Reviews in the Neurosciences, 14, 181-193 (2003)

Statistical Mechanics Beyond the Hopfield Model: Solvable Problems in Neural Network Theory

A.C.C. Coolen and V. Del Prete

Department of Mathematics, King's College London, UK

SYNOPSIS

We present four 'case study' examples of solvable problems in the theory of recurrent neural networks, which are relevant to our understanding of information processing in the brain, but which are also interesting from a purely statistical mechanical point of view, even at the level of simple models (which helps in stimulating interdisciplinary work). The examples concern issues in network dynamics, network connectivity, spike timing and synaptic plasticity.

KEY WORDS

recurrent neural networks, self-programming, synchronization

1. INTRODUCTION

Over the years, a variety of tools from many distinct areas of mathematics and theoretical physics have been used to increase our understanding of information processing in artificial and natural neural networks. In the area of recurrent neural networks with Hebbian-type synapses /20/ a prominent role has been played by the techniques of statistical mechanics, especially following the publication of Hopfield's paper /22/. The strategy of the statistical mechanicist, inspired and reinforced by more than a century of experience and successes in analysing bulk properties of dead matter, is to try to extract from the *microscopic* equations, which govern the firing behaviour of individual neurons in a large recurrent network, a closed set of equations for suitably chosen *macroscopic* quantities (i.e. quantities which involve the states of many neurons). In line with the reductionist philosophy of physics, the models studied along these lines were initially chosen to be as simple as possible, but physicists also realized that due to the complexity of brain structures /4/ it is not clear *a priori* what is the potential for describing real neuronal processes with highly simplified models.

In retrospect one can distinguish two distinct waves of statistical mechanical studies. The first (roughly spanning the period 1985-1990) was initiated mainly by Amit *et al.* /1,2/, and consisted of equilibrium studies of so-called attractor neural networks; these studies were, by their nature, restricted to models which evolve towards an equilibrium state (thereby ruling out all networks which do not evolve towards equilibrium, such as those with non-symmetric synapses and all networks composed of graded-response neurons). The second wave (covering, say, the period 1988-1993) consisted mainly of dynamical studies. In non-equilibrium statistical mechanics there are no *a priori* (and biologically unacceptable) restrictions regarding synaptic symmetry and/or neuron types; one has the possibility in principle to incorporate any biological features suggested by experiment, and still analyse the dynamical behaviour of large populations of interconnected neurons. However, the mathematics is generally more involved, and the solutions for models with realistic neuronal equations are highly non-trivial. For overviews of equilibrium and non-equilibrium statistical mechanical studies of recurrent neural networks, see e.g. the book series /12-14/ or the more recent review papers /7,8/. It would seem fair to say that this interdisciplinary enterprise has over the years been

Reprint address:
Prof. A.C.C. Coolen
Department of Mathematics
King's College London
The Strand
London WC2R 2LS, UK
e-mail: tcoolen@mth.kcl.ac.uk

of benefit to both neuroscience and statistical mechanics.

Since then, a number of statistical mechanicists, interested as they are mainly in developing and applying statistical mechanical theory rather than in neuroscience for its own sake (and often lacking time or appetite for studying biological literature), have moved elsewhere.[1] Those who continue to study the interface of statistical mechanics and neuroscience mostly do so because they have a genuine interest in understanding the brain, and are therefore willing to pay the price of incorporating more biological details into their equations (even at the expense of mathematical solvability). Many computational models have been proposed along this line. Some tried to link network architecture to specific aspects of brain functioning, while keeping the flavour of the original statistical mechanics studies on Hopfield-type models /31/. Others focused on the dynamics of single or multiple neurons, with special emphasis on the importance of the precise timing of individual spikes as a tool to transmit and process information in the brain. All such studies rely heavily on numerical simulations, with little or no attempts at a rigorous mathematical analysis. An exhaustive overview of the general trends in theoretical and computational neuroscience can be found in e.g. /11/. For a recent review focused on models of spiking neurons we refer to /18/.

This paper has been written from a different perspective. We try to show that, just below the surface, there are still many interesting solvable problems in the theory of recurrent neural networks which would seem to be of genuine relevance to neuroscience, but which also pose appealing new challenges to statistical mechanicists (and which lend themselves naturally to their approach), even at the level of relatively simple models. To illustrate this point we discuss four specific case studies (at various stages of progress: ranging from 'solved and published', via 'solved and to be published', to 'in the process of being solved'), which are related to the four basic issues of network dynamics, network connectivity, spike timing and synaptic plasticity.

[1] For instance, in the early 1990s many turned towards learning theory in artificial neural networks, in which the relevant processes concern the dynamics of synapses. See e.g. the review /27/.

2. DYNAMICS: INFORMATION RECALL IN GRADED RESPONSE ATTRACTOR NETWORKS

2.1 Background of the problem

From an early stage onwards several authors have tried to adapt the methods which worked well for systems of two-state (McCulloch-Pitts) neurons /1,2/ to systems with more realistic neuronal equations. One proposal /23/ was to study the following so-called graded-response equations, in which $U_i(t)$ denotes the post-synaptic potential of neuron i, $V_i(t)$ its firing rate, $I_i(t)$ an injected current, and the constants $\{C_i, R_i, \gamma_i\}$ characterize the electro-chemical phenomenology of membranes and transmitters:

$$C_i \frac{d}{dt} U_i = \sum_{j=1}^{N} J_{ij} V_j - \frac{U_i}{R_i} + I_i \qquad V_j = g(\gamma_j U_j) \qquad (1)$$

Here $g(z)$ denotes a monotonic and saturating function (such as $g(z) = \tanh(z)$ or $g(z) = \mathrm{Erf}(z)$, modulo constants), which represents the current-to-frequency transduction in neurons. Models like (1) lack a detailed description of sodium and potassium currents, whose interplay generates the action potential /21/. Yet they retain the main features of leaky integrate-and-fire neurons /38/: the capacitive property of membranes, the leakage current and the summation of the synaptic inputs.

From a mathematical point of view, equations (1) can be shown to have a Lyapunov function as soon as the synapses J_{ij} are symmetric (in which case the system will evolve towards a fixed point). However, surprisingly, even for symmetric synapses one can still not formally apply the methods of equilibrium statistical mechanics; that would have required the above equations to represent a *gradient descent* evolution on the Lyapunov function[2],

[2] A model with continuous variables, such as (1), can be solved (modulo a few exceptions) using equilibrium statistical mechanics if the dynamics takes the form of a gradient descent process, complemented by Gaussian white noise. In such cases the model can be shown to obey 'detailed balance', which means that it has a stationary state without probability currents /42/.

which turns out not to be the case. In /25/ the problem was circumvented by arguing that, for symmetric synapses, the system (1) will evolve towards the (local) minima of the Lyapunov function (just like a gradient descent version would), so that one is allowed to solve the model by taking the zero-noise limit of a corresponding gradient-descent based system. This led, for networks with symmetric Hebbian-type synapses (and in the limit $N\to\infty$), to stationary state phase diagrams which were very similar to those derived in /1,2/ for binary neurons.

It appears that after /25/ no attempt was made to solve the dynamics of graded-response attractor networks (the assumption seems to have been that, due to the absence of detailed balance, this would probably be a very messy and pointless exercise [Kühn R, personal communication]). Note that solving the dynamics of such models is not just a matter of rounding off a job started with the statics. If possible, it allows us to eliminate the restriction to symmetric (i.e. non-realistic) synapses. In fact the dynamics of graded-response attractor networks turns out not only to be relatively easy to solve, but (in contrast to the statics) also to be profoundly different from that of networks with binary neurons, even with symmetric synapses. Full details and mathematical derivations of the summary given below can be found in e.g. /8/, including applications to graded-response networks with non-symmetric synapses.

2.2 The solution and its deliverables

Let us consider the simplest version of (1), where $C_i = R_i = I_i = 1$ and $\gamma_i = \gamma$. However, in recognition of the spiking nature of neuronal communication (which has been lost in the firing rate picture) we add zero-average Gaussian noise sources $\eta_i(t)$:

$$\frac{d}{dt}u_i(t) = \sum_{j=1}^{N} J_{ij}\tanh[\gamma u_j(t)] - u_i(t) + \eta_i(t) \qquad (2)$$

The noise covariance is given by $\langle \eta_i(t)\eta_j(t')\rangle = 2T\delta_{ij}\delta(t-t')$. For systems of the type (2) even the back door exploited in /25/ remains closed: there is no detailed balance, and no scope for applying equilibrium techniques (even if we are only after stationary state properties). For our synapses we choose the standard Hebbian-type recipe

$$J_{ij} = \frac{2}{N}\sum_{\mu=1}^{p}\xi_i^\mu\xi_j^\mu \qquad (3)$$

with $\xi_i^\mu \in \{-1,1\}$ (the stored information, here chosen randomly). We will (for simplicity) assume the number p of stored patterns to remain finite, and will eventually take the limit $N\to\infty$. The macroscopic quantities of interest in such models are the p so-called overlaps m_μ, which measure the degree of recall of the individual patterns:

$$m_\mu = \frac{1}{N}\sum_i \xi_i^\mu \tanh(\gamma u_i) \qquad \mathbf{m} = (m_1,\ldots,m_p) \qquad (4)$$

In non-equilibrium statistical mechanics the objective is to find a set of macroscopic quantities which in the limit $N\to\infty$ obey *closed* dynamical laws. For networks of binary neurons and synapses of the form (3) this role is, for finite p, found to be played by the overlaps (4). In the present case this is no longer true: the time derivative of $\mathbf{m}$ can, even for $N\to\infty$, not be expressed in terms of $\mathbf{m}$.

It turns out that the appropriate quantities to turn to in the case of graded-response networks are the distributions $\rho_\xi(u)$ of the post-synaptic potentials (PSP) in the so-called sub-lattices I_ξ (of size $|I_\xi|$)[3]:

$$\rho_{\boldsymbol{\xi}}(u) = \frac{1}{|I_{\boldsymbol{\xi}}|}\sum_{i\in I_{\boldsymbol{\xi}}}\delta[u-u_i] \qquad (5)$$

These 2^p PSP distributions are for $N\to\infty$ found to evolve according to a so-called time-dependent Ornstein-Uhlenbeck process /42/, for which the solution can be found in the form of Gaussian distributions with time-dependent averages and covariances. This, in turn, allows one to find the desired overlaps (4), which themselves can be written in terms of the PSP distributions (5). In the stationary state this leads to a generalization of the result of /25/ (the latter corresponds to $T=0$):

$$\mathbf{m} = \int\frac{dz}{\sqrt{2\pi}}\,e^{-\frac{1}{2}z^2}\langle\boldsymbol{\xi}\,\tanh[\gamma(\boldsymbol{\xi}\cdot\mathbf{m}+z\sqrt{T})]\rangle_{\boldsymbol{\xi}} \qquad (6)$$

[3] A sub-lattice I_ξ consists of all neurons i with the property $(\xi_i^1, \ldots, \xi_i^p) = \xi$; these neurons will evolve in a very similar way. When p binary patterns have been stored, one has 2^p such sub-lattices.

with the abbreviation $\langle f(\xi)\rangle_{\xi} = \sum_{\xi\in\{-1,1\}^p} f(\xi)$. For $T = 0$ we indeed find an equation which is identical to that found for networks of binary neurons /1/ (confirming the impression in /25/ that differences between attractor networks of binary versus graded-response neurons are minor). Further analysis of (6) for $T > 0$ leads to the phase diagrams and recall amplitudes shown in Figure 1. However, we now also have access to the full macroscopic dynamics. For 'natural' initial conditions (uniform PSPs within sub-lattices at $t = 0$), and an initialization which implies the triggering of one specific pattern ν, one can work out the mathematics further and find that $m_\mu(t) = m(t)\delta_{\mu\nu}$, with the recall amplitude $m(t)$ obeying

$$m(t) = \int \frac{dz}{\sqrt{2\pi}}\, e^{-\frac{1}{2}z^2} \tanh\left[e^{-t}\log[\frac{1+m_0}{1-m_0}]^{\frac{1}{2}} + \gamma\left(\int_0^t ds\; e^{s-t} m(s) + z\sqrt{T(1-e^{-2t})}\right)\right] \quad (7)$$

Examples of numerical solutions of this non-trivial equation are shown in Figure 2. Note that equation (7) is fundamentally different from (and much more interesting than) the type of expression one would have found for binary neurons, viz. the ordinary non-linear differential equation $dm(t)/dt = \tanh[\beta m(t)] - m(t)$. It describes a system with 'memory': the evolution of the recall process (as monitored by m) for times $t > t_1$ does not just depend on $m(t_1)$ (as with networks of binary neurons), but also on the values $m(t)$ with $0 \le t < t_1$. This follows immediately from Figure 2, which shows non-monotonic recall: i.e. the overlap might

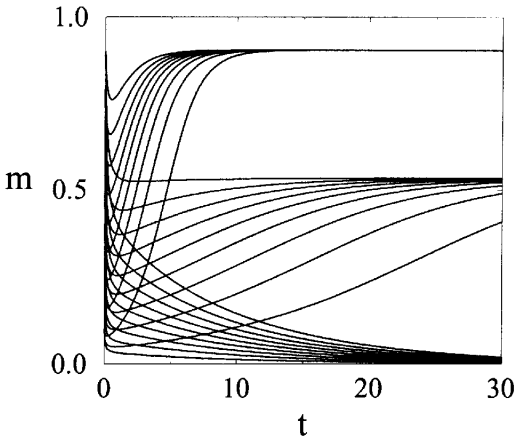

Fig. 2: Overlap evolution in the attractor network with graded-response neurons and $J_{ij} = (2/N)\sum_\mu \xi_i^\mu \xi_j^\mu$, away from saturation. Gain parameter: $\gamma = 4$. Initial conditions: $\rho\eta(u) = \delta[u - k_0\eta\nu]$ (triggering recall of pattern ν, with uniform PSPs within sub-lattices). Lines: overlaps $m = (2/N)\sum_i \xi_i^\nu s_i$ with recalled pattern as functions of time, for $T = 0.25$ (upper set), $T = 0.5$ (middle set) and $T = 0.75$ (lower set), following initial overlaps $m_0 \in \{0.1, 0.2, \ldots, 0.8, 0.9\}$.

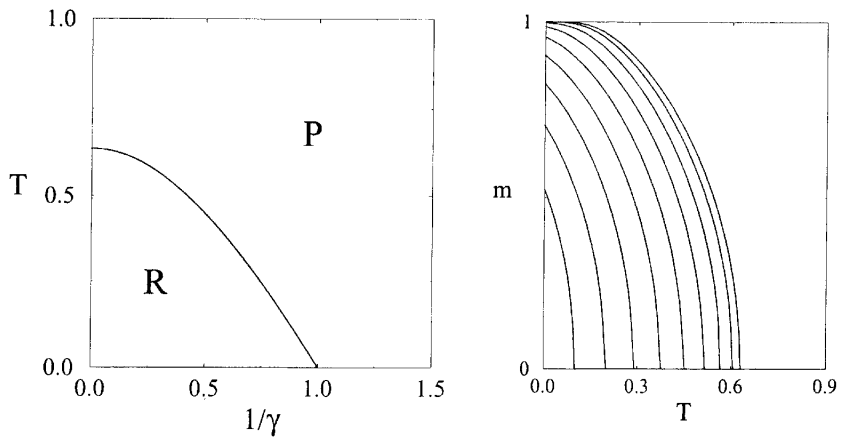

Fig. 1: Left: Phase diagram of the attractor network with graded-response neurons and $J_{ij} = (2/N)\sum_\mu \xi_i^\mu \xi_j^\mu$, away from saturation. P: paramagnetic phase, no recall; R: pattern recall phase. Solid line: separation of the phases (a continuous transition). Right: Asymptotic recall amplitudes $m = (2/N)\sum_i \xi_i^\mu s_i$ of pure states (full recall corresponds to $m = 1$), as functions of the noise level T, for $\gamma^{-1} \in \{0.1, 0.2, \ldots, 0.8, 0.9\}$ (from top to bottom).

initially decrease, yet later reverse the trend and increase towards nearly perfect recall for large times.

We note that the above example corresponds to having made the simplest possible choice $C_i = R_i = I_i = 1$ (mathematically speaking) for the biological parameters in (1). It is not unlikely that one might find an even richer repertoire of behaviour upon making more realistic choices for these parameters.

3. CONNECTIVITY: MODELS WITH NON-UNIFORM INTERACTION RANGES

3.1 Background of the problem

In our second problem we try to move away from the full connectivity which characterized the traditional Hopfield-type models /1,2,22/. Here we are strongly constrained by the competing requirements of biological relevance and mathematical solvability. For instance, solving models in which neurons interact only with nearest neighbours is either boring (in one dimension there cannot be phase transitions) or impossible (in two dimensions we can no longer obtain exact solutions in the presence of external fields). Moreover, from a neuro-anatomical perspective, there is no interest in studying models of auto-associative memory with only short range synapses. In the real brain the distribution of distances spanned by cortico-cortical connections may vary from less than 1 mm to more than 7 mm, and pyramidal cells may have axons which reach the subcortical white substance and subsequently re-enter the cortex elsewhere (see e.g. the discussion in /4/, chapter 26). Another reason for studying the co-operation of short and long range interactions in a network model lies in the modular organization of several cortical areas. For example, studies of the distribution of the inputs to the somatosensory cortex from various types of receptors revealed that vertical columns of cells are activated by distinct receptor subtypes, while all cortical layers within a column respond selectively to the same receptor subtype /29/. Like the somatosensory cortex, the primary visual cortex is also organized into narrow columns, each containing cells tuned to a particular orientation of the visual stimulus /24/. An even more subtle functional subdivision into 'sub-compartments' has recently been found to characterize the primate secondary visual cortex, supporting the notion that the modular organization is area-specific /39/.

Partially motivated by the above experimental evidence, a class of models has been proposed in which the connectivity matrix is a combination of long-range (anti-)Hebbian and nearest neighbour (anti-)Hebbian synapses, $J_{ij} = J_{ij}^{\text{short}} + J_{ij}^{\text{long}}$:

$$J_{ij}^{\text{long}} = \frac{J_\ell}{N}\sum_{\mu=1}^{p}\xi_i^\mu\xi_j^\mu$$

$$J_{ij}^{\text{short}} = \begin{cases} J_s\sum_{\mu=1}^{p}\xi_i^\mu\xi_j^\mu & \text{for } j = i \pm 1 \\ 0 & \text{otherwise} \end{cases} \qquad (8)$$

(with $J_{ii} \to 0$). This choice ensures non-trivial behaviour (the long-range synapses guarantee phase transitions, the short-range synapses induce spatial effects), yet the calculations are feasible (since the short-range synapses act in one spatial dimension), although non-trivial. Below we sketch the main steps and features of the solution of this model; full details can be found in /36,37/.

3.2 The solution and its deliverables

In equilibrium statistical mechanics (which applies here, since the synapses are symmetric), one solves a model by calculating the asymptotic free energy per neuron f, from which the relevant macroscopic quantities can be derived:

$$f = -\lim_{N\to\infty}\frac{1}{\beta N}\log\sum_{\sigma} e^{-\beta H(\sigma)}$$

where $\sigma = (\sigma_1, \ldots, \sigma_N) \in \{-1,1\}^N$ denotes the miscroscopic states of the N binary neurons and $\beta = T^{-1}$ is the inverse noise level in the neuronal dynamics. For the system (8) one has, when p (the number of stored patterns) remains finite:

$$H(\sigma) = -\frac{1}{2}J_\ell N\sum_{\mu=1}^{p}m_\mu^2(\sigma) - J_s\sum_{i=1}^{N}\sigma_i\sigma_{i+1}\sum_{\mu=1}^{p}\xi_i^\mu\xi_{i+1}^\mu$$

with the overlaps

Standard manipulations show that

$$f = \min_{\{m_\mu\}}\left\{\frac{1}{2}J_\ell\sum_{\mu=1}^{p}m_\mu^2 - \lim_{N\to\infty}\frac{1}{\beta N}\log\sum_{\sigma}\left[\prod_{i=1}^{N}T_{\sigma_i\sigma_{i+1}}\right]\right\} \qquad (9)$$

$$T_{\sigma_i\sigma_{i+1}} = e^{\frac{1}{2}\beta J_\ell \sum_\mu m_\mu[\sigma_i\xi_i^\mu+\sigma_{i+1}\xi_{i+1}^\mu]+\beta J_s\sigma_i\sigma_{i+1}\sum_\mu \xi_i^\mu\xi_{i+1}^\mu} \quad (10)$$

The objects (10) are called transfer matrices (see e.g. the textbook /43/). The problem here is that the various $T_{\sigma_i\sigma_{i+1}}$ depend on the *random variables* $\{\xi^\mu, \xi_{i+1}\}$. As a consequence, at this stage the calculation becomes rather technical; one finds that analytical solution requires so-called 1D random-field techniques, based on inspecting the effect of replacing $N \to N + 1$ (i.e. of adding one further neuron).

In terms of behaviour one finds new non-trivial features, absent in models with only long-range synapses. These are generated by the inherent competition between short- and long-range information processing, especially when $J_l > 0$ and $J_s > 0$ (e.g. long-range Hebbian [excitatory] synapses and short-range [inhibitory] anti-Hebbian ones). It should be noted that this particular parameter regime is biologically realistic, since it is believed that inhibitory synapses do indeed act over a shorter range than excitatory ones (see /4/, chapter 15). More specifically:

- There are many new phases and transitions between them, even for small p;
- The non-recalled patterns have a non-negligible impact on the overlap of the recalled one, even for small p;
- The firing rates $\langle\sigma_i\rangle$ can exhibit highly non-trivial and irregular statistics, described by so-called 'Devil's Staircases'.

The model and its solution can also be generalized to include next-nearest neighbour interactions, i.e.

$$J_{ij} = \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu$$

$$\left\{\frac{J_\ell}{N} + J_s^{(1)}[\delta_{i,j+1} + \delta_{i,j-1}] + J_s^{(2)}[\delta_{i,j+2} + \delta_{i,j-2}]\right\} \quad (11)$$

Examples of phase diagram cross-sections for the models (8) and (11) are given in Figures 3 and 4, respectively. Note that the conventional Hopfield model corresponds to $J_s^{(1)} = j_s^{(2)} = 0$, with only a simple continuous transition from non-recall to recall at $\beta J_\ell = 1$. Here much more can happen, in both statics and dynamics (see /36,37/).

4. SPIKE TIMING: COUPLED OSCILLATORS WITH NON-TRIVIAL PHASE PREFERENCES

4.1 Background of the problem

It has become increasingly clear that the timing of spikes communicated between neurons is of relevance in both operation and learning. Unfortunately, spike-based models are much harder to

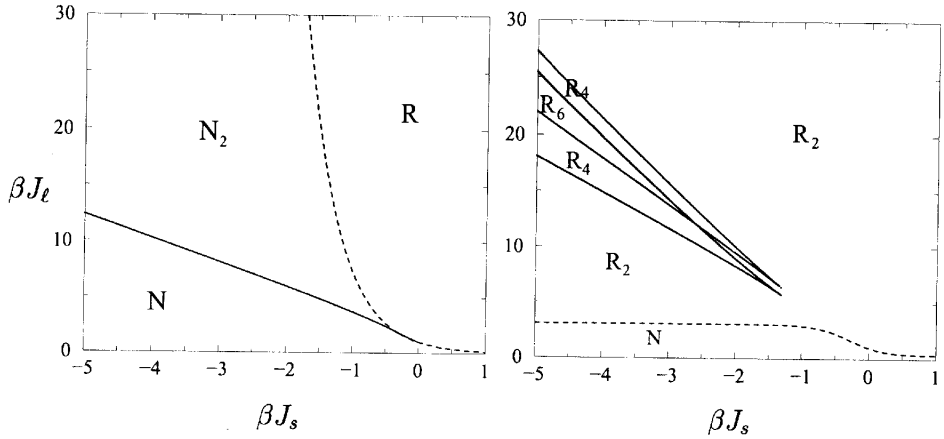

Fig. 3: Phase diagrams of attractor networks, for $p = 1$ (left) and $p = 2$ (right), with Hebbian-type long-range and nearest neighbour synapses in 1-D: $J_{ij} = \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \{J_l / N + J_s[\delta_{i,j+1} + \delta_{i,j-1}]\}$. $\beta = T^{-1}$ (inverse noise level). Phases: R (recall only), N (non-recall only), N$_i$ (i stable recall states, non-recall stable), and R$_i$ (i stable recall states, non-recall unstable). Solid lines indicate discontinuous transitions, dashed lines continuous transitions.

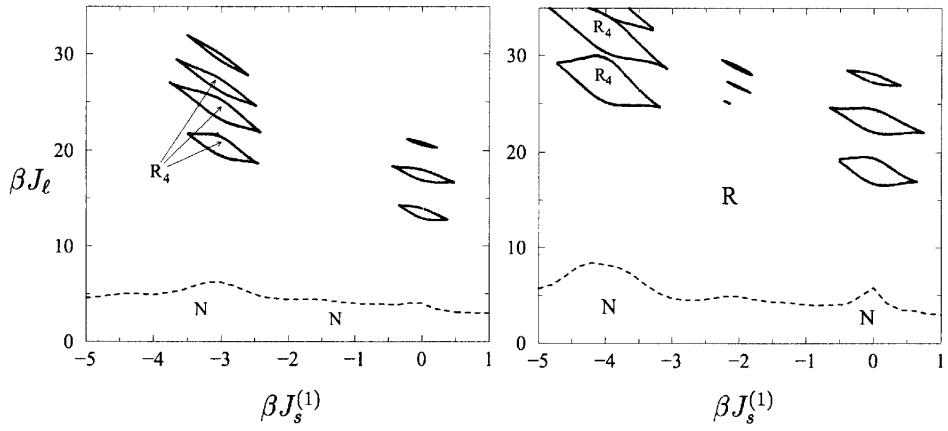

Fig. 4: Phase diagrams of attractor networks, for $p = 5$ and $\beta = T^{-1}$ (inverse noise level), with Hebbian-type long-range, nearest neighbour, and next-nearest neighbour synapses in 1-D:

Left: $J_s^{(2)} = -3$. Right: $J_s^{(2)} = -4$. Phases: R (recall only), ppol N (non-recall only), and R_i (i stable recall states, non-recall unstable). Solid lines indicate discontinuous transitions, dashed lines continuous transitions.

analyze than frequency-based ones. If, for simplicity, we were to build a model in which all firing rates were identical, we would find all timing aspects to be embodied in the mutual phase relations of the oscillating neurons.

Oscillatory activity related to behaviour has been observed in several cortical and subcortical areas in the real brain; examples of coherent oscillatory activity can be found in e.g. the reviews /19,30,34/. Oscillations at the gamma frequency have been proposed as a mechanism to bind different features for the representation of objects in the cortex /35/. Again at the gamma frequency, oscillations in the local field potentials have been observed in the prefrontal motor cortex, which is a primary area in the hierarchy for movement preparation and execution /32/. A number of experiments have been performed in recent years in order to characterize oscillatory activity in the rat hippocampus: large irregular activity, or *sharp waves* /10/, were observed during eating, drinking, grooming and quiet immobility. Rhythmic slow activity at the theta frequency of 5-12 Hz /5/ was observed during postural shifts, walking and running. All this experimental evidence suggests a need for investigating theoretically the possible mechanisms for inducing and synchronizing oscillatory activity in networks.

A convenient phenomenological model with which to study phase relations between coupled oscillators was introduced by Kuramoto /26/, which in its simplest form reads

$$\frac{d}{dt}\phi_i = \sum_{j=1}^{N} J_{ij} \sin[\phi_j - \phi_i] + \eta_i(t) \tag{12}$$

Here $\phi_i \in [-\pi, \pi\}$ (mod 2π) denotes the phase of oscillator i, and $\eta_i(t)$ is a zero-average Gaussian noise source with $\langle \eta_i(t)\eta_j(t')\rangle = 2T\delta_{ij}\delta[t - t']$. One easily convinces oneself that synapses $J_{ij} > 0$ promote configurations where $\phi_i = \phi_j$ (i and j are synchronized), whereas synapses $J_{ij} < 0$ promote configurations where $\phi_i = \phi_j + \pi$ (i and j fire in anti-synchrony).

In the case of symmetric synapses, the deterministic (first) part of the dynamic equations (12) can be written in a gradient descent form, so the process (12) obeys detailed balance, and equilibrium statistical mechanics applies:

$$\frac{d}{dt}\phi_i = -\frac{\partial}{\partial\phi_i} H(\phi) + \eta_i(t) \tag{13}$$

$$H(\phi) = -\sum_{k<\ell}^{N} J_{k\ell} \cos[\phi_k - \phi_\ell]$$

with $\phi = (\phi_1, \ldots, \phi_N)$. As before, solving the model (in the language of statistical mechanics) now

implies calculating the asymptotic free energy per oscillator:

$$f = -\lim_{N\to\infty} \frac{1}{\beta N} \log \int d\phi\, e^{-\beta H(\phi)}$$

In order to appreciate the problem which will be raised shortly, let us quickly review the traditional method with which to calculate f for the simplest model of the type (13): that where $J_{ij} = J/N$ for all $(i<j)$ (with $J>0$). Here we may rewrite

$$H(\phi) = -\frac{J}{N}\sum_{k<\ell}^{N} \cos[\phi_k - \phi_\ell] = -\frac{J}{2N}\left[\sum_{k\ell=1}^{N} \cos[\phi_k - \phi_\ell] - N\right]$$

$$= \frac{J}{2} - \frac{JN}{2}\left[\frac{1}{N}\sum_{k=1}^{N}\cos(\phi_k)\right]^2 - \frac{JN}{2}\left[\frac{1}{N}\sum_{k=1}^{N}\sin(\phi_k)\right]^2$$

Hence we only need to work out the statistics of the simple overall system averages $N^{-1}\sum_i \cos(\phi_i)$ and $N^{-1}\sum_i \sin(\phi_i)$, which leads to the solution (see e.g. /7/)

$$f = \min_{q\geq 0}\left\{\frac{1}{2}Jq^2 - \frac{1}{\beta}\log[2\pi I_0(\beta J q)]\right\} \qquad (14)$$

What, however, if the built-in relative phases are not 0 or π, i.e. if we wish to study coupled oscillators with preferred phase relations intermediate between full synchrony and full anti-synchrony? It is easy to generalize (13) and build this in:

$$H(\phi) = -\sum_{k<\ell} J_{k\ell}\cos[\phi_k - \phi_\ell - \alpha_{k\ell}]$$

Now synapses $J_{ij} > 0$ promote configurations where $\phi_i = \phi_j + \alpha_{ij}$ (and synapses $J_{ij} < 0$ try to avoid this relation). The simplest member of this model class is

$$H(\phi) = -\frac{J}{N}\sum_{k<\ell}\cos[\phi_k - \phi_\ell - \alpha] \qquad (15)$$

$$= -\frac{J}{N}\cos(\alpha)\sum_{k<\ell}\cos[\phi_k - \phi_\ell] - \frac{J}{N}\sin(\alpha)\sum_{k<\ell}\sin[\phi_k - \phi_\ell]$$

Now, unexpectedly, the standard method to rewrite the energy $H(\sigma)$ in terms of simple averages of the type $N^{-1}\sum_k$ no longer works when $\sin(\alpha) \neq 0$; the new terms with $\sin(\phi_k - \phi_\ell)$ resist our attempts to symmetrize $\Sigma_{k<\ell}$ to $\frac{1}{2}\Sigma_{k\neq\ell}$. The reason is that the requirement $\phi_k - \phi_\ell = \alpha$ is symmetric under $k \leftrightarrow \ell$ *only when* $\alpha \in \{0,\pi\}$ ...

4.2 The solution and its deliverables

It thus turns out that even for the apparently simple model (15) we have to find new ways to calculate the free energy f. One such method is described below. It requires first a further generalization of the energy to

$$H(\phi) = -\frac{J}{N}\sum_{k<\ell}^{N}\cos(\phi_k - \phi_\ell - \alpha) - K\sum_{i=1}^{N}\cos(\phi_i - \psi)$$

(i.e. we add external stimuli; these new terms can be removed later). We then inspect the effect of adding one oscillator to the system, $N \to N+1$, on the precursor of the free energy: the so-called partition function Z_N

$$Z_N[K,J] = \int d\phi_N\, e^{-\beta H_N(\phi)}$$

$$f = -\lim_{N\to\infty}\frac{1}{\beta N}\log Z_N[K,J]$$

(where we have added subscripts N to indicate the system size to which we refer). This partition function is found to satisfy the following *closed* recurrent relation:

$$Z_{N+1}[K,J] = \int_{-\pi}^{\pi} d\phi\, e^{\beta K\cos(\phi)}$$

$$Z_N\left[\sqrt{K^2 + \frac{J^2}{(N+1)^2} + \frac{2KJ}{N+1}\cos(\phi+\alpha)}, \frac{NJ}{N+1}\right]$$

with initialization $Z_1[K,J] = 2\pi I_0[\beta K]$ (where $I_n[z]$ denotes the modified Bessel functions). The above relation allows us to find an equation for f. Upon writing

$$f = -\beta^{-1}\log(2\pi) + [\psi + \tfrac{1}{2}K^2\cos(\alpha)]/J$$

this equation takes its simplest form:

$$\frac{\partial\Psi}{\partial J} + \frac{1}{\beta}\log I_0\left[\beta\sqrt{(\partial\Psi/\partial K)^2 + K^2\sin^2(\alpha)}\right] = 0 \qquad (16)$$

with boundary behaviour

$$\Psi = -\frac{1}{2}K^2\cos(\alpha) - (J/\beta)\log I_0[\beta K] + \ldots \qquad (J\to 0)$$

For $\sin(\alpha) = 0$ equation (16) indeed generates the traditional solution (which is not immediately obvious)

$$\Psi_{\alpha\in\{0,\pi\}} = -\frac{1}{2}K^2\cos(\alpha) + J\min_{q\geq 0}\left\{\frac{1}{2}Jq^2 - \frac{1}{\beta}\log I_0[\beta(K+Jq)]\right\}$$

Our task ahead is to solve (16) and extract the physics for arbitrary control parameters. Preliminary numerical work suggests rich static and dynamic behaviour, and phase transitions at $K = 0$ for any α /40/. Next one might inspect non-trivial synapses.

5. SYNAPTIC PLASTICITY: HIERARCHICAL SELF-PROGRAMMING

5.1 Background of the problem

Our final problem concerns models in which neurons and synapses evolve in time simultaneously, according to simple but defensible coupled equations, albeit on widely separated time scales. A computer metaphor would describe such information processing systems as self-programming: the processors (neurons) are running a programme (the synapses) which the processors can themselves rewrite. Such systems are obviously very hard to analyze. Within the area of recurrent neural networks, however, where we are constrained to specific classes of equations, there is hope.

Self-programming recurrent neural networks have been studied mathematically since /15,33/ (the early studies concerned the stability of embedded associative memory 'programmes'). The authors of /15,33/ made the important observation that, provided synapses are much slower than neurons, Hebbian learning can be written as a gradient descent process for the coupled system. The next step /6,9,16,17/ was to add Gaussian white noise to the synaptic equations, converting the self-programming into a process to which equilibrium statistical mechanics could be applied. This resulted in exact equilibrium solutions and phase diagrams, describing transitions between states in which the 'programme' evolves randomly and states in which the programme 'locks' into a stationary configuration, and interesting mathematics. We now try to investigate the possible spontaneous emergence of hierarchically structured programmes in such self-programming networks ('routines', 'sub-routines', 'sub-sub-routines', etc.) by dividing the synapses (which for now are taken to be symmetric, so that we may use equilibrium statistical mechanics) randomly into L different groups I_ℓ (of relative sizes ε_ℓ with $\sum_{\ell=1}^{L}\varepsilon_\ell = 1$), each with their own characteristic time scale τ_ℓ and noise level T_ℓ (representing a hierarchy of increasingly non-volatile programming levels); see Figure 5. We thus get L nested equilibrations of clusters of synapses at time scales $1 << \tau_1 << \tau_2 << \ldots << \tau_L$. The evolution of each synapse J_{ij} is defined as Hebbian-type learning with decay (to limit their values) and zero-average Gaussian noise. A synapse in group I_ℓ thus obeys

$$(i,j) \text{ in } I_\ell: \quad \tau_\ell\frac{d}{dt}J_{ij} = \frac{1}{N}\langle\sigma_i\sigma_j\rangle - \mu_\ell J_{ij} + \eta_{ij}(t)\sqrt{\frac{\tau_\ell}{N}} \tag{17}$$

with noise covariances $\langle\eta_{ij}(t)\,\eta_{k\ell}(t')\rangle = 2T_\ell\delta_{ik}\delta_{j\ell}\delta(t-t')$. The averaging brackets $\langle\ldots\rangle$ in this equation reflect our assumption that synapses are much slower than neurons (so that the neurons can be regarded as in

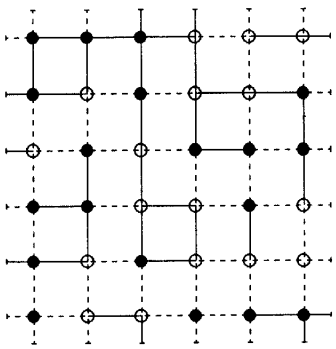

Fig. 5: Illustration of a recurrent self-programming model with $L = 2$. The neurons (active: ∘, inactive: •) evolve on time scales of order 1. The level-1 synapses (solid bonds) evolve on time scales $\tau_1 >> 1$. The level-2 synapses (dashed bonds) evolve on time scales $\tau_2 >> \tau_1$. The bonds are randomly allocated to levels. Our present model differs from this simple picture in two ways: it is fully connected, and we allow for an arbitrary number L of synapse types.

equilibrium, before the synapses change). Our challenge is to solve the above hierarchy of nested stochastic equations.

Expressions like (17) clearly represent a crude simplification (made out of mathematical necessity) of biological reality: while in real neural systems the strength of a synapse indeed depends on the activity of the pre- and post-synaptic neurons, recent experiments have revealed a crucial dependence of synaptic plasticity on the exact timing of pre- and post-synaptic spikes /41/. Rate-based Hebbian learning has been gradually replaced by spike-based learning rules (see /18/, chapter 10).

5.2 The solution and its deliverables

The equilibrium solution of this model (describing the stationary state at the time-scale τ_L of the slowest synapses, full details of which can be found in /41/) is a generalization of that proposed in /6,9,16,17/. It involves L nested partition functions Z_ℓ (one for every time scale), related according to

$$\mathcal{Z}_0[\mathbf{J}^1,\ldots,\mathbf{J}^L] = \sum_{\boldsymbol{\sigma}} e^{-\beta H(\boldsymbol{\sigma},\mathbf{J})}$$

$$\mathcal{Z}_\ell[\mathbf{J}^{\ell+1},\ldots,\mathbf{J}^L] = \int d\mathbf{J}^\ell \; \{\mathcal{Z}_{\ell-1}[\mathbf{J}^\ell,\ldots,\mathbf{J}^L]\}^{m_\ell}$$

$$m_\ell = T_{\ell-1}/T_\ell \qquad m_1 = T/T_1$$

where the vector $\mathbf{J}_\ell$ denotes the collection of dynamic synapses in group ℓ, and with the energy function

$$H(\boldsymbol{\sigma},\mathbf{J}) = -\sum_{i<j} J_{ij}\sigma_i\sigma_j + \frac{1}{2}N\sum_{i<j}\mu_{ij}J_{ij}^2$$

The result of this nesting is a so-called replica theory (see e.g. /28/). If one, furthermore, assumes ergodicity at every time scale τ_ℓ, one can calculate the free energy per neuron f of the system explicitly, via $f = -(\beta N)^{-1}\log Z_L$, resulting in $f = \mathrm{extr}_{\{q1,\ldots,qL\}} f[q_1,\ldots,q_L]$

$$f[q_1,\ldots,q_\ell] = \tfrac{1}{2}\pi_1 q_1 - \frac{1}{\beta}\log 2 + \frac{1}{4}\sum_{\ell=2}^{L} q_\ell^2\pi_\ell\left[\prod_{k=1}^{\ell-1} m_k\right](m_\ell - 1)$$

$$+\tfrac{1}{4}q_1^2\pi_1(m_1-1) - \frac{1}{\beta\prod_{\ell=1}^{L} m_\ell}\log\int Dz_L\left\{\int Dz_{L-1}\left\{\ldots\left\{\int Dz_1\right.\right.\right.$$

$$\left.\left.\left.\left\{\cosh\left[z_L\sqrt{\beta q_L\pi_L} + \sum_{\ell=1}^{L-1} z_\ell\sqrt{\beta(q_\ell\pi_\ell - q_{\ell+1}\pi_{\ell+1})}\right]\right\}^{m_1}\right\}^{m_2}\ldots\right\}^{m_{L-1}}\right\}^{m_L} \tag{18}$$

with $\pi_\ell = \sum_{\ell'=\ell}^{L}(\beta\varepsilon_{\ell'}/\beta_{\ell'}\mu_{\ell'})$, and where the required extremum of (18) is defined as:

$m_\ell > 1$: minimize $f[\ldots]$ with respect to q_ℓ

$m_\ell < 1$: maximize $f[\ldots]$ with respect to q_ℓ

This result is not only relevant in that it allows us to find the possible phases and their transitions, but it is also quite interesting from a statistical mechanical point of view (see /41/ for details).

Of the various results which follow from (18) we will concentrate here, for brevity, mainly on the emerging phases for $L = 2$ and $L = 3$. For the two-level hierarchy $L = 2$ (i.e. two classes of synapses) one finds the following three phases:

P: $q_1 = q_2 = 0$
random motion of neurons and all synapses

SG_1: $q_1 > 0, q_2 = 0$
'locking' of neurons, random motion of all synapses

SG_2: $q_1 > 0, q_2 > 0$
'locking' of neurons and level-1 synapses, random motion of level-2 synapses

For the three-level hierarchy $L = 3$ (i.e. three classes of synapses) one finds the following four phases:

P: $q_1 = q_2 = q_3 = 0$
random motion of neurons and all synapses

SG_1: $q_1 > 0, q_2 = q_3 = 0$
'locking' of neurons, random motion of all synapses

SG_2: $q_1 > 0, q_2 > 0, q_3 = 0$
'locking' of neurons and level-1 synapses, random motion of level-2 and level-3 synapses

SG_2: $q_1 > 0, q_2 > 0, q_3 > 0$
'locking' of neurons and level-1 and level-2 synapses,
random motion of level-3 synapses

Generally, an L-level hierarchy exhibits the phases P, SG_1, SG_2, SG_3, ..., SG_L (with the obvious interpretations). There is a very interesting phenomenology of transitions between these phases, which

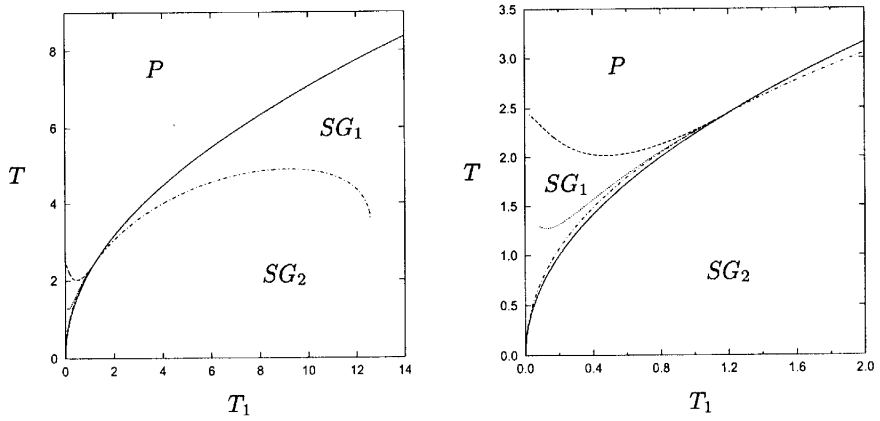

Fig. 6: $L = 2$ phase diagram for $\varepsilon_l/\mu_1 = 1$, $\varepsilon_2/\mu_2 = 2$ and $m_2 = 0.5$ ($m_\ell = T_{\ell-1}/T_\ell$). Right panel: close-up. Continuous transition lines: $T_{\mathrm{p,l}}^{\mathrm{2nd}}$ (solid), $T_{\mathrm{l,2}}^{\mathrm{2nd}}$ (dot-dash); discontinuous transition lines: $T_{\mathrm{p,l}}^{\mathrm{1st}}$ (dashed) and $T_{p,2}^{1st}$ (dotted).

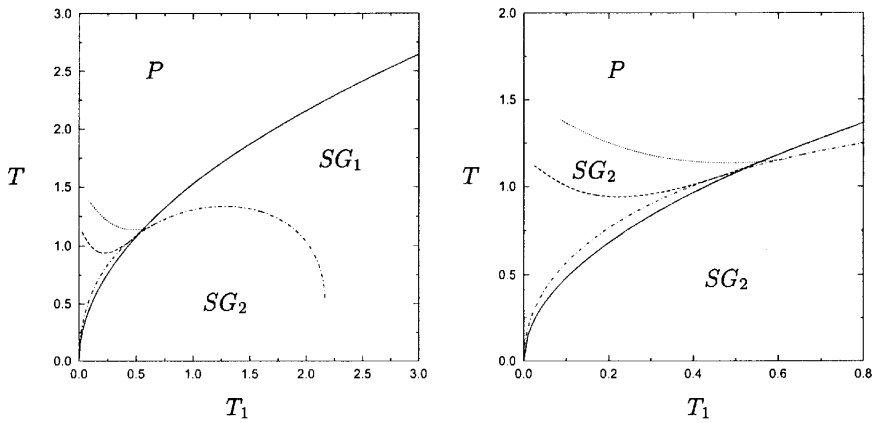

Fig. 7: $L = 2$ phase diagram for $\varepsilon_l/\mu_1 = 1$, $\varepsilon_2/\mu_2 = 2$ and $m_2 = 1.5$ ($m_\ell = T_{\ell-1}/T_\ell$). Right panel: close-up. Continuous transition lines: $T_{\mathrm{p,l}}^{\mathrm{2nd}}$ (solid), $T_{\mathrm{l,2}}^{\mathrm{2nd}}$ (dot-dash); discontinuous transition lines: $T_{\mathrm{p,l}}^{\mathrm{1st}}$ (dashed) and $T_{p,2}^{1st}$ (dotted).

depends in a non-trivial manner on the various noise levels T_ℓ in the system. This can already be appreciated from the phase diagram cross-sections given in Figures 6 (for $m_2 = 0.5$) and 7 (for $m_2 = 1.5$) (more phase diagrams can be found in /41/).

The general picture is that the phase transitions become more discontinuous and rigorous when the values of the noise level ratios $m_\ell = T_{\ell-1}/T_\ell$ increase. For instance, comparison of Figures 6 and 7 illustrates how as the neuronal noise level T is reduced, in contrast to $m_2 < 1$, for $m_2 > 1$ the system goes from the P state (in which processors and synapses all evolve randomly) directly into the SG_2 state (where both neurons and level-1 synapses lock), without any intermediate SG_1 state. Such properties can in fact be proven for arbitrary L /41/.

6. DISCUSSION

In this paper we have presented examples of relatively simple yet hopefully relevant theoretical questions regarding information processing in recurrent neural networks. They concerned the dynamics of graded-response attractor networks,

networks with non-uniform connectivity, coupled oscillators with non-trivial phase relations, and hierarchical self-programming in recurrent networks. All four problems appear to be particularly suited to being studied using the philosophy and mathematical techniques from (equilibrium and non-equilibrium) statistical mechanics; they do not come with a compelling need to involve a great level of biological detail, and all are found to be of interest also when regarded purely as problems in statistical mechanics. Our hidden agenda was to show that fruitful and enjoyable interdisciplinary research at the neuroscience/theoretical physics boundary still does not necessarily require researchers to first become both neuroscientists and theoretical physicists at the same time.

We hope to have demonstrated to neuroscientists that statistical mechanicists have not been standing still since the 1980s, and that their methods and 'tricks' can offer significantly more than just the mathematical solution of conventional Hopfield models (with their binary neurons, symmetric synapses, and full connectivity). The first three case studies in particular would appear to relate directly to many phenomena observed in several areas of the brain, although more work should be done before it is scientifically justified to draw explicit conclusions and predictions on real brain processes from such models. Yet the non-trivial phenomenology observed, combined with feasible and sound mathematical analysis, will hopefully stimulate the interest of experimentalists in this type of research.

Similarly, we hope to have shown to the statistical mechanics community that there continue to be many interesting statistical mechanical problems in the area of neural information processing, which might even stimulate the further development of their own field.

ACKNOWLEDGEMENTS

It is ACCC's pleasure to thank his collaborators Nikos Skantzos and Tatsuya Uezu. VDP would like to thank the EU for financial support (contract QLGA-CT-2001-51056).

REFERENCES

1. Amit DJ, Gutfreund H, Sompolinsky H. Spin-glass models of neural networks. Phys Rev A 1985; 32: 1007-1018.
2. Amit DJ, Gutfreund H, Sompolinsky H. Storing infinite numbers of patterns in a spin-glass model of neural networks. Phys Rev Lett 1985; 55: 1530-1533.
3. Bi G, Poo M. Synaptic modification by correlated activity: Hebb's postulate revisited. Annu Rev Neurosci 2001; 24: 139-166.
4. Braitenberg V, Schutz A. Cortex: Statistics and Geometry of Neuronal Connectivity. New York: Springer, 1998.
5. Buzsaki G. Theta oscillations in the hippocampus. Neuron 2002; 33: 325-340.
6. Caticha N. From quenched to annealed: a study of the intermediate dynamics of disorder. J Phys A 1994; 27: 5501-5507.
7. Coolen ACC. Statistical Mechanics of Recurrent Neural Networks I - Statics. In: Moss F, Gielen S, eds. Handbook of Biological Physics IV. Amsterdam: Elsevier Science, 2001; 531-596.
8. Coolen ACC. Statistical Mechanics of Recurrent Neural Networks II - Dynamics. In: Moss F, Gielen S, eds. Handbook of Biological Physics IV. Amsterdam: Elsevier Science, 2001; 597-662.
9. Coolen ACC, Penney RW, Sherrington D. Dynamics of fast spins and slow interactions in neural networks and spin systems. Phys Rev B 1993; 48: 16116-16118.
10. Csicsvari J, Hirase H, Mamiya A, Buzsaki G. Ensemble patterns of hippocampal CA3-CA1 neurons during sharp wave associated population events. Neuron 2000; 28: 585.
11. Dayan P, Abbott LF. Theoretical Neuroscience. Cambridge, MA: MIT Press, 2001.
12. Domany E, van Hemmen JL, Schulten K, eds. Models of Neural Networks I. Berlin: Springer, 1991.
13. Domany E, van Hemmen JL, Schulten K, eds. Models of Neural Networks II. Berlin: Springer, 1994.
14. Domany E, van Hemmen JL, Schulten K, eds. Models of Neural Networks III. Berlin: Springer, 1995.
15. Dong DW, Hopfield JJ. Dynamic properties of neural networks with adapting synapses. Network 1992; 3: 267-283.
16. Dotsenko V, Franz S, Mézard M. Partial annealing and overfrustration in disordered systems. J Phys A 1994; 27: 2351-2365.
17. Feldman DE, Dotsenko VS. Partially annealed neural networks. J Phys A 1994; 27: 4401-4411.
18. Gerstner W, Kistler WM. Spiking Neurons Models. Cambridge: Cambridge University Press, 2002.
19. Gray CM. The temporal correlation hypothesis of visual feature integration: still alive and well. Neuron 1999; 24: 31-41.

20. Hebb DO. The Organization of Behaviour. New York: Wiley, 1949.
21. Hodgkin AL, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol 1952; 117: 500-544.
22. Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci USA 1982; 79: 2554-2558.
23. Hopfield JJ. Neurons with graded response properties have collective computational properties like those of two-state neurons. Proc Natl Acad Sci USA 1984; 81: 3088-3092.
24. Hubel DH, Wiesel TN, Stryker MP. Anatomical demonstration of orientation columns in macaque monkey. J Comp Neurol 1978; 177: 361-379.
25. Kühn R, Bös S, van Hemmen JL. Statistical mechanics for networks of graded-response neurons. Phys Rev A 1991; 43: 2084-2087.
26. Kuramoto Y. Chemical Oscillations, Waves and Turbulence. Berlin: Springer, 1984.
27. Mace CWH, Coolen ACC. Statistical mechanical analysis of the dynamics of learning in a perceptron. Statistics and Computing 1998; 8: 55-88.
28. Mézard M, Parisi G, Virasoro MA. Spin-Glass Theory and Beyond. Singapore: World Scientific, 1987.
29. Mountcastle VB. Central nervous system in mechanoreceptive sensibility. In: Darian-Smith I, ed. Handbook of Physiology, Section 1: The Nervous System, Vol. III: Sensory Processes. Bethesda, MD: American Physiological Society, 1984; 789-878.
30. Ritz R, Sejnowski TJ. Synchronous oscillatory activity in sensory systems: new vistas on mechanisms. Curr Opin Neurobiol 1997; 7: 536-546.
31. Rolls ET, Treves A. Neural Networks and Brain Function. Oxford: Oxford University Press, 1997.
32. Sanes JN, Donoghue JP. Oscillations in local field potentials of the primate motor cortex during voluntary movement. Proc Natl Acad Sci USA 1993; 90: 4470-4474.
33. Shinomoto S. Memory maintenance in neural networks. J Phys A 1987; 20: L1305-L1309.
34. Singer W. Neuronal synchrony: a versatile code for the definition of relations? Neuron 1999; 24: 49.
35. Singer W, Gray CM. Visual feature integration and the temporal correlation hypothesis. Annu Rev Neurosci 1995; 19: 555-586.
36. Skantzos NS, Coolen ACC. $1+\infty$-dimensional attractor neural networks. J Phys A 2000; 33: 5785-5807.
37. Skantzos NS, Coolen ACC. Attractor modulation and proliferation in $1+\infty$-dimensional neural networks. J Phys A 2001; 34: 929-942.
38. Stein RB. Some models of neuronal variability. Biophys J 1967; 7: 37-68.
39. Ts'o DY, Roe AW, Gilbert CD. A hierarchy of the functional organization for color, form and disparity in primate visual area V2. Vision Res 2001; 41: 1333-1349.
40. Tucker A. MSc Project Report, King's College London, 2001.
41. Uezu T, Coolen ACC. Hierarchical self-programming in recurrent neural networks. J Phys A 2002; 35: 2761-2809.
42. Van Kampen NG. Stochastic Processes in Physics and Chemistry. Amsterdam: North-Holland, 1992.
43. Yeomans JM. Statistical Mechanics of Phase Transitions. Oxford: Clarendon Press, 1992.