INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **34** (2001) 929–942 www.iop.org/Journals/ja PII: S0305-4470(01)16113-5

Attractor modulation and proliferation in $(1 + \infty)$-dimensional neural networks

N S Skantzos and A C C Coolen

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

Received 2 August 2000

Abstract

We extend a recently introduced class of exactly solvable models for recurrent neural networks with competition between one-dimensional nearest-neighbour and infinite-range information processing. We increase the potential for further frustration and competition in these models, as well as their biological relevance, by adding next-nearest-neighbour couplings, and we allow for modulation of the attractors so that we can interpolate continuously between situations with different numbers of stored patterns. Our models are solved by combining mean-field and random-field techniques. They exhibit increasingly complex phase diagrams with novel phases, separated by multiple first- and second-order transitions (dynamical and thermodynamic ones), and, upon modulating the attractor strengths, non-trivial scenarios of phase diagram deformation. Our predictions are in excellent agreement with numerical simulations.

PACS numbers: 8710, 0520

1. Introduction

In real (biological) recurrent neural networks, where information processing is based on the creation and manipulation of attractors, one typically observes an intricate interplay and competition between long-range information processing (via excitatory pyramidal neurons) and short-range information processing (via short-range pyramidal neurons and inhibitory inter-neurons). Studying those properties of such systems which are linked to their spatial structure, using statistical mechanical techniques, requires moving away from the more traditional infinite-range models of attractor neural networks [1, 2]. With the latter objective, an alternative type of attractor neural network was recently proposed and studied [3], in which neurons (represented by Ising spins) are mutually connected by a combination of infinite-range synaptic interactions, and one-dimensional nearest-neighbour interactions. Although real biological network architectures are obviously far more complex, such models, which are still sufficiently simple to be solved exactly, via a combination of mean-field techniques (as in e.g. [2]) and random field techniques (as in e.g. [4–6]), would appear to represent a small but welcome step towards biological reality. Moreover, from a statistical mechanical perspective, the solutions of these models exhibited a remarkably rich behaviour, even in the

0305-4470/01/050929+14$30.00

so-called low-storage regime (where the number of patterns stored in the interactions remains finite in the thermodynamic limit), and particularly in those regions in parameter space where the two types of interaction (long-range versus nearest-neighbour) compete most strongly. The phase diagrams where found to describe a series of regions with different numbers of ergodic components, separated by both second- and first-order transitions (representing various dynamical transitions, in addition to the thermodynamic ones), and to increase dramatically in complexity with the number p of stored patterns.

This paper is devoted to a further exploration and enlargement of the class of models introduced in [3]. We study two orthogonal extensions, each with their own specific objectives. Our first extension is to include also neuronal interactions between next-nearest neighbours in the one-dimensional chain (in addition to the mean-field and nearest-neighbour ones), and to study their impact on the phase diagrams. Here the motivation is, again, partly biological: short-range pyramidal neurons are believed to act on shorter distances than short-range interneurons, and simple models of the type proposed here have indeed been used recently to explain properties of the mammalian visual system [7, 8]. We find, especially when the parameters controlling the new interactions are chosen such as to introduce further competition and frustration into the network, new phases are being created and the complexity of the phase diagram is again significantly increased. Our second extension is primarily motivated by our desire to understand the significant qualitative modification of the phase diagrams as observed in [3] resulting from just a small increase in the number of stored patterns (e.g. $p = 1$ versus 2). More specifically, in contrast to the traditional long-range models, in models with short-range interactions one finds a stronger disruptive effect of non-condensed patterns on the recall of the condensed ones. In order to shed light on such phenomena we extend the original models of [3] by modulating the embedding strengths of the individual stored patterns (and therefore the attractors themselves), as in [9], so that we can smoothly interpolate between, for instance, the $p = 1$ and 2 models. This reveals, as was expected on the basis of the qualitative differences between the $p = 1$ and 2 diagrams, a very complicated and interesting scenario of phase diagram deformation. Both extensions of the original models in [3] introduce technical complications, but these are largely of a quantitative nature, and the extra work needed to again arrive at exact solutions is more than adequately compensated by the richness of the resultant phase diagrams.

2. Definitions

As in [3], each of our extended models is defined as a collection of N binary neuron variables (i.e. Ising spins) $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)$, with $\sigma_i \in \{-1, 1\}$, which evolve in time stochastically and sequentially, following the Glauber-type rule

$$\text{Prob}[\sigma_i(t+1) = \pm 1] = \tfrac{1}{2}[1 \pm \tanh[\beta h_i(\boldsymbol{\sigma}(t))]] \qquad h_i(\boldsymbol{\sigma}) = \sum_{j \neq i} J_{ij}\sigma_j. \tag{1}$$

The parameters J_{ij} represent the synaptic interactions, and the parameter $\beta = 1/T$ controls the amount of stochasticity in the dynamics. If the interaction matrix is symmetric, the process (1) leads to a unique equilibrium state of the Boltzmann type, i.e. with microscopic state probabilities of the form $p_\infty(\boldsymbol{\sigma}) \sim \exp[-\beta H(\boldsymbol{\sigma})]$ and with the conventional Ising Hamiltonian $H(\boldsymbol{\sigma}) = -\sum_{i<j} \sigma_i J_{ij} \sigma_j$. Information processing in such systems is based on the creation and manipulation of attractors in the system's configuration space, by a suitable choice of the spin interactions $\{J_{ij}\}$, which shape the energy landscape. In statistical mechanical terms, the two key aspects of these interactions which determine the analytical solvability or otherwise of the resulting models are (i) the spatial structure defined by the interactions (reflected in which of

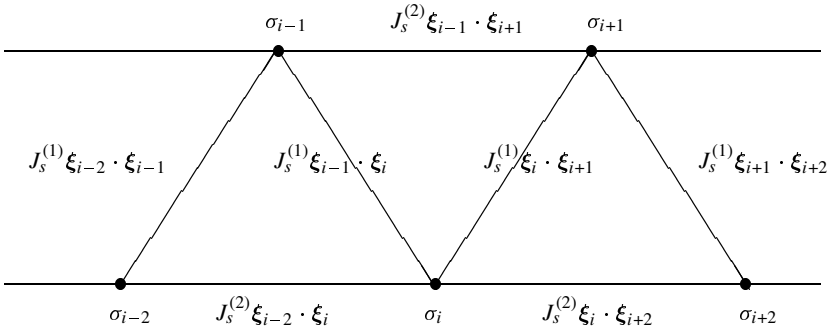

Figure 1. Graphical representation of the spatial structure of model II.

the J_{ij} are non-zero), and (ii) the actual values taken by the non-zero interactions (which will generally be non-trivial, in order to achieve the objective of the creation of specific attractors). For the interaction matrix J_{ij} we now make two different choices, which both generalize the model class of [3], but in qualitatively different ways.

Our first generalization focuses on the values of those interactions which are present, while retaining the mean-field plus nearest-neighbour interaction structure of [3]:

$$\text{model I:}\quad J_{ij}=\sum_{\mu=1}^{p}\left[\frac{J_\mu^\ell}{N}+J_\mu^s(\delta_{j,i+1}+\delta_{j,i-1})\right]\xi_i^\mu\xi_j^\mu \tag{2}$$

in which the components $\xi_i^\mu\in\{-1,1\}$ are all drawn independently at random, with equal probabilities. Neural networks of this type correspond to the result of having stored in a Hebbian-type fashion a set of p binary patterns $\boldsymbol{\xi}_i=(\xi_i^1,\ldots,\xi_i^p)\in\{-1,1\}^N$. The neurons can be thought of as arranged on a one-dimensional array with mean-field interactions between all pairs (i,j) given by $N^{-1}\sum_\mu J_\mu^\ell\xi_i^\mu\xi_j^\mu$, in combination with interactions between nearest neighbours $(i,i+1)$ of strength $\sum_\mu J_\mu^s\xi_i^\mu\xi_{i+1}^\mu$. We will only consider the case where $\lim_{N\to\infty}p/N=0$. The parameters J_μ^s and J_μ^ℓ control the embedding strength of pattern μ in the short- and long-range synapses, with negative values corresponding to the creation of 'repellors' rather than attractors. For uniform embedding strengths, $J_\mu^s=J_s$ and $J_\mu^\ell=J_\ell$ for all μ, we recover [3].

Our second generalization affects the spatial structure of the system, rather than the properties of the attractors (although the latter will be indirectly affected). Here our choice of interactions is

$$\text{model II:}\quad J_{ij}=\left[\frac{J_\ell}{N}+J_s^{(1)}(\delta_{j,i+1}+\delta_{j,i-1})+J_s^{(2)}(\delta_{j,i+2}+\delta_{j,i-2})\right]\sum_{\mu=1}^{p}\xi_i^\mu\xi_j^\mu. \tag{3}$$

In neural networks of type II, the short-range synaptic interactions reach beyond nearest neighbours; here $J_\ell, J_s^{(1)}, J_s^{(2)}\in\mathbb{R}$ control the strengths of long-range, nearest-neighbour and second-nearest-neighbour interactions. Alternatively, in these models the neurons can be thought of as lying on a strip, mutually coupled by infinite range interactions of strength $J_\ell/N\boldsymbol{\xi}_i\cdot\boldsymbol{\xi}_j$, in combination with short-range 'diagonal' interactions of strength $J_s^{(1)}\boldsymbol{\xi}_i\cdot\boldsymbol{\xi}_{i+1}$ and 'edge' interactions of strength $J_s^{(2)}\boldsymbol{\xi}_{i-1}\cdot\boldsymbol{\xi}_{i+1}$ (see figure 1). Note that the models of type II reduce to those in [3] for $J_s^{(2)}=0$. The most relevant observables in our models are the so-called overlap order parameters, defined as $m_\mu(\boldsymbol{\sigma})=N^{-1}\sum_i\xi_i^\mu\sigma_i$, which measure the degree of similarity between the actual network state $\boldsymbol{\sigma}$ and the μth stored pattern.

Due to the presence of short-range interactions in the above models (and, similarly, those of [3]), the solution of even the simplest scenario where $p \ll N$ is already significantly more complicated than solving the standard infinite-range (Hopfield-type, [1,2]) cases. The solution of both models will be based on a suitable adaptation of the random-field techniques of [4].

3. Physics of model I

3.1. Solution via random-field techniques

In order to find the phase diagrams we first isolate the p overlap order parameters by inserting $1 = \int \mathrm{d}\boldsymbol{m}\, \delta[\boldsymbol{m} - \frac{1}{N}\sum_i \sigma_i \boldsymbol{\xi}_i]$ with $\boldsymbol{m} = (m_1, \ldots, m_p)$ and $\boldsymbol{\xi}_i = (\xi_i^1, \ldots, \xi_i^p)$ in the expression for the asymptotic free energy per site $f = -\lim_{N\to\infty}(\beta N)^{-1} \ln Z$, and subsequently replace the delta functions by their integral representations. For model I this leads to

$$\begin{aligned} f &= -\lim_{N\to\infty} \frac{1}{\beta N} \ln \int \mathrm{d}\boldsymbol{m}\, \mathrm{d}\hat{\boldsymbol{m}}\, \mathrm{e}^{-\beta N \phi_N(\boldsymbol{m},\hat{\boldsymbol{m}})} \\ \phi_N(\boldsymbol{m},\hat{\boldsymbol{m}}) &= -\mathrm{i}\hat{\boldsymbol{m}}\cdot\boldsymbol{m} - \frac{1}{2}\sum_\mu J_\mu^\ell m_\mu^2 - \frac{1}{\beta N}\ln R_N(\hat{\boldsymbol{m}}) \end{aligned} \tag{4}$$

where the non-trivial part of the calculation, mainly induced by the short-range interactions, has been concentrated in the term $R_N(\hat{\boldsymbol{m}})$ (we consider non-periodic boundary conditions):

$$\begin{aligned} R_N(\hat{\boldsymbol{m}}) &= \sum_{\boldsymbol{\sigma}} \prod_{i=1}^{N-1} T_{\sigma_i\sigma_{i+1}} \mathrm{e}^{-\mathrm{i}\beta\sum_\mu \hat{m}_\mu \sigma_N \xi_N^\mu} \\ T_{\sigma_i\sigma_{i+1}} &= \mathrm{e}^{-\mathrm{i}\beta\sum_\mu \hat{m}_\mu \sigma_i \xi_i^\mu + \beta\sum_\mu J_\mu^s (\sigma_i\xi_i^\mu)(\sigma_{i+1}\xi_{i+1}^\mu)}. \end{aligned} \tag{5}$$

In the limit $N\to\infty$ the above integral will be evaluated via steepest descent. This results in an expression for f in terms of the relevant saddle-point of the asymptotic form of ϕ_N: $f = \mathrm{extr}_{\boldsymbol{m},\hat{\boldsymbol{m}}} \lim_{N\to\infty} \phi_N(\boldsymbol{m},\hat{\boldsymbol{m}})$. Since the quantity $R(\hat{\boldsymbol{m}})$ does not contain the order parameters $\boldsymbol{m}$, we can immediately take derivatives in (4) with respect to $\boldsymbol{m}$, which allows us to eliminate the conjugate variables $\hat{\boldsymbol{m}}$ via $-\mathrm{i}\hat{m}_\mu = J_\mu^\ell m_\mu$ for all μ. Furthermore we observe that, since for each μ the order parameter m_μ is coupled to the infinite-range embedding strengths J_μ^ℓ, the so-called 'pure-state' ansatz $\boldsymbol{m} = (m, 0, \ldots, 0)$ will automatically render the solution of the model independent of J_μ^ℓ for all $\mu > 1$. From now on we will therefore use the notation $J_1^\ell = J_\ell$. Upon making the pure-state ansatz, the resulting simplifications lead to

$$T_{\sigma_i\sigma_{i+1}} = \mathrm{e}^{\beta J_\ell m \sigma_i \xi_i^1 + \beta\sum_\mu J_\mu^s (\sigma_i\xi_i^\mu)(\sigma_{i+1}\xi_{i+1}^\mu)}. \tag{6}$$

To evaluate (5) we now first define the quantities

$$R_\pm^{(N)}(\boldsymbol{m}) = \sum_{\boldsymbol{\sigma}} \prod_{i=1}^{N-1} T_{\sigma_i\sigma_{i+1}} \mathrm{e}^{\beta J_\ell m \sigma_N \xi_N^1} \delta_{\sigma_N,\pm 1}. \tag{7}$$

These allow us to derive a 2×2 stochastic recurrence relation, mapping $\{R_\pm^{(N-1)}\}$ onto $\{R_\pm^{(N)}\}$:

$$\begin{pmatrix} R_+^{(N)}(\boldsymbol{m}) \\ R_-^{(N)}(\boldsymbol{m}) \end{pmatrix} = \begin{pmatrix} \mathrm{e}^{\beta(J_\ell m\xi_N^1 + \sum_\mu J_\mu^s \xi_{N-1}^\mu \xi_N^\mu)} & \mathrm{e}^{\beta(J_\ell m\xi_N^1 - \sum_\mu J_\mu^s \xi_{N-1}^\mu \xi_N^\mu)} \\ \mathrm{e}^{-\beta(J_\ell m\xi_N^1 + \sum_\mu J_\mu^s \xi_{N-1}^\mu \xi_N^\mu)} & \mathrm{e}^{-\beta(J_\ell m\xi_N^1 + \sum_\mu J_\mu^s \xi_{N-1}^\mu \xi_N^\mu)} \end{pmatrix} \begin{pmatrix} R_+^{(N-1)}(\boldsymbol{m}) \\ R_-^{(N-1)}(\boldsymbol{m}) \end{pmatrix} \tag{8}$$

from which the partition sum of (5) follows as

$$\begin{aligned} -\lim_{N\to\infty}\frac{1}{\beta N}\ln R_N(\boldsymbol{m}) &= -\lim_{N\to\infty}\frac{1}{\beta N}\ln\left[R_+^{(N)}(\boldsymbol{m}) + R_-^{(N)}(\boldsymbol{m})\right] \\ &= -\lim_{N\to\infty}\frac{1}{\beta N}\ln\left\{\begin{pmatrix}1\\1\end{pmatrix}\cdot\left[\prod_{i=2}^{N}\boldsymbol{T}_i\right]\begin{pmatrix} R_+^{(1)}(\boldsymbol{m}) \\ R_-^{(1)}(\boldsymbol{m})\end{pmatrix}\right\}. \end{aligned} \tag{9}$$

The successive matrix multiplications above can be simplified via the use of the following ratio of the conditioned quantities $\{R_+^{(j)}(\boldsymbol{m}),\, R_-^{(j)}(\boldsymbol{m})\}$:

$$k_j = \mathrm{e}^{2\beta m J_\ell \xi_j^1} \frac{R_-^{(j)}(\boldsymbol{m})}{R_+^{(j)}(\boldsymbol{m})}.$$

It now follows from (8) that these numbers k_i are, in turn, generated by the following stochastic process:

$$k_{j+1} = \frac{\mathrm{e}^{-\beta \sum_\mu J_\mu^s \xi_j^\mu \xi_{j+1}^\mu} + k_j \mathrm{e}^{-\beta \sum_\mu J_\mu^s \xi_j^\mu \xi_{j+1}^\mu} \mathrm{e}^{2\beta m J_\ell \xi_j^1}}{\mathrm{e}^{\beta \sum_\mu J_\mu^s \xi_j^\mu \xi_{j+1}^\mu} + k_j \mathrm{e}^{\beta \sum_\mu J_\mu^s \xi_j^\mu \xi_{j+1}^\mu} \mathrm{e}^{2\beta m J_\ell \xi_j^1}}. \tag{10}$$

The stochasticity here is in the pattern components $\{\xi_i^\mu\}$. This allows us to work out the partition sum and express the asymptotic free energy per neuron as $f = \mathrm{extr}_m f(m)$, with

$$f(m) = \frac{1}{2} J_\ell m^2 - \frac{1}{\beta} \int \mathrm{d}k \sum_{\boldsymbol{\xi},\boldsymbol{\xi}' \in \{-1,1\}^p} \rho(k, \boldsymbol{\xi}, \boldsymbol{\xi}') \log\{\mathrm{e}^{\beta \sum_\mu J_\mu^s \xi_\mu \xi'_\mu} + k \mathrm{e}^{-\beta \sum_\mu J_\mu^s \xi_\mu \xi'_\mu} \mathrm{e}^{2\beta J_\ell m \xi_1}\} \tag{11}$$

and

$$\rho(k, \boldsymbol{\xi}, \boldsymbol{\xi}') = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N-1} \delta[k - k_i]\, \delta_{\boldsymbol{\xi},\boldsymbol{\xi}_i} \delta_{\boldsymbol{\xi}',\boldsymbol{\xi}'_{i+1}}. \tag{12}$$

The joint distribution (12), which is the invariant distribution for the process (10) and which can be highly non-trivial [3] (depending on the choice of system parameters, the associated integrated density can take the shape of a devil's staircase), is in practice calculated numerically. In the present case one can in fact simplify matters further by exploiting symmetry properties of $\rho(k, \boldsymbol{\xi}, \boldsymbol{\xi}')$ resulting from the homogeneous distribution assumed for the $\{\xi_i^\mu\}$.

3.2. Phase diagrams and comparison with numerical simulations

We can now extract the macroscopic characteristics of model I by generating the variables $\{k_j; \forall j \leqslant N\}$ numerically (together with the $\{\xi_i^\mu\}$), which leads us to the joint distribution (12), and by subsequently evaluating (numerically) the local minima of the free energy surface (11). We show in figure 2 the resulting phase diagrams, for $\beta J_2^s = \{0, -3/2, -5/2, \beta J_1^s\}$ and for the simplest non-trivial case $p = 2$. In all graphs dashed curves correspond to second-order transitions and solid lines to first-order ones.

Note that for $J_2^s = 0$ (see figure 2, upper left panel) only pattern $\mu = 1$ is effectively embedded in the spin chain, and the phase diagram of the model is identical to that found earlier in [3] for $p = 1$, as it should be. In this diagram we observe, apart from strictly null-recall (N) and recall (R) phases, that there is also a region in which the trivial solution and two non-trivial ones (one with positive and one with negative m) can be locally stable simultaneously (indicated by N_2)[1]. This region corresponds to parameter values for which the two different types of synapse compete most strongly (negative nearest-neighbour interactions versus positive infinite-range ones). It is separated from the recall region by a second-order transition (dashed curve), and from the null-recall region by a first-order transition (solid line); these two lines come together at $\{\beta J_\ell, \beta J_1^s\} = \{\sqrt{3}, -\frac{1}{4}\ln 3\}$. Another benchmark solution of [3] is recovered for the special case of having uniform short-range embedding strengths

[1] From now on, regions which allow for locally stable null-recall solutions will be denoted by N_i, with i indicating the number of simultaneously locally stable recall solutions. Similarly, regions which do not allow for locally stable null-recall solutions will be denoted by R_i.

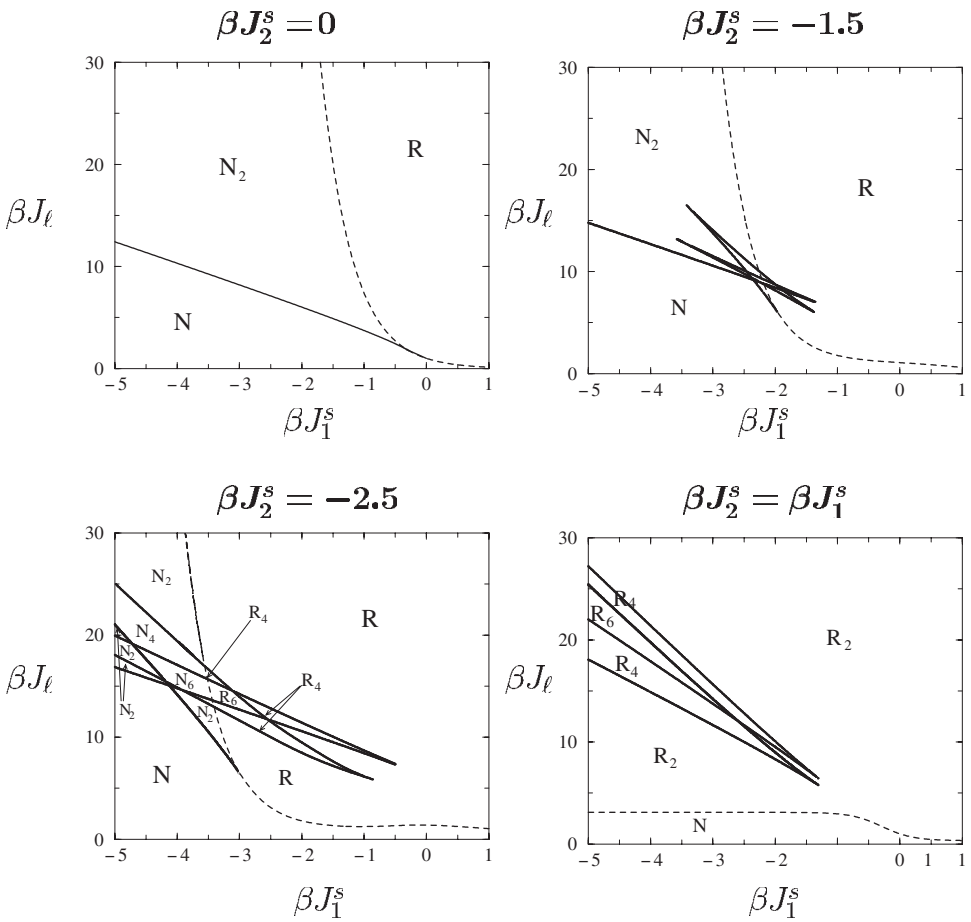

Figure 2. Phase diagram cross-sections of model type I for $p = 2$, upon making the 'pure-state' ansatz for pattern $\mu = 1$, with $J_\ell = J_1^\ell$. The parameters J_1^s and J_2^s control the short-range embedding strength of patterns $\mu = 1$ and 2, respectively, J_ℓ represents the strength of the mean-field interactions and $\beta = T^{-1}$ is the inverse temperature. In the absence of pattern $\mu = 2$ (upper left) or for equally strong short-range embedding strengths (lower right), we recover [3]. Solid/dashed lines denote first/second-order transitions. Regions R and N represent strictly recall or null-recall regions, whereas R_i and N_i correspond to regions where the trivial solution $m = 0$ is (R) or is not (N) locally stable, and with $i \in \{2, 4, 6\}$ giving the number of locally stable $m \neq 0$ solutions.

$J_2^s = J_1^s$ (see figure 2, lower right panel). Here, two 'pairs' of first-order transition lines have appeared, which separate the regions R_4 and R_6 (where $m = 0$ is unstable and where four and six $m \neq 0$ solutions are possible, depending on initial conditions) from the N region. Here the second-order transition line does not touch any of the first-order ones.

The remarkable qualitative difference between the phase diagrams which the aforementioned two special cases produce is striking; in our previous study the physical origin of this difference was not studied. In particular, although the correctness of the solution had been tested in [3] against extensive numerical simulations, it was not at all clear how and why the second-order transition line (dashed curve) would change from the exponentially rising curve (figure 2, upper left) to the one in the lower right corner of figure 2. Our present model generalizes [3], and allows for independent tuning of the short-range embedding strengths:

one can now bring the non-condensed pattern to life in a continuous way. The top right and lower left panels of figure 2, where $J_2^s \neq 0$, show how the system realizes the transition from the $p = 1$ case to the $p = 2$ case (where $J_1^s = J_2^s$). The four panels in the figure can be thought of as different cross-sections of an extended graph in the area $\{\beta J_1^s, \beta J_2^s, \beta J_\ell\}$, which in combination reveal the underlying complexity of the model; due to the competing short- and long-range forces and the high degree of frustration new regions come to life in parameter space, with multiple locally stable overlap solutions. In contrast to the $J_1^s = J_2^s$ phase diagram, where $m = 0$ is unstable everywhere, apart from the strictly null-recall phase, the other two $J_2^s \neq 0$ phase diagrams display regions (N_2, N_4 and N_6) where $m = 0$ coexists as a locally stable state together with multiple locally stable $m \neq 0$ solutions. These latter new phase diagrams appear significantly richer than those found in [3], owing to the breaking of the pattern embedding strength symmetry.

To test and verify our results we have performed extensive simulation experiments. Initial configurations $\boldsymbol{\sigma}(t = 0)$ were chosen randomly, according to

$$p(\boldsymbol{\sigma}(0)) = \prod_i \{\tfrac{1}{2}[1 + m_0]\delta_{\sigma_i(0),\xi_i^1} + \tfrac{1}{2}[1 - m_0]\delta_{\sigma_i(0),-\xi_i^1}\}.$$

In figure 3 we plot the equilibrium value $m_1(t \to \infty)$ of the main order parameter as a function of its initial value $m_1(t = 0)$ (black circles), in order to probe the existence and location of multiple ergodic components. To enable comparison with the theoretically predicted equilibrium values we also show (insets) the dependence of the asymptotic free energy per neuron on the order parameter $m = m_1(t \to \infty)$, as contructed from equations (11) and (16); its local minima are indeed located at those values which are found as allowed equilibrium states in the simulations, given appropriate initial conditions, and within the experimental margin of accuracy. With our system size $N = 1000$, finite-size effects are expected to be of the order of $\mathcal{O}(N^{-\frac{1}{2}}) \approx 0.03$. Our restriction to relatively small system sizes was prompted by the appearance of extremely large equilibration times, due to domain formation. For the case of predominant long-range interactions equilibration was achieved within $\approx 10^4$ flips/spin. For predominant short-range interactions, however, domain formation led to equilibration times which were observed to scale exponentially with the system size, see figure 4. For this reason the observed value of $m_{\rm equi}$ for the ergodic component closest to $m = 0$ in figure 3 appears to differ from the theoretically predicted value $m = 0$ by roughly $0.06 > \mathcal{O}(N^{-\frac{1}{2}})$. In figure 4 we show that for the parameter choice of figure 3 and for $m_1(t = 0) = 0.08$ the system is indeed approaching the predicted state $m_{\rm equil} = 0$, but extremely slowly. Finally, we have also measured the equilibrium overlaps with the non-'condensed' pattern, $m_2(t \to \infty)$, which are seen to remain zero (open diamonds in figure 3), which justifies our 'pure-state' ansatz.

4. Physics of model II

4.1. Solution via random field techniques

Neural network models of type II can be solved analytically using the same techniques as applied to model type I, although here the calculations will be somewhat more elaborate. Upon again making the 'pure-state' ansatz: $\boldsymbol{m} = (m, 0, \ldots, 0)$ and upon eliminating the conjugate order parameters $\hat{\boldsymbol{m}}$ via saddle-point equations, we find that the asymptotic free energy per neuron is given by $f = \mathrm{extr}_m f(m)$, with

$$f(m) = \frac{1}{2} J_\ell m^2 - \lim_{N\to\infty} \frac{1}{\beta N} \ln R_N(m)$$

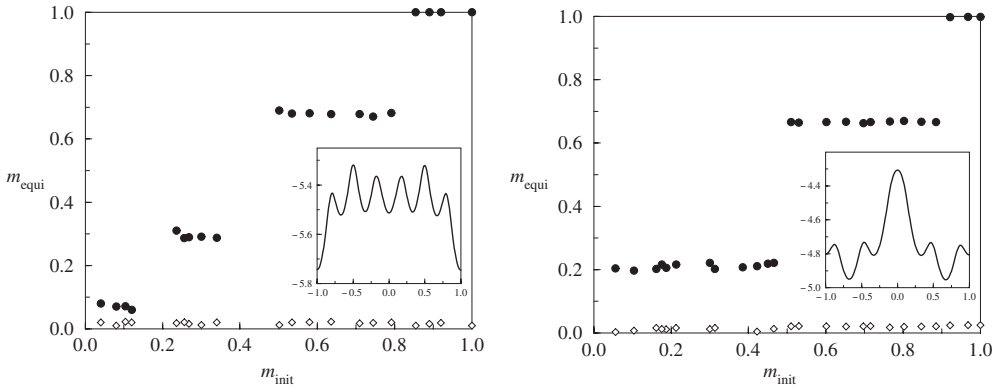

Figure 3. Simulation results for model type I, for system size $N = 1000$ and $p = 2$, showing the equilibrium value (i.e. that obtained after 10 000 iterations per spin) of the 'condensed' overlap $m_{\rm equi} = m_1(t \to \infty)$ as a function of the initial value $m_{\rm init} = m_1(t = 0)$ (solid circles). The initial configurations were drawn at random, subject to the constraint imposed by the required value of $m_1(t = 0)$. The theoretically predicted locations of the ergodic components, as contructed from equation (11), are also shown as local minima of the free energy per neuron in the insets, for comparison. Open diamonds represent the equilibrium values of the non-condensed overlaps $m_2(t)$; they are seen to remain zero, which justifies *a posteriori* the 'pure-state' ansatz. Left plot: $\beta J_2^s = -3.5$, $\beta J_1^s = -4.2$, $\beta J_\ell = 18$. Right plot: $\beta J_1^s = -5.5$, $\beta J_2^s = -3.5$, $\beta J_\ell = 23$.

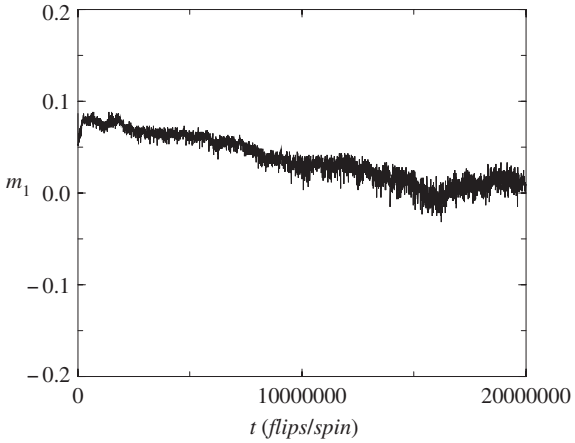

Figure 4. Simulation results for model type I, for system size $N = 1200$ and $p = 2$, showing the 'condensed' overlap m_1 as a function of time (measured in iterations per neuron). The embedding strengths were given by $\beta J_2^s = -3.5$, $\beta J_1^s = -4.2$, $\beta J_\ell = 18$. The relaxation towards zero, following a small (but non-zero) initial value, is seen to be extremely slow, due to domain formation.

where the complicated part of the partition sum is in the last term:

$$R_N(m) = \sum_{\sigma} \left[\prod_{i=1}^{N-2} T_{\sigma_i \sigma_{i+1} \sigma_{i+2}} \right] e^{\beta J_s^{(1)} (\sigma_{N-1}\xi_{N-1})\cdot(\sigma_N \xi_N)} e^{\beta J_\ell m \{\xi_{N-1}\sigma_{N-1} + \xi_N \sigma_N\}}$$
$$T_{\sigma_i \sigma_{i+1} \sigma_{i+2}} = e^{\beta J_s^{(1)} \sum_i (\sigma_i \xi_i)\cdot(\sigma_{i+1}\xi_{i+1}) + \beta J_s^{(2)} \sum_i (\sigma_i \xi_i)\cdot(\sigma_{i+2}\xi_{i+2}) + \beta J_\ell m \sum_i \sigma_i \xi_i} \tag{13}$$

(with open boundary conditions). As in model I we next derive a recurrence relation for conditioned partition sums. For the present model we find that this can be achieved in terms

of the following four quantities:

$$R^{(N)}_{\pm\pm}(\boldsymbol{m}) = \sum_{\boldsymbol{\sigma}} \prod_{i=1}^{N-2} T_{\sigma_i\sigma_{i+1}\sigma_{i+2}} \mathrm{e}^{\beta J_s^{(1)}(\sigma_{N-1}\xi_{N-1})\cdot(\sigma_N\xi_N)} \mathrm{e}^{\beta J_\ell m\{\xi_{N-1}\sigma_{N-1}+\xi_N\sigma_N\}} \delta_{\sigma_{N-1},\pm 1}\ \delta_{\sigma_N,\pm 1}.$$

These are found to be successively generated by a 4×4 linear but stochastic iterative process of the form $\boldsymbol{R}_{N+1} = \boldsymbol{T}_{N+1}\boldsymbol{R}_N$, where $\boldsymbol{R}_j = (R^{(j)}_{++}, R^{(j)}_{+-}, R^{(j)}_{-+}, R^{(j)}_{--})$, the stationary state of which will produce the free energy per neuron. The 4×4 random matrix $\boldsymbol{T}_{N+1}$ can be decomposed further into two coupled 2×2 random matrices:

$$\begin{pmatrix} R^{(N+1)}_{++} \\ R^{(N+1)}_{+-} \end{pmatrix} = \begin{pmatrix} \mathrm{e}^{\beta J_\ell m\xi_{N+1}} & 0 \\ 0 & \mathrm{e}^{-\beta J_\ell m\xi_{N+1}} \end{pmatrix} \begin{pmatrix} L_{N,+} & L_{N,-} \\ L^{-1}_{N,+} & L^{-1}_{N,-} \end{pmatrix} \begin{pmatrix} R^{(N)}_{++} \\ R^{(N)}_{-+} \end{pmatrix} \tag{14}$$

$$\begin{pmatrix} R^{(N+1)}_{-+} \\ R^{(N+1)}_{--} \end{pmatrix} = \begin{pmatrix} \mathrm{e}^{\beta J_\ell m\xi_{N+1}} & 0 \\ 0 & \mathrm{e}^{-\beta J_\ell m\xi_{N+1}} \end{pmatrix} \begin{pmatrix} L^{-1}_{N,-} & L^{-1}_{N,+} \\ L_{N,-} & L_{N,+} \end{pmatrix} \begin{pmatrix} R^{(N)}_{+-} \\ R^{(N)}_{--} \end{pmatrix} \tag{15}$$

where

$$L_{N,\pm} = \mathrm{e}^{\beta(J_s^{(1)}\xi_{N-1}\cdot\xi_N \pm J_s^{(2)}\xi_{N-1}\cdot\xi_{N+1})}.$$

The partition sum in (13) can now be written in terms of the successive multiplication of the random matrices $\boldsymbol{T}$:

$$-\lim_{N\to\infty}\frac{1}{\beta N}\ln R_N = -\lim_{N\to\infty}\frac{1}{\beta N}\ln\left\{\begin{pmatrix}1\\1\\1\\1\end{pmatrix}\cdot\left[\prod_{i=3}^{N}\boldsymbol{T}_i\right]\boldsymbol{R}_2\right\}.$$

Similar to the analysis performed for model I we again define ratios of conditioned partition functions (although here we will need three rather than one):

$$k_j^{(1)} = \mathrm{e}^{-2\beta J_\ell m\xi_j}\frac{R^{(j)}_{++}}{R^{(j)}_{+-}} \qquad k_j^{(2)} = \mathrm{e}^{2\beta J_\ell m\xi_j}\frac{R^{(j)}_{+-}}{R^{(j)}_{-+}} \qquad k_j^{(3)} = \mathrm{e}^{-2\beta J_\ell m\xi_j}\frac{R^{(j)}_{-+}}{R^{(j)}_{--}}.$$

According to (14), (15) these ratios are generated by the following stochastic processes:

$$k^{(1)}_{j+1} = \frac{\mathrm{e}^{\beta J_s^{(2)}\xi_{j-1}\cdot\xi_{j+1}}k_j^{(1)}k_j^{(2)} + \mathrm{e}^{-\beta J_s^{(2)}\xi_{j-1}\cdot\xi_{j+1}}}{\mathrm{e}^{-\beta J_s^{(2)}\xi_{j-1}\cdot\xi_{j+1}}k_j^{(1)}k_j^{(2)} + \mathrm{e}^{\beta J_s^{(2)}\xi_{j-1}\cdot\xi_{j+1}}}\mathrm{e}^{2\beta J_s^{(1)}\xi_{j-1}\cdot\xi_j}$$

$$k^{(2)}_{j+1} = \frac{\mathrm{e}^{-\beta J_s^{(2)}\xi_{j-1}\cdot\xi_{j+1}}k_j^{(1)}k_j^{(2)} + \mathrm{e}^{\beta J_s^{(2)}\xi_{j-1}\cdot\xi_{j+1}}}{\mathrm{e}^{\beta J_s^{(2)}\xi_{j-1}\cdot\xi_{j+1}}k_j^{(2)}k_j^{(3)} + \mathrm{e}^{-\beta J_s^{(2)}\xi_{j-1}\cdot\xi_{j+1}}}k_j^{(3)}\mathrm{e}^{2\beta J_\ell m\xi_j}$$

$$k^{(3)}_{j+1} = \frac{\mathrm{e}^{\beta J_s^{(2)}\xi_{j-1}\cdot\xi_{j+1}}k_j^{(2)}k_j^{(3)} + \mathrm{e}^{-\beta J_s^{(2)}\xi_{j-1}\cdot\xi_{j+1}}}{\mathrm{e}^{-\beta J_s^{(2)}\xi_{j-1}\cdot\xi_{j+1}}k_j^{(2)}k_j^{(3)} + \mathrm{e}^{\beta J_s^{(2)}\xi_{j-1}\cdot\xi_{j+1}}}\mathrm{e}^{-2\beta J_s^{(1)}\xi_{j-1}\cdot\xi_j}.$$

The free energy per neuron can now be expressed in terms of the stationary distribution of this stochastic process:

$$f(m) = \frac{1}{2}J_\ell m^2 - \lim_{N\to\infty}\frac{1}{\beta}\int \mathrm{d}\boldsymbol{k}\sum_{\boldsymbol{\xi},\boldsymbol{\xi}'\in\{-1,1\}^p}\rho(\boldsymbol{k},\boldsymbol{\xi},\boldsymbol{\xi}')\ln\{\mathrm{e}^{\beta J_s^{(2)}\boldsymbol{\xi}\cdot\boldsymbol{\xi}'} + \mathrm{e}^{-\beta J_s^{(2)}\boldsymbol{\xi}\cdot\boldsymbol{\xi}'}\ k^{(2)}k^{(3)}\} \tag{16}$$

with

$$\rho(\boldsymbol{k},\boldsymbol{\xi},\boldsymbol{\xi}') = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N-3}\delta[\boldsymbol{k}-\boldsymbol{k}_i]\delta_{\boldsymbol{\xi},\boldsymbol{\xi}_i}\delta_{\boldsymbol{\xi}',\boldsymbol{\xi}'_{i+2}} \tag{17}$$

and $\boldsymbol{k}_i = (k_i^{(1)}, k_i^{(2)}, k_i^{(3)})$.

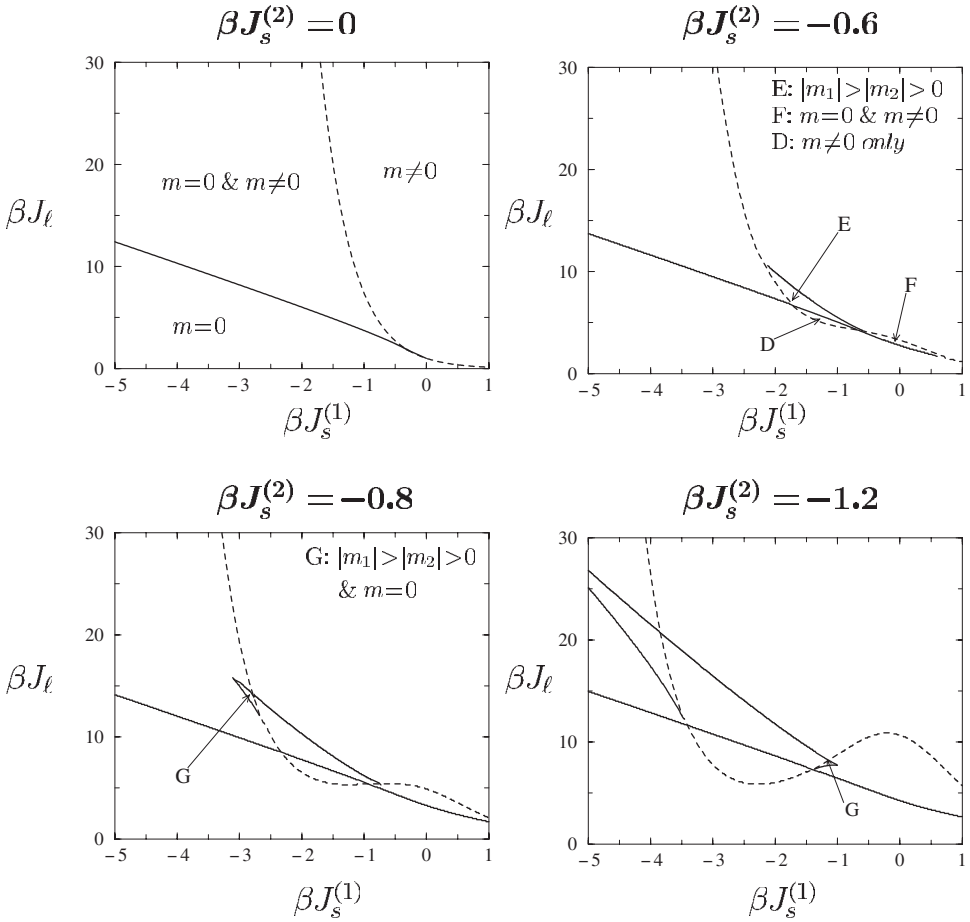

Figure 5. Phase diagram cross-sections of model type II for $p = 1$, with $J_s^{(1)}$, $J_s^{(2)}$ and J_ℓ denoting nearest-neighbour, next-nearest-neighbour and long-range embedding strengths, respectively. In the absence of next-nearest-neighbour interactions (upper left) we recover the model of [3], whereas for $J_s^{(2)} \neq 0$ new regions appear in the phase diagram, with different numbers of simultaneously locally stable solutions for the 'overlap' order parameter m. Solid curves denote first-order transitions; dashed curves denote second-order ones.

4.2. Phase diagrams and comparison with numerical simulations

Numerical evaluation of the energy surface defined by (16) leads to the phase diagrams shown in figures 5 and 6, which describe the cases $p = 1$ and 5, respectively. They are drawn in the $\{\beta J_s^{(1)}, \beta J_\ell\}$ plane, for four different values of the next-nearest-neighbour embedding strength $J_s^{(2)}$. In all phase diagrams the solid lines represent continuous (second-order) phase transitions, whereas the dashed curves correspond to discontinuous (first-order) ones. In the absence of next-nearest-second-neighbour interactions, i.e. for $J_s^{(2)} = 0$ (upper left graph in figure 5), our model reduces to that of [3]. For $J_s^{(2)} > 0$ one finds no new phase regimes, compared with the $J_s^{(2)} = 0$ case; the two transition lines of the $J_s^{(2)} = 0$ phase diagram are found to simply move towards $\beta J_s^{(1)} = \infty$. However, as soon as $J_s^{(2)} < 0$, frustration effects become more important, with new regions appearing in the phase diagram as a result. In the upper right graph of figure 5, where $\beta J_s^{(2)} = -0.6$, we observe that three new regions have been created: region D (with

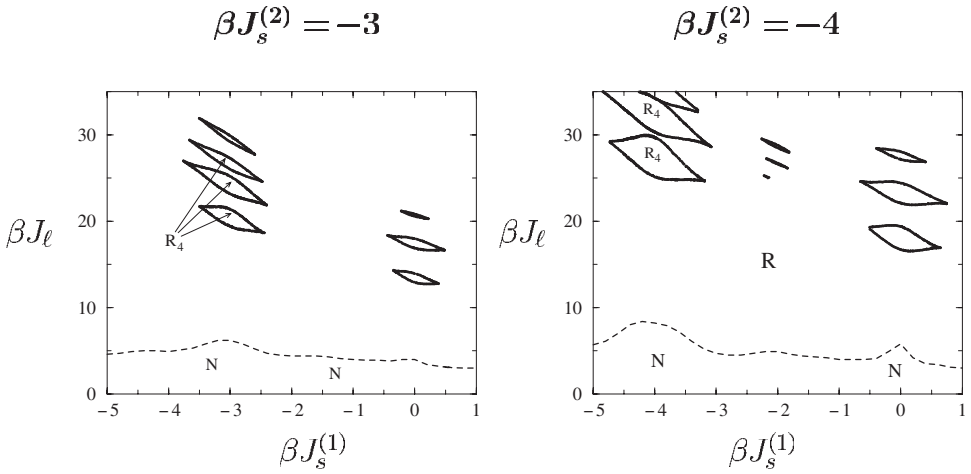

Figure 6. Phase diagram cross-sections for model type II, for $p = 5$ and upon making the 'pure-state' ansatz, with $J_s^{(1)}$, $J_s^{(2)}$ and J_ℓ denoting nearest-neighbour, next-nearest-neighbour and long-range embedding strengths, respectively. The two graphs show typical results for the $p > 1$ phase phenomenology. The 'islands' correspond to regions with four simultaneously locally stable states (two with positive m, and two with negative m). The structural differences between the $p = 1$ and $p > 1$ diagrams can be understood upon modulating the embedding strengths of the stored patterns, as with model type I.

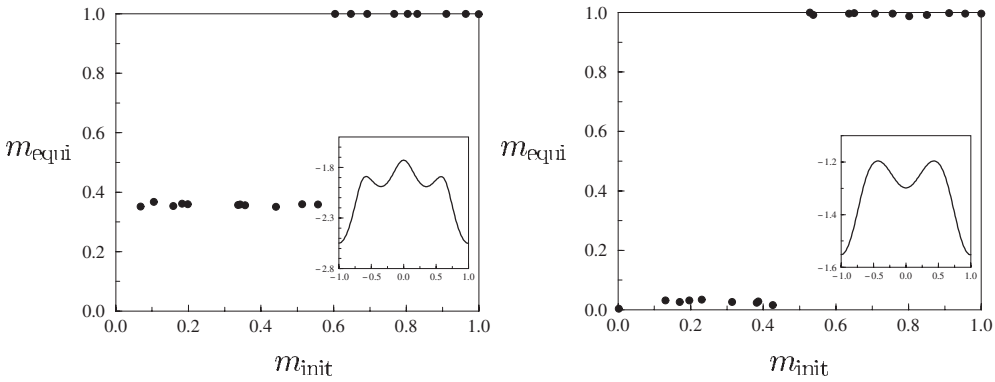

Figure 7. Simulation results for model II, for system size $N = 1000$ and $p = 1$, showing the equilibrium value of the 'condensed' overlap $m_{\mathrm{equi}} = m_1(t \to \infty)$ as a function of the initial value $m_{\mathrm{init}} = m_1(t = 0)$ (solid circles). The initial configurations were drawn at random, subject to the constraint imposed by the required value of $m_1(t = 0)$. The theoretically predicted locations of the ergodic components, as constructed from equation (16), are also shown as local minima of the free energy per neuron in the insets, for comparison. Left plot: $\beta J_s^{(1)} = -2.5$, $\beta J_s^{(2)} = -1.2$, $\beta J_\ell = 12.5$. Right plot: $\beta J_s^{(1)} = -0.5$, $\beta J_s^{(2)} = -1.2$, $\beta J_\ell = 6.5$.

$m \neq 0$), region F (where $m = 0$ and $m \neq 0$ are simultaneously locally stable) and region E (with two positive and two negative locally stable $m \neq 0$ states). In the lower left graph, where $\beta J_s^{(2)} = -0.8$, we see that, in addition to the previously created regions, a further new region G comes to life, where the trivial state as well as four $m \neq 0$ ones are all simultaneously locally stable. For $p > 1$, first-order transition lines are found to emerge as boundaries of 'islands' in the $\{\beta J_s^{(1)}, \beta J_\ell\}$ plane, where four $m \neq 0$ solutions are simultaneously locally

stable. For increasingly negative values of $J_s^{(2)}$ these islands expand in size, and at some point start overlapping, which creates additional new regions. In figure 6 we show typical phase diagrams for the case $p = 5$. Extensive numerical work also shows that increasing the number of patterns p further leads to the appearance of further transition lines in the phase diagrams. This is due to the explicit dependence of the free energy per neuron, as defined by equation (16), on p. Unlike the more conventional long-range Hopfield-type networks [1,2], where after the 'pure-state' ansatz has been made the macroscopic observables have become independent of p, in the present model the short-range interactions ensure that thermal fluctuations around a pure state will always induce non-negligible p-dependent interference on the recall of the pure state.

To test our results we have again performed extensive numerical simulation experiments, the results of which are shown in figure 7. The initial states were drawn similar to those in the simulations of model type I. We plot the equilibrium overlap oder parameter $m_1(t \to \infty)$ as a function of its initial value $m_1(t = 0)$, to probe different ergodic components, and compare the locations of these components (in terms of the associated values of $m_1(t \to \infty)$) with the theoretical prediction by also showing (inset graphs) the asymptotic free energy per site as constructed from equations (11) and (16). The agreement between the two is quite satisfactory, and well within the error margin given by the finite-size effects (with $N = 1000$ these are estimated at $\mathcal{O}(N^{-\frac{1}{2}}) \approx 0.03$).

5. Discussion

In this paper we presented an equilibrium statistical mechanical analysis of a generalized family of recurrent neural network models, with information stored in the form of attractors in the neuronal state space, but with one-dimensional spatial structure and competition between short- and long-range information processing. We have solved two specific classes of problems. In the first class, patterns are embedded in both the long-range and nearest-neighbour interactions of the neuronal chain, but with pattern-dependent embedding strengths (similar to [9]). This generalizes a previous study [3], where all embedding strengths were independent of the pattern labels. The breaking of the previous embedding strength symmetry is found to yield significantly richer phase diagrams, and, moreover, serves to elucidate the remarkable structural differences which were observed (but not understood) in [3] between the phase diagrams for the two simplest cases $p = 1$ and 2. In our second class of models, which is a qualitatively different generalization of the models in [3], our neurons are equipped with next-nearest-neighbour interactions (in addition to the long-range and the nearest-neighbour ones), which increases significantly the potential for frustration and competition, given appropriate choices of the various pattern embedding strengths. We have been able to solve our models exactly, dealing with the random transfer matrix multiplications in the relevant partition sums (generated by the short-range interactions) using suitable adaptations of the random-field techniques presented in [4]. Alternatively, one could solve our present models using the random-field techniques of [5], which provides an independent theoretical test of our solution: we have carried out this test, and found full agreement (see appendix). For both model types we found surprisingly rich phase diagrams, with qualitatively distinct topologies, and interesting scenarios of phase diagram deformation when appropriate control parameters are varied. Extensive numerical simulation experiments support our theoretical results convincingly, in terms of the appearance and location of the multiple ergodic components in phase space.

Note that we have concentrated in this paper on the analysis of the phenomenology of dynamical phase transitions, i.e. we have concerned ourselves with the *local stability* of

extremal points of the free energy, written as a function of the main order parameter, rather than with the actual value of the free energy in the various locally stable states. In the case of large recurrent neural networks one is simply not interested in thermodynamic transitions, since the timescales relevant to their operation as associative memories are much smaller than the escape times of locally stable states (which diverge with the system size); the locally stable states, and their domains of attraction, determine the relevant physical properties. The models and methods in this paper can be adapted in a straightforward manner to cover systems with non-binary neuron states, or other types of one-dimensional architecture; the inclusion of more distant short-range interactions, however, will lead to higher-dimensional random transfer matrices, and the calculations will become more involved. Alternatively, one could turn to models with synchronous rather than sequential dynamics.

Long-range models, as in [1,2], have been of immense value in shaping our understanding of information processing in attractor neural networks, but are far removed from biological reality. Our present study emphasizes once more the richness of attractor neural networks with spatial structure, and their analytical solvability (at least, within the context of $(1+\infty)$-dimensional models). Not only can exact solutions be obtained beyond the familiar infinite-range models, but one can also generalize the relatively simple but solvable $(1+\infty)$-dimensional models of [3], increasing again (albeit with small steps) their biological relevance. This allows one to investigate further (quantitatively) the significant impact of simple forms of spatial structure on information processing via the manipulation of attractors in recurrent neural networks.

Appendix. Comparison with Rujan's solution

An alternative method to calculate the partition sums (5) and (13) is provided by the random-field technique of [5]. This requires the result of each of the individual spin summations to be written in the form

$$\text{model I:}\quad \sum_{\sigma_{j-1}} T_{\sigma_{j-1}\sigma_j} = \mathrm{e}^{h_j(\xi^\mu_{j-1},\xi^\mu_j)\sigma_j + L_j(\xi^\mu_{j-1},\xi^\mu_j)} \tag{18}$$

for some $\{h_j, L_j\}$, for $\sigma_j \in \{-1, 1\}$. This transformation allows us to evaluate for $N \to \infty$ the non-trivial part of (4)

$$\begin{aligned} -\lim_{N\to\infty}\frac{1}{\beta N}\ln R(\boldsymbol{m}) &= -\lim_{N\to\infty}\frac{1}{2\beta N}\sum_i \ln\left[4\cosh\left[\beta\sum_\mu (J^s_\mu \xi^\mu_i \xi^\mu_{i+1}) + \beta J_\ell m \xi^1_i + h_i\right]\cosh\right. \\ &\quad \left.\times\left[\beta\sum_\mu (J^s_\mu \xi^\mu_i \xi^\mu_{i+1}) + \beta J_\ell m \xi^1_i - h_i\right]\right] \end{aligned}$$

where in the thermodynamic limit we will assume that the asymptotic distribution of the stochastic variables $\{h_i\}$ is uniquely generated by the process

$$h_{i+1} = \frac{1}{2}\ln\frac{\cosh[\beta\sum_\mu (J^s_\mu \xi^\mu_i \xi^\mu_{i+1}) + \beta J_\ell m \xi^1_i + h_i]}{\cosh[\beta\sum_\mu (J^s_\mu \xi^\mu_i \xi^\mu_{i+1}) + \beta J_\ell m \xi^1_i - h_i]} \tag{19}$$

becomes stationary.

In a similar fashion one can perform for $N \to \infty$ the partition sum in (13) requiring

$$\text{model II:}\quad \sum_{\sigma_{j-1}} T_{\sigma_{j-1}\sigma_j\sigma_{j+1}} = \mathrm{e}^{h^{(1)}\sigma_j\sigma_{j+1} + h^{(2)}\sigma_j + h^{(3)}\sigma_{j+1} + L_{j-1}} \tag{20}$$

to be true for $\sigma_j, \sigma_{j+1} \in \{-1, 1\}$. This allows us to write

$$-\lim_{N\to\infty} \frac{1}{\beta N} \ln R(m) = -\lim_{N\to\infty} \frac{1}{4\beta N} \sum_i \ln\{2^4\, \Omega_{++}^{(i)}\, \Omega_{+-}^{(i)}\, \Omega_{-+}^{(i)} \Omega_{--}^{(i)}\}$$

where the quantities $\Omega_{\lambda\lambda'}^{(i)}$ are obtained iteratively as functions of three stochastically evolving variables $\{h_{j-1}^{(1)}, h_{j-1}^{(2)}, h_{j-2}^{(3)}; \forall\ j \leqslant i\}$:

$$\Omega_{\lambda\lambda'}^{(i)} = \cosh[(\beta J_s^{(1)} \boldsymbol{\xi}_i \cdot \boldsymbol{\xi}_{i+1} + h_{i-1}^{(1)} \theta(\mathrm{i}-2))\lambda + \beta J_s^{(2)} \boldsymbol{\xi}_i \cdot \boldsymbol{\xi}_{i+2} \lambda' + \beta J_\ell m \xi_i + h_{i-1}^{(2)} \theta(\mathrm{i}-2) + h_{i-2}^{(3)} \theta(\mathrm{i}-3)]$$

where $\theta(j) = 1$ if $j \geqslant 0$ and $\theta(j) = 0$ otherwise, and with

$$h_i^{(1)} = \frac{1}{4} \ln \frac{\Omega_{++}^{(i)} \Omega_{--}^{(i)}}{\Omega_{+-}^{(i)} \Omega_{-+}^{(i)}} \qquad h_i^{(2)} = \frac{1}{4} \ln \frac{\Omega_{++}^{(i)}\, \Omega_{+-}^{(i)}}{\Omega_{-+}^{(i)} \Omega_{--}^{(i)}} \qquad h_i^{(3)} = \frac{1}{4} \ln \frac{\Omega_{++}^{(i)}\, \Omega_{-+}^{(i)}}{\Omega_{+-}^{(i)} \Omega_{--}^{(i)}}. \tag{21}$$

Numerical iteration of the processes (19) and (21) and subsequently evaluation of the asymptotic free energies and of the order parameters shows excellent agreement with the results found earlier for models I and II.

References

[1] Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554–8
[2] Amit D J, Gutfreund H and Sompolinsky H 1995 *Phys. Rev.* A **32** 1007–18
[3] Skantzos N S and Coolen A C C 2000 *J. Phys. A: Math. Gen.* **33** 5785–807
[4] Bruinsma R and Aeppli G 1983 *Phys. Rev. Lett.* **50** 1494–7
[5] Rujan P 1978 *Physica* A **91** 549–62
[6] Grinstein G and Mukamel D 1983 *Phys. Rev.* B **27** 4503–6
[7] Ben-Yishai R, Bar-Or R L and Sompolinsky H 1995 *Proc. Natl Acad. Sci. USA* **93** 3844–8
[8] Douglas R J, Koch C, Mahowald M A, Martin K A C and Suarez H 1995 *Science* **269** 981–5
[9] Viana L 1988 *J. Physique* **49** 167–74