\langle\langle Z^n \rangle\rangle - 1] \tag{16}$$

where

$$Z^n = \mathrm{Tr}_{\{V^\gamma\}} \exp\left(-\beta \sum_\gamma H^\gamma\right) \tag{17}$$

is the partition function for n identical replicas of the system, each labelled by the index $\gamma = 1, \ldots, n$. The average over the quenched variables η is performed in two stages (Amit *et al* 1987a): first over the infinite number of uncondensed patterns, then over a finite set of condensed ones, which retain their discrete nature. The calculation, sketched in appendix 2, is performed in the approximation of replica symmetry. In fact, our interest will inherently be directed to parameter regions in which replica-symmetry-breaking effects are expected to be small. After all, we are trying to stay away from freezing.

The result, after taking the $n \to 0$ limit, is

$$g = -\frac{1}{\beta}\left\langle\left\langle \int \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} \ln 2\cosh\frac{\beta}{2}\left[\sum_\mu \frac{\eta^\mu}{a}\hat{x}^\mu - \Delta + z\sqrt{\frac{2\rho}{\beta}}\right]\right\rangle\right\rangle$$
$$+\frac{1}{2}\left(\Delta - \sum_\mu \hat{x}^\mu\right)(1-2\nu) - \rho\hat{y} + \frac{1}{2}\sum_\mu (\hat{x}^\mu)^2$$
$$+\frac{\alpha}{2\beta}\left[\ln\left(1+\beta\frac{(1-a)}{a}\hat{y}\right) - \frac{\hat{y}+\nu}{[a/\beta(1-a)]+\hat{y}}\right] \tag{18}$$

where the index μ runs over condensed patterns only, and the parameters $\hat{x}^\mu$, Δ, $\hat{y}$, ρ satisfy the saddle-point equations

$$\hat{x}^\lambda + \nu = \frac{1}{2} + \frac{1}{2}\left\langle\left\langle \frac{\eta^\lambda}{a}\int \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} \tanh\frac{\beta}{2}\left[\sum_\mu \frac{\eta^\mu}{a}\hat{x}^\mu - \Delta + z\sqrt{\frac{2\rho}{\beta}}\right]\right\rangle\right\rangle$$
$$\nu = \frac{1}{2} + \frac{1}{2}\left\langle\left\langle \int \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} \tanh\frac{\beta}{2}\left[\sum_\mu \frac{\eta^\mu}{a}\hat{x}^\mu - \Delta + z\sqrt{\frac{2\rho}{\beta}}\right]\right\rangle\right\rangle$$
$$\hat{y} = -\frac{1}{2}\left\langle\left\langle \int \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} \frac{z}{\sqrt{2\beta\rho}} \tanh\frac{\beta}{2}\left[\sum_\mu \frac{\eta^\mu}{a}\hat{x}^\mu - \Delta + z\sqrt{\frac{2\rho}{\beta}}\right]\right\rangle\right\rangle \tag{19}$$
$$\rho = \frac{\alpha}{2\beta}\frac{\hat{y}+\nu}{2\beta([a/\beta(1-a)]+\hat{y})^2}.$$

The parameter Δ is essentially a chemical potential fixing the overall mean rate, which is constrained to be ν, in the large-p limit, as explained above. The parameter ρ is related to the mean fluctuations of the overlaps with the uncondensed patterns.

5. Phase diagram for $\alpha \to 0$

Next, we analyse (19) in the limit $\alpha \to 0$, despite the fact that the limit $p \to \infty$ has been taken. This implies that p increases more slowly than N. In this case the analysis simplifies considerably and its results provide a reasonable approximation to the finite α case.

The character of the solutions of (19) as $\alpha \to 0$ depends on whether or not $\rho \to 0$ in this limit. If $\rho \to 0$, the 'slow' noise, due to the storage of an infinite number of patterns (Amit *et al* 1987a), is negligible compared with the 'fast' noise T. At sufficiently high noise levels T, this is the case. The free energy becomes

$$g = -\frac{1}{\beta}\left\langle\left\langle \ln 2\cosh\frac{\beta}{2}\left[\sum_\mu \frac{\eta^\mu}{a}\hat{x}^\mu - \Delta\right]\right\rangle\right\rangle + \frac{1}{2}\left(\Delta - \sum_\mu \hat{x}^\mu\right)(1-2\nu) + \frac{1}{2}\sum_\mu (\hat{x}^\mu)^2 \tag{20}$$

and the saddle-point equations reduce to

$$\hat{x}^\lambda + \nu = \frac{1}{2} + \frac{1}{2}\left\langle\left\langle \frac{\eta^\lambda}{a}\tanh\frac{\beta}{2}\left[\sum_\mu \frac{\eta^\mu}{a}\hat{x}^\mu - \Delta\right]\right\rangle\right\rangle$$
$$\nu = \frac{1}{2} + \frac{1}{2}\left\langle\left\langle \tanh\frac{\beta}{2}\left[\sum_\mu \frac{\eta^\mu}{a}\hat{x}^\mu - \Delta\right]\right\rangle\right\rangle \tag{21}$$
$$\hat{y} = -\frac{1}{4} + \frac{1}{4}\left\langle\left\langle \tanh^2\frac{\beta}{2}\left[\sum_\mu \frac{\eta^\mu}{a}\hat{x}^\mu - \Delta\right]\right\rangle\right\rangle$$

for which a set of simple solutions emerges.

5.1. The ergodic region

The simplest solution is when no pattern condenses. Then

$$\Delta = 2T \tanh^{-1}(1-2\nu) \qquad \hat{y} = -\nu(1-\nu) \tag{22}$$

and the free energy is

$$g = T[\nu \ln \nu + (1-\nu)\ln(1-\nu)]. \tag{23}$$

The structure is transparent: no pattern is selected and the system is maximally disordered. The mean rate is constrained by the inhibition, and every neuron has a probability ν to be in its firing state, i.e.

$$x = \sum_i \langle V_i \rangle / N = \nu \qquad y = \sum_i \langle V_i \rangle^2 / N = \nu^2. \tag{24}$$

The free energy is just the entropy term associated with this disordered state.

This solution does not exist for all noise levels. It is destabilised at the critical level

$$T = T_c \equiv \frac{(1-a)}{a}\nu(1-\nu). \tag{25}$$

Note that in the noise range in which the ergodic solution exists, i.e. above this critical level, the mean rate does not depend on the noise, while one would expect that, as $T \to \infty$, $\langle V \rangle \to \frac{1}{2}$. This comes about because in the limit $p \to \infty$, T must be much greater than p for the entropy term to dominate and shift x towards $\frac{1}{2}$.

5.2. Retrieval

The next type of solution is a retrieval solution for a single pattern. This solution has one overlap different from ν; we choose it to be $\mu = 1$. Its value is obtained by taking the average over the η in the first two of equations (21), and combining them to find an equation for $\hat{x}^1$ alone:

$$\hat{x}^1 = 2aT\left[\tanh^{-1}\left(1-2\nu+\frac{2a\hat{x}^1}{(1-a)}\right) - \tanh^{-1}(1-2\nu-2\hat{x}^1)\right]. \tag{26}$$

The parameters Δ and $\hat{y}$ are determined by $\hat{x}^1$ to be

$$\begin{aligned} \Delta &= 2T\left[\tanh^{-1}\left(1-2\nu+\frac{2a\hat{x}^1}{(1-a)}\right)\right] \\ \hat{y} &= -\nu(1-\nu)+\frac{a}{(1-a)}(\hat{x}^1)^2. \end{aligned} \tag{27}$$

The mean activity in the network is still ν (it is fixed in the $p \to \infty$ limit), but the activity of those neurons that are active in the first pattern has been enhanced, at the expense of the activity $\bar{x}^1$ of the quiescent ones

$$\begin{aligned} x^1 &\equiv \sum_i \eta_i^1 \langle V_i \rangle / Na = \nu + \hat{x}^1 \\ \bar{x}^1 &\equiv \sum_i (1-\eta_i^1)\langle V_i \rangle / N(1-a) = \nu - \hat{x}^1 \frac{a}{(1-a)}. \end{aligned} \tag{28}$$

Within the two groups, however, there is still maximal disorder. This can be seen from the correlation parameter

$$y \equiv \sum_i \langle V_i \rangle^2 / N = a(x^1)^2 + (1-a)(\bar{x}^1)^2 \tag{29}$$

as explained in § 3. The free energy has once more a transparent form

$$g = Ta[x^1 \ln x^1 + (1-x^1)\ln(1-x^1)] + T(1-a)[\bar{x}^1 \ln \bar{x}^1 + (1-\bar{x}^1)\ln(1-\bar{x}^1)] - (\hat{x}^1)^2/2 \tag{30}$$

as a sum of entropy and energy terms. The entropy is reduced with respect to the disordered solution, because the system is now partitioned into the two subsystems of active and passive neurons, but there is a lower energy (the last term of (30)) due to the emergence of a correlated activity pattern.

These are the *retrieval* solutions. Whenever the system flows into the basin of attraction of one such solution, we shall say that the corresponding pattern has been associatively retrieved by the stimulus, which is represented in the initial configuration. Retrieval solutions exist only below a critical noise level T_R (see figure 1). At T_R non-trivial solutions of (26) first appear, and they correspond to local minima of the free energy. At a slightly lower level $T_M > T_c$ they become the absolute minima of the free energy. They are destabilised at a level $T_G < T_c$. In the interval T_G-T_R (whose extension depends on a and ν; see figure 1) the activity x^1 of the 'on' neurons is greater than the mean activity ν but smaller than ν/a (when $x^1 \to \nu/a$, $\bar{x}^1$, which has to be positive, approaches 0). Figure 1 shows, for sample values of a and ν, that x^1

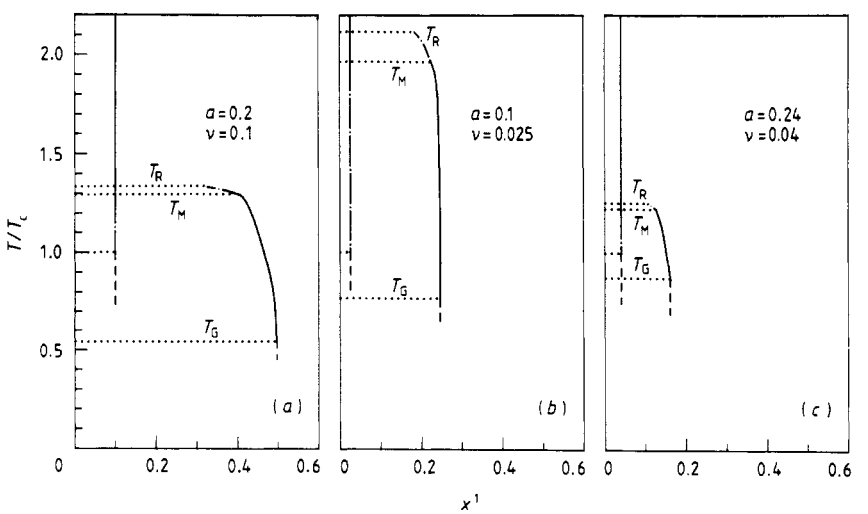

Figure 1. Activity levels in the disordered and retrieval solutions as a function of the noise level. The analytic result for $x^1(T)$, in the limit $a \to 0$, is plotted for three different choices of the parameters a and ν. Noise levels T are scaled by $T_c(a, \nu)$. In the disordered solutions $x^1(T) = x(T) = \nu$ is represented by vertical lines, while in the retrieval solutions there is some mild dependence on the noise level. On the full lines the retrieval solutions are the absolute minima of the free energy. On the chain lines they are local minima. Retrieval solutions appear at the critical level $T_R(a, \nu)$ and become global minima at $T_M(a, \nu)$. Broken lines indicate the freezing occurring below the spin-glass transitions at T_c (disordered solution) and at T_G (retrieval solution).

does not depend strongly on the noise level. The equations determining T_R, T_M and T_G are given in appendix 3.

5.3. Spin-glass effects

At T_G the retrieval phases undergo a *spin-glass* transition. The disordered phase also undergoes a spin-glass transition, but at the higher noise level T_c. The valleys of the free-energy landscape split into a multitude of sub-valleys, representing different microscopic realisations of states characterised by the same order parameters. Formally, $\rho \neq 0$, and (29) is no longer satisfied. This means that some of the active neurons start firing at higher rates than others, until at very low noise levels all neurons 'freeze', either in their quiescent or in their active state. We are not interested in this low-noise regime, and so this analysis will stop here.

5.4. Spurious states

There are also other solutions of the saddle-point equations, besides retrieval and disordered ones. For example, one can identify symmetric n-mixture solutions, which represent equal mixtures of n activity patterns (see Amit *et al* (1985a) for the analogous type of solution in the Hopfield model). They appear only below the noise level at which retrieval solutions exist. Compact expressions for the equations satisfied by this family of solutions can be derived, and are reported in appendix 4. In general, we shall consider all solutions, with more than a single condensed pattern, as *spurious* states. Initial configurations that happen to be in their basin of attraction, will fail to produce unambiguous retrieval of a pattern. However, these solutions are not always stable. A calculation of the eigenvalues controlling their stability shows that the sign of these eigenvalues depends on a, ν and T.

In those cases in which such solutions are stable local minima of the free energy, their importance in the dynamics of the network depends on the size of their basins of attraction, and on the barriers separating them from deeper local minima. Although we are not able to evaluate these features analytically, simulations indicate that the 'spurious' basins, when they exist, are always limited in extent and are surrounded by low free-energy barriers. This is a favourable situation, since the retrieval states are the global minima of the free energy in a wide noise range, while spurious states are all but irrelevant to the dynamics. These features depend, though, on the detailed form (equation (6)) we chose for the inhibition. An inhibition depending only on the mean excitatory activity x, would make the system more susceptible to spurious states.

5.5. Overall picture for $\alpha = 0$

The conclusion is that at high levels of noise the network is disordered (ergodic), characterised by a mean activity ν, and its state is devoid of any information. In an intermediate noise window ergodicity is broken, and the dynamics is dominated by attractors with a simple structure. The system tends to choose one of the stored activity patterns (the one closest to the initial activity distribution), and two mean firing frequencies emerge, one for the neurons active in that pattern and one for the passive ones. In this window the network behaves as envisaged in the scenario of § 3. Finally, if the fast, 'thermal' noise is too low, each single neurons is frozen in one of its two

states, and the system is clamped in one of the many possible spin-glass configurations, producing very high spike rates.

6. Finite-α and finite-p corrections

The results obtained in the previous section were derived in the limit $p \to \infty$ and $\alpha \equiv p/N \to 0$. Before comparing them with the behaviour of actual systems of finite size, it is important to evaluate the effects of relaxing these two simplifying assumptions. Assuming that the limit $N \to \infty$ is a good approximation, one can distinguish two cases, and treat them separately: (1) Extensive loading, i.e. $p = \alpha N$, with finite α. (2) Low loading, i.e. p finite as $N \to \infty$.

6.1. The effects of extensive storage

Equations (19) are valid for arbitrary α, provided replica symmetry holds, and they can be studied numerically to find, for example, the shifts in the order parameters and in the critical noise levels which occur for $\alpha \neq 0$. One can also go back to the original replica theory and study the onset of replica-symmetry breaking. Here we shall limit ourselves to the appearance of non-uniform firing rates, at finite α. In particular, we shall concentrate on the retrieval phase. In other words, at finite α, during the retrieval of a pattern, the neurons which should be active in that pattern have different rates. The same is true for passive ones. These are spin-glass effects, which manifest themselves in the deviation of y, the analogue of the Edwards-Anderson order parameter, from its uncorrelated value, given by (29).

Let us consider a retrieval solution in the noise range in which it is stable, and expand (19) for small α. This way we obtain Taylor expansions for $\hat{x}^1$, Δ and $\hat{y}$, whose zeroth-order terms are the ones previously discussed. The terms of $O(\alpha)$ violate (29), which we write in the form

$$y = a[(x^1)^2 + (\sigma^1)^2] + (1-a)[(\bar{x}^1)^2 + (\bar{\sigma}^1)^2] \tag{31}$$

where x^1 and $\bar{x}^1$ are the mean activities of active and passive neurons, respectively, correct to $O(\alpha)$. We have denoted the remaining corrections by σ^1 and $\bar{\sigma}^1$. They are quantities of $O(\sqrt{\alpha})$. Note that the correction is interpreted as a sum of two terms. One, proportional to a, is interpreted as corresponding to the contribution of the active neurons. The other, proportional to $(1-a)$, as the contribution of the passive neurons. What this implies is that not all neurons active in the retrieved pattern will have the same probability to fire, nor will the passive ones. The firing rate distribution, which for $\alpha = 0$ consisted of two delta functions centred at x^1 and $\bar{x}^1$ respectively, now consists of two broader peaks. If α is small, the two peaks can be approximated by Gaussian distributions, whose widths, $2\sigma^1$ and $2\bar{\sigma}^1$, are proportional to $\sqrt{\alpha}$. Explicitly one finds, to this order

$$\sigma^1 = x^1(1-x^1)\sqrt{2\beta\rho} = \frac{x^1(1-x^1)}{[a/\beta(1-a)] + y - \nu}\sqrt{y\alpha} \tag{32}$$

for the half-width of the peak of active neurons. The same expression, with $\bar{x}^1$ replacing x^1, gives the width of the rate distribution for passive ones.

These are rather wide peaks, especially in the lower part of the allowed noise range. In fact, the denominator in (32) approaches zero as $T \to T_G$, and the width of the peaks diverges. The lack of uniformity in the firing rates, that results from the 'slow' noise generated by the random overlaps with the increasing number of stored patterns, is a realistic feature of the model. However, too broad distributions imply that the firing rates of the most active neurons approach the maximal rate. This, in practice, is the most stringent constraint on the loading α of the network. The larger α, the wider the distributions and the higher the fraction of active neurons that fires at high rates.

6.2. The model at finite p

It is straightforward to write the mean-field equations for the case when p remains finite as $N \to \infty$. They are

$$x^\mu = \frac{1}{2} + \frac{1}{2}\left\langle\left\langle \frac{\eta^\mu}{a} \tanh \frac{\beta}{2} \sum_\lambda \frac{\eta^\lambda}{a}\left[x^\lambda - \frac{(\Sigma_\rho x^\rho)^2}{\nu p^2}\right]\right\rangle\right\rangle. \tag{33}$$

The problem is that the solutions to these equations have no simple expression for finite p. As p becomes very large, it is possible to substitute smooth distributions for the parameters depending on p discrete quantities, and that is why, in fact, the present model is simple to analyse only in the limit $p \to \infty$.

Some understanding of the effects of a large but finite p can be obtained by expanding (33) in $1/p$. Consider, for example, a retrieval solution, in which one of the overlaps, say x^1, has a higher value than the others, which are equal. To perform the average over the η^μ for $\mu > 1$, one approximates the binomial distribution of $\Sigma_{\mu=2}^{p} \eta^\mu$ by a Gaussian distribution of mean $a(p-1)$ and square width $a(1-a)(p-1)$, and then proceeds to solve for the two unknowns x^1 and $x^{\mu>1}$, both taken as Taylor series in $1/p$. This way one can calculate the first-order correction to the overall activity of the network, which leads to

$$x = \nu + \frac{1}{p}\left[T \ln \frac{1-\bar{x}^1}{\bar{x}^1} + \nu - x^1\right] + O(1/p^2). \tag{34}$$

In the disordered state one finds

$$x = \nu + \frac{T}{p} \ln \frac{1-\nu}{\nu} + O(1/p^2) \tag{35}$$

which can also be found by minimising the free energy obtained by adding to the energy of uncorrelated configurations (equation (13)) the entropy of the disordered state (equation (23))

$$g = p(-\tfrac{1}{2}x^2 + (1/3\nu)x^3) + T[x \ln x + (1-x)\ln(1-x)]. \tag{36}$$

Equation (35) shows that for finite p the overall mean activity does depend on the noise level, and grows smoothly with the noise. One could also expand in the inverse parameter, p/T, to find that as $T \to \infty$, $x \to \frac{1}{2}$. Corrections can then be computed in the framework of a 'high-temperature' expansion.

At finite p there are fluctuations in the local field acting upon single neurons, and the activity distribution consists once more of a pair of broad peaks. The half-width

of the peak for active neurons can be found, from the same expansion in $1/p$, to be

$$\sigma^1 = x^1(1-x^1)\left(\frac{1-a}{ap}\right)^{1/2} \ln\frac{1-\bar{x}^1}{\bar{x}^1} + O(1/p). \tag{37}$$

This is again a rather strong effect: considering specific values for x^1, $\bar{x}^1$ and a, one realises that p has to be very large to suppress the fluctuations in the firing rates.

7. Numerical simulations

The model described by the Hamiltonian of (7) has been simulated for a variety of sizes N, storage loads p, sparse coding parameters a, inverse inhibition strengths ν and noise levels T. We have chosen $0 < a < \frac{1}{2}$ and $0 < \nu < a$. Typical values for N and p were $N \approx 5000$–$50\,000$, $p \approx 50$–200. The neurons were updated sequentially with the probability given by (2). A 'time cycle' was taken to be the time required for updating the complete set of neurons. The initial configuration of the network was either chosen at random, assigning an equal probability to be 'on' for each neuron; or it was chosen to be correlated with one or more of the stored patterns, by assigning different probabilities to be initially 'on' to neurons that were active or passive in those patterns.

At high noise levels the system is ergodic: it wanders freely in configuration space. The overall mean activity is regulated by the parameter ν, and for sufficiently large N and p, and moderate T, it approaches the value ν itself. What is meant by 'high' noise level depends on a and ν; to get a rough estimate, it is useful to compare T with the critical level T_c of (25). When T is very large (of the order of p), the mean activity approaches $\frac{1}{2}$. The correlation parameter y approaches the square of the mean activity when the noise level is that high, indicating a virtual absence of correlations. The network relaxes to its mean *spatial* activity equilibrium value with 1-2 time cycles, from any initial configuration. This then remains as its mean temporal rate of activity.

At very low noise levels, $T \ll T_c$, the evolution of the network shows a marked dependence on the initial configuration. Once the mean activity has approached, in a few time cycles, its asymptotic value, most neurons stop changing state. Some fraction of them is always 'on', and the rest are 'off', and the state chosen by each neuron is strongly correlated with its state in the initial configuration. A small number of neurons still changes state now and then, but this number eventually vanishes if the noise level is set to zero.

7.1. Retrieval

The interesting behaviour occurs at intermediate noise levels, of the order of T_c. Retrieval states appear, and play the central role in the dynamics of the systems. When T is slightly above T_c, individual neurons keep changing their states under the influence of fast noise, but the overlaps with the embedded patterns stabilise in a few time cycles at definite values, around which they fluctuate mildly. Retrieval occurs when one of these overlaps is high and the others are low, as in figure 2(*b*). Note, in the example in the figure, that the high overlap fluctuates around 0.15, which is rather close to $\nu/a = 0.2$, while the other overlaps fluctuate around 0.04, close to $\nu = 0.05$.

Starting from the high-noise region characterised by full ergodicity, and performing simulations with decreasing levels of noise, one finds a rather abrupt change of

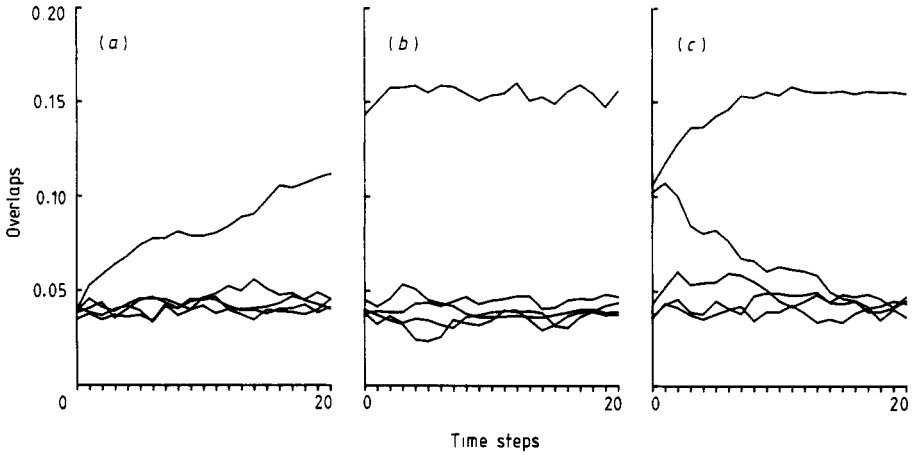

Figure 2. The overlaps (activity levels) of five patterns selected out of a total of 180, in three simulations performed on a system of 16 000 neurons, with $a = 0.25$, $\nu = 0.05$ and $T = 0.16$ ($T/T_c = 1.12$). (*a*) Random initial configuration with mean activity ν; (*b*) initial configuration ordered to overlap with a single pattern; (*c*) initial configuration with equal overlaps with two patterns as in a symmetric 2-mixture. See text.

behaviour at a critical level $T \simeq 1.5\text{-}2\ T_c$ (the precise value depends on a and ν). Strong initial correlations with one of the patterns are no longer washed away during the evolution. In fact, most initial configurations are attracted toward one of the retrieval states. The value of the higher overlap, corresponding to the retrieved pattern, is always of the order of ν/a, as in the example above, while the other overlaps have roughly the same value as the mean overall activity. If one of the patterns has a significantly higher overlap with the initial configuration, its retrieval state will be the one selected by the dynamics. If the initial configuration is uncorrelated with any of the patterns, the network sometimes continues to evolve in a completely disordered fashion, with roughly equal overlaps, and a mean activity regulated by the strength of the inhibition. In other cases a fluctuation increases the overlap with a specific pattern, which is then enhanced until it stabilises at the value typical of the retrieval state.

Figure 2 exemplifies this behaviour. Note that this phenomenon takes place despite the fact that the disordered state is a *stable* solution in the thermodynamic limit. Still, in figure 2(*a*), the fluctuations due to the finite size of the network are strong enough to make the system flow towards a randomly selected retrieval state, which corresponds to a global minimum of the free energy. This behaviour indicates a strong reduction of the free-energy barriers, which are naively expected to be of $O(N)$.

In figure 2(*c*) the initial configuration is a symmetric mixture of two patterns. In the thermodynamic limit, for the parameters given, this solution is *unstable.* In fact, the system flows, faster than in the case of figure 2(*a*), to the retrieval state associated with one of the two privileged patterns.

Inspecting the distribution of firing rates among individual neurons, one finds a marked non-uniformity (figure 3). For relatively small networks, a fraction of the neurons is essentially always 'on'. To produce firing distributions characterised by two well defined peaks around a mean rate for active and passive neurons, one has to simulate large networks, so that both p and N/p are large. The reason for this lack of uniformity is the slow noise discussed above. The expressions derived in the

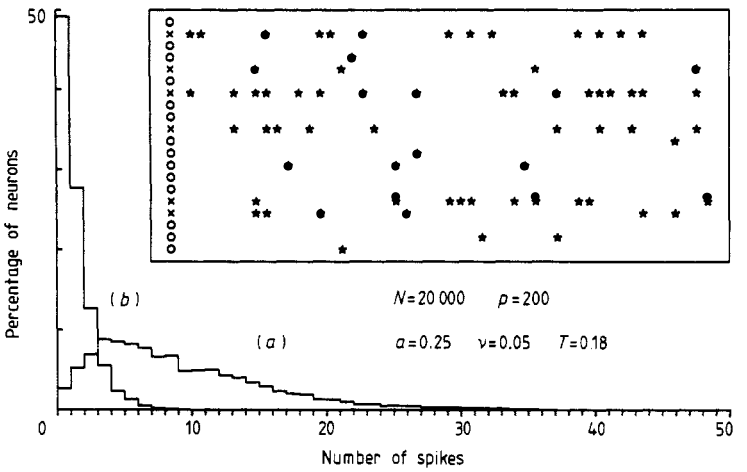

Figure 3. Result of a sample simulation lasting 50 time cycles, with the parameters indicated. Initial configuration ordered according to pattern 1. Distribution of the number of spikes emitted by (*a*) active and (*b*) passive neurons in pattern 1. Inset: spikes emitted by 20 randomly chosen neurons over the 50 time cycles. The first column indicates whether the neuron was active (×) or passive (○) in pattern 1.

preceding section (equations (32) and (37)) account quantitatively for the half-width of the peaks.

As the noise level is decreased below T_c, there is no abrupt transition to the very low-noise behaviour. Rather, at first the distribution of firing rates gradually flattens, until it begins to show a peak corresponding to neurons active at every time cycle, and another peak of neurons that freeze in the quiescent state. This is accompanied by a time evolution of the overlaps which shows an increasing dependence on their initial value, and the $p-1$ 'low' overlaps stabilise at widely different values, depending on the mutual correlations of the stored patterns among themselves and with the initial configuration.

Figure 3 also shows the actual dynamical evolution of 20 neurons randomly selected from a network of 20 000. The spiking activity appears to proceed in a disordered fashion, except that neurons that are 'on' in the pattern are more active than those that are 'off': the former fire between 4 and 17 times in the 50 cycles, while the latter at most (in one case out of 14) emit 3 spikes. Note that if we assume a time cycle of 2.5 ms, the simulation extends for 125 ms, and a neuron that is active 10 times (for example, the second 'on' neuron from the bottom of the picture) is firing at a rate of $80\ s^{-1}$. Most 'off' neurons appear to fire either never or just once, so that their mean firing rate is of a few spikes per second.

8. Discussion and conclusions

The chief aim of the present work has been to address the problem of low spiking rates. The model proposed above illustrates a possible solution. The main idea is that to stabilise low levels of activity for the individual neurons, in a way that does not depend on a fine tuning of the model's parameters, one has to have an inhibitory

contribution to the input of the neurons, that dominates the excitatory contribution when the overall activity level becomes too high. The particular way in which this idea is realised in the present model, is by taking the excitation to grow linearly, as usual, with the mean activity, while the inhibition grows *quadratically*. This is just one possible choice that reproduces the effect. A different choice has been made, for example, in a model recently proposed (Rubin and Sompolinsky 1988) that addresses the same problem. There the inhibition is taken to be linear, while the excitation essentially does not grow with the mean activity. This is obtained by assuming a *negative* threshold for excitatory neurons, which prompts them to fire even in the absence of any input, while as soon as the network is activated, the linear inhibitory feedback takes control of the firing rates.

We argue that the quadratic inhibition can represent in an effective way both the different dynamical characteristics of inhibitory neurons and the non-linear operation of inhibitory synapses. These aspects were absent in earlier simple models. We propose a way of incorporating them that solves the low-rates problem, while the network remains simple to understand.

The particular choice of couplings, written in terms of a set of stored patterns, is made in order to keep the model in a familiar conceptual framework, and to ensure that it be amenable to a comprehensive analytical study. There are two disadvantages: the structure of the inhibitory couplings looks rather artificial, and the model functions properly only with very large networks. Neither is an essential flaw. Consider, for example, the requirement that p be large. Its role is to prevent excessive rates by some neurons, as indicated by the width of the activity distribution (equation (37)). The hyperactive neurons are those that, in the random assignment of quenched patterns, 'belong' to fewer patterns, i.e. $\Sigma_\mu \eta_i^\mu < pa$. The average inhibition acting upon these neurons is lower, and the reduction outweighs the corresponding reduction in the excitation. The effect is a fluctuation that becomes negligible as $p \to \infty$, but for moderate p values the highest firing rates are indeed rather high (see figure 3, where 0.5% of the active neurons fire more than 30 out of 50 time cycles). However, in a more comprehensive model that encompasses learning, one may conceive a compensatory effect that gradually strengthens the couplings of intensely firing neurons, thereby lowering their mean activity through enhanced inhibition. This would remove the requirement of very large p, which might thus disappear naturally when enlarging the scope of the theory.

Clearly, any reformulation of the standard Hopfield model will have to deal with a host of questions which have already found answers in the previous context. Among those are issues of structured data sets, of retrieval of temporal sequences etc. This applies to the present model as well and points to a great deal of work that has to be done.

Acknowledgments

We are indebted to Professor M Abeles for impressing the gravity of the rates problem upon us and to Professor H Sompolinsky, E Gardner, L Abbott, B Derrida, M Mezard, J J Hopfield and N Rubin for fruitful discussions and remarks. The intimate contact with biologists we owe to the Institute for Advanced Studies at the Hebrew University in Jerusalem. The work of DJA has been supported by a grant from the US—Israel Binational Science Foundation.

Appendix 1

We derive here equation (14). For a fixed configuration

$$\left\langle\left\langle \sum_{\mu} (x^{\mu} - x)^2 \right\rangle\right\rangle = \frac{1}{N^2}\left\langle\left\langle \sum_{\mu}\sum_{i,j}\left(\frac{\eta_i^{\mu}}{a} - 1\right)\left(\frac{\eta_j^{\mu}}{a} - 1\right) V_i V_j \right\rangle\right\rangle$$
$$= \frac{1}{N^2}\left\langle\left\langle \sum_{\mu}\sum_{i}\left(\frac{\eta_i^{\mu}}{a} - 1\right)^2 V_i^2 \right\rangle\right\rangle = \alpha\frac{1-a}{a}x. \tag{A1.1}$$

Appendix 2

We indicate here the main lines of the calculation used to derive the expression (18) from (16). One starts by introducing the subtracted overlap parameters through delta functions, so that

$$Z^n = \left(\frac{Na}{2\pi}\right)^{pn} \int \mathrm{d}t^{\mu\gamma}\,\mathrm{d}\hat{x}^{\mu\gamma}\,\mathrm{Tr}_{\{V^{\gamma}\}} \exp \mathrm{i} \sum_{\mu,\gamma} t^{\mu\gamma}\left(Na\hat{x}^{\mu\gamma} + Na\nu - \sum_i \eta_i^{\mu} V_i^{\gamma}\right)$$
$$\times \exp \beta N \sum_{\gamma}\left[\frac{\Sigma_{\mu}(\hat{x}^{\mu\gamma}+\nu)^2}{2} - \frac{[\Sigma_{\mu}(\hat{x}^{\mu\gamma}+\nu)]^3}{3\nu p^2} - \frac{[\Sigma_{\mu}(\hat{x}^{\mu\gamma}+\nu)]}{2Na}\right] \tag{A2.1}$$

where γ is the replica index and $t^{\mu\gamma}$ is the Lagrange multiplier which imposes the definitions of the order parameters.

Then one assumes that only a finite number s of patterns can condense, i.e. have a finite subtracted overlap $\hat{x}^{\mu}$ as $N \to \infty$, and averages 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

$$\hat{y}^{\gamma\delta} = \frac{1}{N}\sum_i V_i^{\gamma} V_i^{\delta} - \nu \qquad \gamma \neq \delta$$
$$\hat{x}^{\gamma} = \hat{y}^{\gamma\gamma} = \frac{1}{N}\sum_i V_i^{\gamma} - \nu \tag{A2.2}$$

via delta functions. Denoting by Y the matrix with elements

$$Y_{\gamma\delta} = \hat{y}^{\gamma\delta} + \nu, \ \gamma \neq \delta \qquad Y_{\gamma\gamma} = \hat{x}^{\gamma} + \nu \tag{A2.3}$$

one obtains

$$\langle\langle Z^n \rangle\rangle = \left(\frac{Na}{2\pi}\right)^{pn}\left(\frac{N}{2\pi}\right)^{[n(n+1)/2]}\left(\frac{2\pi}{Na(1-a)}\right)^{[n(p-s)/2]} \int \mathrm{d}\hat{x}^{\mu\gamma}\,\mathrm{d}t^{\sigma\gamma}\,\mathrm{d}t^{\gamma}\,\mathrm{d}\hat{x}^{\gamma}\,\mathrm{d}y^{\gamma\delta}\,\mathrm{d}r^{\gamma\delta}$$
$$\times \exp \beta N \sum_{\gamma}\left[\frac{\Sigma_{\mu}(\hat{x}^{\mu\gamma}+\nu)^2}{2} - \frac{[\Sigma_{\mu}(\hat{x}^{\mu\gamma}+\nu)]^3}{3\nu p^2} - \frac{[\Sigma_{\mu}(\hat{x}^{\mu\gamma}+\nu)]}{2Na}\right]$$
$$\times \exp -\left[\frac{p-s}{2}\mathrm{Tr}_{\gamma} \ln Y + \frac{Na}{2(1-a)}\sum_{\gamma,\delta,\lambda}(\hat{x}^{\lambda\gamma} - \hat{x}^{\gamma}) Y^{-1}_{\gamma\delta}(\hat{x}^{\lambda\delta} - \hat{x}^{\delta})\right]$$
$$\times \mathrm{Tr}_{\{V^{\gamma}\}}\left\langle\left\langle \exp \mathrm{i}\left[\sum_{\sigma,\gamma} t^{\sigma\gamma}\left(Na\hat{x}^{\sigma\gamma} + Na\nu - \sum_i \eta_i^{\mu} V_i^{\gamma}\right)\right]\right.\right.$$
$$\left.\left.\times \exp \mathrm{i}\left[\sum_{\gamma} t^{\gamma}\left(N\hat{x}^{\gamma} + N\nu - \sum_i V_i^{\gamma}\right) + \sum_{(\gamma,\delta)} r^{\gamma\delta}\left(Ny^{\gamma\delta} + N\nu - \sum_i V_i^{\gamma} V_i^{\delta}\right)\right]\right\rangle\right\rangle \tag{A2.4}$$

(the index σ runs over the condensed patterns only, the index λ over the uncondensed ones).

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

$$\chi^{\gamma}=\frac{1}{p}\sum_{\mu}\hat{x}^{\mu\gamma}-\hat{x}^{\gamma} \tag{A2.5}$$

to be able to perform the integrals over $\hat{x}^{\mu\gamma}$ for $\mu > s$. Considering that the χ and $\hat{x}$ are fluctuations of order $O(1/\sqrt{p})$ and $O(1/p)$ respectively, one can also integrate over them. Then $\langle\langle Z^n\rangle\rangle$ can be evaluated at the saddle point as $\exp-(n\beta Ng)$ where g is given by terms which remain finite as p, $N\to\infty$:

$$\begin{aligned} g=&-\frac{1}{n}\sum_{\gamma}\sum_{\sigma}\frac{(\hat{x}^{\sigma\gamma})^2}{2}+\frac{\alpha}{2\beta n}\operatorname{Tr}_{\gamma}\ln\left(1-\beta\frac{1-a}{a}Y\right)\\ &-\frac{\mathrm{i}}{\beta n}\left[\sum_{\sigma,\gamma}t^{\sigma\gamma}(a\hat{x}^{\sigma\gamma}+a\nu)+\sum_{\gamma}t^{\gamma}\nu+\sum_{(\gamma,\delta)}r^{\gamma\delta}(\hat{y}^{\gamma\delta}+\nu)\right]\\ &-\frac{1}{\beta n}\left\langle\left\langle\ln\operatorname{Tr}_{\{V^{\gamma}\}}\exp-\mathrm{i}\left[\sum_{\sigma,\gamma}t^{\sigma\gamma}\eta^{\sigma}V^{\gamma}+\sum_{\gamma}t^{\gamma}V^{\gamma}+\sum_{(\gamma,\delta)}r^{\gamma\delta}V^{\gamma}V^{\delta}\right]\right\rangle\right\rangle. \end{aligned} \tag{A2.6}$$

One then uses a replica-symmetry ansatz and takes the limit $n\to 0$. Using the saddle point equations for the $\hat{x}^{\sigma}$ to eliminate the t^{σ}, and writing $\rho=-\mathrm{i}r/\beta$, $\Delta=\mathrm{i}(t-r)/\beta$ one arrives at (18) and (19).

Appendix 3

The critical noise levels T_{R}, T_{M} and T_{G} are obtained (in the $\alpha\to 0$ limit) by solving a system of two equations in the variables $\hat{x}^1$ and T. The first equation is the same in all three cases, and it is the relation that determines the overlap of the selected pattern as a function of the noise levels, i.e. equation (26):

$$q(\hat{x}^1)=\hat{x}^1-2aT\left[\tanh^{-1}\left(1-2\nu+\frac{2a\hat{x}^1}{(1-a)}\right)-\tanh^{-1}(1-2\nu-2\hat{x}^2)\right]=0 \tag{A3.1}$$

where we have denoted with q the quantity that must vanish in order to satisfy the relation.

The second equation of the system that yields the critical noise level T_{R}, at which non-zero solutions of (A3.1) first appear, is obtained by requiring the vanishing of the derivative

$$q'(\hat{x}^1)=0 \tag{A3.2}$$

which is where the local minimum disappears.

To determine the critical level T_M, at which the retrieval state becomes the global minimum of the free energy, we add as second equation the requirement that the free energies of the disordered and retrieval states be equal,

$$g_{ds}(T) = g_{rs}(\hat{x}^1(T), T) \tag{A3.3}$$

where g_{ds} is given by (23), and g_{rs} by (30).

Finally, to determine T_G, the level at which, in the $\alpha \to 0$ limit, the retrieval state undergoes a spin-glass transition, one imposes the divergence of the spin-glass susceptibility. This coincides with the vanishing of the denominator in the last of equations (19), and it can be written, by using the expression for y in the retrieval state (27) and the value of T_c (equation (25)) as

$$T - T_c + (\hat{x}^1)^2 = 0. \tag{A3.4}$$

Appendix 4

In a symmetric solution, n patterns condense, and their overlap parameters have the same value. The mean activity of an individual neuron depends on whether the neuron is active in all the n condensed patterns, or just in $n-1$ of them, or in $n-2$, and so on up to neurons that are not active in any of those patterns. One can decompose the quenched averages in (21) as sums of contributions from all these classes of neurons. Denoting by

$$x_k = \tfrac{1}{2} + \tfrac{1}{2}\tanh(\beta/2)(k\hat{x}^\sigma/a - \Delta) \tag{A4.1}$$

the mean activity of neurons active in k of the n patterns, ($\nu + \hat{x}^\sigma$ is the overlap with any of the n patterns) one has to solve the constraints for the overall mean activity and for the overlap with each of the patterns in the mixture, i.e. respectively,

$$\begin{aligned} \nu &= \sum_{k=0}^{n} \binom{n}{k} a^k (1-a)^{n-k} x_k \\ \nu + \hat{x}^\sigma &= \sum_{k=0}^{n-1} \binom{n-1}{k} a^k (1-a)^{n-1-k} x_{k+1}. \end{aligned} \tag{A4.2}$$

Equations (49) and (50) can be solved for the $n+2$ unknowns x_k, $\hat{x}^\sigma$, Δ, and non-trivial solutions exist in certain regions of parameter space.

In the intermediate noise range in which the symmetric mixtures exist they are not always stable. The stability of these solutions is determined by the eigenvalue which controls the flow into a retrieval state. This eigenvalue is denoted by $\lambda_{n\to 1}$. It can be computed from the matrix of the second derivatives of the free energy. It is found to be

$$\lambda_{n\to 1} = 1 - \frac{T_c}{T} \sum_{k=0}^{n-2} \binom{n-2}{k} a^k (1-a)^{n-2-k} \frac{x_{k+1}(1-x_{k+1})}{\nu(1-\nu)} \tag{A4.3}$$

and its sign depends on a, ν and T.

References

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

Amit D J 1987 *Proc. Tuebingen Symp. on the Physics of Structure Formation, 1986* ed W Guttinger and G Dangelmayr (Berlin: Springer) p 2

—— 1989 *Modelling Brain Function* (Cambridge: Cambridge University Press) in press

Amit D J, Gutfreund H and Sompolinsky H 1985a *Phys. Rev.* A **32** 1007

—— 1985b *Phys. Rev. Lett.* **55** 1530

—— 1987a *Ann. Phys., NY* **173** 30

—— 1987b *Phys. Rev.* A **35** 2293

Amit D J and Treves A 1989 *Associative memory neural networks with low temporal spiking rates* Rome University preprint No 640

Anderson R A and Mountcastle V B 1983 *J. Neurosci.* **3** 532

Bower J 1988 Private communication

Buhmann J, Divko R and Schulten K 1988 *Associative memory with high information content* Munich Technical University preprint

Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167

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

Fatt P and Katz B 1953 *J. Physiol.* **121** 374

Glauber R J 1963 *J. Math. Phys.* **4** 294

Goldberg M E and Bruce C J 1985 *Vision Res.* **25** 471

Gutfreund H 1987 unpublished

Hebb D O 1949 *The Organization of Behaviour* (New York: Wiley)

Hopfield J J 1982 *Proc. Natl Acad. Sci., USA* **79** 2554

—— 1984 *Proc. Natl Acad. Sci., USA* **81** 3088

Miyashita Y and Chang H S 1988 *Nature* **331** 68

Peretto P and Niez J J 1986 *IEEE Transactions SMC* **16** 73

Rubin N and Sompolinsky H 1989 *Neural networks with low local firing rates* Racah Institute preprint

Segev I and Parnas I 1983 *Biophys. J.* **41** 41

Segev I and Rall W 1987 *Synaptic Function* ed G Edelman *et al* (New York: Wiley) p 605

Sherrington D and Kirkpatrick S 1978 *Phys. Rev.* B **17** 4384

Shinomoto S 1987 *Biol. Cybern.* **57** 197

Sompolinsky H 1987 *Proc. Heidelberg Coll. on Glassy Dynamics, 1986* ed I Morgenster and I L van Hemmen (Berlin: Springer) p 485

Sur M, Wall J T and Kaas J H 1984 *J. Neurophysiol.* **51** 724

Tsodyks M V and Feigelman M V 1988 *Europhys. Lett.* **6** 101

Willshaw D J, Buneman O P and Longuet-Higgins H C 1969 *Nature* **222** 960