IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

Feed-forward chains of recurrent attractor neural networks near saturation

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

1996 J. Phys. A: Math. Gen. 29 7855

(http://iopscience.iop.org/0305-4470/29/24/011)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 171.66.16.71
The article was downloaded on 02/06/2010 at 04:07

Please note that terms and conditions apply.

J. Phys. A: Math. Gen. **29** (1996) 7855–7866. Printed in the UK

Feed-forward chains of recurrent attractor neural networks near saturation

A C C Coolen† and L Viana‡

† Department of Mathematics, King's College, University of London, Strand, London WC2R 2LS, UK

‡ Lab. de Ensenada, Instituto de Física, UNAM, A Postal 2681, 22800 Ensenada, BC, México

Received 2 July 1996

Abstract. We perform a stationary state replica analysis for a layered network of Ising spin neurons, with recurrent Hebbian interactions within each layer, in combination with strictly feed-forward Hebbian interactions between successive layers. This model interpolates between the fully recurrent and symmetric attractor network studied by Amit *et al*, and the strictly feed-forward attractor network studied by Domany *et al*. Due to the absence of detailed balance, it is as yet solvable only in the zero-temperature limit. The built-in competition between two qualitatively different modes of operation, feed-forward (ergodic within layers) versus recurrent (non-ergodic within layers), is found to induce interesting phase transitions.

1. Introduction

There are two main classes of solvable models for neural networks, functioning as associative information processing devices. The first class consists of models in which the matrix of interactions between the neurons is symmetric [1–3]. This ensures detailed balance, so that equilibrium statistical mechanics applies. The second class of models are equipped with neuron interactions which are non-symmetric, but are of a strictly feed-forward nature. Such models can be constructed explicitly, by arranging the neurons in feed-forward layers [4–6], or implicitly, by introducing an extreme dilution such that on finite time-scales the systems behave effectively in a tree-like manner [7, 8]. The latter models can no longer be studied within equilibrium statistical mechanics; here it is the strict feed-forward nature of the interactions which enables a dynamical solution. Even the more recent studies of layered attractor networks all rely for the solution on the connectivity being either strictly symmetric (e.g. [9, 10]) or strictly feed-forward (e.g. [11]).

In this paper we study a dual model, where there is a controllable competition between the highly non-ergodic recurrent information processing, typical for symmetric systems, and the feed-forward processing typical for tree-like structures, where individual layers are ergodic. We consider chains of equally large neural layers, each consisting of Ising spin neurons with recurrent interactions, in combination with strictly feed-forward interactions between successive layers. All allowed interactions are of a Hebbian type; this choice guarantees meaningful information processing, induces familiar order parameters, and creates convenient benchmarks, in that in the two extreme limits of fully recurrent and fully feed-forward connectivity the familiar results of [2, 3] and [4–6] must be recovered. Each individual layer can be seen as a recurrent attractor network with association cues provided by (thermally fluctuating) external fields. These fields, however, contain contributions from

0305-4470/96/247855+12$19.50 © 1996 IOP Publishing Ltd

non-nominated patterns, and therefore contain the disorder variables to be averaged out (which already appear in the recurrent interactions), in contrast to the situation with most other attractor models with external fields, such as [12–15] (with the exception of [16], where the disorder averaging is indeed similar to the one in the present paper).

Our motivation for studying this type of model is threefold. From a statistical mechanical point of view the model is of interest simply because its interaction matrix is neither symmetric nor strictly feed-forward, so that the conventional routes to a solution cannot be followed, and since it interpolates between two systems with quite different properties. Secondly, because the model can none the less be solved (at least in the zero-temperature limit), it could serve as a convenient future playground for testing new formalisms aimed at calulating stationary state properties of non-symmetric neural network models, such as [17]. Thirdly, from a biological point of view, both recurrent and feed-forward information processing have specific advantages and disadvantages, and especially in the peripherical brain regions one therefore usually finds both interaction types present. Recent experimental evidence even suggests that the balance between recurrent and feed-forward processing in these regions is actively controlled by neuromodulators [18]. Yet, as far as we are aware, no dual models have been solved so far.

This paper is organized as follows. After having defined our model, we analyse its zero temperature stationary state, using replica theory. The resulting saddle-point equations are solved upon making the replica symmetry (RS) ansatz, leaving two control parameters: α (the information storage level) and ω (a parameter reflecting the balance between recurrent and feed-forward operation). We then study the various types of phase transitions exhibited by the model: (i) the saturation (or storage capacity) transition, (ii) simple (RS) ergodicity breaking transitions, and (iii) complex (i.e. replica symmetry breaking, RSB) ergodicity-breaking transitions. All analytical results are confirmed by numerical simulations.

2. Model definition and solution

2.1. Model definition

Our model system is a feed-forward chain of L recurrent layers, each of which consists of N Ising spin neurons. The variable $\sigma_i^\ell \in \{-1, 1\}$ denotes the state (non-firing/firing) of neuron i in layer ℓ, and the collective state of the N neurons in layer ℓ is written as $\boldsymbol{\sigma}^\ell = (\sigma_1^\ell, \ldots, \sigma_N^\ell)$. The dynamics is the usual Glauber-type sequential stochastic alignment of the neurons σ_i^ℓ to local fields h_i^ℓ, to which here only neurons from within layer ℓ (via recurrent interactions) and neurons from the previous layer $\ell - 1$ (via feed-forward interactions) are allowed to contribute (see figure 1):

$$\begin{aligned} &\text{Prob}[\sigma_i^\ell \to -\sigma_i^\ell] = \tfrac{1}{2}[1 - \tanh(\beta \sigma_i^\ell h_i^\ell(\boldsymbol{\sigma}^\ell, \boldsymbol{\sigma}^{\ell-1}))] \\ &h_i^\ell(\boldsymbol{\sigma}^\ell, \boldsymbol{\sigma}^{\ell-1}) = \sum_j J_{ij}^\ell \sigma_j^\ell + \sum_j W_{ij}^\ell \sigma_j^{\ell-1}. \end{aligned} \tag{1}$$

At each time step the candidate neuron (i, ℓ) to be updated is drawn at random from $\{1, \ldots, N\} \times \{1, \ldots, L\}$. The parameter $\beta = T^{-1}$ controls the amount of noise, with $T = 0$ corresponding to deterministic alignment (although the order of updates still remains random).

The $2L \times N^2$ synaptic interactions are defined as the result of the network having learned $p = \alpha N$ patterns with a Hebbian-type rule:

$$J_{ij}^\ell = \frac{J_0}{N}[1 - \delta_{ij}] \sum_{\mu=1}^{p} \xi_i^{\mu,\ell} \xi_j^{\mu,\ell} \qquad W_{ij}^\ell = \frac{J}{N} \sum_{\mu=1}^{p} \xi_i^{\mu,\ell} \xi_j^{\mu,\ell-1} \tag{2}$$

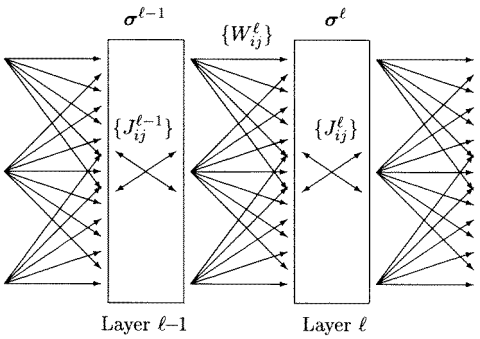

Figure 1. Two adjacent modules in the L-layer chain, with their associated microscopic variables and parameters: $\boldsymbol{\sigma}^{\ell} \in \{-1, 1\}^N$ (neuron states in layer ℓ), $\{J_{ij}^{\ell}\}$ (symmetric recurrent interactions in layer ℓ), and $\{W_{ij}^{\ell}\}$ (feed-forward interactions from layer $\ell - 1$ to layer ℓ).

with $\xi_i^{\mu,\ell} \in \{-1, 1\}$ denoting component i in layer ℓ of pattern μ ($\mu = 1, \ldots, p$). Each pattern represents a specific microscopic neural configuration in the chain as a whole. For simplicity all pattern components are drawn independently at random from $\{-1, 1\}$, and N is assumed to be large (eventually we will take the limit $N \to \infty$). The two parameters (J_0, J) (one of which will become redundant in the limit $T \to 0$) control the relative strength of the two interaction types. The only exception to (2) is the first layer, where by definition we have to set $W_{ij}^1 = 0$ $(\forall ij)$, and where we have to distinguish between two modes of operation: (i) free relaxation of layer 1, following a specific initialization which serves as the recall cue, and (ii) so-called 'clamped' operation, where the recall cue is provided by the state vector $\boldsymbol{\sigma}^1$ itself, which is stationary and specified externally.

For $J = 0$ our model reduces to a collection of L decoupled symmetric attractor networks of the Hopfield [1] type, which can be solved in equilibrium using equilibrium statistical mechanics [2, 3]. Such systems are known to have a storage capacity of $\alpha_c \sim 0.138$ (for $T = 0$, in RS approximation), and non-trivial ergodicity breaking (RSB) for sufficiently low temperatures. For $J_0 = 0$, on the other hand, we have feed-forward interactions only. Now the local fields are found to have Gaussian probability distributions in the stationary state, enabling the derivation of recurrent relations for the values of order parameters in subsequent layers [4–6]. Here one finds a storage capacity of $\alpha_c \sim 0.269$ (for $T = 0$) and no RSB at any temperature. The solution of the present model will have to represent a marriage of these two extremes, such that the equations first derived in [2] and [4] follow as special cases, upon taking the limits $J \to 0$ and $J_0 \to 0$, respectively.

Away from saturation, for $\alpha = p/N \to 0$, the dynamics of this model can be solved easily using existing techniques (for a review see [19]), leading to a set of coupled differential equations for a small number of macroscopic observables. In order to solve the model near saturation (i.e. for $\alpha > 0$), we will exploit the fact that, due to the strictly feed-forward nature of the inter-layer interactions, at $T = 0$ all layers will eventually go to a stationary state; for each layer ℓ there will be a stage after which the input from layer $\ell - 1$ is stationary. This enables an equilibrium statistical mechanical replica analysis at least in the $T \to 0$ limit (for $T > 0$ one has to take thermal fluctuations in the external fields into account, which destroy the Boltzmann form of the stationary states of the individual layers). For technical reasons we will take the limit $T \to 0$ after the limit $N \to \infty$, a commonly taken step which in the present case, however, need not be as harmless as for models obeying detailed balance (it is justified *a posteriori* by the agreement between theory and numerical simulations).

2.2. Replica analysis of the stationary state

We analyse the stationary state for a given layer ℓ, upon assuming stationary inputs from layer $\ell-1$. Given this assumption, the dynamics (1) will evolve towards equilibrium, characterized by

$$p_\infty(\boldsymbol{\sigma}^\ell) \sim \mathrm{e}^{-\beta H(\boldsymbol{\sigma}^\ell)} \qquad H(\boldsymbol{\sigma}^\ell) = -\tfrac{1}{2}\sum_{ij}\sigma_i^\ell J_{ij}^\ell \sigma_j^\ell - \sum_{ij}\sigma_i^\ell W_{ij}^\ell \sigma_j^{\ell-1}. \tag{3}$$

The associated thermal averages are written as $\langle\cdots\rangle$. We introduce a macroscopic description in terms of the so-called overlaps between the system state and the stored patterns, and make the usual ansatz that in equilibrium only a finite number k of the patterns are condensed, which (due to permutation symmetry with respect to pattern indices) we can take to be $\mu = 1,\ldots,k$:

$$m_{\mu,\ell}(\boldsymbol{\sigma}^\ell) = \frac{1}{N}\sum_i \xi_i^{\mu,\ell}\sigma_i^\ell \qquad \begin{cases} \mu\in\{1,\ldots,k\}: & \langle m_{\mu,\ell}^2(\boldsymbol{\sigma}^\ell)\rangle = \mathcal{O}(1) \\ \mu\in\{k+1,\ldots,p\}: & \langle m_{\mu,\ell}^2(\boldsymbol{\sigma}^\ell)\rangle = \mathcal{O}(1/N). \end{cases} \tag{4}$$

We assume the free energy per spin $f = -\frac{1}{\beta N}\log Z$ of the system described by (3) to be self-averaging for $N\to\infty$ with respect to the realization of the non-condensed patterns $\mu > k$ (which play the role of 'frozen disorder'), so that we can simplify the calculation of the free energy by averaging over these patterns. This property can also be rigorously proven. To simplify the pattern average we use the identity $\log Z = \lim_{n\to 0}\frac{1}{n}[Z^n - 1]$, giving

$$-\beta f = \lim_{n\to 0}\frac{1}{nN}\left[\langle Z^n\rangle_{\mathrm{patt}} - 1\right] \qquad Z = \sum_{\boldsymbol{\sigma}\in\{-1,1\}^N}\mathrm{e}^{-\beta H(\boldsymbol{\sigma})}. \tag{5}$$

Upon writing the pattern overlaps of the previous layer $\ell-1$ (which are stationary by assumption) as $\tilde{m}_\mu = \frac{1}{N}\sum_i \xi_i^{\mu,\ell-1}\sigma_i^{\ell-1}$, we can write the key quantity in (5) for integer values of n in the following form:

$$\langle Z^n\rangle_{\mathrm{patt}} = \mathrm{e}^{-\frac{1}{2}\alpha\beta J_0 n}\sum_{\boldsymbol{\sigma}^1\in\{-1,1\}^N}\cdots\sum_{\boldsymbol{\sigma}^n\in\{-1,1\}^N}\langle \mathrm{e}^{\beta N\sum_{\mu=1}^p\sum_{\alpha=1}^n[\frac{1}{2}J_0 m_{\mu,\ell}^2(\boldsymbol{\sigma}^\alpha)+Jm_{\mu,\ell}(\boldsymbol{\sigma}^\alpha)\tilde{m}_\mu]}\rangle_{\mathrm{patt}}. \tag{6}$$

The condensed contribution to (6) (the terms $\mu\leqslant k$ in the summation over pattern indices) is linearized by a Gaussian transformation:

$$\mathrm{e}^{\beta N\sum_{\mu\leqslant k}\sum_\alpha[\frac{1}{2}J_0 m_\mu^2(\boldsymbol{\sigma}^\alpha)+J\tilde{m}_\mu m_\mu(\boldsymbol{\sigma}^\alpha)]} = \left[\frac{\beta J_0 N}{2\pi}\right]^{\frac{nk}{2}}\int\left[\prod_{\mu\leqslant k}\prod_\alpha \mathrm{d}m_\mu^\alpha\right]\mathrm{e}^{\sum_{\mu\leqslant\ell}\sum_\alpha[-\frac{1}{2}\beta J_0 N(m_\mu^\alpha)^2+\beta\sum_i\sigma_i^\alpha\xi_i^{\mu,\ell}(J_0 m_\mu^\alpha+J\tilde{m}_\mu)]}. \tag{7}$$

The uncondensed contribution to (6) (the terms $\mu > k$ in the summation over pattern indices) is first linearized by a Gaussian transformation and then averaged over the frozen disorder, i.e. the pattern components $\xi_i^{\mu,\ell}$ with $\mu > k$, giving

$$\langle \mathrm{e}^{\beta N\sum_{\mu>k}\sum_\alpha[\frac{1}{2}J_0 m_\mu^2(\boldsymbol{\sigma}^\alpha)+J\tilde{m}_\mu m_{\mu,\ell}(\boldsymbol{\sigma}^\alpha)]}\rangle_{\mathrm{patt}} = \prod_{\mu>k}\int \mathrm{D}\boldsymbol{z}\,\mathrm{e}^{\sum_i\log\cosh\sum_\alpha[(\frac{\beta J_0}{N})^{\frac{1}{2}}z_\alpha\sigma_i^\alpha+\beta J\tilde{m}_\mu\sigma_i^\alpha]}$$

with the Gaussian measure $\mathrm{D}\boldsymbol{z} = \prod_\alpha[(2\pi)^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}z_\alpha^2}\,\mathrm{d}z_\alpha]$. Using the property $\tilde{m}_\mu^2 = \mathcal{O}(N^{-1})$ for $\mu > k$, we expand $\log\cosh(x) = \frac{1}{2}x^2 + \mathcal{O}(x^4)$. Inserting an integral representation of

unity to isolate the spin glass order parameters $q_{\alpha\beta}=\frac{1}{N}\sum_i \sigma_i^\alpha\sigma_i^\beta$, and repeatedly forgetting about terms which are of vanishing order either for $n\to 0$ or for $N\to\infty$, then leads to:

$$\langle \mathrm{e}^{\beta N\sum_{\mu>k}\sum_\alpha[\frac{1}{2}J_0 m_\mu^2(\boldsymbol{\sigma}^\alpha)+Jm_\mu(\boldsymbol{\sigma}^\alpha)\tilde{m}_\mu]}\rangle_{\mathrm{patt}}$$
$$=\int \mathrm{d}\boldsymbol{q}\,\mathrm{d}\hat{\boldsymbol{q}}\,\mathrm{e}^{N\sum_{\alpha\beta}q_{\alpha\beta}[\mathrm{i}\hat{q}_{\alpha\beta}+\frac{1}{2}(\beta J)^2\sum_{\mu>k}\tilde{m}_\mu^2]-\mathrm{i}\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sum_i\sigma_i^\alpha\sigma_i^\beta}\prod_{\mu>k}\int \mathrm{D}\boldsymbol{z}\,\mathrm{e}^{\frac{1}{2}\beta J_0\boldsymbol{z}\cdot\boldsymbol{q}\boldsymbol{z}+\sum_\alpha z_\alpha x_\alpha^\mu}$$
$$=\int \mathrm{d}\boldsymbol{q}\,\mathrm{d}\hat{\boldsymbol{q}}\,\mathrm{e}^{N\sum_{\alpha\beta}q_{\alpha\beta}[\mathrm{i}\hat{q}_{\alpha\beta}+\frac{1}{2}(\beta J)^2\sum_{\mu>k}\tilde{m}_\mu^2]-\mathrm{i}\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sum_i\sigma_i^\alpha\sigma_i^\beta-\frac{1}{2}p\log\det\boldsymbol{\Lambda}+\frac{1}{2}\sum_{\alpha\beta}\sum_{\mu>k}x_\alpha^\mu(\boldsymbol{\Lambda})^{-1}_{\alpha\beta}x_\beta^\mu} \tag{8}$$

with the abbreviation $x_\alpha^\mu=\beta J(\beta J_0 N)^{1/2}\tilde{m}_\mu\sum_\beta q_{\alpha\beta}$, and where $\boldsymbol{\Lambda}$ is an $n\times n$ matrix with components $\Lambda_{\alpha\beta}=\delta_{\alpha\beta}-\beta J_0 q_{\alpha\beta}$. We finally substitute the results (7), (8) into equation (6), and perform the remaining summations over the spin variables $\{\sigma_i^\alpha\}$. The free energy per spin (5) is subsequently obtained by interchanging the limits $N\to\infty$ and $n\to 0$ and by performing the remaining integral by the method of steepest descent, leading to the final result:

$$f=\frac{1}{2}\alpha J_0-\beta^{-1}\log 2-\lim_{n\to 0}\frac{1}{\beta n}\,\mathrm{extr}\left\{\mathrm{i}\sum_{\alpha\beta}\hat{q}_{\alpha\beta}q_{\alpha\beta}-\tfrac{1}{2}\beta J_0\sum_\alpha\sum_{\mu\leqslant k}[m_\mu^\alpha]^2\right.$$
$$-\tfrac{1}{2}\alpha\log\det[\mathbb{1}-\beta J_0\boldsymbol{q}]+\tfrac{1}{2}\beta^2J^2\Big[\sum_{\mu>k}\tilde{m}_\mu^2\Big]\sum_{\alpha\beta\gamma}q_{\alpha\beta}[\mathbb{1}-\beta J_0\boldsymbol{q}]^{-1}_{\beta\gamma}$$
$$\left.+\sum_{\boldsymbol{\xi}}\log\sum_{\boldsymbol{\sigma}}\mathrm{e}^{\beta\sum_\alpha\sigma^\alpha\sum_{\mu\leqslant k}[J_0m_\mu^\alpha+J\tilde{m}_\mu]\xi_\mu-\mathrm{i}\sum_{\alpha\beta}\sigma_\alpha\hat{q}_{\alpha\beta}\sigma_\beta}\right\} \tag{9}$$

with $\boldsymbol{\xi}\in\{-1,1\}^k$ and $\boldsymbol{\sigma}\in\{-1,1\}^n$. The extremum in (9) refers to variation of the parameters $\{\hat{q}_{\alpha\beta},q_{\alpha\beta},m_\mu^\alpha\}$, and is defined as the analytical continuation for $n\to 0$ of the saddle point which for $n\geqslant 1$ minimizes f. The physical meaning of the parameters m_μ^α can be deduced, for instance by taking the derivative of (5) with respect to $\tilde{m}_\mu$, which gives $\lim_{n\to 0}\frac{1}{n}\sum_\alpha m_\mu^\alpha=\frac{1}{N}\sum_i\xi_i^{\mu,\ell}\langle\sigma_i^\ell\rangle=\langle m_{\mu,\ell}(\boldsymbol{\sigma}^\ell)\rangle$.

2.3. Replica-symmetric solution

We have obtained an expression for the free energy per neuron f in terms of the parameters $q_{\alpha\beta}$, $\hat{q}_{\alpha\beta}$ and m_μ^α. In order to take the limit $n\to 0$ in (9) and obtain an explicit solution, we will make the replica symmetry (RS) ansatz for the relevant saddle-point, i.e.

$$\begin{aligned}\hat{q}_{\alpha\beta}&=\tfrac{1}{2}\mathrm{i}\alpha\beta^2[R\delta_{\alpha\beta}+r(1-\delta_{\alpha\beta})]\\ q_{\alpha\beta}&=\delta_{\alpha\beta}+q(1-\delta_{\alpha\beta})\\ m_\mu^\alpha&=m_\mu.\end{aligned} \tag{10}$$

Substitution of (10) into (9) and linearization of the exponent involving the spin variables, allows us to perform the remaining spin summations and take the limit $n\to 0$, which gives

$$f_{\mathrm{RS}}=\frac{1}{2}\alpha[J_0+\beta r(1-q)]+\frac{1}{2}J_0\sum_{\mu\leqslant k}m_\mu^2+\frac{\alpha}{2\beta}\left[\log[1-\beta J_0(1-q)]-\frac{\beta J_0 q}{1-\beta J_0(1-q)}\right]$$
$$-\frac{1}{2}\Big[\sum_{\mu>\ell}\tilde{m}_\mu^2\Big]\frac{\beta J^2(1-q)}{1-\beta J_0(1-q)}$$
$$-\frac{1}{\beta}\left\langle\int \mathrm{D}z\,.\log 2\cosh\beta\Big[\sum_{\mu\leqslant k}\xi_\mu(J_0m_\mu+J\tilde{m}_\mu)+z\sqrt{\alpha r}\Big]\right\rangle_{\boldsymbol{\xi}} \tag{11}$$

where $\mathrm{D}z = (2\pi)^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}z^2}\,\mathrm{d}z$ and $\langle G[\boldsymbol{\xi}]\rangle_{\boldsymbol{\xi}} = 2^{-k}\sum_{\boldsymbol{\xi}\in\{-1,1\}^k} G[\boldsymbol{\xi}]$. Taking the derivative of the free energy with respect to the control parameter J_0 in addition yields the following identity:

$$\sum_{\mu>k}\langle m_{\mu,\ell}^2(\boldsymbol{\sigma}^\ell)\rangle = \alpha - \sum_{\mu\leqslant k}\langle m_{\mu,\ell}^2(\boldsymbol{\sigma}^\ell)\rangle - 2\frac{\partial f_{\mathrm{RS}}}{\partial J_0}. \tag{12}$$

By solving the saddle point equations $\partial f_{\mathrm{RS}}/\partial m_\mu = \partial f_{\mathrm{RS}}/\partial q = \partial f_{\mathrm{RS}}/\partial r = 0$, from which we can eliminate $\sum_{\mu>k}\tilde{m}_\mu^2$ by applying (12) to layer $\ell-1$, we finally arrive at a set of recurrent equations which relate the values of order parameters in subsequent layers. Upon adapting our notation accordingly, these recurrent relations can be written in their most natural form, $(\boldsymbol{m}, q, r) \to (\boldsymbol{m}', q', r')$:

$$\boldsymbol{m}' = \left\langle \boldsymbol{\xi}\int \mathrm{D}z\ \tanh\beta\left[\boldsymbol{\xi}\cdot(J_0\boldsymbol{m}' + J\boldsymbol{m}) + z\sqrt{\alpha r'}\right]\right\rangle_{\boldsymbol{\xi}} \tag{13}$$

$$q' = \left\langle \int \mathrm{D}z\ \tanh^2\beta\left[\boldsymbol{\xi}\cdot(J_0\boldsymbol{m}' + J\boldsymbol{m}) + z\sqrt{\alpha r'}\right]\right\rangle_{\boldsymbol{\xi}} \tag{14}$$

$$r'[1-\beta J_0(1-q')]^2 - J_0^2 q' = \beta^2 J^2 r(1-q)^2 - J^2 q + \frac{J^2(1+q)}{1-\beta J_0(1-q)}. \tag{15}$$

The macroscopic state of every layer ℓ is now characterized by its value for the order parameter set $(\boldsymbol{m}, q, r)$, with $\boldsymbol{m} = (m_1, \ldots, m_k)$, and with the (familiar) physical meaning $m_\mu = \frac{1}{N}\sum_i \xi_i^{\mu,\ell}\langle\sigma_i^\ell\rangle$ and $q = \frac{1}{N}\sum_i\langle\sigma_i^\ell\rangle^2$. The only exceptions to (13)–(15) arise in the context of the special status of the first layer. If this first layer is clamped into a randomly selected configuration $\boldsymbol{\sigma}^1$ with a prescibed vector of the condensed overlaps $\boldsymbol{m}$, then (15) is to be replaced by

$$r'[1-\beta J_0(1-q')]^2 - J_0^2 q' = J^2. \tag{16}$$

If, on the other hand, the first layer is allowed free relaxation towards equilibrium, then (13)–(15) do hold, but with the macroscopic state $(\boldsymbol{m}, q, r)$ of the first layer solved from the set corresponding to the $J = 0$ situation, where inter-layer interactions are absent:

$$\boldsymbol{m} = \left\langle \boldsymbol{\xi}\int \mathrm{D}z\ \tanh\beta\left[J_0\boldsymbol{\xi}\cdot\boldsymbol{m} + z\sqrt{\alpha r}\right]\right\rangle_{\boldsymbol{\xi}} \tag{17}$$

$$q = \left\langle \int \mathrm{D}z\ \tanh^2\beta\left[J_0\boldsymbol{\xi}\cdot\boldsymbol{m} + z\sqrt{\alpha r}\right]\right\rangle_{\boldsymbol{\xi}} \tag{18}$$

$$r = \frac{q J_0^2}{[1-\beta J_0(1-q)]^2}. \tag{19}$$

In the limit $T \to 0$, the equations (13)–(19) constitute the solution of our model. We will now analyse their consequences and validate their predictions with simulation experiments. We eliminate the parameter redundancy at $T = 0$ by putting $J_0 = \frac{1}{2}[1+\omega]$ and $J = \frac{1}{2}[1-\omega]$, with $\omega \in [-1, 1]$.

3. Phase transitions

3.1. Saturation transition in infinitely long chains

We first calculate the information storage capacity α_c for (infinitely) long chains. A stationary situation is reached along the chain when $(\boldsymbol{m}, q, r) = (\boldsymbol{m}', q', r')$ in (13)–(15),

giving

$$\boldsymbol{m} = \left\langle \boldsymbol{\xi} \int \mathrm{D}z \, \tanh\beta\left[\boldsymbol{\xi}\cdot\boldsymbol{m} + z\sqrt{\alpha r}\right]\right\rangle_{\boldsymbol{\xi}} \tag{20}$$

$$q = \left\langle \int \mathrm{D}z \, \tanh^2\beta\left[\boldsymbol{\xi}\cdot\boldsymbol{m} + z\sqrt{\alpha r}\right]\right\rangle_{\boldsymbol{\xi}} \tag{21}$$

$$r = \frac{(1-\omega)^2 + q(1+\omega)^2 - 2\omega\beta q(1+\omega)(1-q)}{4[1-\frac{1}{2}\beta(1+\omega)(1-q)][1-\beta(1+\omega)(1-q)+\omega\beta^2(1-q)^2]}. \tag{22}$$

Upon taking the limit $T \to 0$ and concentrating on pure states, where $m_\mu = m\delta_{\mu\lambda}$, we can perform the averages and integrations, substitute $x = m/\sqrt{2\alpha r}$, and simply follow the procedure described in [3] to reduce the above set of equations (20)–(22) to a single transcendental equation (with $m = \mathrm{erf}(x)$):

$$x\sqrt{2\alpha} = \frac{\mathrm{erf}(x) - \frac{2x}{\sqrt{\pi}}\mathrm{e}^{-x^2}}{\sqrt{\frac{1}{2}(1+\omega^2)}} \left\{ \frac{\left[\mathrm{erf}(x) - \frac{2\omega x}{\sqrt{\pi}}\mathrm{e}^{-x^2}\right]\left[\mathrm{erf}(x) - \frac{(1+\omega)x}{\sqrt{\pi}}\mathrm{e}^{-x^2}\right]}{\left[\mathrm{erf}(x) - \frac{2x}{\sqrt{\pi}}\mathrm{e}^{-x^2}\right]\left[\mathrm{erf}(x) - \frac{\omega^2+\omega}{\omega^2+1}\frac{2x}{\sqrt{\pi}}\mathrm{e}^{-x^2}\right]} \right\}^{\frac{1}{2}}. \tag{23}$$

This equation is to be solved numerically. The storage capacity α_c is the value for α for which the non-zero solutions of (23) vanish, resulting in figure 2 (left panel). For $\omega = -1$ equation (23) reduces to the results of [4–6]; for $\omega = 1$ it reduces to the results of [2, 3], as it should. Somewhat surprisingly, the largest storage capacity is obtained for $\omega \sim -0.12$, giving $\alpha_c \sim 0.317$. We support these analytical results with numerical simulations on chains with $L = 60$ layers of $N = 900$ neurons each, for $T = 0$ and $\omega = 0$. Figure 2 (right panel) shows the stationary values of the condensed overlap m as a function of the layer number. For $\alpha \in \{0.26, 0.28\}$ the desired state $m \sim 1$ appears stable, for $\alpha \in \{0.33, 0.35\}$ the state $m \sim 1$ is unstable, whereas for $\alpha \in \{0.29, 0.30\}$ we appear to be close to the critical value, with $m \sim 1$ appearing stable initially, but eventually destabilizing further down along the chain. This is in reasonable agreement with the theory, which predicts $\alpha_c \sim 0.314$ for $\omega = 0$, if we take finite size effects into account.

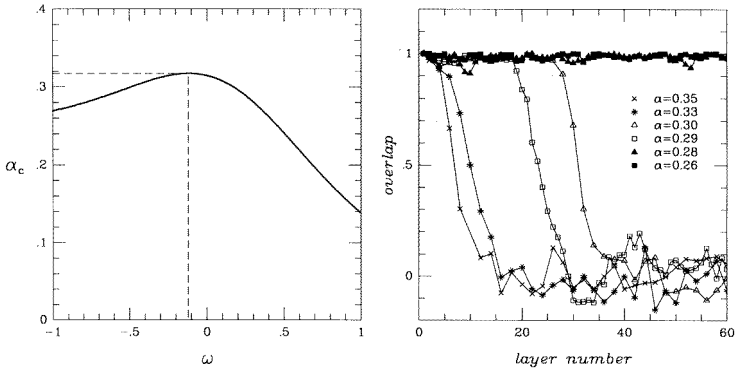

Figure 2. Left panel—theory: storage capacity α_c for long chains at $T = 0$ (broken line: location of the optimum). Right panel—numerical simulations ($N = 900$): stationary (pure state) overlap m as a function of layer number ℓ, for $T = 0$ and $\omega = 0$.

3.2. Simple ergodicity breaking transitions

Next we turn to transitions marking the appearance of multiple (replica-symmetric) coexistent stable states. We restrict ourselves to pure states, i.e. in each layer $m_\mu = m\delta_{\mu,\lambda}$ for some λ and some layer-dependent m, and to the junction between layer $\ell = 1$ and layer $\ell = 2$, where the first such transitions are expected. Insertion of pure state solutions, and taking the limit $T \to 0$ in (13)–(15) again allows us to perform the averages and integrations. Substitution of $y = [J_0 m' + Jm]/\sqrt{2\alpha r'}$ and, in the case where the first layer evolves freely, of $x = J_0 m/\sqrt{2\alpha r}$, now leads to the expression $m' = \mathrm{erf}(y)$, in which y is the solution of the following transcendental equation:

$$F(y) = y\sqrt{2\alpha}\left[1+\rho\left(\frac{1-\omega}{1+\omega}\right)^2\right]^{\frac{1}{2}} - m\left[\frac{1-\omega}{1+\omega}\right] \qquad F(y) = \mathrm{erf}(y) - \frac{2y}{\sqrt{\pi}}\mathrm{e}^{-y^2}. \tag{24}$$

The value of ρ in (24) depends on whether or not the input layer is clamped into a state, and, in the case of free relaxation, on the actual value of m:

$$\begin{aligned} &\text{Input clamped: } m \text{ given, } \rho = 1 \\ &\text{Free relaxation: } m = \mathrm{erf}(x),\ F(x) = x\sqrt{2\alpha},\ \rho = [\mathrm{erf}(x)/F(x)]^2. \end{aligned} \tag{25}$$

The simplest situation, $m = 0$, already clearly demonstrates the distinction between the two modes of operation (input clamped versus free relaxation). In the clamped case, where $\rho = 1$, a rescaling of the storage ratio α maps (24) onto the equations describing the model of [2], and solutions with $y \neq 0$ (i.e. $m' \neq 0$) must therefore bifurcate at

$$\alpha_{\mathrm{bif}}(\omega) = \frac{\alpha_{\mathrm{bif}}(1)}{\sqrt{1+\left(\frac{1-\omega}{1+\omega}\right)^2}} \sim \frac{0.138}{\sqrt{1+\left(\frac{1-\omega}{1+\omega}\right)^2}}. \tag{26}$$

On the other hand, for free relaxation of the first layer we find $\rho = \infty$, so $y = 0$ (i.e. $m' = 0$ is the only solution of (24)). In the first case the imposed $m = 0$ state of layer one is like a paramagnetic state; in the second case the $m = 0$ state is of a spin-glass type.

For arbitrary m we obtain the condition for new solutions of (24) to bifurcate by derivation of (24) with respect to y. This gives a new equation, to be solved simultaneously with (24); a simple transformation allows us to rewrite the resulting pair as

$$\alpha = \frac{8}{\pi}y^4\mathrm{e}^{-2y^2}\left[1+\rho\left(\frac{1-\omega}{1+\omega}\right)^2\right]^{-1} \qquad m\left[\frac{1-\omega}{1+\omega}\right] = G(y) \tag{27}$$

(to be solved with the appropriate values for (m, ρ), given in (25)), and with

$$G(y) = \frac{2y}{\sqrt{\pi}}(1+2y^2)\mathrm{e}^{-y^2} - \mathrm{erf}(y). \tag{28}$$

It is clear from (27) that bifurcations of new solutions can only occur for sufficiently small α. For clamped input operation, where the input overlap m is an independent parameter, numerical solution of (27) results in bifurcation lines in the (ω, α) plane, shown in figure 3. Left and above the full curves are the regions with a unique stable state, and at the other side of the full curves (near $\omega = 1$) multiple stable states exist. The bottom-right panel shows the result of combining all regions with multiple stable states, for $m \in [0, 1]$: left of the full curve there is a single stable state in the second layer, irrespective of m, right of the full curve the input overlaps $m \in [0, 1]$ can be found which give rise to multiple stable states in the second layer. These results are again supported by numerical simulations: figure 4 shows the relation between initial and (almost) final overlaps in layer 2, for $N = 12\,000$ and $m = 1$,

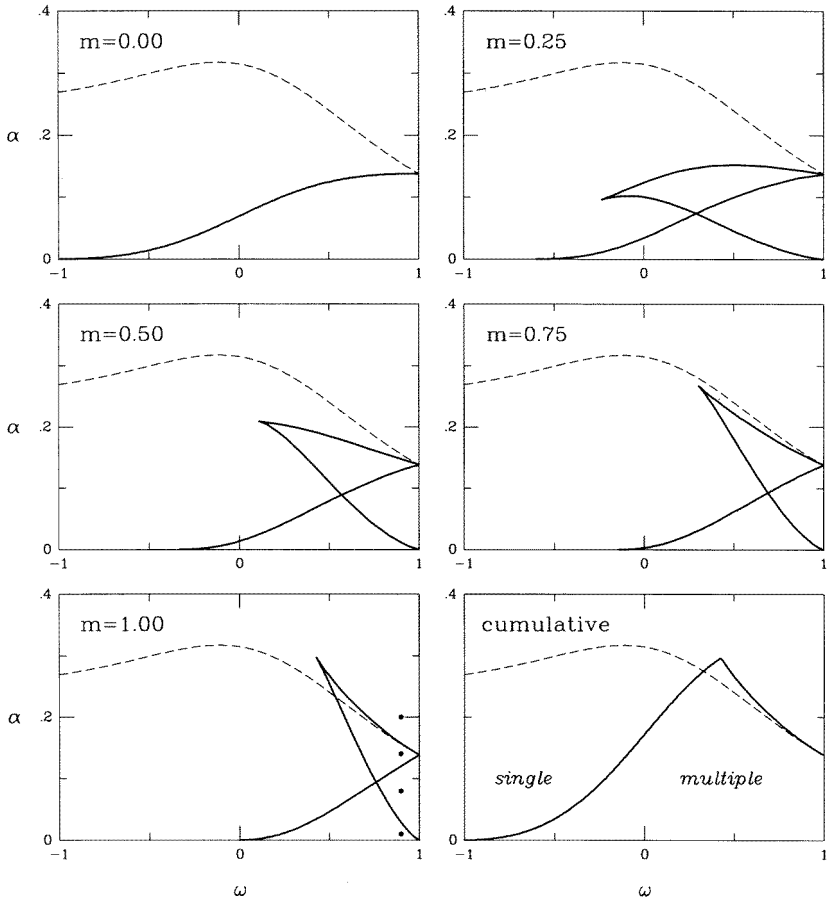

Figure 3. Examples of bifurcation lines, marking the appearance of multiple pure (and replica-symmetric) stable states in layer 2, for clamped operation. Left of and above the full curves: a single stable state; close to the $\omega = 1$ line: multiple stable states. Broken curves: the long-chain storage capacity α_c. Circles in $m = 1.00$ panel: control parameters (ω, α) used in numerical simulations (described below). Bottom-right panel: boundary of the cumulative region where multiple stable states exist (obtained by combining the results for input overlaps $m \in [0, 1]$).

and the four (ω, α) combinations indicated with circles in figure 3. The number of stable states predicted (figure 3) are $(2, 3, 2, 1)$, for the four cases $\alpha = (0.01, 0.08, 0.14, 0.20)$, respectively. This is in perfect agreement with the simulation results of figure 4, where for each graph the number of stable states is the number of discontinuities plus one. For a system in equilibrium one expects horizontal line segments between the discontinuities; apparently for $\alpha \in \{0.08, 0.14, 0.20\}$ the state at $t = 200$ is not an equilibrium state yet (due to the large relaxation times involved).

For free relaxation of the first layer, a non-zero m, being a solution of (25), is a monotonically decreasing function of α with $m(\alpha) \geqslant 0.966$. The bifurcation lines obtained by solving numerically (25) and (27) are now found to be practically indistinguishable from those of the clamped case with $m = 1$ (see figure 3), provided $\alpha \leqslant \alpha_c(\omega = 1) \sim 0.138$. For $\alpha > \alpha_c(\omega = 1)$ we are back at $m = 0$, where no bifurcations were found to be allowed.

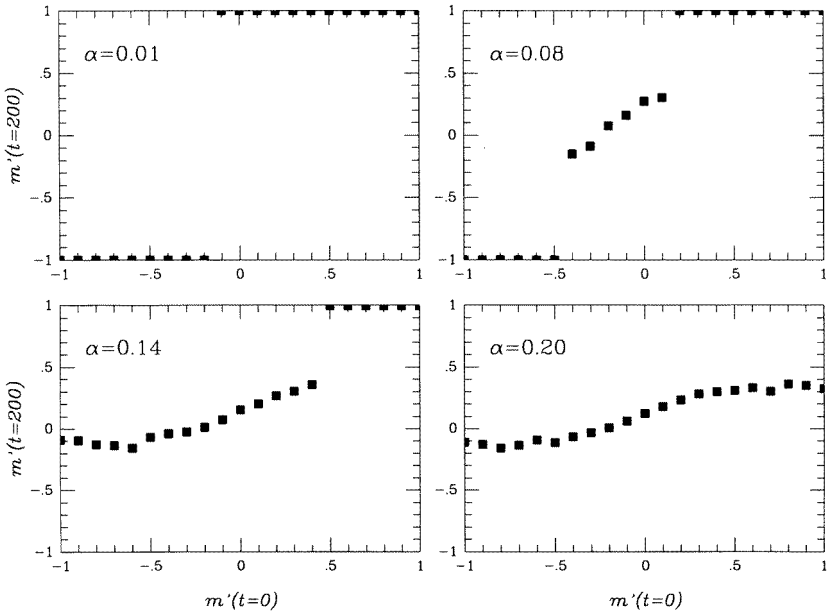

Figure 4. Simulation results ($N = 12\,000$), indicating the number of (pure) stable states in layer 2. Each picture shows the overlap m' at time $t = 200$ versus the overlap m' at $t = 0$, for $\omega = 0.9$ and input overlap $m = 1$. The α values correspond to the circles in figure 3.

3.3. The AT instability

The third and final type of transition to be analysed is the AT-line [20], where the relevant saddle-point of (9) ceases to be replica-symmetric. Note that the fully recurrent limit $\omega \to 1$ of the present model exhibits broken replica symmetry (RSB) near $T = 0$ [2], whereas the $\omega \to -1$ limit does not [4], which hints at the possible existence of an RSB transition at some critical value $-1 < \omega_{\rm AT} < 1$ for $T = 0$. In order to perform a bifurcation analyis *à la* [20], we go back to equation (9), and work out the RSB saddle point equations:

$$m_\alpha = \left\langle \frac{\sum_{\boldsymbol\sigma} \sigma_\alpha {\rm e}^{\beta \sum_\alpha \sigma^\alpha \sum_{\mu\leqslant k}[J_0 m_\mu^\alpha + J\tilde m_\mu]\xi_\mu - {\rm i}\sum_{\alpha\beta}\sigma_\alpha \hat q_{\alpha\beta}\sigma_\beta}}{\sum_{\boldsymbol\sigma} {\rm e}^{\beta \sum_\alpha \sigma^\alpha \sum_{\mu\leqslant k}[J_0 m_\mu^\alpha + J\tilde m_\mu]\xi_\mu - {\rm i}\sum_{\alpha\beta}\sigma_\alpha \hat q_{\alpha\beta}\sigma_\beta}} \right\rangle_\xi \tag{29}$$

$$q_{\alpha\beta} = \left\langle \frac{\sum_{\boldsymbol\sigma} \sigma_\alpha\sigma_\beta {\rm e}^{\beta \sum_\alpha \sigma^\alpha \sum_{\mu\leqslant k}[J_0 m_\mu^\alpha + J\tilde m_\mu]\xi_\mu - {\rm i}\sum_{\alpha\beta}\sigma_\alpha \hat q_{\alpha\beta}\sigma_\beta}}{\sum_{\boldsymbol\sigma} {\rm e}^{\beta \sum_\alpha \sigma^\alpha \sum_{\mu\leqslant k}[J_0 m_\mu^\alpha + J\tilde m_\mu]\xi_\mu - {\rm i}\sum_{\alpha\beta}\sigma_\alpha \hat q_{\alpha\beta}\sigma_\beta}} \right\rangle_\xi \tag{30}$$

$$\hat q_{\alpha\beta} = \frac{1}{2}{\rm i}\beta^2 J^2 x_\alpha x_\beta \left[\sum_{\mu>k} \tilde m_\mu^2\right] + \frac{1}{2}{\rm i}\alpha\beta J_0 \frac{\int {\rm D}\boldsymbol z\, z_\alpha z_\beta {\rm e}^{\frac{1}{2}\beta J_0 \boldsymbol z\cdot\boldsymbol q\boldsymbol z}}{\int {\rm D}\boldsymbol z\, {\rm e}^{\frac{1}{2}\beta J_0 \boldsymbol z\cdot\boldsymbol q\boldsymbol z}} \tag{31}$$

with $\boldsymbol\sigma \in \{-1,1\}^n$, $\boldsymbol z \in \Re^n$, ${\rm D}\boldsymbol z = (2\pi)^{-n/2}{\rm e}^{-\frac{1}{2}z^2}\,{\rm d}\boldsymbol z$, and $x_\alpha = \sum_\gamma [\mathbb{1} - \beta J_0 \boldsymbol q]^{-1}_{\alpha\gamma}$. Following [20] we now consider the so-called replicon fluctuations around the RS solution (13)–(15):

$$q_{\alpha\beta} \to \delta_{\alpha\beta} + q(1-\delta_{\alpha\beta}) + \delta q_{\alpha\beta} \qquad \hat q_{\alpha\beta} \to \tfrac{1}{2}{\rm i}\alpha\beta^2[R\delta_{\alpha\beta} + r(1-\delta_{\alpha\beta})] + \delta\hat q_{\alpha\beta} \tag{32}$$

with the properties $|\delta q_{\alpha\beta}| \ll 1$, $\delta q_{\alpha\beta} = \delta q_{\beta\alpha}$, $\sum_\alpha \delta q_{\alpha\beta} = 0$, and $\delta q_{\alpha\alpha} = 0$. Working out the (coupled) variations in (30), (31) due to (32), and requiring such variations to lead to an instability for $n \to 0$ (a massless mode in first-order perturbation theory) gives, after a

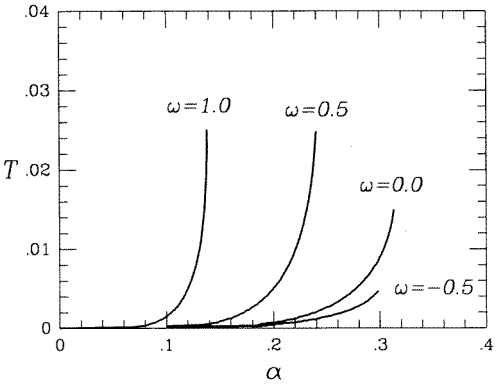

Figure 5. Location of the continuous replica-symmetry-breaking transition (AT line) in the (α, T) plane. For $\omega = -1$ the AT line collapses onto the line $T = 0$.

modest amount of algebra, the condition

$$1 = \frac{\alpha(\beta J_0)^2}{[1-\beta J_0(1-q)]^2}\left\langle \int \mathrm{D}z\, \cosh^{-4}\beta\left[\boldsymbol{\xi}\cdot(J_0\boldsymbol{m}+J\tilde{\boldsymbol{m}})+z\sqrt{\alpha r}\right]\right\rangle_{\boldsymbol{\xi}}. \tag{33}$$

The RS solution (13)–(15) is stable if the right-hand side of (33) is smaller than 1. For $J_0 = 0$ (i.e. for the feed-forward model of [4–6]) clearly no RSB occurs. For $J_0 > 0$ (i.e. for $\omega > -1$) we have to solve (33) numerically, in combination with the RS saddle-point equations (13)–(15). For stationary states in large chains (i.e. $(\boldsymbol{m}, q, r,) = (\boldsymbol{m}', q', r')$ and $L \to \infty$), and pure states $m_\mu = m\delta_{\mu\lambda}$, the solution thus obtained is shown in figure 5 for $\omega \in \{-0.5, 0, 0.5, 1.0\}$ (for $\omega = -1$ the AT line collapses onto the line $T = 0$). Our first conclusion is that for $T = 0$ the RSB transition occurs at $\omega = -1$; any finite fraction of recurrence in the interactions apparently destabilizes the replica-symmetric solution. Secondly, although we cannot interpret the $T > 0$ data in figure 5 directly in terms of the operation of the present model, the fact that, as we move away from the fully recurrent case $\omega = 1$, the noise level T at which replica symmetry breaks, decreases monotonically, suggests that we can be confident that for the present model, as for the fully recurrent case, replica symmetry breaking effects will be of a minor quantitative nature only. This, in fact, is borne out by the agreement observed between simulations and the RS solution.

4. Discussion

In this paper we have analysed a simple layered Ising spin neural network model, with a controllable competition between recurrent and feed-forward information processing. This model interpolates between the fully recurrent and symmetric attractor model of [2, 3] and the strictly feed-forward model of [4–6]. At zero noise level and in a stationary state, the model can be solved analytically near saturation, using replica theory (where we have restricted ourselves to the replica-symmetric ansatz), in spite of its interaction matrix being neither symmetric nor strictly feed-forward (which are the features on which analysis usually relies). This property also turns it into a nice toy model for testing new tools for tackling the stationary states of neural network models without detailed balance.

In two extreme limits, fully recurrent and fully feed-forward operation, respectively, the results of [2, 3] and [4–6] are recovered correctly, as they should. In the intermediate regime, the built-in competition between recurrent operation (which is highly non-ergodic within

individual layers) versus feed-forward operation (which is ergodic within layers), is reflected in a non-trivial way in various types of transitions. These describe saturation breakdown, simple ergodicity-breaking involving pure states, and the AT instability [20] with respect to replica-symmetry-breaking. The largest storage capacity is found to be $\alpha_c \sim 0.317$, which is obtained for a specific balance of the two types of interactions. Replica symmetry turns out to break down as soon as one moves away from the strictly feed-forward limit, i.e. for any finite fraction of recurrence in the interactions.

Our results, which are supported by numerical simulations, might also play a role in describing the competition between recurrent and feed-forward operation in the peripherical regions of the brain (upon suitable quantitative adaptation of model details), where architectures similar to the one studied in this paper can be found.

Acknowledgments

The authors wish to thank the National University of Mexico (UNAM) (DGAPA Project IN100294) and the J S McDonnel Foundation (Visiting Fellowship, LV) for support.

References

[1] Hopfield J J 1982 *Proc. Natl Acad. Sci., USA* **79** 2554
[2] Amit D J, Gutfreund H and Sompolinsky H 1985 *Phys. Rev. Lett.* **55** 1530
[3] Amit D J, Gutfreund H and Sompolinsky H 1987 *Ann. Phys.* **173** 30
[4] Domany E, Meir R and Kinzel W 1986 *Europhys. Lett.* **2** 175
[5] Meir R and Domany E 1988 *Phys. Rev.* A **37** 608
[6] Derrida B and Meir R 1988 *Phys. Rev.* A **38** 3116
[7] Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167
[8] Fassnacht C and Zippelius A 1991 *Network* **2** 63
[9] Kurchan J, Peliti L and Saber M 1994 *J. Physique* I **4** 1627
[10] Englisch H, Mastropietro V and Tirozzi B 1995 *J. Physique* I **5** 85
[11] Wong K Y M, Campbell C and Sherrington D 1995 *J. Phys. A: Math. Gen.* **28** 1603
[12] Rau A and Sherrington D 1990 *Europhys. Lett.* **11** 499
[13] Rau A, Sherrington D and Wong K Y M 1991 *J. Phys. A: Math. Gen.* **24** 313
[14] Engel A, Bouten M, Komoda A and Serneels R 1990 *Phys. Rev.* A **42** 4998
[15] Yau H W and Wallace D J 1991 *J. Phys. A: Math. Gen.* **24** 5639
[16] Engel A, English H and Schütte A 1989 *Europhys. Lett.* **8** 393
[17] Coolen A C C and Sherrington D 1993 *Physica* **200A** 602
[18] Verschure P F M J Private communication
[19] Coolen A C C and Sherrington D 1993 *Mathematical Approaches to Neural Networks* ed J G Taylor (Amsterdam: North-Holland) p 293
[20] de Almeida J R L and Thouless D J 1978 *J. Phys. A: Math. Gen.* **11** 983