Willshaw model: Associative memory with sparse coding and low firing rates

D. Golomb, N. Rubin, and H. Sompolinsky
Racah Institute of Physics, The Hebrew University of Jerusalem, Jerusalem, 91904 Israel
(Received 8 June 1989)

The Willshaw model of associative memory, implemented in a fully connected network with stochastic asynchronous dynamics, is studied. In addition to Willshaw's learning rule, the network contains uniform synaptic inhibition, of relative strength K, and negative neural threshold $-\theta, \theta > 0$. The P stored memories are sparsely coded. The total number of *on* bits in each memory is Nf, where f is much smaller than 1 but much larger than $\ln N/N$. Mean-field theory of the system is solved in the limit where $C \equiv \exp(-f^2 P)$ is finite. Memory states are stable (at zero temperature), as long as $C > h_0 \equiv K - 1 + \theta$ and $h_0 > 0$. When $C < h_0$ or $h_0 < 0$, P retrieval phases, highly correlated with the memory states, exist. These phases are only partially frozen at low temperature, so that the full memories can be retrieved from them by averaging over the dynamic fluctuations of the neural activity. In particular, when $h_0 < 0$ the retrieval phases at low temperatures correspond to freezing of most of the population in a quiescent state while the rest are active with a time average that can be significantly smaller than the saturation level. These features resemble, to some extent, the observed patterns of neural activity in the cortex, in experiments of short-term memory tasks. The maximal value of P for which stable retrieval phases exist, scales as $f^{-3}/|\ln f|$ for $f \gg 1/\ln N$, and as $f^{-2}\ln(Nf/|\ln f|)$ for $f \ll 1/\ln N$. Numerical simulations of the model with $N = 1000$ and $f = 0.04$ are presented. We also discuss the possible realization of the model in a biologically plausible architecture, where the inhibition is provided by special *inhibitory* neurons.

I. INTRODUCTION

A. Neutral network models and biology

Recent interest in neural network models stems partly from their superficial resemblance to biological neural systems. However, relatively little progress has been made so far, in making a more concrete connection between the simplified models and the known facts concerning the architecture and operation of real neural systems. Here we focus on two problems that arise when simple recurrent network models of associative memory are compared with neural systems in the cortex.

In most of the simple neural network models, excitation and inhibition play identical roles. The model learning rules store information by modifications of both excitatory and inhibitory synapses. Although experimental support of Hebb-like synaptic plasticity has been accumulated, most of the available evidence concerns excitatory synapses, and not inhibitory ones.[1-4] Although this does not rule out the existence of similar changes in inhibitory synapses, it does motivate the study of models in which information is stored only in excitatory synapses.

A somewhat related issue is that neutrons in the cortex are believed to be either excitatory, namely, neurons that send out through their synaptic junctions only excitatory signals, or inhibitory. Although exceptions to this, so-called Dale's law, are known in other nervous systems, there is an accumulating evidence that Dale's law, in the form defined here, does hold in the cortex.[5,6] This suggests that the two types of neurons may have different roles in the computational function of the system. In contrast, most simple models consists of neurons that send out both excitatory and inhibitory signals.

The second difficulty is concerned with the neural firing activity. The main paradigm of computation in fully connected recurrent networks (i.e., networks with strong internal feedback) has been *computation by convergence to an attractor.*[7-9] The outcome of the computation is encoded in the persistent levels of neural activities, which are interpreted as representing firing rates of biological neurons. Recent studies considered also network models with stable limit cycles.[10,11] In these persistent patterns of neural activities, a significant fraction of the population exhibits activity levels that are close to their saturation value. For biology this implies that occasional firing at frequencies close to saturation rates should be observed during the execution of computational processes. These *bursts* of action potentials should last at least for several microscopic time units, to allow for a meaningful retrieval of information. However, recordings of neural activity in the cortex during various short-term memory tasks show that the firing rates of individual neurons fluctuate in time with a time average which is very low compared to the saturation rates.[12-14] Typically, the activity of a large fraction of the recorded neurons remains at the extremely low background level of about 3–5 spikes per second. Other neurons do exhibit enhancement in their activity level which may persist for 10 sec or more. However, the enhanced rates are still relatively low, with a long-time average that is in the range of 20–50 spikes per second. Bursts of activity with frequencies in the range of several hundred Hz, that last for at least a few milliseconds, are extremely rare.[12]

It is important to distinguish between firing rates measured by averaging over a population of neurons and the

rates measured locally. Low global firing rates can be achieved even in simple neural networks, e.g., by introducing a sufficiently strong inhibition. In fact, several recent models have been studied that possess stable states with very low global activity.[15–21] In the context of associative memory these states represent sparsely coded memories. Nevertheless, the activity levels of the active neurons are close to saturation, although their number is small. In fact, stabilizing persistent patterns of activity with low *local* firing rates in highly connected recurrent networks is rather difficult, regardless of the details of the considered models.[22–24]

B. The Willshaw model

Motivated by the above-mentioned issues we have studied in detail a generalization of one of the earliest neural network models of associative memory, proposed some 20 years ago by Willshaw.[21,25] The model stores memories in excitatory synapses using an extremely simple version of *Hebb rules*. The synaptic efficacy J_{ij} between the jth (presynaptic) neuron and the ith (postsynaptic) neuron is 1 if the two neurons are simultaneously active in at least one of the patterns. Otherwise, J_{ij} is zero. Let us denote by $[V_i^\mu]$, $i=1,\ldots,N$, $\mu=1,\ldots,P$ the P memories that are stored in the network. Each memory consists of an N-bit vector, with $V_i^\mu=0,1$. The synaptic matrix is given by

$$J_{ij}=\Theta\left[\sum_{\mu=1}^{P} V_i^\mu V_j^\mu\right], \tag{1.1}$$

where $\Theta(x)=1$ for $x>0$ and zero otherwise. Willshaw's original network was a fully synchronized system with a simple two-layer architecture. Here we study an implementation of the model in a fully connected recurrent network with asynchronous dynamics. A similar implementation has been studied recently, in simulations and in hardware, mainly by Thakoor *et al.*[26]

The Willshaw model has several attractive features. It offers an extremely simple way of storing sparsely coded memories, although it exhibits a poor performance with regards to random, *uncorrelated* memories. Let us assume that V_i^μ are random except for the constraints that $\sum_{i=1}^{N} V_i^\mu = Nf$, where $f \ll 1$. Then, the memories are stable states of the network, provided that the threshold value is appropriately chosen, for

$$P \le -\frac{1}{f^2}\ln(1-N^{-1/Nf}) . \tag{1.2}$$

This holds in the limit of $f\to 0$ and $N\to\infty$. Of course one has to take into account the fact that the information content of each memory decreases with decreasing f as

$$\begin{aligned} I/P &= -N[f\ln f+(1-f)\ln(1-f)] \\ &\approx -Nf\ln f . \end{aligned} \tag{1.3}$$

As was shown by Willshaw, Buneman, and Longuet-Higgins,[21] the maximal information capacity is achieved when $f=\ln N/(N\ln 2)$. Using Eq. (1.2) one obtains in this limit $I_{\max}=N^2\ln 2$ which is 69% of the information-theoretic bound $I=N^2$.

A recurrent network with the learning rule of Eq. (1.1) may posses *partially ordered* low-temperature phases in which part of the population is quiescent while the rest are active at levels that are substantially lower than saturation. This unusual behavior resembles to some extent the above-mentioned *low firing rates* observed in neural activity in the cortex. These phases exist only when a uniform inhibition, i.e., a constant negative term, is added to Eq. (1). The resultant synaptic matrix is no longer purely excitatory, and in particular, Dale's law is violated. Thus biological plausibility requires that the uniform synaptic inhibitory component should not be interpreted as representing a direct inhibitory coupling. Rather, we interpret it as an effective inhibitory interaction between excitatory neurons, that results from the activity of inhibitory neons.

Finally, Willshaw's model, implemented in a fully connected recurrent network, has some interesting features from the statistical mechanical point of view. The frustration,[27] which gives rise to multiplicity of stable states, does not result from a spin-glass-like mixture[28] of negative and positive bonds. Here it is generated by the competition between a distribution of positive bonds and external fields. Adding global inhibition creates an additional source of frustration which is spatially uniform. The above-mentioned partially ordered phases are to a certain extent similar to disordered phases with an extensive zero temperature entropy, that are seen in many short- and long-range uniformly frustrated systems.[28]

C. The present work

In this paper we present a statistical mechanical study of a fully connected network in which sparsely coded memories are stored according to Eq. (1). A uniform neural threshold and a uniform synaptic inhibition are added. The focus of the present study differs from that of other studies of Willshaw's model. Previous studies[21] focused on values of thresholds and f which yield optimal capacity. Here we study the robustness of the model by allowing a wide range of values of threshold and inhibition strength as well as by introducing stochastic noise in the form of temperature. Second, the extreme limit of sparseness, $f\propto \ln N/N$, although optimal for information storage, is uninteresting for biological modeling. For instance, within the cortex, a reasonable estimate of the size of a highly connected network yields N in the order of 10^4 neurons. For this value of N, $f\sim\ln N/N$ would imply that only about 10 neurons are active in a given memory state. Here we study the more realistic case in which f vanishes not faster than N^{-x}, $0<x<1$. In addition, special attention is devoted here to parameter regimes where the memories are not stable but partially ordered low-temperature phases exist which are highly correlated with the memories. Lastly, we address the question whether synaptic inhibition via special inhibitory neurons is equivalent to a direct inhibitory coupling. Preliminary accounts of the results concerning low firing rates were reported elsewhere.[29] A different associative memory model that exhibits low local neural activities has been studied by Amit and Treves.[30]

In Sec. II we define the model and present an analytical study of its equilibrium properties using mean-field theory (MFT). Section III is devoted to an analysis of the corrections to the MFT and their consequences. Results of simulations of the model are presented in Sec. IV. In Sec. V we discuss a modified model in which memory is stored in a purely excitatory network that is coupled to an ensemble of inhibitory neurons. The main results are summarized and discussed in Sec. VI.

II. THE GENERALIZED WILLSHAW MODEL—MEAN-FIELD THEORY

A. The model

We consider a network of N binary neurons in which P sparse memories are stored. The memories are N-bit vectors denoted by V_i^μ, $\mu=1,\ldots,P$; $i=1,\ldots,N$. The V_i^μ take the values 0 and 1 at random subject to the constraints $\sum_i V_i^\mu = Nf$, where $f \ll 1$. The memories are encoded in the synaptic efficacies according to Willshaw's rule:

$$J_{ij} = \frac{1}{Nf}\Theta\left[\sum_{\mu=1}^{P} V_i^\mu V_j^\mu\right], \tag{2.1}$$

where $\Theta(x)=1$ if $x>0$; otherwise it is zero. The synaptic efficacies contain also a uniform inhibitory component denoted by $-K(Nf)^{-1}$, $K>0$. In addition, the neurons have a *negative* uniform threshold denoted by $-\theta, \theta>0$. Thus the instantaneous local field of the ith neuron is given by

$$h_i = \sum_j J_{ij} V_j - \frac{K}{Nf}\sum_j V_j + \theta . \tag{2.2}$$

The neural states V_j assume the values 1 for an active state and 0 for a passive one. The normalization of the coupling constants has been chosen so that when the total activity of the network $V=\sum_i V_i/N$ is of order f all the three terms in Eq. (2.2) are of order unity. We will work in the parameter range

$$0 < f\theta < K-1 . \tag{2.3}$$

This ensures that the states where all neurons are active or all are passive are not stable.

The network is assumed to evolve according to an asynchronous stochastic dynamics with a noise level which is denoted by a "temperature" T.[9] Since the synaptic matrix is symmetric the long-time behavior of the system can be described by the following energy function:

$$H = -\tfrac{1}{2}\sum_{i,j} J_{ij} V_i V_j + \frac{K}{2Nf}\left[\sum_i V_i\right]^2 - \theta\sum_i V_i . \tag{2.4}$$

Obviously the behavior of the system depends on the distribution of J_{ij}, i.e., on the number of stored patterns P and their activity level f. The average of J_{ij} is given by

$$\langle\langle J_{ij}\rangle\rangle = \frac{1}{Nf}(1-C) , \tag{2.5}$$

where C is the fraction of zero bonds and is given, in the limit $P\to\infty$ and $f\to 0$, by

$$C = e^{-Pf^2} . \tag{2.6}$$

In this section we will study the limit where C *remains finite* as $f\to 0$, i.e., P is proportional to f^{-2}. In this limit, a finite fraction of the bonds is zero, and the system is described by a simple MFT in the limit $N\to\infty$. It should be emphasized that although we assume that f is small, we do consider the extreme case of f being proportional to $\ln N/N$. Instead, we deal with the more "realistic" case where f vanishes as $1/\ln N$ or as N^{-x}, $0<x<1$.

B. Mean-field theory

In the limit of finite C and $N\to\infty$, the fluctuations of J_{ij} can be neglected. Anticipating a big overlap of the state of the system with one of the memories, say V_i^1, the coupling constants J_{ij} can be replaced by their average over the pattern V_i^μ, $\mu>1$. The probability that $J_{ij}=0$ if $V_i^1=0$ is $(1-f^2)^{P-1}\approx C$. Hence J_{ij} can be replaced by

$$J_{ij}^0 = \frac{1}{Nf}\left\langle\!\!\left\langle \Theta\left[V_{ij}^1 V_j^1 + \sum_{\mu\geq 2} V_i^\mu V_j^\mu\right]\right\rangle\!\!\right\rangle = \frac{1}{Nf}(CV_i^1 V_j^1 + 1 - C) , \tag{2.7}$$

where $\langle\!\langle\ \rangle\!\rangle$ refers to averaging over all the patterns with $\mu>1$. The ensemble-averaged free energy per neuron F is given by

$$-\beta NF = \left\langle\!\!\left\langle \ln \mathrm{Tr}_{\{V_i\}} \exp\left[\frac{\beta}{2}\sum_{i,j} J_{ij} V_i V_j - \frac{\beta K}{2Nf}\sum_{i,j} V_i V_j - \beta\theta\sum_i V_i\right]\right\rangle\!\!\right\rangle . \tag{2.8}$$

Substituting Eq. (2.7) for J_{ij} in Eq. (2.4) and using standard saddle-point methods, the free energy can be written as

$$F(V^+,V) = \tfrac{1}{2}C(V^+)^2 - \tfrac{1}{2}(K+C-1)(V/f)^2 - T\ln\{1+\exp\beta[CV^+ - (K+C-1)(V/f)+\theta]\} - \frac{T}{f}\ln\{1+\exp\beta[-(K+C-1)(V/f)+\theta]\} . \tag{2.9}$$

The order parameters V^+ and V are

$$V^+ \equiv \frac{1}{Nf}\sum_i V_i^1 \langle V_i\rangle , \tag{2.10}$$

$$V \equiv \frac{1}{N}\sum_i \langle V_i\rangle , \tag{2.11}$$

where $\langle V_i\rangle$ is the thermal average of the neuron activity.

We will call the neurons with $V_i^1=1$ the *on* neurons, and those with $V_i^1=0$ the *off* neurons. This classification is, of course, different for each memory. The total average activity can be written as $V=f(V^+ +(1-f)V^-)$ where

$$V^- \equiv \frac{1}{Nf(1-f)} \sum_i (1-V_i^1)\langle V_i \rangle \ . \tag{2.12}$$

These order parameters are determined by

$$\frac{\partial F}{\partial V^+} = \frac{\partial F}{\partial V} = 0 \ , \tag{2.13}$$

which yields the following simple set of mean-field equations:

$$V^+ = \frac{1}{1+e^{-\beta h^+}} \ , \tag{2.14}$$

$$fV^- = \frac{1}{1+e^{-\beta h^-}} \ , \tag{2.15}$$

where h^+ is the local field on the *on* neurons, and h^- is the local field on the *off* neurons. They are given by

$$h^+ = (1-K)V^+ + (1-C-K)V^- + \theta \ , \tag{2.16}$$

$$h^- = h^+ - CV^+ \ . \tag{2.17}$$

The meaning of these equations is that an on neuron is influenced, through the J_{ij}, by all the on neurons as well as by a fraction $1-C$ of the off neurons. On the other hand, each of the off neurons "feels" the influence of a fraction $1-C$ of all the neurons. Note that according to Eqs. (2.14) and (2.15) the local fields within each of the two populations are uniform. This holds only in the limit $N \to \infty$ and $f \to 0$. In the following sections we analyze the solutions of the mean-field equations, Eqs. (2.14)–(2.17). We first discuss the case of positive h_0, where

$$h_0 = \theta + 1 - K \tag{2.18}$$

is the field on the on neurons in a memory state. Later, we will address the case of $h_0 < 0$.

C. Analytical solution of the mean-field equations for $f \to 0$

1. $h_0 > 0$

In this paragraph we present the asymptotic form of the solutions of the mean-field equations, in the limit $f \to 0$. Two types of solutions exist. One consists of $V^+ \gg fV^-$, i.e., the activity level, per neuron, of the on neurons is much bigger than that of the off neurons. This solution is termed a *retrieval state*. The other type is a *symmetric state* where $V^+ \approx fV^-$, implying the total loss of information regarding the memory. In both solutions, the activity level averaged over the whole network V is low, i.e., of order f. For this to hold, the temperature must scale as

$$T = \bar{T}/(-\ln f) \ , \tag{2.19}$$

where $\bar{T}$ is of order 1, as will be shown below. We limit our study to this regime. When T becomes of order 1, V becomes of order 1 implying a behavior that is far from the one that characterizes the memory states.

The retrieval state: $T=0$. Substituting $V^+=1$, $V^-=0$ in Eqs. (2.14)–(2.17) at $T=0$ one easily finds that the memory state is stable at zero temperature as long as

$$C > h_0 > 0 \ . \tag{2.20}$$

This sets the maximum capacity of the model in the thermodynamic limit. However, retrieval states exist also for $C < h_0$. They are characterized by the vanishing of h^- while $h^+ = h^- + CV^+ \approx C > 0$. Thus, in this regime V^+ remains 1 but the activity of the neurons with $V_i^1=0$ is no longer zero, but equals fV^- where

$$V^- = \frac{h_0 - C}{K + C - 1}, \quad C < h_0 \ . \tag{2.21}$$

The fact that the activity of this population, at $T=0$, is neither 0 nor 1 implies that the off neurons do not saturate even at zero T, but keep fluctuating in time, due to the vanishing of their local field h^-, Eq. (2.17). The time average of their activity per neuron is given by Eq. (2.21).

This interpretation of the retrieval phase is supported by the calculation of the entropy per neuron, $S = -\partial F/\partial T$. From Eq. (2.9) one obtains that the entropy per neuron of the retrieval phase is finite in the limit of $T \to 0$,

$$S = -f(V^+ \ln V^+ + V^- \ln V^-) \approx -fV^- \ln V^- \ , \tag{2.22}$$

where V^- is given by Eq. (2.21). This type of retrieval state is different from those encountered in the Hopfield model and its variations.[9] The retrieval states in those models are stable configurations that contain a small fraction of erroneous bits.

The retrieval state: $\bar{T} > 0$. When $C > \bar{T} + h_0$ the effect of nonzero $\bar{T}$ is exponentially small, and the local fields are essentially the same as at $\bar{T}=0$, i.e., $h^+ = h_0$, $h_- = h_0 - C < -\bar{T}$. From Eqs. (2.14)–(2.17) one obtains the following results:

$$V^+ \approx 1 - f^{\bar{\beta} h_0} \ , \tag{2.23}$$

$$fV^- \approx f^{\bar{\beta}(C - h_0)} \ , \tag{2.24}$$

where $\bar{\beta} = -\beta/|\ln f|$. This solution exists as long as $\bar{T} < C - h_0$. We call this regime the *strong memory* regime. When $C < \bar{T} + h_0$, V^+ is still close to 1 but V^- is of order 1, implying that the contribution of the activity of the off neurons to the total activity, V, is of the same order as that of the on neurons. This is the *weak memory* regime. The value of V^- is such that $h^- = -\bar{T}$ yielding

$$V^+ \approx 1 - f^{\bar{\beta} C - 1} \ , \tag{2.25}$$

$$V^- = \frac{\bar{T} - C + h_0}{K + C - 1} \ . \tag{2.26}$$

Note that as $\bar{T} \to 0$ h^- vanishes and V^- approaches the limit of Eq. (2.21). The transition from the strong memory to the weak memory behavior, at $\bar{T} = C - h_0$, is not a sharp transition, except at $\bar{T}=0$. The retrieval

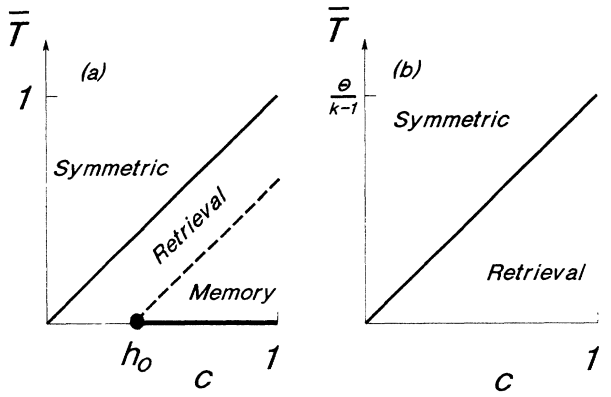

FIG. 1. The phase diagram of the mean-field theory in the limit $f \to 0$. (a) $h_0 > 0$. The thick horizontal line marks the regime where the memory states are stable. (b) $h_0 < 0$.

state exists up to $\bar{T} = \bar{T}_c$,

$$\bar{T}_c = C \ . \tag{2.27}$$

Above this temperature the correlations with the individual memories are lost and the state of the system is characterized as a symmetric state.

The symmetric state. In addition to the above solution there is a stable solution in which $fV^- \approx V^+ \approx O(f)$, i.e., the activity of the on and off neurons is roughly equal. The mean-field equation for V^+ becomes

$$V^+ \approx f^{\bar{\beta}[(K+C-1)V^+/f-\theta]} \tag{2.28}$$

yielding

$$fV^- \approx V^+ \approx \frac{\bar{T}+\theta}{K+C-1} \ . \tag{2.29}$$

This solution exists and is stable for all $\bar{T} > 0$ and C. It is the only solution when $\bar{T} > C$. The phase diagram of the mean-field theory in the limit $f \to 0$ is shown in Fig. 1(a).

2. $h_0 < 0$

When the local field on the on neurons in a memory state is negative this state is, of course, unstable. Nevertheless, at low $\bar{T}$ there are retrieval states characterized by a big difference in the activity levels of the on and off neurons, as in the case $C < h_0$ above. However, here the off neurons are quiescent at $T = 0$, whereas the on neurons are only partially active. Thus $h^+ = 0$, and V^+ is given by

$$V^+ = \frac{\theta}{K-1} \ . \tag{2.30}$$

The local field on the off neurons is $h^- = CV^+$. Hence this state (with $V^- \approx 0$) exists for $\bar{T} < \bar{T}_c$ where

$$\bar{T}_c = \frac{C\theta}{K-1} \ . \tag{2.31}$$

In addition to the retrieval states, there exists also a symmetric phase at all $\bar{T} > 0$ with the same activities as in the case $h_0 > 0$, i.e., Eqs. (2.28) and (2.29). The phase diagram for this case is shown in Fig. 1(b).

D. Numerical solution of the mean-field equations

The analytical solutions of the mean-field equations of the previous paragraph are valid in the $f \to 0$ limit. Mean-field solutions for nonzero f can be obtained only numerically. Of course one has to keep in mind that the mean-field equations themselves are valid only in the $f \to 0$. However, finite f corrections to the mean-field

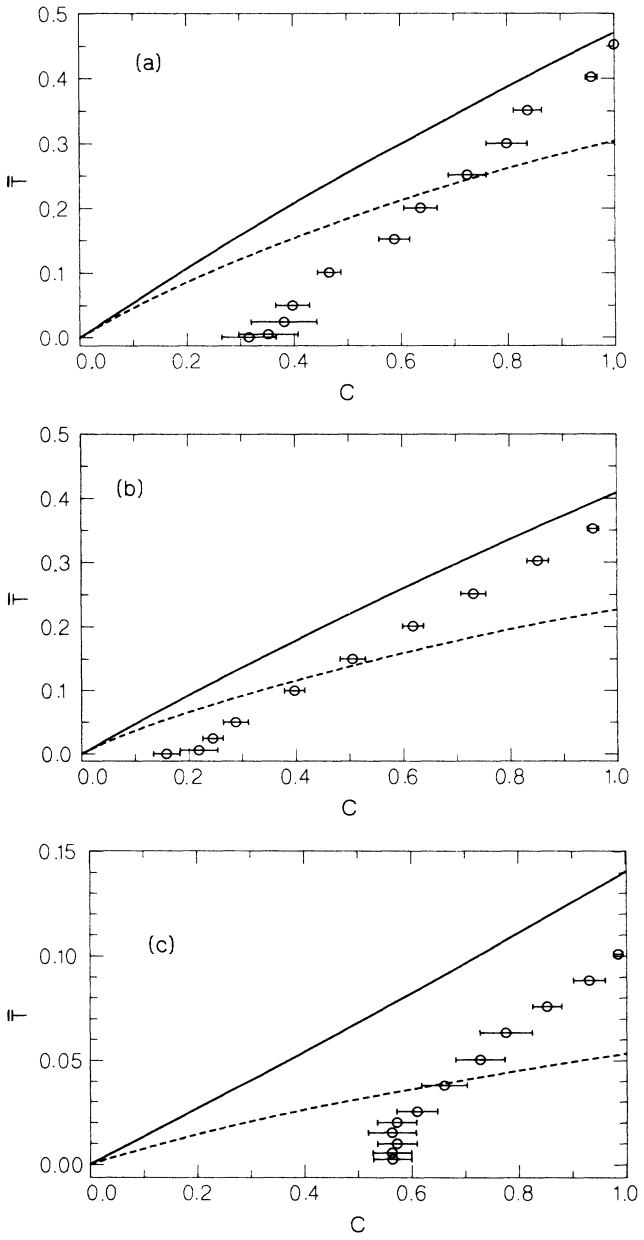

FIG. 2. Numerical solutions of the MFT and simulations with $f = 0.04$, $N = 1000$, $K = 2$. Below the solid lines, the retrieval phase is stable acccording to MFT. The symmetric phase exists above the dashed lines. Dots represent the results of simulations for the value of C below which the retrieval phase exists. The errors, denoted by the bars, result from the uncertainty in defining exactly the transition, as well as from sample to sample fluctuations. (a) $h_0 = 0.75$. (b) $h_0 = 0.25$. (c) $h_0 = -0.75$.

equations, which will be discussed in the following section, are in general smaller than the finite f corrections to the mean-field *solutions.*

Studying the mean-field equations, Eqs. (2.14)–(2.17), at finite small values of f, one finds that most of the features of the phase diagram, presented in Fig. 1, are not modified. The main differences are concerned with the symmetric state. At finite f the values of V^+ and fV^- are not identical but differ by a small amount. At $\bar{T}$ of order 1 this has only a minor effect since both V^+ and fV^- are of order f. However at low $\bar{T}$, the small difference $h^+ - h^- = CV^+$ is sufficient to generate large differences in the activity of the two populations. Thus the symmetric state disappears at low temperatures below a critical temperature, $\bar{T}$ of order f. The results of the numerical solution of the mean-field equations for $f = 0.04$ are presented in Fig. 2, for positive and negative h_0. Note that the critical line for the appearance of retrieval phases is approximately linear but the slope is substantially less than that predicted by Eq. (2.27) and Eq. (2.31). The fact that there are relatively large finite f corrections to that slope can be already observed by examining the *leading* corrections to the $f \to 0$ results for $\bar{T}_c$. This analysis yields

$$\bar{T}_c = \frac{C}{\left[1 + \frac{\ln|\ln f|}{|\ln f|}\right]} . \tag{2.32}$$

III. CORRECTIONS TO THE MFT

The MFT predicts the existence of retrieval phases for all values of C. Obviously, this should break down for sufficiently small values of C, i.e., for sufficiently large P. To evaluate the capacity of the system one has to consider the effects of both finite f and finite N corrections to mean-field theory.

A. Corrections to the local fields

We define δJ_{ij} as the deviations of the synaptic efficacies J_{ij} from their mean-field value, Eq. (2.7):

$$\delta J_{ij} = J_{ij} - J_{ij}^0 . \tag{3.1}$$

They are given by

$$\delta J_{ij} = \frac{1}{Nf}(1 - V_i^1 V_j^1)(\Theta_{ij} - C) , \tag{3.2}$$

where:

$$\Theta_{ij} = \Theta\left[\sum_{\mu=2}^{P} V_i^\mu V_j^\mu\right] . \tag{3.3}$$

The local fields on the neurons of the on and off populations (relative to pattern 1) can be written as

$$h_i^\pm = h^\pm + \delta h_i^\pm , \tag{3.4}$$

where $h^\pm$ are the local fields of the mean-field theory, i.e., Eqs. (2.16) and (2.17), and $\delta h_i^\pm$ are the fluctuations,

$$\delta h_i^\pm = h_i^\pm - h^\pm = \sum_j \delta J_{ij}^\pm \langle V_j \rangle . \tag{3.5}$$

For large N, $\delta h_i^\pm$ can be treated as Gaussian variables with zero mean and variance $\Delta^{\pm 2} \equiv \langle\langle \delta h_i^{\pm 2} \rangle\rangle$. The variance of h_i^+ is given by

$$\Delta^{+2} = \left\langle\!\!\left\langle \sum_{j \neq k} (1 - V_j^1)(1 - V_k^1)\langle V_j \rangle \langle V_k \rangle \frac{1}{N^2 f^2}(\Theta_{ij} - C)(\Theta_{ik} - C) \right\rangle\!\!\right\rangle + \left\langle\!\!\left\langle \sum_i (1 - V_j^1)\langle V_j \rangle^2 \frac{1}{N^2 f^2}(\Theta_{ij} - C)^2 \right\rangle\!\!\right\rangle . \tag{3.6}$$

The first term in Eq. (3.6) represents finite f corrections due to correlations between J_{ij} and J_{ik}. In the Appendix it is shown that

$$\langle\langle \Theta_{ij}\Theta_{il} \rangle\rangle_c = \left\langle\!\!\left\langle \Theta\left[\sum_\mu V_i^\mu V_j^\mu\right]\Theta\left[\sum_{\mu'} V_i^{\mu'} V_l^{\mu'}\right] \right\rangle\!\!\right\rangle_c \approx C^2 |\ln C| f , \tag{3.7}$$

where $\langle xy \rangle_c \equiv \langle xy \rangle - \langle x \rangle \langle y \rangle$. The second term represents fluctuations in J_{ij}^2. These spin-glass-like fluctuations[28] are, in the present case, finite N corrections, since their total contribution to Eq. (3.6) is of the order of $1/Nf$. In fact, defining

$$q^- = \frac{1}{N}\sum_j (1 - V_j^1)\langle V_j \rangle^2 \tag{3.8}$$

one obtains

$$\Delta^{+2} = \langle\langle \delta h_i^{+2} \rangle\rangle = fC^2 |\ln C| (V^-)^2 + \frac{C(1-C)q^-}{Nf^2} . \tag{3.9}$$

In the same way we find

$$\Delta^{-2} \equiv \langle\langle \delta h_i^{-2} \rangle\rangle = fC^2 |\ln C| (V^+ + V^-)^2 + \frac{C(1-C)q}{Nf^2} , \tag{3.10}$$

where

$$q = \frac{1}{N}\sum_j \langle V_j \rangle^2 . \tag{3.11}$$

In the presence of these quenched fluctuations, the thermal activities of the neurons are no longer homogeneous. In particular, the spatial-averaged activity of the off population is given by

$$fV^- = \int_{-\infty}^{\infty} e^{-z^2/2} \frac{dz}{\sqrt{2\pi}\{1 + \exp[-\beta(h^- + z\Delta^-)]\}} . \tag{3.12}$$

At $T=0$ one obtains

$$fV^- = \mathrm{erfc}(-h^-/\sqrt{2}\Delta^-) \ . \tag{3.13}$$

Similar expressions hold for the activity of the on population.

B. Capacity of memory states

The above results can be used to calculate the effect of fluctuations on the stability of the memory states at $\bar{T}=0$. In a memory state, $V^-=0$, $V^+=1$, $q^-=0$, and $q=f$, implying that $\Delta^+=0$ but $\Delta^->0$. As C decreases the *noise* in h_i^- eventually causes one of the off neurons to fire. The probability that at least one off neuron will fire is finite if $fV^- \gg 1/N$. This implies through Eq. (3.13) that the limit of stability of the memory states is given by

$$\frac{h^{-2}}{\Delta^{-2}} = \frac{(C-h_0)^2}{fC^2|\ln C| + C(1-C)/(Nf)} \tag{3.14a}$$

$$= 2/\ln N \ . \tag{3.14b}$$

In analyzing the effects of fluctuations it is useful to distinguish between two regimes. One is the case of *large f*, defined by $f \gg 1/\ln N$. In this regime the dominant contributions to the fluctuations are the first terms in Eqs. (3.9) and (3.10), namely, the fluctuations due to correlations. In the *small f* regime, $f \ll 1/\ln N$, the dominant terms are the spin-glass fluctuations, namely, the second terms in these equations. In the small f regime, Eq. (3.14b) reduces to

$$C = \frac{h_0}{1-\sqrt{2f\ln N|\ln C|}} \ . \tag{3.15}$$

This represents only a small correction to the mean-field result $C=h_0$; see Eq. (2.20). In fact, Eq. (3.15) holds even in the case of $f \propto 1/\ln N$. When f is much larger than $1/\ln N$ the critical value of C approaches unity, implying that, in the large f regime, the memory states are never stable. This, however, does not invalidate the above mean-field results since even in this case, the *fraction* of errors, i.e., the fraction of off neurons that fire, is not finite (as $N \to \infty$) until C reaches the neighborhood of h_0.

C. Capacity of retrieval states, $h_0 > 0$

According to the MFT, even when the memory states are not stable there are retrieval phases highly correlated with the memories. This holds also when fluctuations are taken into account. However, it is important to note that at low temperatures, $\bar{T} \ll \Delta^-$, the fluctuations *pin* the local activities of the neurons. At finite C, the freezing of the neural fluctuations sets in at

$$\bar{T} \approx \sqrt{f} \ . \tag{3.16}$$

In particular, at $\bar{T}=0$, the retrieval phases are frozen states, with $V_i^+=1$, while $V_i^-=0$, or 1 according to the values of h_i^-. Although the dynamic nature of the phases changes drastically by the fluctuations, the population activity levels are only slightly modified. To leading order, one has (at $T=0$) $V^+ \approx 1$ while V^- is approximately given by Eq. (2.21). The zero T value of h^- differs from zero. Its (small) value is determined by the consistency of Eqs. (3.13) and (2.21) yielding

$$h^- \approx -\sqrt{2\ln|f|}\Delta^- \ . \tag{3.17}$$

On the other hand, $h^+ = h^- + CV^+ \approx h^- + C$ must be positive, hence h^- must be greater than $-C$. Thus the retrieval state is stable only for C larger than that given by

$$C = \sqrt{2\ln|f|}\Delta^- \ . \tag{3.18}$$

When C decreases below this value, the local fields on the on populations acquire negative values leading to a drastic reduction in the activity of this population and to destabilization of the phase. To determine the actual limits on C, implied by the above equation, we distinguish again between the large f and the small f regimes. In the large f regime, $\Delta^{-2} \approx fC^2|\ln C|(1+V^-)^2$, where V^- is given by Eq. (2.21); see Eq. (3.10). Hence Eq. (3.18) reduces to

$$Pf^2 = -\ln C < \frac{A}{f|\ln f|} \ , \quad f \gg 1/\ln N \tag{3.19}$$

where $A=[1+h_0/(K-1)]^2/2$. In the small f regime, $\Delta^{-2} \approx C(1+V^-)/Nf$ [see Eq. (3.10)], yielding

$$C > A\frac{|\ln f|}{Nf}, \quad f \ll 1/\ln N \tag{3.20}$$

where $A=2[1+h_0/(K-1)]$. Note that the capacity increases as h_0 decreases.

D. Capacity of retrieval states, $h_0 < 0$

When $h_0 < 0$ the retrieval states are characterized, in the MFT, by $V^-=0$, whereas V^+ has a value less than unity, Eq. (2.30), reflecting the partial saturation of the system. To leading order, V^+ is still given by Eq. (2.30), implying that $h^+ \approx 0$, and $h^- \approx -CV^+$. The value of V^- is determined by Eq. (3.13). Self-consistency requires that $V^- \ll 1$, otherwise the activity of the on neurons will be suppressed. Solving Eq. (3.13), using Eq. (3.10), we find that there is a self-consistent solution for

$$Pf^2 = -\ln C < \frac{1}{2f|\ln f|}, \quad f \gg 1/\ln N \tag{3.21}$$

in the large f regime, and for

$$C > A\frac{|\ln f|}{Nf}, \quad f \gg 1/\ln N \tag{3.22}$$

where $A=2/V^+=2(K-1)/\theta$ in the small f regime.

IV. NUMERICAL SIMULATIONS

We have carried out numerical simulations of the generalized Willshaw model, Eq. (2.4), with $N=1000$, $f=0.04$, and $K=2$. The memories were chosen at random subject to the constraint $\sum_i^N V_i^\mu = Nf$. The network evolves according to a finite temperature single-spin-flip sequential dynamics, with a random order of updatings, starting from the state $V_i = V_i^1$. The total activity of the

on and off neurons was computed for different values of P and θ. At zero temperature and small P, $V^+=1$ and $V^-=0$ as expected. When P increases, a few wrong bits appear. In most cases, some of the off neurons start to fire, i.e., V^- becomes greater than 0. Only at larger values of P, a reduction in V^+ is observed.

A. Capacity of memory states

The capacity for the stability of the memory states for different values of h_0 is shown in Fig. 3. For large h_0, these results are in good agreement with the analytical results (solid line). The latter were calculated by equating the right-hand side (RHS) of Eq. (3.13) to $1/N$, using Eq. (3.14a) for h^-/Δ^-. This yields a slightly better agreement with the simulations than using the asymptotic result, Eq. (3.14b). When h_0 decreases, the difference between the numerical and analytical results increases. This happens because the analytical results are based on a Gaussian approximation for the distribution of the local fields on the off neurons; see Eq. (3.12). This approximation is valid in the limit $N\to\infty$ and finite h_0. However, for a fixed finite N, the local fields that cause an off neuron to fire get close to the "tail" of the local fields' distribution as $h_0\to 0$, and the Gaussian approximation is invalidated.

B. Phase diagram

Above the capacity for the stability of memory states one observes a phase characterized by $V^+\approx 1$, $fV^-\approx 0$ similar to the retrieval phase predicted by the MFT. We have measured the values of P and $\bar{T}$ below which the retrieval phases exist, for three values of h_0: 0.25, 0.75, and -0.75. Above these values of P and $\bar{T}$, the system settles in a state characterized by $V^+\approx fV^-$ both of the order f. This state is similar to the symmetric state of the MFT. The results are presented in Fig. 2, together with the theoretical predictions. The theoretical lines are based on numerical solution of the mean-field equations, Eqs. (2.14)–(2.17), with $f=0.04$. For a given $\bar{T}$, the critical value of P below which retrieval phases appear is always lower than the value predicted by the mean-field result. This discrepancy is smaller as the temperature increases. Note that at high $\bar{T}$ the capacity increases with h_0 as expected from the mean-field results.

The major deviations from the mean-field results occurs at low temperatures. In particular, fluctuations destabilize the retrieval states below some critical value of C, while the mean-field line extends to $C\to 0$. This effect has been discussed in the preceding section. The asymptotic estimates of Eqs. (3.21) and (3.22) for the limit of stability of the retrieval phases at zero T are, in fact, of the same order of magnitude as the numerical values. Unfortunately, a precise quantitative comparison is difficult to make, as the various limits assumed by the theoretical analysis are not justified for $N=1000$ and $f=0.04$.

C. Low firing rates

As Fig. 2(c) indicates, the simulation results for $h_0=-0.75$ are in good agreement with the theory. To check whether the retrieval phases are indeed unfrozen, we have measured the local firing rates. The results are shown in Fig. 4, where the histogram of V_i, averaged over two time windows (Δt), are displayed. The peak near zero corresponds to the off population, which remains completely quiescent. The level of activity, averaged over the whole population of on neurons, is 0.27 (for $\Delta t=50$), and 0.275 (for $\Delta t=200$). These numbers agree well with the mean-field prediction $V^+=0.25$ for the parameters used in the simulations. Note that the width of the histogram (of the on neurons) shrinks from $\delta V=0.106$, in the case of $\Delta t=50$ to $\delta V=0.054$ for $\Delta t=200$. This is consistent with the expectation that the individual neurons fluctuate at random, essentially independent of each other, except for the global constraint on the average activity.

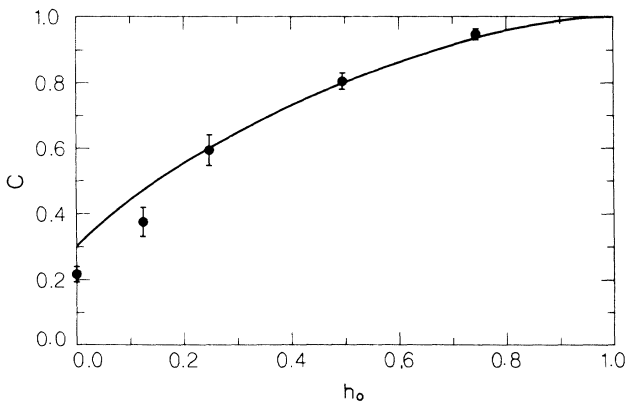

FIG. 3. Capacity of memory states for $f=0.04$, $N=1000$, $K=2$. The theoretical results, represented by the solid line, have been derived by equating the RHS of Eq. (3.13) to $1/N$ using the value of h^-/Δ^- as given by Eq. (3.14a). The points are the simulation results.

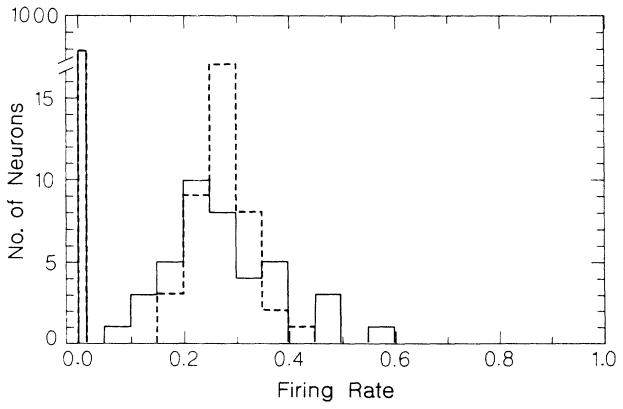

FIG. 4. Histogram of the local firing rates, obtained from simulations with $f=0.04$, $N=1000$, $K=2$, and $h_0=-0.75$. The histogram represents averages over 50 time steps (solid curve) and 200 time steps (dashed curve). The peak near zero corresponds to the 960 off sites.

D. Spurious states

So far we have discussed simulations that start from one of the memories. In order to check the existence of spurious states and the basins of attraction of the memories, we have performed simulations starting from random initial conditions that have an overall activity of $f=0.04$. The overlaps of the state of the network with all the stored patterns are computed. The simulations were done for $P=50$ ($C=0.92$) and varying temperature, and for $\bar{T}=0$ and varying P. In both cases h_0 was 0.5. For $P=50$, the network always reaches a memory state at zero temperature. When the temperature increases, the system settles in one of the retrieval phases. Above $\bar{T}=0.35$, the system settles in a state which has small overlaps with many patterns, i.e., a mixed state. The typical overlap of this state with individual memories is small, typically of the order of f, similar to the symmetric phase. This is consistent with the MFT, according to which at $\bar{T}>0.255$ the symmetric phase exists. Increasing P at $\bar{T}=0$, we find that the system converges to the symmetric phase for $P>220$, i.e., $C<0.7$. Note that at this value of h_0, retrieval phases are stable at $\bar{T}=0$ up to $C=0.26$. This implies that symmetric phases are stable at zero temperature before the retrieval phases lose their stability. In conclusion, at low temperature and small P, spurious states, if they exist at all, have very small basins of attraction. In large P or $\bar{T}$, spurious symmetric phases appear besides this retrieval phase.

V. NETWORK WITH EXCITATORY AND INHIBITORY NEURONS

So far, we have discussed the properties of the symmetric model, Eq. (2.4), which violates Dale's law, as mentioned in the Introduction. In this section we modify the model so that it contains two groups of neurons. Memories are stored, using Willshaw's rule, Eq. (2.1), in the excitatory synaptic connections between *excitatory* neurons. Inhibition is provided through the coupling of these neurons with *inhibitory* neurons.

We assume for simplicity that there are N neurons of each type. In addition to the Willshaw connection matrix, there is a uniform excitatory interaction from the excitatory to the inhibitory populations, of strength J/Nf. The inhibitory neurons have a uniform mutual inhibitory coupling, $-J'/Nf$, and they inhibit the excitatory neurons with a uniform coupling, $-K/Nf$. Thus the role of the inhibitory neurons is to serve as sources of a uniform inhibitory feedback for the excitatory population. The network architecture is depicted in Fig. 5. For simplicity we choose the parameters $J=J'=1$ and assume that the threshold of the inhibitory neurons is zero. The threshold of the excitatory neurons is $-\theta, \theta>0$. The activities of the excitatory (inhibitory) neurons are denoted by V_i (U_i).

The time evolution of the system is governed by a stochastic Markov process of single spin flips. The transition rates are given by (Ref. 31) $w(V_i)=\frac{1}{2}[1+(2V_i-1)\tanh(\frac{1}{2}\beta h_i)]$. Let us denote by $\langle V_i\rangle$ and $\langle U_i\rangle$ the activities of the excitatory and inhibitory neurons, respectively, averaged over the stochastic noise. The dynamic equations of the local activities are

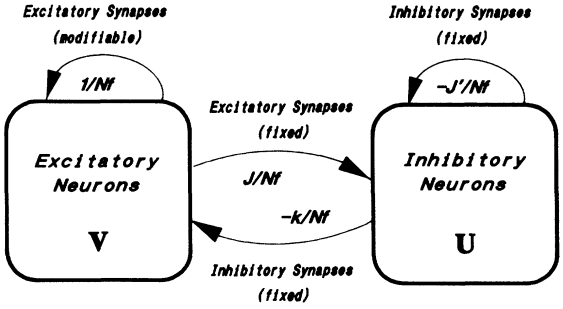

FIG. 5. A network with a biologically plausible architecture. There are two neural populations, excitatory neurons and inhibitory neurons. Information is stored on the excitatory neurons only, via Willshaw's learning rule. The uniform coupling strengths of each population to itself and to the other population are indicated in the figure.

$$\frac{\partial}{\partial t}\langle V_i\rangle=-\langle V_i\rangle+\langle g(\beta h_i^{\mathrm{ex}})\rangle\ , \tag{5.1}$$

$$\frac{\partial}{\partial t}\langle U_i\rangle=-\langle U_i\rangle+\langle g(\beta h_i^{\mathrm{in}})\rangle\ , \tag{5.2}$$

where $g(x)=1/[1+\exp(-x)]$. The fields on the excitatory neurons are

$$h_i^{\mathrm{ex}}(t)=\sum_{j=1}^{N}J_{ij}V(t)_j-f^{-1}KU(t)+\theta\ , \tag{5.3}$$

where $U\equiv N^{-1}\sum_i U_i$. The fields on the inhibitory neurons are uniform and equal

$$h_i^{\mathrm{in}}(t)=[V(t)-U(t)]/f\ . \tag{5.4}$$

We assume that the initial state is correlated with only one memory, say, with $\mu=1$. In the limit of $N\to\infty$, $f\to 0$, and finite C, one obtains, in a manner similar to the derivation of Sec. II B, the following mean-field equations:

$$\frac{\partial}{\partial t}V^{+}=-V^{+}+g(\beta h^{\mathrm{ex}+})\ , \tag{5.5}$$

$$\frac{\partial}{\partial t}V^{-}=-V^{-}+g(\beta h^{\mathrm{ex}-})\ , \tag{5.6}$$

$$\frac{\partial}{\partial t}U=-U+g(\beta h^{\mathrm{in}})\ , \tag{5.7}$$

where V^{+},V^{-} are defined as in Eqs. (2.10) and (2.12). The mean fields are given by

$$h^{\mathrm{ex}+}=V^{+}+(1-C)V^{-}-KU/f+\theta\ , \tag{5.8}$$

$$h^{\mathrm{ex}-}=(1-C)(V^{+}+V^{-})-KU/f+\theta\ , \tag{5.9}$$

and $h^{i}=(V-U)/f$ as above.

Equations (5.5)–(5.9) may have time-periodic solutions. We focus here on low-temperature properties of the stationary solutions, which read

$$V^{+}=g(\beta h^{\mathrm{ex}+}) \tag{5.10}$$

and similarly for fV^- and U. As in previous sections, we choose parameter range $0<f\theta<K-1$ which ensures that the average activity of the excitatory population, V, is of order f. At zero temperature, the only self-consistent stationary solution is that in which the $U=V$. In other words, the inhibitory population never saturates and its time-averaged activity matches the average activity of the excitatory population. Substituting this equality in Eqs. (5.5)–(5.9), one obtains the same mean-field equations as Eqs. (2.14)–(2.17). We have checked the dynamic stability of the stationary solutions by linearizing Eqs. (5.5)–(5.7). We find that all the stationary solutions are stable at all the parameter values that they exist, i.e., for $h_0<0$.

The above mean-field results imply that in the limit of large N, a system with an indirect inhibitory feedback through inhibitory neurons will have the same stationary phases as that with direct inhibitory couplings. This includes the phases with low firing rates, for $h_0<0$. However, numerical simulations of this network with $N=500$ revealed that the phases with low firing rate are, most of the time, unstable. The problem lies in the fact that the indirect inhibitory feedback is too slow to suppress effectively fluctuation in the level of activity of the excitatory neurons. We will discuss the implications of this problem in the following section.

VI. SUMMARY AND DISCUSSION

A. Summary of results

The extended Willshaw model, Eq. (2.4), has a simple behavior in the limit $N\to\infty$, $f\to 0$, and

$$P\propto f^{-2}\ . \tag{6.1}$$

The temperature range where the main phase transitions occur is given by

$$T=\bar{T}/|\ln f|\ , \tag{6.2}$$

where $\bar{T}=O(1)$. The quantity that determines the stability of the memory states is the local field generated in a memory state on the *on* neurons. This field is $h_0=\theta+1-K$. When $h_0>0$ memory states are stable (at $T=0$) for

$$P<\frac{|\ln h_0|}{f^2}\ . \tag{6.3}$$

This is valid provided that $f\ll 1/\ln N$.

For general values of T and $C\equiv\exp(-Pf^2)$ there are three phases.

(1) *Memory phase.* The fraction of errors is extremely small even at $\bar{T}>0$. The activity per neuron of the on neutrons, V^+, is nearly 1, whereas that of the off population, fV^-, is much smaller than f. This phase exists only for $C>h_0>0$.

(2) *Retrieval phase.* In this phase, for $h_0>C>0$, V^+ is still close to 1. The level of activity (per neuron) of the off neuron is $fV^-\approx O(1)$. The low-temperature limit of V^- is

$$V^-=\frac{h_0-C}{K-1+C}\ . \tag{6.4}$$

For $h_0<0$, V^- is close to zero (even at finite $\bar{T}$) while V^+ is nonzero even at low $\bar{T}$, where it is given by

$$V^+=\frac{\theta}{K-1}\ . \tag{6.5}$$

At all finite $\bar{T}$ the results (6.4) and (6.5) represent not only population averages but also the time average of the fluctuating activities of the individual neurons. The transition from the memory to retrieval phase is sharp only at $\bar{T}=0$, i.e., at $C=h_0$.

(3) *Symmetric phase.* Here, $V^+\approx fV^-$ so that the correlation between the state and individual memories is small. This phase exists at all $C,\bar{T}>0$.

The above results are valid when C and $\bar{T}$ are finite as $f\to 0$. When this is not the case, finite f corrections are important. At finite f, the local fields within the on and off populations are not uniform. Two sources of fluctuations exist. One is the correlation between different bonds converging on the same neurons. These are dominant in the large f regime, $f\gg 1/\ln N$. The other source is spin-glass-like fluctuations in the values of individual bonds. They are the dominant fluctuations in the small f regime, $f\ll 1/\ln N$. The fluctuations set limits on the value of P for which retrieval phases exist. The limits are of the forms

$$P<\frac{A}{f^3|\ln f|},\quad f\gg 1/\ln N \tag{6.6}$$

and

$$P<\frac{1}{f^2}\ln\left[\frac{Nf}{|\ln f|}\right],\quad f\ll 1/\ln N \tag{6.7}$$

where the constant A depends on h_0 and K; see Eqs. (3.19)–(3.22). Finally, at very low $\bar{T}$, the spatial fluctuations freeze the thermal fluctuations of the retrieval phases, leading to local activities that are either quiescent or saturated. For finite C, this freezing sets in when $\bar{T}\approx\sqrt{f}$. This effect is particularly strong in the case of $h_0>0$.

B. Discussion

In this paper we have studied the generalized Willshaw model in the realistic case $f\gg\ln N/N$. In addition, the threshold θ was not confined to the optimal value $\theta=K-1$ which implies $h_0=0$. Under these circumstances, the model is robust, as demonstrated by our finite temperature results. However, in this regime of parameters, the capacity, given by Eqs. (6.6) and (6.7), is inferior to other models of storing sparsely coded memories in recurrent neural networks. In particular, the model of Tsodyks and Feigel'man,[17] which is also based on simple Hebb rules, has a capacity that is close to the optimal limit[16]

$$P=\frac{N}{2f|\ln f|}\ , \tag{6.8}$$

We believe that the essential difference between their model and the present one is that the Hebb rule adopted by Tsodyks and Feigel'man induces also *suppression* of the synaptic efficacies, whereas Willshaw's rule, Eq. (1.1), uses only enhancement of synaptic efficacies. In biology, both long-term potentiation[1-4] (LTP) of synapses and long-term depotentiation[3] (LTD) have been observed. The superior performance of the model of Tsodyks and Feigel'man suggests that both mechanisms are essential for an effective learning.

The numerical simulations, reported in Sec. IV, confirm qualitatively the theoretical results. However, a quantitative comparison is difficult to make, as the asymptotic limits regarding f and N, assumed in the theory, are hardly achieved in the simulations with $N=1000$ and $f=0.04$. This difficulty probably exists also in simulations of other models of sparse coding.

One of the motivations for studying this model has been the issue of local firing rates, discussed in the Introduction. Indeed, we have shown that when $h_0<0$, the off neurons are quiescent, whereas the activities of the on neurons fluctuate in time with an average given by Eq. (6.5). This average can be made small by, e.g., decreasing the magnitude of θ. Note that although θ represents a *negative* neuronal threshold, it does not imply that the *intrinsic* neural thresholds are necessarily negative. A negative threshold can be realized biologically by uniform excitatory inputs to the neurons from, e.g., other parts of the cortex or the thalamus. This excitatory input controls the mode of operation of the network. When it becomes sufficiently weak so that the effective neural threshold is positive, the network settles in a quiescent state, and memory retrieval is blocked.

In relating the present model to biology, several problems remain. First, the dynamic fluctuations in the local activities are sensitive to quenched noise. For instance, quenched fluctuations in the values of the local thresholds will cause freezing of the local activities, at temperatures that are low relative to the width of these fluctuations. This freezing implies that the neurons will be forced to be either in a quiescent state or close to saturation. Secondly, in our model, the existence of stable, partially saturated phases depends on the type of dynamics. In particular, in a fully synchronized parallel dynamics, uniform inhibition will generate oscillations with relatively high frequencies, at least *at low temperatures.* Thus it is important to study the behavior of this model in a more biologically realistic dynamics.

Finally, our analysis of the asymmetric network shows that in order to stabilize low firing rates by indirect inhibition, through *excitatory* neurons, one has to consider sizes significantly bigger than $N=1000$. It should be noted that we have assumed, in Sec. V, that both the excitatory and inhibitory synapses have the same time constant. Low firing rate can be stabilized by direct inhibition if the time constants of the inhibitory synapses are assumed to be substantially shorter relative to the excitatory ones. At present, it is unclear whether such an assumption is compatible with the available data on synaptic transmission in the cortex.[24] To conclude, understanding the function of the cortex in terms of simple recurrent neural network models is still an open challenge.

ACKNOWLEDGMENTS

We thank M. Abeles for bringing the problem of low firing rates to our attention, and for many helpful discussions. We are grateful to J. J. Hopfield for illuminating discussions, and in particular for stimulating our interest in the Willshaw model. Helpful discussions with D. J. Amit and A. Treves are acknowledged. The research is partially supported by the U.S.-Israel Binational Science Foundation

APPENDIX

In this appendix, we evaluate the correlation $\langle J_{ij}J_{ik}\rangle$, $j\neq k$, where

$$J_{ij}\equiv\Theta\left[\sum_{\mu=1}^{P}V_i^\mu V_j^\mu\right]. \tag{A1}$$

The quantity $1-\langle J_{ij}J_{ik}\rangle$ is the probability that the product $J_{ij}J_{ik}$ is zero. The probability that J_{ij} is zero is $(1-f^2)^P$, and the same for J_{ik}. The probability that both vanish is $(\mathcal{N}_{\rm prob})^P$, where $\mathcal{N}_{\rm prob}$ is the probability that both $V_i^\mu V_j^\mu$ and $V_i^\mu V_k^\mu$ are zero. Clearly, $\mathcal{N}_{\rm prob}=1-f+f(1-f)^2=1-2f^2+f^3$. Therefore

$$1-\langle J_{ij}J_{ik}\rangle=2(1-f^2)^P-(1-2f^2+f^3)^P. \tag{A2}$$

The connected correlation is

$$\begin{aligned}\langle J_{ij}J_{ik}\rangle_c&\equiv\langle J_{ij}J_{ik}\rangle-\langle J_{ij}\rangle\langle J_{ik}\rangle\\&=(1-2f^2+f^3)^P-(1-2f^2+f^4)^P.\end{aligned} \tag{A3}$$

When $f\ll 1$ this can be approximated by

$$\langle J_{ij}J_{ik}\rangle_c\approx Pf^3e^{-2Pf^2}=C^2|\ln C|f. \tag{A4}$$

For finite C, the connected correlation goes to 0 when $f\to 0$.

Using a similar method, it can be shown that

$$\langle J_{ij}J_{ik}J_{il}\rangle_c=C^3|\ln C|f^2. \tag{A5}$$

[1]S. R. Kelso, A. H. Ganong, and T. H. Brown, Proc. Natl. Acad, Sci. U.S.A. **83**, 6326 (1986).
[2]G. V. diPrisco, Prog. Neurobiol. **22**, 89 (1984).
[3]W. B. Levy, in *Synaptic Modification, Neural Selectivity and Nervous System Organization,* edited by W. B. Levy, J. A. Anderson, and S. Lehmkuhle (Lawrence Erlbaum Associates, Hillsdale, NJ, 1985).
[4]T. H. Brown, P. F. Chapman, E. W. Kairiss, and C. L. Keenan, Science **242**, 724 (1988).
[5]V. Breitenberg, *On the Texture of Brains* (Springer, Berlin,

1977); in *Brain Theory,* edited by G. Palm and A. Aersten (Springer, Berlin, 1986).
[6]A. Peters and E. G. Jones, in *The Cerebral Cortex,* edited by A. Peters and E. G. Jones (Plenum, New York, 1984), Vol. 1, Chap. 4.
[7]W. A. Little, Math. Biosci. **19**, 101 (1974).
[8]J. J. Hopfield, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982); J. J. Hopfield and D. W. Tank, Science **233**, 625 (1986).
[9]D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. A **32**, 1007 (1985); Ann. Phys. (N.Y.) **173**, 30 (1987).
[10]D. Kleinfeld and H. Sompolinsky, Biophys. J. **54**, 1039 (1988).
[11]H. Sompolinsky and I. Kanter, Phys. Rev. Lett. **57**, 2861 (1986).
[12]M. Abeles, *Local Cortical Circuits* (Springer-Verlang, Berlin, 1982); (private communication).
[13]Y. Miyashita and H. S. Chang, Nature (London) **331**, 68 (1988).
[14]J. M. Fuster and J. P. Jervey, J. Neurosci. **2**, 361 (1982).
[15]D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. A **35**, 2293 (1987).
[16]E. Gardner, J. Phys. A **21**, 257 (1988).
[17]M. V. Tsodyks and M. V. Feigel'man, Europhys. Lett. **6**, 101 (1988).
[18]M. V. Tsodyks, Europhys. Lett. **7**, 203 (1988).
[19]J. Buhman, V. Divko, and K. Schulten, Phys. Rev. A **39**, 2689 (1989).
[20]M. R. Evans, J. Phys. A **22**, 2103 (1989).
[21]D. J. Willshaw, O. P. Buneman, and H. C. Longuet-Higgins, Nature (London) **222**, 960 (1969).
[22]J. S. Griffiths, Biophys. J. **3**, 299 (1963).
[23]S. Pantilat, Ph.D. thesis, The Hebrew University, Jerusalem, 1985, (unpublished).
[24]M. Abeles, *Corticonics: Neural Circuits of the Cerebral Cortex* (Cambridge University Press, Cambridge, England, in press).
[25]G. Palm, Biol. Cybernetics **36**, 19 (1980).
[26]A. P. Thakoor *et al.,* Appl. Opt. **26**, 5085 (1987).
[27]G. Thoulouse, Commun. Phys. **2**, 115 (1977).
[28]K. Binder and A. P. Young, Rev. Mod. Phys. **58**, 801 (1986).
[29]N. Rubin and H. Sompolinsky, Europhys. Lett. **10**, 465 (1989).
[30]D. J. Amit and A. Treves, Proc. Natl. Acad. Sci. U.S.A. **86**, 7871 (1989); A. Treves and D. J. Amit, J. Phys. A **22**, 2205 (1989).
[31]R. J. Glauber, J. Math. Phys. **4**, 294 (1963).