Retrieval properties of attractor neural networks that obey Dale's law using a self-consistent signal-to-noise analysis

Anthony N. Burkitt *

Research School of Information Sciences and Engineering,
Australian National University,
Canberra, ACT 0200, Australia

Abstract. *The recently proposed self-consistent signal-to-noise analysis is applied to a current–rate dynamics attractor network of excitatory neurons with a Hebbian synaptic matrix. The effect of inhibitory interneurons is included by a term modeling their effective inhibition that depends upon both the level of activity of the excitatory neurons and the stored patterns. The low rate attractor structure is identified, and at low loading the network retrieves single patterns with uniform low rates without errors, and is stable to the admixture of additional patterns. The self-consistent signal-to-noise method enables the analysis of the network properties with an extensive number of patterns, and the results are compared with simulations. The method allows the identification of the fixed point structure of networks for which there is no Lyapanov function, and hence for which mean-field techniques cannot be used. This analysis is shown to provide a powerful and straightforward way to analyse the properties of networks with neuronal specificity, low spike rates and synaptic noise, as well as incorporating the effects of random asymmetric synaptic dilution and limited analog synaptic depth in a natural way. The simulations show that the network properties are very robust both to errors in the stimulus and to the stimulus strength and duration.*

1. Introduction

Replica symmetric mean-field theory [1, 2, 3] has been found to be a powerful method for studying the properties of attractor neural networks. It has enabled the detailed study of such network characteristics as the capacity, quality of retrieval, and robustness to noise, and can also be applied to analogue networks [4, 5, 6]. However the mean-field approach is limited in its applicability to networks for which a Lyapanov function exists. This limitation does not hold for the self-consistent signal-to-noise analysis of analogue networks by Shiino and Fukai [7], which provides a method for studying the properties of analogue neural networks that is also applicable to networks with asymmetric synaptic matrices. This method, which involves a renormalization of the signal part of the local field, reproduces the results of the replica symmetric mean-field calculations [4, 5, 6] for those networks for which such an analysis is possible.

In this paper the self-consistent signal-to-noise analysis is extended to the study of analogue networks of excitatory neurons whose connections are given by a Hebbian synaptic

*email: Anthony.Burkitt@anu.edu.au

matrix and where the effect of the inhibitory interneurons is modeled by an effective inhibition that depends upon the level of activity of the excitatory neurons. The network thus consists of *specific* neurons (i.e., the individual neurons are either excitatory or inhibitory), in accord with Dale's law, in contrast to standard Hopfield-type models, which do not possess such specificity. The inhibitory neurons, whose effects are incorporated by an effective inhibition term, serve to control the activity of the excitatory neurons through a feedback process. The details of the model, in which the neuronal dynamics is described in terms of the firing rate and afferent current of each neuron, is given in the following section. The particular form of the inhibition that is chosen is contrasted with the effective inhibition of Hopfield-type models. The self-consistent signal-to-noise method as applied to analogue networks is presented in section 3. This is compared with the mean-field approach in section 3.2. The solution of the fixed-point equations is presented in section 4, both for the case of finitely and extensively many patterns. The effects of random asymmetric synaptic dilution and limited analog synaptic depth can be incorporating in the self-consistent signal-to-noise analysis in a natural way, as is shown in section 5. The results of numerical simulations are presented in section 6, and section 7 contains the discussion.

2. The model

The neural dynamics is described in terms of two continuous variables: the firing rate ν_i and the afferent current I_i of each (excitatory) neuron:

$$\mathcal{T}\frac{dI_i(t)}{dt} = -I_i(t) + \sum_{j(j\neq i)}^{N} \mathcal{T} J_{ij}\nu_j(t) - B\left(\{\eta_j^\mu\}, \{\nu_j(t)\}\right) \tag{1}$$

where $\mathcal{T}$ is the time decay constant, J_{ij} is the synaptic matrix, N is the number of neurons, and $\{\eta_j^\mu\}$ are the stored patterns. The function B represents the effective inhibition of the inhibitory interneurons. Amit and Tsodyks [8, 9] (hereafter referred to as ATI and ATII respectively) investigated a network of noisy integrate-and-fire neurons and showed that the description in terms of individual spikes can be converted into an effective description in terms of current dynamics driven by spike rates. The afferent current induces a spike rate described by a neural gain function, thus providing a closed set of dynamical equations.

The form of the neural gain function that Amit and Tsodyks (ATI) found is given by:

$$\phi(I) = \left\{1 + \tau\frac{\sqrt{\pi}}{\sigma}\int_0^\theta dz \exp\left(\frac{(z-I-\mu)^2}{\sigma^2}\right)\left[\operatorname{erf}\left(\frac{z-I-\mu}{\sigma}\right)+1\right]\right\}^{-1} \tag{2}$$

where τ is the time constant for the membrane depolarization, θ is the threshold, and μ and σ are the mean and RMS of the afferent noise. These two equations provide a complete description of the dynamics of the network that is deterministic at the level of rates and currents. They showed that a network described by this current–rate dynamics can lead to stationary (reverberating) activity states with low spike rates and that the various approximations involved in this description are quite realistic in typical cortical conditions [8].

The synaptic matrix is chosen as:

$$J_{ij} = \frac{1}{\mathcal{T}Nf}\sum_\mu^p \eta_i^\mu\eta_j^\mu \tag{3}$$

where the η_i^μ, $\mu = 1,..,p$ are the random N-bit memorized patterns on the N excitatory neurons, of which Nf were ones (foreground) and $N(1-f)$ were zeros (background), and where f is the sparsity of the patterns. Information in this model is stored only on the synapses between excitatory neurons, in accord with the common view in neuroanatomy [10].

The effect of the inhibitory neurons upon the excitatory neurons is given by the function B, which may be described in terms of the activation of the excitatory neurons and parameterized in the following way:

$$\begin{aligned} B\left(\{\eta_j^\mu\},\{\nu_j\}\right) &= \sum_\mu \left\{ \eta_i^\mu B_1\left(\frac{1}{N}\sum_j^N \nu_j\right) + f B_2\left(\frac{1}{N}\sum_j^N \frac{\eta_j^\mu}{f}\nu_j\right) \right. \\ &\qquad \left. \eta_i^\mu B_3\left(\frac{1}{N}\sum_j^N \frac{\eta_j^\mu}{f}\nu_j\right) + f B_4\left(\frac{1}{N}\sum_j^N \nu_j\right) \right\} \end{aligned} \tag{4}$$

In general, the inhibition could also consist of higher-order combinations of these terms, as investigated, for example, by Amit and Treves [11, 12, 13].

The Hopfield model [1], which usually allows neurons to have both excitatory and inhibitory synaptic connections, may be interpreted as having Hebbian excitatory synaptic connections together with an effective inhibition. The form of the effective inhibition $B_{Hopfield}$ in the case where the Hopfield model is formulated for sparse patterns that take the values $\{0,1\}$ [9] is:

$$B_{Hopfield}\left(\{\eta_j^\mu\},\{\nu_j\}\right) = \sum_\mu \left\{ \eta_i^\mu \frac{1}{N}\sum_j^N \nu_j + f\frac{1}{N}\sum_j^N \frac{\eta_j^\mu}{f}\nu_j - f\frac{1}{N}\sum_j^N \nu_j \right\} \tag{5}$$

The third term on the right of this expansion is an excitation and may be considered as either a non-Hebbian contribution to the excitatory synaptic matrix or as the effect of mutual inhibition among the inhibitory interneurons. The Hopfield network thus exhibits a highly structured effective inhibition, involving a linear combination of both pattern-specific and non-specific contributions, which has the advantage that the overall effective synaptic matrix is symmetric and therefore enables mean-field calculations to be performed, as discussed further in section 3.2.

A number of neurophysiological studies have been carried out to investigate the relationship between long-term potentiation (LTP) and the activity of interneurons, and in particular the enhancement of interneuronal output during LTP [14, 15, 16]. A recent study by Ouardouz and Lacaille [17] seems to indicate that LTP of excitatory synapses may selectively occur in hippocampal interneurons, and that this LTP resembles that found in hippocampal CA1 pyramidal cells. It is therefore of considerable interest to investigate the effect of different forms of effective inhibition. The particular type of inhibition investigated here is that given by:

$$B\left(\{\eta_j^\mu\},\{\nu_j\}\right) = \sum_\mu \eta_i^\mu \frac{1}{N}\sum_j^N \nu_j \tag{6}$$

Such a form of effective inhibition could possibly arise as a result of LTP involving inhibitory interneurons, however a study of the mechanisms by which it could be produced lies outside the scope of the present investigations. This particular choice of effective inhibition is both simple and effective in facilitating the network retrieval of the stored patterns, as will be shown. The purpose of choosing this particular form of effective inhibition is to investigate, using The self-consistent signal-to-noise analysis, a simple form of the effective inhibition that can not be investigated by mean-field techniques. The self-consistent signal-to-noise analysis can likewise be used to study the generalized Willshaw model [18]. It is also of interest to consider the case when the inhibition contains random noise, which may be done using this approach (section 5).

Using the usual notation [9], the arrival of the network at an attractor is characterized by a set of neurons remaining quiescent while the remaining neurons fire at a substantially higher rate for some macroscopic time. The quality of retrieval of the patterns is characterized by the similarity of the firing rates ν_i to the stored patterns η_i^μ:

$$\begin{aligned} m_\mu^+ &= \frac{1}{fN}\sum_i^N \eta_i^\mu \langle \nu_i \rangle \\ m_\mu^- &= \frac{1}{(1-f)N}\sum_i^N (1-\eta_i^\mu)\langle \nu_i \rangle \\ m_\mu &= m_\mu^+ - m_\mu^- = \frac{1}{f(1-f)N}\sum_i^N (\eta_i^\mu - f)\langle \nu_i \rangle \end{aligned} \tag{7}$$

where m_μ^+ and m_μ^- are the mean firing rates of the neurons in the foreground and background respectively. In addition to m_μ there is also the average firing rate of the neurons:

$$m_0 = \frac{1}{N}\sum_i^N \langle \nu_i \rangle \tag{8}$$

The rates are measured as the number of spikes per absolute refractory period, which can be expressed in units of spikes s^{-1} by multiplying by $1/\tau \approx 500 s^{-1}$ with the absolute refractory period τ of 2 ms.

The fixed-points of the dynamical equation for the neurons are given by:

$$I_i = \sum_j^N \mathcal{T} J_{ij} \nu_j - B\left(\{\eta_j^\mu\}, \{\nu_j\}\right) \tag{9}$$

The associative storage of the patterns and their recall is then a result of the cooperative dynamics of the network, in which the activity of the excitatory neurons is controlled by the inhibitory neurons through a feedback process. One of the most important questions in this formulation is the domain of validity of the fixed points. The absence of a Lyapanov function, except for particular forms of the effective inhibition such as $B_{Hopfield}$, means that a mean-field analysis is not possible and the network may exhibit a rich dynamical behaviour including oscillations and chaotic behaviour. However, studies involving the asymmetrical dilution of the Hopfield model have shown that the fixed points are remarkably robust to the introduction of the asymmetry [3], particularly at low loadings. In order to examine this issue more closely, numerical simulations have been carried out (section 6).

3. Self-consistent signal-to-noise analysis

The basis of this analysis is the systematic splitting of the I_i into signal and noise parts [7]. Using the form of effective inhibition, equation(6), the local current I_i may be expressed in terms of the overlaps as:

$$I_i = \sum_\mu \eta_i^\mu \left(m_\mu^+ - m_0\right) - \alpha\phi(I_i) = (1-f)\sum_\mu \eta_i^\mu m_\mu - \alpha\phi(I_i) \tag{10}$$

which will be solved to give:

$$\nu_i = \phi(I_i) = \hat{\phi}\left((1-f)\sum_\mu \eta_i^\mu m_\mu\right) \tag{11}$$

where the argument of $\hat{\phi}$ is denoted as $\hat{I}$. A systematic split of $\hat{I}$ into signal and noise components is now made. We are interested in retrieval solutions in which $m_\mu = O(1)$ for the condensed patterns $\mu = 1, ..., p_0$. The noise term comes from the sum over the μ-indices for the $p - p_0$ uncondensed patterns (hereafter labeled $\tilde{\mu}$) for which $m_{\tilde{\mu}} = O(1/\sqrt{N})$. Expanding $\hat{I}$ up to first order in $m_{\tilde{\mu}}$:

$$\hat{I}_i = (1-f)\sum_\mu \eta_i^\mu m_\mu + (1-f)\sum_{\tilde{\mu}} \eta_i^{\tilde{\mu}} m_{\tilde{\mu}} \tag{12}$$

Expanding $\hat{\phi}$ to first order in $m_{\tilde{\mu}}$:

$$m_{\tilde{\mu}} = \frac{1}{f(1-f)N}\sum_i (\eta_i^{\tilde{\mu}} - f)\hat{\phi}(\hat{I}_i(\tilde{\mu})) + \frac{(1-f)}{fN}\sum_i \eta_i^{\tilde{\mu}} m_{\tilde{\mu}} \hat{\phi}'(\hat{I}_i(\tilde{\mu})) \tag{13}$$

where the notation $\hat{I}_i(\tilde{\mu})$ indicates that the argument is that in equation(12) without the $\tilde{\mu}$ contribution. Replacing $1/N\sum_j ...$ by an average $\langle ... \rangle$ over the η's, this becomes:

$$m_{\tilde{\mu}} = \frac{1}{K}\left\{\frac{1}{f(1-f)N}\sum_i (\eta_i^{\tilde{\mu}} - f)\hat{\phi}(\hat{I}_i(\tilde{\mu}))\right\} \tag{14}$$

with

$$K = 1 - (1-f)\langle \hat{\phi}'(\hat{I}_i(\tilde{\mu}))\rangle \tag{15}$$

Substituting this into the expression for I_i:

$$I_i = (1-f)\sum_\mu \eta_i^\mu m_\mu - \alpha\phi(I_i) + \frac{(1-f)}{KfN}\sum_{\tilde{\mu}} \eta_i^{\tilde{\mu}} \hat{\phi}(\hat{I}_i(\tilde{\mu})) + \bar{z} \tag{16}$$

where $\bar{z}$ is given by

$$\bar{z} = \frac{1}{KfN}\sum_{\tilde{\mu}} \eta_i^{\tilde{\mu}} \sum_{j(j\neq i)} (\eta_j^{\tilde{\mu}} - f)\hat{\phi}(\hat{I}_j(\tilde{\mu})) \tag{17}$$

The term $\hat{\phi}(\hat{I}_i(\tilde{\mu}))$ can now safely be replaced by $\phi(I_i)$. As in [7], the term $\bar{z}$ is a sum of almost uncorrelated random variables, with $\langle \bar{z} \rangle = 0$ and

$$\langle \bar{z}^2 \rangle = \frac{\alpha(1-f)}{K^2}\langle \phi(I_i)^2 \rangle \tag{18}$$

which enables this noise contribution to be written as a Gaussian distribution with zero mean and variance $\sigma^2 = \langle \bar{z}^2 \rangle$:

$$D(\bar{z}) = \frac{d\bar{z}}{\sqrt{2\pi}\sigma}\exp\frac{-\bar{z}^2}{2\sigma^2} \tag{19}$$

The local afferent current $Y \equiv \phi(I)$ is therefore implicitly determined as a function of $\bar{z}$ by

$$Y_\eta(\bar{z}) \quad = \quad \phi\left((1-f)\sum_\mu \eta^\mu m_\mu - \alpha Y_\eta(\bar{z}) + \alpha\frac{(1-f)}{K}Y_\eta(\bar{z}) + \bar{z}\right) \tag{20}$$

where the μ-sum is over the uncondensed patterns. The expressions for the overlaps then become

$$\begin{aligned} m_\mu &= \frac{1}{f(1-f)}\langle(\eta^\mu - f)Y_\eta(\bar{z})\rangle_{\eta,\bar{z}} \\ m_0 &= \langle Y_\eta(\bar{z})\rangle_{\eta,\bar{z}} \end{aligned} \tag{21}$$

with

$$\begin{aligned} \sigma^2 &= \alpha\frac{(1-f)}{K^2}\langle Y_\eta^2(\bar{z})\rangle_{\eta,\bar{z}} \\ K &= 1-(1-f)\langle\frac{d}{d\bar{z}}Y_\eta(\bar{z})\rangle_{\eta,\bar{z}} \end{aligned} \tag{22}$$

Introducing the change of variables:

$$\sqrt{\alpha r} \quad = \quad \sigma \quad , \quad z = \frac{\bar{z}}{\sigma} \tag{23}$$

$$q \quad = \quad \frac{K^2\sigma^2}{\alpha(1-f)} \quad , \quad U = \frac{1-K}{1-f} \tag{24}$$

a more familiar form of the equations is obtained:

$$\begin{aligned} m_\mu &= \frac{1}{f(1-f)}\langle(\eta^\mu - f)Y_\eta(z)\rangle_{\eta,z} \\ m_0 &= \langle Y_\eta(z)\rangle_{\eta,z} \\ q &= \langle Y_\eta(z)^2\rangle_{\eta,z} \\ U\sqrt{\alpha r} &= \langle zY_\eta(z)\rangle_{\eta,z} \end{aligned} \tag{25}$$

where $Y_\eta(z)$ is given by the implicit relationship

$$Y_\eta(z) = \phi\left((1-f)\sum_\mu \eta^\mu m_\mu + \alpha\Gamma Y_\eta(z) + z\sqrt{\alpha r}\right) \tag{26}$$

with

$$\begin{aligned} \Gamma &= \frac{(1-f)U - f}{1-(1-f)U} \\ r &= q\frac{(1-f)}{(1-(1-f)U)^2} \end{aligned} \tag{27}$$

3.1. Including $1/p$ corrections

It is straightforward to include $1/p$ corrections in this approach. The choice of inhibition above, namely $B_1(x) = x$ may be extended as:

$$B_1(x) = x + \frac{1}{fp} b_1(x) \tag{28}$$

The analysis proceeds as before, but with the additional term in the expression for $\hat{I}_i$, equation (12):

$$\hat{I}_i = (1-f)\sum_{\mu} \eta_i^{\mu} m_{\mu} + (1-f)\sum_{\tilde{\mu}} \eta_i^{\tilde{\mu}} m_{\tilde{\mu}} - b_1(m_0) \tag{29}$$

The implicit equation for the local afferent current, equation (26), becomes:

$$Y_{\eta}(z) = \phi\left((1-f)\sum_{\mu} \eta^{\mu} m_{\mu} - b_1(m_0) + \alpha\Gamma Y_{\eta}(z) + z\sqrt{\alpha r}\right) \tag{30}$$

Such a term may serve to help regulate the functioning of the network and dampen large fluctuations.

3.2. Comparison with mean-field theory

The Hopfield-type inhibition, equation (5), allows the network dynamics to be written with a synaptic matrix of the form

$$J_{ij}^{H} = \frac{1}{\mathcal{T}Nf}\sum_{\mu}^{p} (\eta_i^{\mu} - f)(\eta_j^{\mu} - f) \tag{31}$$

The dynamical equation for $I_i(t)$ (equation 1) can be written in the form

$$\mathcal{T}\frac{dI_i(t)}{dt} = -\frac{\partial H}{\partial \nu_i} \tag{32}$$

with the straightforward Lyapanov function

$$H = -\frac{\mathcal{T}}{2}\sum_{i,j}^{N} J_{ij}^{H} \nu_i \nu_j + \sum_{j}^{N} G(\nu_j) \tag{33}$$

where $G(\nu) = \int_0^{\nu} I(\nu')d\nu'$. The model can therefore be analysed using mean-field techniques [2, 4, 7]. The free energy and the partition function are defined as usual by:

$$F_N(\beta) = -\frac{1}{\beta N}\ln Z \quad , \quad Z = \int \prod_{i=1}^{N} d\nu_i \exp(-\beta H) \tag{34}$$

where the ν-integrals are over the interval $\{0,1\}$. This is exactly the model investigated by Amit and Tsodyks [9], apart from a factor of $(1-f)$ in the synaptic matrix J_{ij}^{H}. The free energy per neuron may be evaluated by the replica method and the minima are found by taking the limit $\beta \to \infty$, in which ν is determined as the solution of an implicit equation. The resulting expressions [9] are exactly those obtained by the self-consistent signal-to-noise analysis.

4. Solution of the fixed-point equations

4.1. *Fixed-point equations for $\alpha \to 0$*

The equations for the fixed-point take a particularly simple form when the number of patterns p is kept fixed, as N increases (i.e., $\alpha = p/N \to 0$ as $N \to \infty$) since in this limit ν_η becomes independent of z:

$$\nu_\eta = \phi\left((1-f)\sum_\mu \eta_\mu m_\mu\right) \tag{35}$$

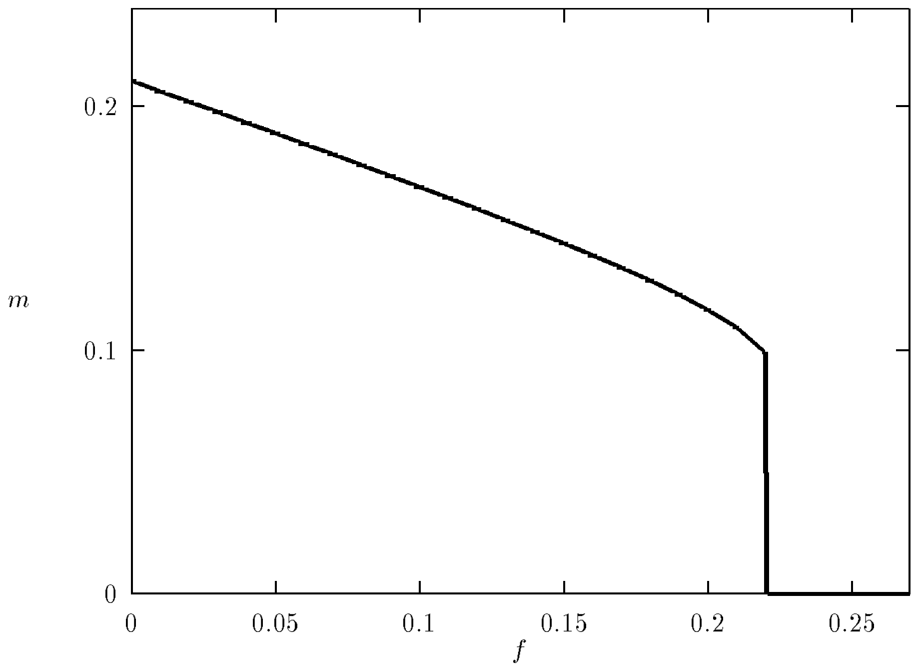

Figure 1: The overlap m for the retrieval of a single pattern as a function of the sparseness f, for $\alpha = 0$. The overlap gives the firing rate as the number of spikes per absolute refractory period.

Consider first the retrieval of a single pattern (i.e., with $m^\mu = m\delta^{\mu,1}$):

$$m = \phi((1-f)m) - \phi(0) \tag{36}$$

This expression simply represents the difference between the (uniform) rates of the neurons in the foreground (the first term) and in the background (the $\phi(0)$ term). The argument of the first term on the righthand side is the afferent current in the foreground, which is uniform across the foreground neurons. The neurons in the background receive no afferent current and have a firing rate determined solely by the level of afferent noise.

The resulting value of the overlap m for the retrieval of a single pattern for the gain function, equation (2), is plotted in figure 1 as a function of the sparsity f, with neural parameters as used in ATII (i.e., $\tau = 4.0$, $\theta = 0.25$, $\nu = 0.2$, and $\sigma = 0.0125$). As in the case examined in ATII, when the network retrieves a stored pattern (with $m \neq 0$)

there are essentially no retrieval errors, i.e., all the neurons that correspond to ones in the pattern have a rate of firing equal to m, while those corresponding to zeros do not fire (because the current afferent on a background neuron is negative and it therefore induces a negligible firing rate). Likewise, at the values of the neural parameters used here there are no simple, symmetric spurious states consisting of mixtures of two or three patterns.

4.2. Fixed-point equations in the $p \to \infty$ limit

The self-consistent signal-to-noise analysis expressions, equations (26), for $\alpha > 0$ may be solved numerically and the results are presented here for the same set of parameter values as used previously. The mean firing rates for the neurons in the foreground (m_+) and background (m_-) are plotted in figure 2 as a function of the loading α. The mean rate m_+ for neurons in the foreground of the retrieved pattern remains at a relatively constant value (approx. 0.19 for this particular set of neural parameters, corresponding to a firing rate of 95 spikes/sec) up to a loading of approx. 0.10, and the mean rate m_- of the background neurons remains extremely low. For larger values of α the rate m_+ falls and m_- increases, with the two rates coalescing as in ATII. The results indicate that there is a clearly defined critical capacity at which there is a sharp transition between a retrieval phase and a non-retrieval phase. Figure 2 contains the results for a value $f = 0.05$ of the sparseness and is representative of the results for $f \leq 0.22$ (the fixed-point equations are computationally intensive to solve).

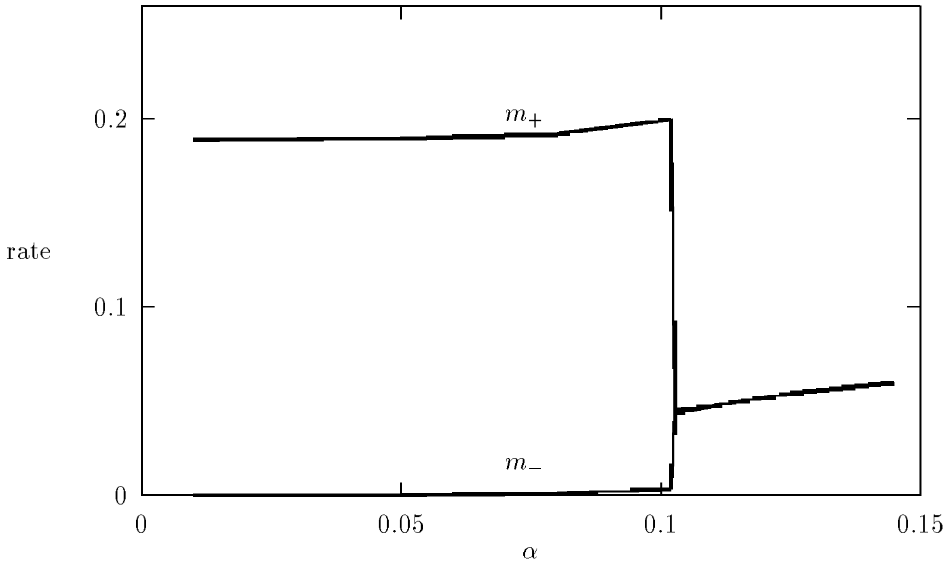

Figure 2: The self-consistent signal-to-noise solutions for the firing rates of neurons in the foreground (m_+) and background (m_-) versus α for $f = 0.05$. Neural parameters are as in ATII (an overlap of $m = 0.2$ corresponds to a firing rate of 100 spikes/sec).

5. Synaptic noise and synaptic dilution

It is straightforward within the framework of the self-consistent signal-to-noise analysis to consider the deviation of the synapses from their Hebbian values. Assuming that such deviations are random, the synaptic efficacies have the form

$$T_{ij} = J_{ij} + \epsilon_{ij} \tag{37}$$

where J_{ij} is the standard Hebbian synaptic matrix and ϵ_{ij} is a random variable which is assumed to be uncorrelated with the stored patterns. These random variables are assumed to have a Gaussian distribution with mean and mean square deviation given by:

$$[\epsilon_{ij}] = 0 \qquad , \qquad [\epsilon_{ij}^2] = \frac{\epsilon^2}{N} \tag{38}$$

where the square brackets denote averaging over the distribution of the noise. The resulting change in the afferent current is

$$I_i = (1-f)\sum_{\mu} \eta_i^{\mu} m_{\mu} - \alpha\phi(I_i) + \sum_{j(j\neq i)} \epsilon_{ij}\nu_j \tag{39}$$

The analysis proceeds exactly as before, except that the expression for the noise term, $\bar{z}$, contains the additional term:

$$\bar{z}_{new} \quad = \quad \bar{z}_{old} + \sum_{j(j\neq i)} \epsilon_{ij}\phi(I_j) \tag{40}$$

and thus

$$\langle \bar{z}^2 \rangle_{new} \quad = \quad \langle \bar{z}^2 \rangle_{old} + \epsilon^2 \langle \phi(I)^2 \rangle \tag{41}$$

The effect of this new source of noise in the network can therefore be incorporated into a new noise distribution with Gaussian width:

$$\sigma_{new}^2 = \sigma_{old}^2 + \epsilon^2 q \tag{42}$$

The self-consistent signal-to-noise analysis equations (26) remain unchanged, while the equation for $Y_\eta(z)$ becomes

$$Y_\eta(z) = \phi\left((1-f)\sum_{\mu} \eta^{\mu} m_{\mu} + \alpha\Gamma Y_\eta(z) + z\sqrt{\alpha r + \epsilon^2 q}\right) \tag{43}$$

In the case where mean-field theory can be applied (i.e., using $B_{Hopfield}$) the result agrees exactly with that obtained by Sompolinsky [19]. The straightforward derivation of the result here is in stark contrast to the *tour de force* of the mean-field calculation. Nevertheless, all the results of Sompolinsky's analysis can be applied to the present model to enable the study of clipped synapses and dilution of synapses. It is very pleasing that these effects, which are important for the biological plausibility of the model, can be incorporated into the self-consistent signal-to-noise analysis in such a straightforward manner and that the results of the Sompolinsky analysis can be used without the necessity of the complex associated derivation.

6. Numerical simulations

The numerical simulations of the network dynamics serve a number of important purposes. Firstly it serves to check the numerical values obtained by the fixed-points of the self-consistent signal-to-noise analysis. Secondly it serves to check the finite N effects in the analysis, which is only exact in the $N \to \infty$ limit. Thirdly, and perhaps most importantly for the system being analysed here, it enables the stability of the fixed-point to be investigated, since there is no longer a Lyapunov function guaranteeing the convergence of the network dynamics to a fixed-point solution.

Since the equations (1) being simulated are completely deterministic differential equations in which the neuronal noise is incorporated into the gain function, equation (2), the simulations are really determined by the initial conditions and the synaptic couplings. The initial conditions are determined by the stimulus, which consisted of incoming external currents of intensity 0.2 presented during the first ten steps of the integration. The simulations were carried out on a network of 1000 neurons with a varying number of stored patterns, from 10 to 800, $\alpha = 0.01 - 0.80$, and with a sparsity of $f = 0.05$. The neural parameters were those used in ATII, i.e., $\tau = 4.0$, $\theta = 0.25$, $\nu = 0.2$, and $\sigma = 0.0125$, and likewise the same integration step was used (0.2 of the time constant of the neuron).

One of the stored patterns was chosen as the stimulus, and an external current was imposed upon the foreground neurons of this pattern. As in ATII, at each storage level simulations were carried out with five sets of patterns, and for each set it was run five times, each with a stimulus based on a different stored pattern. A number of additional simulations were also carried out to investigate the error correction, the effect of $1/p$ corrections, and the stability of the network to variation of the stimulus strength and duration, as well as other parameters, and the results of these will be discussed shortly.

The most striking feature of the simulations was that the network reached a stable state in every simulation carried out, with no chaotic or oscillatory trajectories being observed. Moreover, when the network retrieved a single pattern, the convergence rate was very rapid. At large loadings the network tended to retrieve an increasing proportion of mixture states, in which not only the overlap with the stimulus pattern was high, but also the overlap with one (or more) other patterns tended to increase.

The results of the simulations agree with the self-consistent signal-to-noise analysis results for small loadings. The simulations, however, do not show the discontinuous jump in the retrieval behaviour that the self-consistent signal-to-noise analysis results show as the loading is increased above $\alpha_c \simeq 0.10$. The results of the numerical simulations for the overlap of the retrieved patterns is plotted in figure 3 for a number of different loadings α. Also plotted (dashed curve) is the percentage of stimuli for which the corresponding stored pattern was successfully retrieved. The variation of the rate is also indicated by the error bars, which denote the standard deviation of the results. As observed in ATII, there is no sharp transition at the loading α_c as in figure 2, rather there is a gradual decrease in the fraction of singly retrieved patterns. However, in contrast to the picture of ATII, the stimuli for which there is no single retrieved pattern almost invariably result in mixture states, in which the overlap with two (and sometimes more) patterns increases to the level associated with the target retrieved pattern.

The question that then arises is whether this difference in the results above $\alpha_c \simeq 0.10$ between the self-consistent signal-to-noise analysis and the numerical simulations is due to the finite size of the lattices used in the simulations. Since the self-consistent signal-

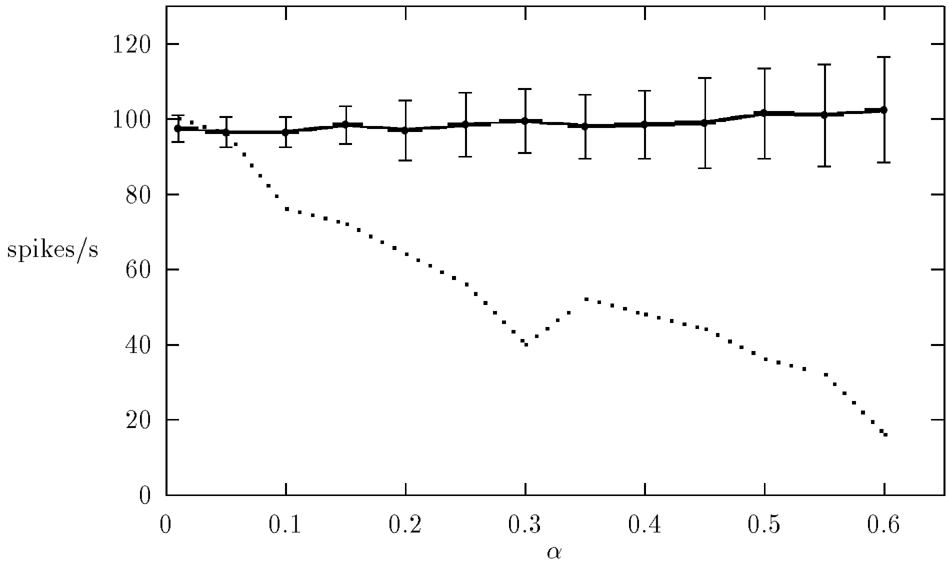

Figure 3: Results from the numerical simulation of the network with $f = 0.05$ and neural parameters as given in the text. The error bars show the standard deviation in the rates at each α. The dashed line shows the percentage of stimuli which resulted in the successful retrieval of the corresponding single stored pattern (a firing rate of 100 spikes/sec corresponds to a an overlap of $m = 0.2$).

to-noise analysis indicates that there is a first-order transition to a non-retrieving state as α is increased above $\alpha_c \simeq 0.10$, is would be expected that the distribution of final overlaps (m) would show a double peak structure, as seen in [2]. Numerical simulations at $\alpha = 0.15$ with lattice sizes ranging from $N = 500$ to $N = 2000$, as plotted in figure 4, failed to show any evidence of such a two-peaked structure. Moreover, the fraction of simulations for which a single pattern was successfully retrieved also increased from 0.55 for $N = 500$ to 0.67 for $N = 1000$ and to 0.76 for $N = 2000$. Further, a comparison of the systematic finite size effects at $\alpha = 0.05$ and $\alpha = 0.15$ failed to reveal any evidence of a discontinuity in the overlap between these two values of the loading. These results suggest that the self-consistent signal-to-noise analysis does not capture the retrieval properties of the network above $\alpha_C \simeq 0.10$. In the study ATII they also did not observe the sharp transition predicted by mean-field theory between a retrieval phase and a freezing phase. They postulated that the discrepancy may be due to finite-size effects, but did not study the question systematically. As with the results presented here, they observed a gradual decrease in the fraction of retrieved stimuli as the loading α increases.

The sensitivity of the results to the distribution of ones in the patterns has also been investigated. As pointed out in ATII, when the sparsity f is taken as the mean fraction of a random distribution (rather then the fixed fraction of ones in the patterns, as done here and in ATII), this may introduce strong finite-size effects. In the simulations it was observed that this led to an increased number of runs resulting in either mixture states or the quiescent state.

The network retrieval was fairly insensitive to the amplitude and duration of the afferent stimulus–current, both of which were varied over a range of values without appreciable

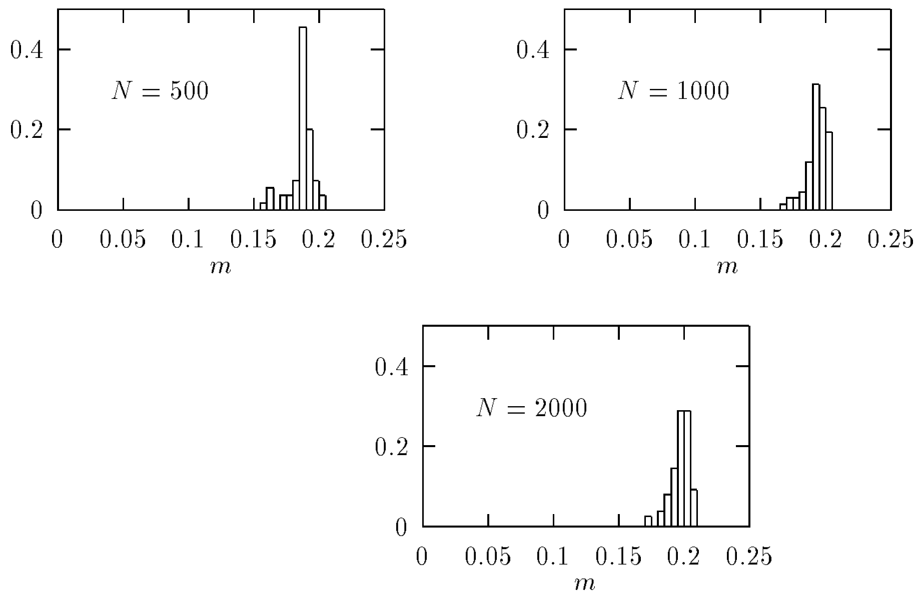

Figure 4: Histogram of the fraction of simulations that retrieve a single pattern as a function of the resulting overlap for $\alpha = 0.15$ with 500, 1000 and 2000 neurons.

effect upon the results. The retrieval was also rather insensitive to errors in the stimulus and simulations with varying levels of noise in both the foreground and background neurons were carried out. The results were very similar to those of ATII: the foreground neurons which received no external afferent current nevertheless increased activation, and background neurons which received an afferent current soon reduced their activation after the external afferent current ceased.

7. Discussion

The self-consistent signal-to-noise analysis provides a new method for analysing the behaviour of networks of neurons that have asymmetric synaptic matrices. The replica symmetric mean-field approach can only be applied to networks with symmetric synaptic matrices, although the two methods result in identical equations for such network models. The advantage of the self-consistent signal-to-noise analysis is that it enables the investigation of a much wider class of networks to which the replica method can not be applied, and in particular models that incorporate more biologically desirable features. Moreover, it provides a very different framework to the replica method, and one that is more readily comprehensible in terms of the origin and interpretation of the resulting variables. The analysis presented in this study is of a network in which the neuronal dynamics is described by the afferent current and firing rate of each neuron and which consists of coupled excitatory and inhibitory neurons, i.e., a network which obey Dale's law. The excitatory neurons are connected by a Hebbian synaptic matrix and the effect of the inhibitory interneurons is modeled by an effective interaction that depends upon both the level of activity of the excitatory neurons and the stored patterns. This generalizes the standard Hopfield-type of synaptic matrix for sparse patterns, which is widely used in mean-field network studies but which in effect incorporates a very particular form of inhibition that is motivated more by mathematical than biological considerations.

The results of the present study using an asymmetric form of the effective synaptic matrix show many similarities to the mean-field results of ATII: the network stabilizes selective low-rate attractors, the simulation results agree with theory for low loading, and the simulation results depart from theoretical predictions above saturation. The two sets of results also show an insensitivity to details of the stimulus and the ability to correct errors in the stimulus. Both approaches face a number of similar questions, such as the inherent adequacy of the current–rate representation, the level of noise in the gain function that is tolerable for this analysis, and the question of the minimal rates that make sense within this framework. However, there are a number of differences between the two sets of results, including the somewhat higher rates in the problem studied here and the appearance of mixture states at large loadings.

The approach adopted here of considering the effective interaction rather than modeling the dynamics of the inhibitory interneurons directly has been adopted in a number of studies [11, 12, 20]. This approach neglects the time course of the interneurons and assumes that the inhibitory interneurons operate very fast, so that their activity instantaneously reflects the activity of the excitatory neurons. Although this clearly involves an approximation to the neurophysiological situation, it may nevertheless partially capture the shorter transmission times for inhibitory signals that results from the more local nature of inhibitory neurons.

The spread of activation that was observed at large loading is a not unexpected consequence of the type of inhibition considered here. This effect may be counterbalanced by an increase in the inhibition, and two methods for incorporating such enhanced inhibition into this model may be adopted. The first method is to consider introducing additional terms in the effective inhibition, such as a non-linear contribution or a pattern non-specific inhibitory feedback, i.e., the term B_4 in equation (5) that depends only upon the average level of activation of the excitatory neurons. The second method is via $1/p$ corrections, as outlined in section 3.1, and preliminary simulations in which $b_1(m_0)$ is taken as linear do indeed indicate that there is both a significantly reduction in the proportion of mixture states and the resultant stable retrieval states have rates below those presented above (and closer to the values of ATII). More detailed investigations of these and related questions are currently under way.

It has been shown that it is straightforward within the framework of the self-consistent signal-to-noise analysis to consider the deviation of the synapses from their Hebbian values. This was done by assuming that the deviations are random, and that they may be therefore included as a new source of noise whose effect is to increase the width of the Gaussian distribution of the noise. This results in exactly the equations obtained by Sompolinsky [19], thus enabling the study of clipped synapses and dilution of synapses, and the straightforward derivation of the result here contrasts with that of the replica symmetric mean-field calculation. Such calculations are currently under way and will be compared with the results of numerical simulations. Likewise, the study of the effect of eliminating synapses is being investigated. Other features that deserve further investigation are related to finite size effects, such as the sensitivity to the number of neurons in the pattern, fluctuations in the rates and the stability of the retrieved patterns.

An important feature of the results presented here is that the effective inhibition considered here, which results in an asymmetric effective synaptic matrix, does not appear to introduce additional types of attractors, such as periodic or chaotic attractors. Thus the fixed-point equations of the self-consistent signal-to-noise analysis provides an accurate analysis of the network properties at low loadings, and it can guide our understanding of the network properties at higher loadings. This enables more realistic networks to be investigated, not only using numerical simulations but also maintaining some analytic control of the network properties.

Acknowledgement

I am indebted to A. Coop for the references to work concerning the relationship between long-term potentiation and the activity of interneurons. The helpful comments by the referees are also acknowledged.

References

[1] J.J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities", *Proc. Natl. Sci.* **79** (1982) 2554–2558.

[2] D.J. Amit, H. Gutfreund, and H. Sompolinsky, "Statistical mechanics of neural networks near saturation", *Ann. Phys., NY* **173** (1987) 30–67.

[3] D.J. Amit, *Modeling Brain Function: The world of attractor neural networks*, (Cambridge University Press, 1989).

[4] R. Kühn, "Statistical mechanics for networks of analog neurons", In I. Garrido, *Statistical mechanics of neural networks*, (Springer, Berlin, 1990).

[5] R. Kühn, S. Bös, and J.L. van Hemmen, "Statistical mechanics for networks of graded-response neurons", *Phys. Rev.* **A43** (1991) 2084–2087.

[6] M. Shiino and T. Fukai, "Replica-symmetric theory of the nonlinear analogue neural networks", *J. Phys.* **A23** (1990) L1009–L1017.

[7] M. Shiino and T. Fukai, "Self-consistent signal-to-noise analysis and its application to analogue neural networks with asymmetric connections", *J. Phys.* **A25** (1992) L375–L381.

[8] D.J. Amit and M.V. Tsodyks, "Quantitative study of attractor neural network retrieving at low spike rates: I.", *Network* **2** (1991) 259–273.

[9] D.J. Amit and M.V. Tsodyks, "Quantitative study of attractor neural network retrieving at low spike rates: II.", *Network* **2** (1991) 275–294.

[10] E.R. Kandel and J.R. Schwartz, *Principles of Neural Science*, (Elsevier, Amsterdam, 1985).

[11] D.J. Amit and A. Treves, "Associative memory neural network with low temporal spiking rates", *Proc. Natl. Acad. Sci.* **86** (1989) 7871–7875.

[12] A. Treves and D.J. Amit, "Low firing rates: an effective Hamiltonian for excitatory neurons", *J. Phys.* **A22** (1989) 2205–2226.

[13] A. Treves, "Threshold-linear formal neurons in auto-associative nets", *J. Phys.* **A23** (1990) 2631–2650.

[14] H.L. Haas and G. Rose, "Long-term potentiation of excitatory synaptic transmission in the rat hippocampus: the role of inhibitory processes", *J. Physiol.* **329** (1982) 541–552.

[15] J.S. Taube and P.A Schwartzkroin, "Intracellular recording from hippocampal CA1 interneurons before and after development of long-term potentiation", *Brain Research* **419** (1987) 32–38.

[16] A. Stelzer, G. Simon, G. Kovacs, and R. Rai, "Synaptic disinhibition during maintenance of long-term potentation in the CA1 hippocampal subfield", *Proc. Natl. Acad. Sci.* **91** (1994) 3058–3062

[17] M. Ouardouz and J.-C. Lacaille, "Mechanisms of selective long-term potentation of excitatory synapses in stratum oriens/alveus interneurons of rat hippocampal slices", *J. Neurophysiol.* **73** (1995) 810–819.

[18] A.N. Burkitt, "Solution of the generalized Willshaw model using the self-consistent signal-to-noise analysis", in preparation.

[19] H. Sompolinsky, "Neural networks with nonlinear synapses and a static noise", *Phys. Rev. A* **34** (1986) 2571–2574.

[20] A.N. Burkitt, "Attractor neural networks with excitatory neurons and fast inhibitory interneurons at low spike rates", *Network* **5** (1994) 437–448.