CHAPTER 14

Statistical Mechanics of Recurrent Neural Networks I – Statics

A.C.C. COOLEN

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

Handbook of Biological Physics
Volume 4, edited by F. Moss and S. Gielen

Contents

1. Introduction

Statistical mechanics deals with large systems of stochastically interacting microscopic elements (particles, atomic magnets, polymers, etc.). The strategy of statistical mechanics is to abandon any ambition to solve models of such systems at the microscopic level of individual elements, but to use the microscopic laws to calculate equations describing the behavior of a suitably chosen set of *macroscopic* observables. The toolbox of statistical mechanics consists of methods to perform this reduction from the microscopic to a macroscopic level, which are all based on efficient ways to do the bookkeeping of probabilities. The experience and intuition that have been built up over the last century tells us what to expect, and serves as a guide in finding the macroscopic observables and in seeing the difference between relevant mathematical subtleties and irrelevant ones. As in any statistical theory, clean and transparent mathematical laws can be expected to emerge only for large (preferably infinitely large) systems. In this limit one often encounters phase transitions, i.e. drastic changes in the system's macroscopic behavior at specific values of global control parameters.

Recurrent neural networks, i.e. neural networks with synaptic feedback loops, appear to meet the criteria for statistical mechanics to apply, provided we indeed restrict ourselves to large systems. Here the microscopic stochastic dynamical variables are the firing states of the neurons or their membrane potentials, and one is mostly interested in quantities such as average state correlations and global information processing quality, which are indeed measured by macroscopic observables. In contrast to layered networks, one cannot simply write down the values of successive neuron states for models of recurrent neural networks; here they must be solved from (mostly stochastic) coupled dynamic equations. Under special conditions ('detailed balance'), which usually translate into the requirement of synaptic symmetry, the stochastic process of evolving neuron states leads towards an equilibrium situation where the microscopic state probabilities are known, and where the techniques of *equilibrium statistical mechanics* can be applied in one form or another. The equilibrium distribution found, however, will not always be of the conventional Boltzmann form. For nonsymmetric networks, where the asymptotic (stationary) statistics are not known, dynamical techniques from *nonequilibrium statistical mechanics* are the only tools available for analysis. The 'natural' set of macroscopic quantities (or 'order parameters') to be calculated can be defined in practice as the smallest set which will obey closed deterministic equations in the limit of an infinitely large network.

Being high-dimensional nonlinear systems with extensive feedback, the dynamics of recurrent neural networks are generally dominated by a wealth of attractors (fixed-point attractors, limit-cycles, or even more exotic types), and the

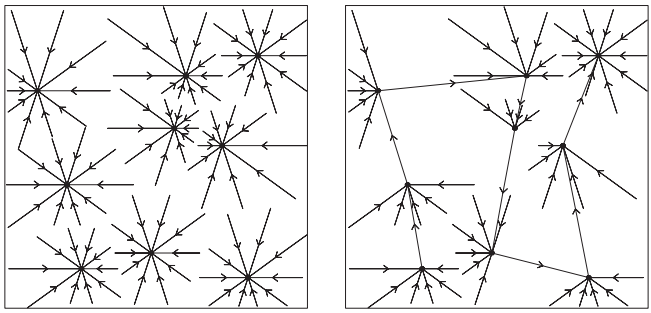

Fig. 1. Information processing by recurrent neural networks through the creation and manipulation of attractors in state space. Patterns stored: the microscopic states •. If the synapses are symmetric we will generally find that the attractors will have to be fixed-points (left picture). With non-symmetric synapses, the attractors can also be sequences of microscopic states (right picture).

practical use of recurrent neural networks (in both biology and engineering) lies in the potential for creation and manipulation of these attractors through adaptation of the network parameters (synapses and thresholds). Input fed into a recurrent neural network usually serves to induce a specific initial configuration (or firing pattern) of the neurons, which serves as a cue, and the 'output' is given by the (static or dynamic) attractor which has been triggered by this cue. The most familiar types of recurrent neural network models, where the idea of creating and manipulating attractors has been worked out and applied explicitly, are the so-called attractor neural networks for associative memory, designed to store and retrieve information in the form of neuronal firing patterns and/or sequences of neuronal firing patterns. Each pattern to be stored is represented as a microscopic state vector. One then constructs synapses and thresholds such that the dominant attractors of the network are precisely the pattern vectors (in the case of static recall), or where, alternatively, they are trajectories in which the patterns are successively generated microscopic system states. From an initial configuration (the 'cue', or input pattern to be recognized) the system is allowed to evolve in time autonomously, and the final state (or trajectory) reached can be interpreted as the pattern (or pattern sequence) recognized by network from the input (see Fig. 1). For such programs to work one clearly needs recurrent neural networks with extensive 'ergodicity breaking': the state vector will during the course of the dynamics (at least on finite time-scales) have to be confined to a restricted region of state space (an 'ergodic component'), the location of which is to depend strongly on the initial conditions. Hence our interest will mainly be in systems with many attractors. This, in turn, has implications at a theoretical/mathematical level: solving models of recurrent neural networks with extensively many attractors requires advanced tools from disordered systems theory, such as replica theory (statics) and generating functional analysis (dynamics). It will turn out that a crucial issue is whether or not the synapses are symmetric. Firstly, synaptic asymmetry is found to

rule out microscopic equilibrium, which has implications for the mathematical techniques which are available: studying models of recurrent networks with non-symmetric synapses requires solving the dynamics, even if one is only interested in the stationary state. Secondly, the degree of synaptic asymmetry turns out to be a deciding factor in determining to what extent the dynamics will be glassy, i.e. extremely slow and nontrivial, close to saturation (where one has an extensive number of attractors).

In this paper (on statics) and its sequel (on dynamics) I will discuss only the statistical mechanical analysis of neuronal firing processes in recurrent networks with static synapses, i.e. network operation as opposed to network learning. I will also restrict myself to networks with either full or randomly diluted connectivity, the area in which the main progress has been made during the last few decades. Apart from these restrictions, the text aims to be reasonably comprehensive and self-contained. Even within the confined area of the operation of recurrent neural networks a truly impressive amount has been achieved, and many of the constraints on mathematical models which were once thought to be essential for retaining solvability but which were regrettable from a biological point of view (such as synaptic symmetry, binary neuron states, instantaneous neuronal communication, a small number of attractors, etc.) have by now been removed with success. At the beginning of the new millennium we know much more about the dynamics and statics of recurrent neural networks than ever before. I aim to cover in a more or less unified manner the most important models and techniques which have been launched over the years, ranging from simple symmetric and non-symmetric networks with only a finite number of attractors, to the more complicated ones with an extensive number, and I will explain in detail the techniques which have been designed and used to solve them.

In the present paper I will first discuss and solve various members of the simplest class of models: those where all synapses are the same. Then I turn to the Hopfield model, which is the archetypical model to describe the functioning of symmetric neural networks as associative memories (away from saturation, where the number of attractors is finite), and to a coupled oscillator model storing phase patterns (again away from saturation). Next I will discuss a model with Gaussian synapses, where the number of attractors diverges, in order to introduce the so-called replica method, followed by a section on the solution of the Hopfield model near saturation. I close this paper with a guide to further references and an assessment of the past and future deliverables of the equilibrium statistical mechanical analysis of recurrent neural networks.

2. Definitions and properties of microscopic laws

In this section I define the most common microscopic models for recurrent neural networks, I show how one can derive the corresponding descriptions of the stochastic evolution in terms of evolving state probabilities, and I discuss some fundamental statistical mechanical properties.

2.1. Stochastic dynamics of neuronal firing states

2.1.1. Microscopic definitions for binary neurons

The simplest nontrivial definition of a recurrent neural network is that where N binary neurons $\sigma_i \in \{-1, 1\}$ (in which the states '1' and '-1' represent firing and rest, respectively) respond iteratively and synchronously to post-synaptic potentials (or local fields) $h_i(\boldsymbol{\sigma})$, with $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)$. The fields are assumed to depend linearly on the instantaneous neuron states:

Parallel:

$$\sigma_i(\ell + 1) = \mathrm{sgn}[h_i(\boldsymbol{\sigma}(\ell)) + T\eta_i(\ell)], \quad h_i(\boldsymbol{\sigma}) = \sum_j J_{ij}\sigma_j + \theta_i. \tag{1}$$

The stochasticity is in the independent random numbers $\eta_i(\ell) \in \Re$ (representing threshold noise), which are all drawn according to some distribution $w(\eta)$. The parameter T is introduced to control the amount of noise. For $T = 0$ the process (1) is deterministic: $\sigma_i(\ell + 1) = \mathrm{sgn}[h_i(\boldsymbol{\sigma}(\ell))]$. The opposite extreme is choosing $T = \infty$, here the system evolution is fully random. The external fields θ_i represent neural thresholds and/or external stimuli, J_{ij} represents the synaptic efficacy at the junction $j \to i$ ($J_{ij} > 0$ implies excitation, $J_{ij} < 0$ inhibition). Alternatively we could decide that at each iteration step ℓ only a single randomly drawn neuron σ_{i_ℓ} is to undergo an update of the type (1):

Sequential:

$$\begin{aligned} &i \neq i_\ell : \sigma_i(\ell + 1) = \sigma_i(\ell) \\ &i = i_\ell : \sigma_i(\ell + 1) = \mathrm{sgn}[h_i(\boldsymbol{\sigma}(\ell)) + T\eta_i(\ell)] \end{aligned} \tag{2}$$

with the local fields as in (1). The stochasticity is now both in the independent random numbers $\eta_i(\ell)$ (the threshold noise) and in the site i_ℓ to be updated, drawn randomly from the set $\{1, \ldots, N\}$. For simplicity we assume $w(-\eta) = w(\eta)$, and define

$$g[z] = 2\int_0^z \mathrm{d}\eta\, w(\eta) : \; g[-z] = -g[z], \quad \lim_{z\to\pm\infty} g[z] = \pm 1, \quad \frac{\mathrm{d}}{\mathrm{d}z} g[z] \geqslant 0$$

Popular choices for the threshold noise distributions are

$$\begin{aligned} &w(\eta) = (2\pi)^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}\eta^2} : \; g[z] = \mathrm{Erf}[z/\sqrt{2}], \\ &w(\eta) = \tfrac{1}{2}[1 - \tanh^2(\eta)] : \; g[z] = \tanh(z). \end{aligned}$$

2.1.2. From stochastic equations to evolving probabilities

From the microscopic Eqs. (1) and (2), which are suitable for numerical simulations, we can derive an equivalent but mathematically more convenient description in terms of microscopic state probabilities $p_\ell(\boldsymbol{\sigma})$. Eqs. (1) and (2) state that, if the system state $\boldsymbol{\sigma}(\ell)$ is given, a neuron i to be updated will obey

$$\mathrm{Prob}[\sigma_i(\ell + 1)] = \tfrac{1}{2}[1 + \sigma_i(\ell + 1) g[\beta h_i(\boldsymbol{\sigma}(\ell))]] \tag{3}$$

with $\beta = T^{-1}$. In the case (1) this rule applies to all neurons, and thus we simply get $p_{\ell+1}(\boldsymbol{\sigma}) = \prod_{i=1}^{N} \frac{1}{2}[1 + \sigma_i\, g[\beta h_i(\boldsymbol{\sigma}(\ell))]]$. If, on the other hand, instead of $\boldsymbol{\sigma}(\ell)$ only the probability distribution $p_\ell(\boldsymbol{\sigma})$ is given, this expression for $p_{\ell+1}(\boldsymbol{\sigma})$ is to be averaged over the possible states at time ℓ:

Parallel:

$$p_{\ell+1}(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\sigma}'} W[\boldsymbol{\sigma};\boldsymbol{\sigma}']p_\ell(\boldsymbol{\sigma}'), \quad W[\boldsymbol{\sigma};\boldsymbol{\sigma}'] = \prod_{i=1}^{N} \frac{1}{2}[1 + \sigma_i g[\beta h_i(\boldsymbol{\sigma}')]]. \tag{4}$$

This is the standard representation of a Markov chain. Also the sequential process (2) can be formulated in terms of probabilities, but here expression (3) applies only to the randomly drawn candidate i_ℓ. After averaging over all possible realizations of the sites i_ℓ we obtain:

$$p_{\ell+1}(\boldsymbol{\sigma}) = \frac{1}{N}\sum_i \left\{ \left[\prod_{j\neq i} \delta_{\sigma_j,\sigma_j(\ell)} \right] \frac{1}{2}[1 + \sigma_i\, g[\beta h_i(\boldsymbol{\sigma}(\ell))]] \right\}$$

(with the Kronecker symbol: $\delta_{ij} = 1$ if $i = j$, $\delta_{ij} = 0$ otherwise). If, instead of $\boldsymbol{\sigma}(\ell)$, the probabilities $p_\ell(\boldsymbol{\sigma})$ are given, this expression is to be averaged over the possible states at time ℓ, with the result:

$$p_{\ell+1}(\boldsymbol{\sigma}) = \frac{1}{N}\sum_i \frac{1}{2}[1 + \sigma_i\, g[\beta h_i(\boldsymbol{\sigma})]]p_\ell(\boldsymbol{\sigma}) + \frac{1}{N}\sum_i \frac{1}{2}[1 + \sigma_i\, g[\beta h_i(F_i\boldsymbol{\sigma})]]p_\ell(F_i\boldsymbol{\sigma})$$

with the state-flip operators $F_i\Phi(\boldsymbol{\sigma}) = \Phi(\sigma_1, \ldots, \sigma_{i-1}, -\sigma_i, \sigma_{i+1}, \ldots, \sigma_N)$. This equation can again be written in the standard form $p_{\ell+1}(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\sigma}'} W[\boldsymbol{\sigma};\boldsymbol{\sigma}']p_\ell(\boldsymbol{\sigma}')$, but now with the transition matrix

Sequential:

$$W[\boldsymbol{\sigma};\boldsymbol{\sigma}'] = \delta_{\boldsymbol{\sigma},\boldsymbol{\sigma}'} + \frac{1}{N}\sum_i \left\{ w_i(F_i\boldsymbol{\sigma})\delta_{\boldsymbol{\sigma},F_i\boldsymbol{\sigma}'} - w_i(\boldsymbol{\sigma})\delta_{\boldsymbol{\sigma},\boldsymbol{\sigma}'} \right\}, \tag{5}$$

where $\delta_{\boldsymbol{\sigma},\boldsymbol{\sigma}'} = \prod_i \delta_{\sigma_i,\sigma_i'}$ and

$$w_i(\boldsymbol{\sigma}) = \tfrac{1}{2}[1 - \sigma_i \tanh[\beta h_i(\boldsymbol{\sigma})]]. \tag{6}$$

Note that, as soon as $T > 0$, the two transition matrices $W[\boldsymbol{\sigma};\boldsymbol{\sigma}']$ in (4) and (5) both describe *ergodic* systems: from any initial state $\boldsymbol{\sigma}'$ one can reach any final state $\boldsymbol{\sigma}$ with nonzero probability in a finite number of steps (being one in the parallel case, and N in the sequential case). It now follows from the standard theory of stochastic processes (see e.g. [1,2]) that in both cases the system evolves towards a unique stationary distribution $p_\infty(\boldsymbol{\sigma})$, where all probabilities $p_\infty(\boldsymbol{\sigma})$ are nonzero.

2.1.3. From discrete to continuous times

The above processes have the (mathematically and biologically) less appealing property that time is measured in discrete units. For the sequential case we will now

assume that the *duration* of each of the iteration steps is a continuous random number (for parallel dynamics this would make little sense, since all updates would still be made in full synchrony). The statistics of the durations is described by a function $\pi_\ell(t)$, defined as the probability that at time t precisely ℓ updates have been made. Upon denoting the previous discrete-time probabilities as $\hat{p}_\ell(\boldsymbol{\sigma})$, our new process (which now includes the randomness in step duration) will be described by

$$p_t(\boldsymbol{\sigma}) = \sum_{\ell \geqslant 0} \pi_\ell(t) \hat{p}_\ell(\boldsymbol{\sigma}) = \sum_{\ell \geqslant 0} \pi_\ell(t) \sum_{\boldsymbol{\sigma}'} W^\ell[\boldsymbol{\sigma}; \boldsymbol{\sigma}'] p_0(\boldsymbol{\sigma}')$$

and time has become a continuous variable. For $\pi_\ell(t)$ we make the Poisson choice $\pi_\ell(t) = \frac{1}{\ell!}(\frac{t}{\Delta})^\ell \mathrm{e}^{-t/\Delta}$. From $\langle \ell \rangle_\pi = t/\Delta$ and $\langle \ell^2 \rangle_\pi = t/\Delta + t^2/\Delta^2$ it follows that Δ is the average duration of an iteration step, and that the relative deviation in ℓ at a given t vanishes for $\Delta \to 0$ as $\sqrt{\langle \ell^2 \rangle_\pi - \langle \ell \rangle_\pi^2}/\langle \ell \rangle_\pi = \sqrt{\Delta/t}$. The nice properties of the Poisson distribution under temporal derivation allow us to derive:

$$\Delta \frac{\mathrm{d}}{\mathrm{d}t} p_t(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\sigma}'} W[\boldsymbol{\sigma}; \boldsymbol{\sigma}'] p_t(\boldsymbol{\sigma}') - p_t(\boldsymbol{\sigma}).$$

For sequential dynamics we choose $\Delta = \frac{1}{N}$ so that, as in the parallel case, in one time unit each neuron will on average be updated once. The master equation corresponding to (5) acquires the form

$$\frac{\mathrm{d}}{\mathrm{d}t} p_t(\boldsymbol{\sigma}) = \sum_i \{ w_i(F_i \boldsymbol{\sigma}) p_t(F_i \boldsymbol{\sigma}) - w_i(\boldsymbol{\sigma}) p_t(\boldsymbol{\sigma}) \}. \tag{7}$$

The $w_i(\boldsymbol{\sigma})$ (6) now play the role of *transition rates*. The choice $\Delta = \frac{1}{N}$ implies $\sqrt{\langle \ell^2 \rangle_\pi - \langle \ell \rangle_\pi^2}/\langle \ell \rangle_\pi = \sqrt{1/Nt}$, so we will still for $N \to \infty$ no longer have uncertainty in where we are on the t axis.

2.1.4. Microscopic definitions for continuous neurons

Alternatively, we could start with continuous neuronal variables σ_i (representing e.g. firing frequencies or oscillator phases), where $i = 1, \ldots, N$, and with stochastic equations of the form

$$\sigma_i(t + \Delta) = \sigma_i(t) + \Delta f_i(\boldsymbol{\sigma}(t)) + \sqrt{2\,T\Delta}\,\xi_i(t). \tag{8}$$

Here we have introduced (as yet unspecified) deterministic state-dependent forces $f_i(\boldsymbol{\sigma})$, and uncorrelated Gaussian distributed random forces $\xi_i(t)$ (the noise), with $\langle \xi_i(t) \rangle = 0$ and $\langle \xi_i(t) \xi_j(t') \rangle = \delta_{ij} \delta_{t,t'}$. As before, the parameter T controls the amount of noise in the system, ranging from $T = 0$ (deterministic dynamics) to $T = \infty$ (completely random dynamics). If we take the limit $\Delta \to 0$ in (8) we find a Langevin equation (with a continuous time variable):

$$\frac{\mathrm{d}}{\mathrm{d}t} \sigma_i(t) = f_i(\boldsymbol{\sigma}(t)) + \eta_i(t). \tag{9}$$

This equation acquires its meaning only as the limit $\Delta \to 0$ of (8). The moments of the new noise variables $\eta_i(t) = \xi_i(t)\sqrt{2T/\Delta}$ in (9) are given by $\langle \eta_i(t) \rangle = 0$ and

$\langle \eta_i(t)\eta_j(t')\rangle = 2T\delta_{ij}\delta(t-t')$. This can be derived from the moments of the $\xi_i(t)$. For instance:

$$\langle \eta_i(t)\eta_j(t')\rangle = \lim_{\Delta\to 0}\frac{2\,T}{\Delta}\langle \xi_i(t)\xi_j(t')\rangle = 2\,T\delta_{ij}\lim_{\Delta\to 0}\frac{1}{\Delta}\delta_{t,t'} = 2\,TC\delta_{ij}\delta(t-t').$$

The constant C is found by summing over t', before taking the limit $\Delta\to 0$, in the above equation:

$$\int \mathrm{d}t'\langle \eta_i(t)\eta_j(t')\rangle = \lim_{\Delta\to 0} 2\,T\sum_{t'=-\infty}^{\infty}\langle \xi_i(t)\xi_j(t')\rangle = 2\,T\delta_{ij}\lim_{\Delta\to 0}\sum_{t'=-\infty}^{\infty}\delta_{t,t'} = 2\,T\delta_{ij}.$$

Thus $C=1$, which indeed implies $\langle \eta_i(t)\eta_j(t')\rangle = 2T\delta_{ij}\delta(t-t')$. More directly, one can also calculate the moment generating function

$$\left\langle \exp\left(\mathrm{i}\int \mathrm{d}t\sum_i \psi_i(t)\eta_i(t)\right)\right\rangle = \lim_{\Delta\to 0}\prod_{i,t}\int\frac{\mathrm{d}z}{\sqrt{2\,\pi}}\exp\left(-\frac{1}{2}z^2+\mathrm{i}z\psi_i(t)\sqrt{2\,T\Delta}\right)$$
$$= \lim_{\Delta\to 0}\prod_{i,t}\mathrm{e}^{-T\Delta\psi_i^2(t)} = \mathrm{e}^{-T\int \mathrm{d}t\,\sum_i \psi_i^2(t)}. \qquad (10)$$

2.1.5. From stochastic equations to evolving probabilities

A mathematically more convenient description of the process (9) is provided by the Fokker–Planck equation for the microscopic state probability density $p_t(\boldsymbol{\sigma}) = \langle \delta[\boldsymbol{\sigma}-\boldsymbol{\sigma}(t)]\rangle$, which we will now derive. For the discrete-time process (8) we expand the δ-distribution in the definition of $p_{t+\Delta}(\boldsymbol{\sigma})$ (in a distributional sense):

$$p_{t+\Delta}(\boldsymbol{\sigma}) - p_t(\boldsymbol{\sigma}) = \left\langle \delta\left[\boldsymbol{\sigma}-\boldsymbol{\sigma}(t)-\Delta\mathbf{f}(\boldsymbol{\sigma}(t))-\sqrt{2T\Delta}\boldsymbol{\xi}(t)\right]\right\rangle - \langle\delta[\boldsymbol{\sigma}-\boldsymbol{\sigma}(t)]\rangle$$
$$= -\sum_i\frac{\partial}{\partial\sigma_i}\left\langle\delta[\boldsymbol{\sigma}-\boldsymbol{\sigma}(t)]\left[\Delta f_i(\boldsymbol{\sigma}(t))+\sqrt{2\,T\Delta}\xi_i(t)\right]\right\rangle$$
$$+ T\Delta\sum_{ij}\frac{\partial^2}{\partial\sigma_i\partial\sigma_j}\langle\delta[\boldsymbol{\sigma}-\boldsymbol{\sigma}(t)]\xi_i(t)\xi_j(t)\rangle + \mathcal{O}(\Delta^{\frac{3}{2}}).$$

The variables $\boldsymbol{\sigma}(t)$ depend only on noise variables $\xi_j(t')$ with $t'<t$, so that for any function A: $\langle A[\boldsymbol{\sigma}(t)]\xi_i(t)\rangle = \langle A[\boldsymbol{\sigma}(t)]\rangle\langle\xi_i(t)\rangle = 0$, and $\langle A[\boldsymbol{\sigma}(t)]\xi_i(t)\xi_j(t)\rangle = \delta_{ij}\langle A[\boldsymbol{\sigma}(t)]\rangle$. As a consequence:

$$\frac{1}{\Delta}[p_{t+\Delta}(\boldsymbol{\sigma})-p_t(\boldsymbol{\sigma})] = -\sum_i\frac{\partial}{\partial\sigma_i}\langle\delta[\boldsymbol{\sigma}-\boldsymbol{\sigma}(t)]f_i(\boldsymbol{\sigma}(t))\rangle$$
$$+ T\sum_i\frac{\partial^2}{\partial\sigma_i^2}\langle\delta[\boldsymbol{\sigma}-\boldsymbol{\sigma}(t)]\rangle + \mathcal{O}(\Delta^{\frac{1}{2}})$$
$$= -\sum_i\frac{\partial}{\partial\sigma_i}[p_t(\boldsymbol{\sigma})f_i(\boldsymbol{\sigma})] + T\sum_i\frac{\partial^2}{\partial\sigma_i^2}p_t(\boldsymbol{\sigma}) + \mathcal{O}(\Delta^{\frac{1}{2}})$$

By taking the limit $\Delta \to 0$ we then arrive at the Fokker–Planck equation:

$$\frac{\mathrm{d}}{\mathrm{d}t} p_t(\boldsymbol{\sigma}) = -\sum_i \frac{\partial}{\partial \sigma_i} [p_t(\boldsymbol{\sigma}) f_i(\boldsymbol{\sigma})] + T \sum_i \frac{\partial^2}{\partial \sigma_i^2} p_t(\boldsymbol{\sigma}). \tag{11}$$

2.1.6. Examples: graded response neurons and coupled oscillators

In the case of graded response neurons the continuous variable σ_i represents the membrane potential of neuron i, and (in their simplest form) the deterministic forces are given by $f_i(\boldsymbol{\sigma}) = \sum_j J_{ij} \tanh[\gamma \sigma_j] - \sigma_i + \theta_i$, with $\gamma > 0$ and with the θ_i representing injected currents. Conventional notation is restored by putting $\sigma_i \to u_i$. Thus equation (9) specializes to

$$\frac{\mathrm{d}}{\mathrm{d}t} u_i(t) = \sum_j J_{ij} \tanh[\gamma u_j(t)] - u_i(t) + \theta_i + \eta_i(t). \tag{12}$$

One often chooses $T = 0$ (i.e. $\eta_i(t) = 0$), the rationale being that threshold noise is already assumed to have been incorporated via the nonlinearity in (12).

In our second example the variables σ_i represent the phases of coupled neural oscillators, with forces of the form $f_i(\boldsymbol{\sigma}) = \sum_j J_{ij} \sin(\sigma_j - \sigma_i) + \omega_i$. Individual synapses J_{ij} now try to enforce either pair-wise synchronization ($J_{ij} > 0$) or pair-wise antisynchronization ($J_{ij} < 0$), and the ω_i represent the natural frequencies of the individual oscillators. Conventional notation dictates $\sigma_i \to \phi_i$, giving

$$\frac{\mathrm{d}}{\mathrm{d}t} \phi_i(t) = \omega_i + \sum_j J_{ij} \sin[\phi_j(t) - \phi_i(t)] + \eta_i(t). \tag{13}$$

2.2. Synaptic symmetry and Lyapunov functions

2.2.1. Noise-free symmetric networks of binary neurons

In the deterministic limit $T \to 0$ the rules (1) for networks of synchronously evolving binary neurons reduce to the deterministic map

$$\sigma_i(\ell + 1) = \mathrm{sgn}[h_i(\boldsymbol{\sigma}(\ell))]. \tag{14}$$

It turns out that for systems with symmetric interactions, $J_{ij} = J_{ji}$ for all (ij), one can construct a Lyapunov function, i.e. a function of $\boldsymbol{\sigma}$ which during the dynamics decreases monotonically and is bounded from below (see e.g. [3]):

Binary & Parallel:

$$L[\boldsymbol{\sigma}] = -\sum_i |h_i(\boldsymbol{\sigma})| - \sum_i \sigma_i \theta_i. \tag{15}$$

Clearly $L \geqslant -\sum_i [\sum_j |J_{ij}| + |\theta_i|] - \sum_i |\theta_i|$. During iteration of (14) we find:

$$
\begin{aligned}
L[\boldsymbol{\sigma}(\ell+1)] - L[\boldsymbol{\sigma}(\ell)] &= -\sum_i |h_i(\boldsymbol{\sigma}(\ell+1))| + \sum_i \sigma_i(\ell+1)\left[\sum_j J_{ij}\sigma_j(\ell) + \theta_i\right] \\
&\quad - \sum_i \theta_i[\sigma_i(\ell+1) - \sigma_i(\ell)] \\
&= -\sum_i |h_i(\boldsymbol{\sigma}(\ell+1))| + \sum_i \sigma_i(\ell) h_i(\boldsymbol{\sigma}(\ell+1)) \\
&= -\sum_i |h_i(\boldsymbol{\sigma}(\ell+1))|[1 - \sigma_i(\ell+2)\sigma_i(\ell)] \leqslant 0
\end{aligned}
$$

(where we used (14) and $J_{ij} = J_{ji}$). So L decreases monotonically until a stage is reached where $\sigma_i(\ell+2) = \sigma_i(\ell)$ for all i. Thus, with symmetric interactions this system will in the deterministic limit always end up in a limit cycle with period $\leqslant 2$. A similar result is found for networks with binary neurons and sequential dynamics. In the limit $T \to 0$ the rules (2) reduce to the map

$$
\sigma_i(\ell+1) = \delta_{i,i_\ell}\operatorname{sgn}[h_i(\boldsymbol{\sigma}(\ell))] + [1 - \delta_{i,i_\ell}]\sigma_i(\ell) \tag{16}
$$

(in which we still have randomness in the choice of site to be updated). For systems with symmetric interactions and without self-interactions, i.e. $J_{ii} = 0$ for all i, we again find a Lyapunov function:

Binary & Sequential:

$$
L[\boldsymbol{\sigma}] = -\frac{1}{2}\sum_{ij} \sigma_i J_{ij}\sigma_j - \sum_i \sigma_i\theta_i. \tag{17}
$$

This quantity is bounded from below: $L \geqslant -\frac{1}{2}\sum_{ij}|J_{ij}| - \sum_i |\theta_i|$. Upon calling the site i_ℓ selected for update at step ℓ simply i, the change in L during iteration of (16) can be written as:

$$
\begin{aligned}
L[\boldsymbol{\sigma}(\ell+1)] - L[\boldsymbol{\sigma}(\ell)] &= -\theta_i[\sigma_i(\ell+1) - \sigma_i(\ell)] \\
&\quad - \frac{1}{2}\sum_k J_{ik}[\sigma_i(\ell+1)\sigma_k(\ell+1) - \sigma_i(\ell)\sigma_k(\ell)] \\
&\quad - \frac{1}{2}\sum_j J_{ji}[\sigma_j(\ell+1)\sigma_i(\ell+1) - \sigma_j(\ell)\sigma_i(\ell)] \\
&= [\sigma_i(\ell) - \sigma_i(\ell+1)]\left[\sum_j J_{ij}\sigma_j(\ell) + \theta_i\right] \\
&= -|h_i(\boldsymbol{\sigma}(\ell))|[1 - \sigma_i(\ell)\sigma_i(\ell+1)] \leqslant 0.
\end{aligned}
$$

Here we used (16), $J_{ij} = J_{ji}$, and absence of self-interactions. Thus L decreases monotonically until $\sigma_i(t+1) = \sigma_i(t)$ for all i. With symmetric synapses, but without diagonal terms, the sequentially evolving binary neurons system will in the deterministic limit always end up in a stationary state.

2.2.2. Noise-free symmetric networks of continuous neurons

One can derive similar results for models with continuous variables. Firstly, in the deterministic limit the graded response equations (12) simplify to

$$\frac{\mathrm{d}}{\mathrm{d}t}u_i(t) = \sum_j J_{ij}\tanh[\gamma u_j(t)] - u_i(t) + \theta_i. \tag{18}$$

Symmetric networks again admit a Lyapunov function (there is no need to eliminate self-interactions):

Graded response:

$$L[\mathbf{u}] = -\frac{1}{2}\sum_{ij} J_{ij}\tanh[\gamma u_i]\tanh[\gamma u_j] + \sum_i \left[\gamma\int_0^{u_i}\mathrm{d}v\, v[1-\tanh^2[\gamma v]] - \theta_i\tanh[\gamma u_i]\right]$$

Clearly $L \geqslant -\frac{1}{2}\sum_{ij}|J_{ij}| - \sum_i|\theta_i|$ (the term in $L[\mathbf{u}]$ with the integral is nonnegative). During the noise-free dynamics (18) one can use the identity $\partial L/\partial u_i = -\gamma[1-\tanh^2[\gamma u_i]](\mathrm{d}u_i/\mathrm{d}t)$, valid only when $J_{ij} = J_{ji}$, to derive

$$\frac{\mathrm{d}}{\mathrm{d}t}L = \sum_i \frac{\partial L}{\partial u_i}\frac{\mathrm{d}u_i}{\mathrm{d}t} = -\gamma\sum_i[1-\tanh^2[\gamma u_i]]\left[\frac{\mathrm{d}}{\mathrm{d}t}u_i\right]^2 \leqslant 0.$$

Again L is found to decrease monotonically, until $\mathrm{d}u_i/\mathrm{d}t = 0$ for all i, i.e. until we are at a fixed-point.

Finally, the coupled oscillator equations (13) reduce in the noise-free limit to

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_i(t) = \omega_i + \sum_j J_{ij}\sin[\phi_j(t) - \phi_i(t)]. \tag{19}$$

Note that self-interactions J_{ii} always drop out automatically. For symmetric oscillator networks, a construction of the type followed for the graded response equations would lead us to propose

Coupled oscillators:

$$L[\boldsymbol{\phi}] = -\frac{1}{2}\sum_{ij} J_{ij}\cos[\phi_i - \phi_j] - \sum_i \omega_i\phi_i. \tag{20}$$

This function indeed decreases monotonically, due to $\partial L/\partial\phi_i = -\mathrm{d}\phi_i/\mathrm{d}t$:

$$\frac{\mathrm{d}}{\mathrm{d}t}L = \sum_i \frac{\partial L}{\partial\phi_i}\frac{\mathrm{d}\phi_i}{\mathrm{d}t} = -\sum_i\left[\frac{\mathrm{d}}{\mathrm{d}t}\phi_i\right]^2 \leqslant 0.$$

In fact (19) describes gradient descent on the surface $L[\boldsymbol{\phi}]$. However, due to the term with the natural frequencies ω_i the function $L[\boldsymbol{\phi}]$ is not bounded, so it cannot be a Lyapunov function. This could have been expected; when $J_{ij} = 0$ for all (i,j), for instance, one finds continually increasing phases $\phi_i(t) = \phi_i(0) + \omega_i t$. Removing the

ω_i, in contrast, gives the bound $L \geqslant -\sum_j |J_{ij}|$. Now the system must go to a fixed-point. In the special case $\omega_i = \omega$ (N identical natural frequencies) we can transform away the ω_i by putting $\phi(t) = \tilde{\phi}_i(t) + \omega t$, and find the relative phases $\tilde{\phi}_i$ to go to a fixed-point.

2.3. Detailed balance and equilibrium statistical mechanics

2.3.1. Detailed balance for binary networks

The results obtained above indicate that networks with symmetric synapses are a special class. We now show how synaptic symmetry is closely related to the detailed balance property, and derive a number of consequences. An ergodic Markov chain of the form (4) and (5), i.e.

$$p_{\ell+1}(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\sigma}'} W[\boldsymbol{\sigma}; \boldsymbol{\sigma}'] p_\ell(\boldsymbol{\sigma}'). \tag{21}$$

is said to obey detailed balance if its (unique) stationary solution $p_\infty(\boldsymbol{\sigma})$ has the property

$$W[\boldsymbol{\sigma}; \boldsymbol{\sigma}'] p_\infty(\boldsymbol{\sigma}') = W[\boldsymbol{\sigma}'; \boldsymbol{\sigma}] p_\infty(\boldsymbol{\sigma}) \quad \text{for all } \boldsymbol{\sigma}, \boldsymbol{\sigma}'. \tag{22}$$

All $p_\infty(\boldsymbol{\sigma})$ which satisfy (22) are stationary solutions of (21), this is easily verified by substitution. The converse is not true. Detailed balance states that, in addition to $p_\infty(\boldsymbol{\sigma})$ being stationary, one has *equilibrium*: there is no net probability current between any two microscopic system states.

It is not a trivial matter to investigate systematically for which choices of the threshold noise distribution $w(\eta)$ and the synaptic matrix $\{J_{ij}\}$ detailed balance holds. It can be shown that, apart from trivial cases (e.g. systems with self-interactions only) a Gaussian distribution $w(\eta)$ will not support detailed balance. Here we will work out details only for the choice $w(\eta) = \frac{1}{2}[1 - \tanh^2(\eta)]$, and for $T > 0$ (where both discrete systems are ergodic). For parallel dynamics the transition matrix is given in (4), now with $g[z] = \tanh[z]$, and the detailed balance condition (22) becomes

$$\frac{e^{\beta \sum_i \sigma_i h_i(\boldsymbol{\sigma}')} p_\infty(\boldsymbol{\sigma}')}{\prod_i \cosh[\beta h_i(\boldsymbol{\sigma}')]} = \frac{e^{\beta \sum_i \sigma_i' h_i(\boldsymbol{\sigma})} p_\infty(\boldsymbol{\sigma})}{\prod_i \cosh[\beta h_i(\boldsymbol{\sigma})]} \quad \text{for all } \boldsymbol{\sigma}, \boldsymbol{\sigma}'. \tag{23}$$

All $p_\infty(\boldsymbol{\sigma})$ are nonzero (ergodicity), so we may safely put $p_\infty(\boldsymbol{\sigma}) = e^{\beta[\sum_i \theta_i \sigma_i + K(\boldsymbol{\sigma})]} \prod_i \cosh[\beta h_i(\boldsymbol{\sigma})]$, which, in combination with definition (1) simplifies the detailed balance condition to:

$$K(\boldsymbol{\sigma}) - K(\boldsymbol{\sigma}') = \sum_{ij} \sigma_i [J_{ij} - J_{ji}] \sigma_j' \quad \text{for all } \boldsymbol{\sigma}, \boldsymbol{\sigma}'. \tag{24}$$

Averaging (24) over all possible $\boldsymbol{\sigma}'$ gives $K(\boldsymbol{\sigma}) = \langle K(\boldsymbol{\sigma}') \rangle_{\boldsymbol{\sigma}'}$ for all $\boldsymbol{\sigma}$, i.e. K is a constant, whose value follows from normalizing $p_\infty(\boldsymbol{\sigma})$. So, if detailed balance holds the equilibrium distribution must be:

$$p_{\rm eq}(\boldsymbol{\sigma}) \sim {\rm e}^{\beta \sum_i \theta_i \sigma_i} \prod_i \cosh[\beta h_i(\boldsymbol{\sigma})]. \tag{25}$$

For symmetric systems detailed balance indeed holds: (25) solves (23), since $K(\boldsymbol{\sigma}) = K$ solves the reduced problem (24). For nonsymmetric systems, however, there can be no equilibrium. For $K(\boldsymbol{\sigma}) = K$ condition (24) becomes $\sum_{ij} \sigma_i \big[J_{ij} - J_{ji}\big] \sigma'_j = 0$ for all $\boldsymbol{\sigma}, \boldsymbol{\sigma}' \in \{-1, 1\}^N$. For $N \geqslant 2$ the vector pairs $(\boldsymbol{\sigma}, \boldsymbol{\sigma}')$ span the space of all $N \times N$ matrices, so $J_{ij} - J_{ji}$ must be zero. For $N = 1$ there simply exists no non-symmetric synaptic matrix. In conclusion: for binary networks with parallel dynamics, interaction symmetry implies detailed balance, and vice versa.

For sequential dynamics, with $w(\eta) = \frac{1}{2}[1 - \tanh^2(\eta)]$, the transition matrix is given by (5) and the detailed balance condition (22) simplifies to

$$\frac{{\rm e}^{\beta \sigma_i h_i(F_i \boldsymbol{\sigma})} p_\infty(F_i \boldsymbol{\sigma})}{\cosh[\beta h_i(F_i \boldsymbol{\sigma})]} = \frac{{\rm e}^{-\beta \sigma_i h_i(\boldsymbol{\sigma})} p_\infty(\boldsymbol{\sigma})}{\cosh[\beta h_i(\boldsymbol{\sigma})]} \quad \text{for all } \boldsymbol{\sigma} \text{ and all } i.$$

Self-interactions J_{ii}, inducing $h_i(F_i \boldsymbol{\sigma}) \neq h_i(\boldsymbol{\sigma})$, complicate matters. Therefore we first consider systems where all $J_{ii} = 0$. All stationary probabilities $p_\infty(\boldsymbol{\sigma})$ being nonzero (ergodicity), we may write:

$$p_\infty(\boldsymbol{\sigma}) = \exp\left(\beta \left[\sum_i \theta_i \sigma_i + \frac{1}{2} \sum_{i \neq j} \sigma_i J_{ij} \sigma_j + K(\boldsymbol{\sigma}) \right] \right). \tag{26}$$

Using relations like $\sum_{k \neq l} J_{kl} F_i(\sigma_k \sigma_l) = \sum_{k \neq l} J_{kl} \sigma_k \sigma_l - 2\sigma_i \sum_{k \neq i} [J_{ik} + J_{ki}] \sigma_k$ we can simplify the detailed balance condition to $K(F_i \boldsymbol{\sigma}) - K(\boldsymbol{\sigma}) = \sigma_i \sum_{k \neq i} [J_{ik} - J_{ki}] \sigma_k$ for all $\boldsymbol{\sigma}$ and all i. If to this expression we apply the general identity $[1 - F_i] f(\boldsymbol{\sigma}) = 2\sigma_i \langle \sigma_i f(\boldsymbol{\sigma}) \rangle_{\sigma_i}$ we find for $i \neq j$:

$$[F_j - 1][F_i - 1] K(\boldsymbol{\sigma}) = -2\, \sigma_i \sigma_j \big[J_{ij} - J_{ji}\big] \quad \text{for all } \boldsymbol{\sigma} \text{ and all } i \neq j.$$

The left-hand side is symmetric under permutation of the pair (i, j), which implies that the interaction matrix must also be symmetric: $J_{ij} = J_{ji}$ for all (i, j). We now find the trivial solution $K(\boldsymbol{\sigma}) = K$ (constant), detailed balance holds and the corresponding equilibrium distribution is

$$p_{\rm eq}(\boldsymbol{\sigma}) \sim {\rm e}^{-\beta H(\boldsymbol{\sigma})}, \quad H(\boldsymbol{\sigma}) = -\frac{1}{2} \sum_{i \neq j} \sigma_i J_{ij} \sigma_j - \sum_i \theta_i \sigma_i. \tag{27}$$

In conclusion: for binary networks with sequential dynamics, but without self-interactions, interaction symmetry implies detailed balance, and vice versa. In the case of self-interactions the situation is more complicated. However, here one can still show that nonsymmetric models with detailed balance must be pathological,since the requirements can be met only for very specific choices for the $\{J_{ij}\}$.

2.3.2. Detailed balance for networks with continuous neurons

Let us finally turn to the question of when we find microscopic equilibrium (stationarity without probability currents) in continuous models described by a Fokker–

Planck equation (11). Note that (11) can be seen as a continuity equation for the density of a conserved quantity: $\frac{\mathrm{d}}{\mathrm{d}t}p_t(\boldsymbol{\sigma}) + \sum_i \frac{\partial}{\partial\sigma_i}J_i(\boldsymbol{\sigma},t) = 0$. The components $J_i(\boldsymbol{\sigma},t)$ of the current density are given by

$$J_i(\boldsymbol{\sigma},t) = \left[f_i(\boldsymbol{\sigma}) - T\frac{\partial}{\partial\sigma_i}\right]p_t(\boldsymbol{\sigma})$$

Stationary distributions $p_\infty(\boldsymbol{\sigma})$ are those which give $\sum_i \frac{\partial}{\partial\sigma_i}J_i(\boldsymbol{\sigma},\infty) = 0$ (divergence-free currents). Detailed balance implies the stronger statement $J_i(\boldsymbol{\sigma},\infty) = 0$ for all i (zero currents), so $f_i(\boldsymbol{\sigma}) = T\partial\log p_\infty(\boldsymbol{\sigma})/\partial\sigma_i$, or

$$f_i(\boldsymbol{\sigma}) = -\partial H(\boldsymbol{\sigma})/\partial\sigma_i, \quad p_\infty(\boldsymbol{\sigma}) \sim \mathrm{e}^{-\beta H(\boldsymbol{\sigma})} \tag{28}$$

for some $H(\boldsymbol{\sigma})$, i.e. the forces $f_i(\boldsymbol{\sigma})$ must be conservative. However, one can have conservative forces without a normalizable equilibrium distribution. Just take $H(\boldsymbol{\sigma}) = 0$, i.e. $f_i(\boldsymbol{\sigma},t) = 0$: here we have $p_{\mathrm{eq}}(\boldsymbol{\sigma}) = C$, which is not normalizable for $\boldsymbol{\sigma} \in \Re^N$. For this particular case Eq. (11) is solved easily: $p_t(\boldsymbol{\sigma}) = [4\pi Tt]^{-N/2}\int \mathrm{d}\boldsymbol{\sigma}' p_0(\boldsymbol{\sigma}')\mathrm{e}^{-[\boldsymbol{\sigma}-\boldsymbol{\sigma}']^2/4Tt}$, so the limit $\lim_{t\to\infty} p_t(\boldsymbol{\sigma})$ indeed does not exist. One can prove the following (see e.g. [4]). If the forces are conservative and if $p_\infty(\boldsymbol{\sigma}) \sim \mathrm{e}^{-\beta H(\boldsymbol{\sigma})}$ is normalizable, then it is the unique stationary solution of the Fokker–Planck equation, to which the system converges for all initial distributions $p_0 \in L^1[\Re^N]$ which obey $\int_{\Re^N}\mathrm{d}\boldsymbol{\sigma}\ \mathrm{e}^{\beta H(\boldsymbol{\sigma})}p_0^2(\boldsymbol{\sigma}) < \infty$.

Assessing when our two particular model examples of graded response neurons or coupled oscillators obey detailed balance has thus been reduced mainly to checking whether the associated deterministic forces $f_i(\boldsymbol{\sigma})$ are conservative. Note that conservative forces must obey

$$\text{for all } \boldsymbol{\sigma}, \text{ for all } i \neq j : \partial f_i(\boldsymbol{\sigma})/\partial\sigma_j - \partial f_j(\boldsymbol{\sigma})/\partial\sigma_i = 0. \tag{29}$$

In the graded response equations (18) the deterministic forces are $f_i(\mathbf{u}) = \sum_j J_{ij}\tanh[\gamma u_j] - u_i + \theta_i$. Here $\partial f_i(\mathbf{u})/\partial u_j - \partial f_j(\mathbf{u})/\partial u_i = \gamma\{J_{ij}[1-\tanh^2[\gamma u_j]-J_{ji}[1-\tanh^2[\gamma u_i]\}$. At $\mathbf{u} = \mathbf{0}$ this reduces to $J_{ij} - J_{ji} = 0$, i.e. the interaction matrix must be symmetric. For symmetric matrices we find away from $\mathbf{u} = \mathbf{0}$: $\partial f_i(\mathbf{u})/\partial u_j - \partial f_j(\mathbf{u})/\partial u_i = \gamma J_{ij}\{\tanh^2[\gamma u_i] - \tanh^2[\gamma u_j]\}$. The only way for this to be zero for any $\mathbf{u}$ is by having $J_{ij} = 0$ for all $i \neq j$, i.e. all neurons are disconnected (in this trivial case the system (18) does indeed obey detailed balance). Network models of interacting graded-response neurons of the type (18) apparently never reach equilibrium, they will always violate detailed balance and exhibit microscopic probability currents. In the case of coupled oscillators (13), where the deterministic forces are $f_i(\boldsymbol{\phi}) = \sum_j J_{ij}\sin[\phi_j - \phi_i] + \omega_i$ one finds the left-hand side of condition (29) to give $\partial f_i(\boldsymbol{\phi})/\partial\phi_j - \partial f_j(\boldsymbol{\phi})/\partial\phi_i = [J_{ij} - J_{ji}]\cos[\phi_j - \phi_i]$. Requiring this to be zero for any $\boldsymbol{\phi}$ gives the condition $J_{ij} = J_{ji}$ for any $i \neq j$. We have already seen that symmetric oscillator networks indeed have conservative forces: $f_i(\boldsymbol{\phi}) = -\partial H(\boldsymbol{\phi})/\partial\phi_i$, with $H(\boldsymbol{\phi}) = -\frac{1}{2}\sum_{ij}J_{ij}\cos[\phi_i - \phi_j] - \sum_i \omega_i\phi_i$. If in addition we choose all $\omega_i = 0$ the function $H(\boldsymbol{\sigma})$ will also be bounded from below, and, although $p_\infty(\boldsymbol{\phi}) \sim \mathrm{e}^{-\beta H(\boldsymbol{\phi})}$ is still not normalizable on $\boldsymbol{\phi} \in \Re^N$, the full 2π-periodicity of the function $H(\boldsymbol{\sigma})$ now allows us to identify $\phi_i + 2\pi \equiv \phi_i$ for all i, so that now

$\boldsymbol{\phi} \in [-\pi, \pi]^N$ and $\int \mathrm{d}\boldsymbol{\phi}\ \mathrm{e}^{-\beta H(\boldsymbol{\phi})}$ does exist. Thus symmetric coupled oscillator networks with zero natural frequencies obey detailed balance. In the case of nonzero natural frequencies, in contrast, detailed balance does not hold.

2.3.3. Equilibrium statistical mechanics

The above results establish the link with equilibrium statistical mechanics (see e.g. [5,6]). For binary systems with symmetric synapses (in the sequential case: without self-interactions) and with threshold noise distributions of the form $w(\eta) = \frac{1}{2}[1 - \tanh^2(\eta)]$, detailed balance holds and we know the equilibrium distributions. For sequential dynamics it has the Boltzmann form (27) and we can apply standard equilibrium statistical mechanics. The parameter β can formally be identified with the inverse 'temperature' in equilibrium, $\beta = T^{-1}$, and the function $H(\boldsymbol{\sigma})$ is the usual Ising spin Hamiltonian. In particular we can define the partition function Z and the free energy F:

$$p_{\mathrm{eq}}(\boldsymbol{\sigma}) = \frac{1}{Z}\mathrm{e}^{-\beta H(\boldsymbol{\sigma})}, \quad H(\boldsymbol{\sigma}) = -\frac{1}{2}\sum_{i\neq j}\sigma_i J_{ij}\sigma_j - \sum_i \theta_i\sigma_i, \tag{30}$$

$$Z = \sum_{\boldsymbol{\sigma}} \mathrm{e}^{-\beta H(\boldsymbol{\sigma})}, \quad F = -\beta^{-1}\log Z. \tag{31}$$

The free energy can be used as the generating function for equilibrium averages. Taking derivatives with respect to external fields θ_i and interactions J_{ij}, for instance, produces $\langle\sigma_i\rangle = -\partial F/\partial\theta_i$ and $\langle\sigma_i\sigma_j\rangle = -\partial F/\partial J_{ij}$, whereas equilibrium averages of arbitrary state variables $f(\boldsymbol{\sigma})$ can be obtained by adding suitable generating terms to the Hamiltonian: $H(\boldsymbol{\sigma}) \to H(\boldsymbol{\sigma}) + \lambda f(\boldsymbol{\sigma})$, $\langle f\rangle = \lim_{\lambda\to 0}\partial F/\partial\lambda$.

In the parallel case (25) we can again formally write the equilibrium probability distribution in the Boltzmann form [7] and define a corresponding partition function $\tilde{Z}$ and a free energy $\tilde{F}$:

$$p_{\mathrm{eq}}(\boldsymbol{\sigma}) = \frac{1}{Z}\mathrm{e}^{-\beta\tilde{H}(\boldsymbol{\sigma})}, \quad \tilde{H}(\boldsymbol{\sigma}) = -\sum_i \theta_i\sigma_i - \frac{1}{\beta}\sum_i \log 2\cosh[\beta h_i(\boldsymbol{\sigma})], \tag{32}$$

$$\tilde{Z} = \sum_{\boldsymbol{\sigma}} \mathrm{e}^{-\beta\tilde{H}(\boldsymbol{\sigma})}, \quad \tilde{F} = -\beta^{-1}\log\tilde{Z}, \tag{33}$$

which again serve to generate averages: $\tilde{H}(\boldsymbol{\sigma}) \to \tilde{H}(\boldsymbol{\sigma}) + \lambda f(\boldsymbol{\sigma})$, $\langle f\rangle = \lim_{\lambda\to 0}\partial\tilde{F}/\partial\lambda$. However, standard thermodynamic relations involving derivation with respect to β need no longer be valid, and derivation with respect to fields or interactions generates different types of averages, such as

$$-\partial\tilde{F}/\partial\theta_i = \langle\sigma_i\rangle + \langle\tanh[\beta h_i(\boldsymbol{\sigma})]\rangle, \quad -\partial\tilde{F}/\partial J_{ii} = \langle\sigma_i\tanh[\beta h_i(\boldsymbol{\sigma})]\rangle,$$

$$i \neq j: -\partial\tilde{F}/\partial J_{ij} = \langle\sigma_i\tanh[\beta h_j(\boldsymbol{\sigma})]\rangle + \langle\sigma_j\tanh[\beta h_i(\boldsymbol{\sigma})]\rangle.$$

One can use $\langle\sigma_i\rangle = \langle\tanh[\beta h_i(\boldsymbol{\sigma})]\rangle$, which can be derived directly from the equilibrium equation $p_{\mathrm{eq}}(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\sigma}'} W[\boldsymbol{\sigma};\boldsymbol{\sigma}']\,p_{\mathrm{eq}}(\boldsymbol{\sigma}')$, to simplify the first of these identities.

A connected network of graded-response neurons can never be in an equilibrium state, so our only model example with continuous neuronal variables for which we can set up the equilibrium statistical mechanics formalism is the system of coupled oscillators (13) with symmetric synapses and absent (or uniform) natural frequencies ω_i. If we define the phases as $\phi_i \in [-\pi, \pi]$ we have again an equilibrium distribution of the Boltzmann form, and we can define the standard thermodynamic quantities:

$$p_{\rm eq}(\boldsymbol{\phi}) = \frac{1}{Z} e^{-\beta H(\boldsymbol{\phi})}, \quad H(\boldsymbol{\phi}) = -\frac{1}{2}\sum_{ij} J_{ij}\cos[\phi_i - \phi_j], \tag{34}$$

$$Z = \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} d\boldsymbol{\phi}\ e^{-\beta H(\boldsymbol{\phi})}, \quad F = -\beta^{-1}\log Z. \tag{35}$$

These generate equilibrium averages in the usual manner. For instance $\langle \cos[\phi_i - \phi_j]\rangle = -\partial F/\partial J_{ij}$, whereas averages of arbitrary state variables $f(\boldsymbol{\phi})$ follow, as before, upon introducing suitable generating terms: $H(\boldsymbol{\phi}) \to H(\boldsymbol{\phi}) + \lambda f(\boldsymbol{\phi})$, $\langle f\rangle = \lim_{\lambda\to 0} \partial F/\partial\lambda$.

In this chapter we restrict ourselves to symmetric networks which obey detailed balance, so that we know the equilibrium probability distribution and equilibrium statistical mechanics applies. In the case of sequential dynamics we will accordingly not allow for the presence of self-interactions.

3. Simple recurrent networks with binary neurons

3.1. Networks with uniform synapses

We now turn to a simple toy model to show how equilibrium statistical mechanics is used for solving neural network models, and to illustrate similarities and differences between the different dynamics types. We choose uniform infinite-range synapses and zero external fields, and calculate the free energy for the binary systems (1) and (2), parallel and sequential, and with threshold noise distribution $w(\eta) = \frac{1}{2}[1 - \tanh^2(\eta)]$:

$$J_{ij} = J_{ji} = J/N \ (i \neq j), \quad J_{ii} = \theta_i = 0 \quad \text{forall } i.$$

The free energy is an extensive object, $\lim_{N\to\infty} F/N$ is finite. For the models (1) and (2) we now obtain:

Binary and sequential:

$$\lim_{N\to\infty} F/N = -\lim_{N\to\infty} (\beta N)^{-1}\log \sum_{\boldsymbol{\sigma}} e^{\beta N\left[\frac{1}{2}Jm^2(\boldsymbol{\sigma})\right]}$$

Binary and parallel:

$$\lim_{N\to\infty} \tilde{F}/N = -\lim_{N\to\infty} (\beta N)^{-1}\log \sum_{\boldsymbol{\sigma}} e^{N[\log 2\cosh[\beta J m(\boldsymbol{\sigma})]]}$$

with the average activity $m(\boldsymbol{\sigma}) = \frac{1}{N}\sum_k \sigma_k$. We have to count the number of states $\boldsymbol{\sigma}$ with a prescribed average activity $m = 2n/N - 1$ (n is the number of neurons i with $\sigma_i = 1$), in expressions of the form

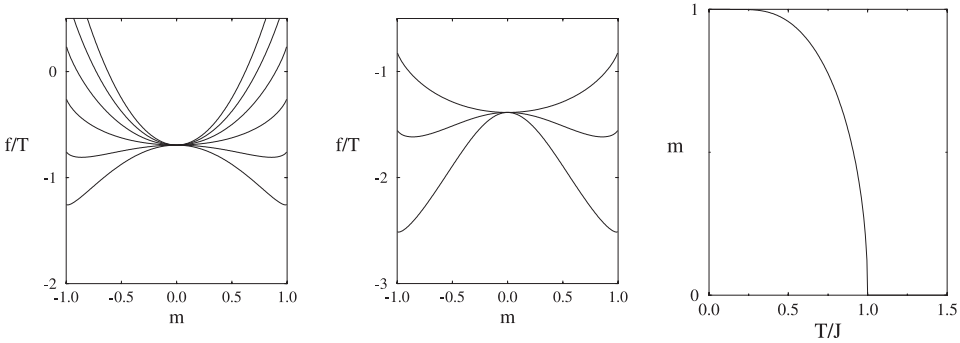

Fig. 2. The functions $f_{\rm seq}(m)/T$ (left) and $f_{\rm par}(m)/T$ (middle) for networks of binary neurons and uniform synapses, and for different choices of the re-scaled interaction strength J/T ($T=\beta^{-1}$). Left picture (sequential dynamics): $J/T=-\frac{5}{2},-\frac{3}{2},-\frac{1}{2},\frac{1}{2},\frac{3}{2},\frac{5}{2}$ (from top to bottom). Middle picture (parallel dynamics): $J/T=\pm\frac{5}{2},\pm\frac{3}{2},\pm\frac{1}{2}$ (from top to bottom, here the free energy is independent of the sign of J). The right picture gives, for $J>0$, the location of the nonnegative minimum of $f_{\rm seq}(m)$ and $f_{\rm par}(m)$ (which is identical to the average activity in thermal equilibrium) as a function of T/J. A phase transition to states with nonzero average activity occurs at $T/J=1$.

$$\begin{aligned}\frac{1}{N}\log\sum_{\boldsymbol{\sigma}}\mathrm{e}^{NU[m(\boldsymbol{\sigma})]}&=\frac{1}{N}\log\sum_{n=0}^{N}\binom{N}{n}\mathrm{e}^{NU[2n/N-1]}\\&=\frac{1}{N}\log\int_{-1}^{1}\mathrm{d}m\,\mathrm{e}^{N[\log 2-c^*(m)+U[m]]}\lim_{N\to\infty}\frac{1}{N}\log\sum_{\boldsymbol{\sigma}}\mathrm{e}^{NU[m(\boldsymbol{\sigma})]}\\&=\log 2+\max_{m\in[-1,1]}\{U[m]-c^*(m)\}\end{aligned}$$

with the entropic function $c^*(m)=\frac{1}{2}(1+m)\log(1+m)+\frac{1}{2}(1-m)\log(1-m)$. In order to get there we used Stirling's formula to obtain the leading term of the factorials (only terms which are exponential in N survive the limit $N\to\infty$), we converted (for $N\to\infty$) the summation over n into an integration over $m=2n/N-1\in[-1,1]$, and we carried out the integral over m via saddle-point integration (see e.g. [8]). This leads to a saddle-point problem whose solution gives the free energies:

$$\lim_{N\to\infty}F/N=\min_{m\in[-1,1]}f_{\rm seq}(m),\quad \beta f_{\rm seq}(m)=c^*(m)-\log 2-\frac{1}{2}\beta Jm^2, \tag{36}$$

$$\lim_{N\to\infty}\tilde{F}/N=\min_{m\in[-1,1]}f_{\rm par}(m),\quad \beta f_{\rm par}(m)=c^*(m)-2\log 2-\log\cosh[\beta Jm]. \tag{37}$$

The functions to be minimized are shown in Fig. 2. The equations from which to solve the minima are easily obtained by differentiation, using $\frac{\mathrm{d}}{\mathrm{d}m}c^*(m)=\tanh^{-1}(m)$. For sequential dynamics we find

Binary and sequential:

$$m = \tanh[\beta J m] \tag{38}$$

(the so-called Curie–Weiss law). For parallel dynamics we find

$$m = \tanh[\beta J \tanh[\beta J m]].$$

One finds that the solutions of the latter equation again obey a Curie–Weiss law. The definition $\hat{m} = \tanh[\beta|J|m]$ transforms it into the coupled equations $m = \tanh[\beta|J|\hat{m}]$ and $\hat{m} = \tanh[\beta|J|m]$, from which we derive $0 \leqslant [m - \hat{m}]^2 = [m - \hat{m}][\tanh[\beta|J|\hat{m}] - \tanh[\beta|J|m]] \leqslant 0$. Since $\tanh[\beta|J|m]$ is a monotonically increasing function of m, this implies $\hat{m} = m$, so

Binary & Parallel:

$$m = \tanh[\beta|J|m]. \tag{39}$$

Our study of the toy models has thus been reduced to analyzing the nonlinear equations (38) and (39). If $J \geqslant 0$ (excitation) the two types of dynamics lead to the same behavior. At high noise levels, $T > J$, both minimization problems are solved by $m = 0$ (see Fig. 2), describing a disorganized (paramagnetic) state. This can be seen upon writing the right-hand side of (38) in integral form:

$$m^2 = m \tanh[\beta J m] = \beta J m^2 \int_0^1 \mathrm{d}z [1 - \tanh^2[\beta J m z]] \leqslant \beta J m^2.$$

So $m^2[1 - \beta J] \leqslant 0$, which gives $m = 0$ as soon as $\beta J < 1$. A phase transition occurs at $T = J$ (a bifurcation of nontrivial solutions of (38)), and for $T < J$ the equations for m are solved by the two nonzero solutions of (38), describing a state where either all neurons tend to be firing ($m > 0$) or where they tend to be quiet ($m < 0$). This becomes clear when we expand (38) for small m: $m = \beta J m + \mathcal{O}(m^3)$, so precisely at $\beta J = 1$ one finds a de-stabilization of the trivial solution $m = 0$, together with the creation of (two) stable nontrivial ones (see also Fig. 2). Furthermore, using the identity $c^*(\tanh x) = x \tanh x - \log \cosh x$, we obtain from (36) and (37) the relation $\lim_{N\to\infty} \tilde{F}/N = 2 \lim_{N\to\infty} F/N$. For $J < 0$ (inhibition), however, the two types of dynamics give quite different results. For sequential dynamics the relevant minimum is located at $m = 0$ (the paramagnetic state). For parallel dynamics, the minimization problem is invariant under $J \to -J$, so the behavior is again of the Curie–Weiss type (see Fig. 2 and Eq. (39)), with a paramagnetic state for $T > |J|$, a phase transition at $T = |J|$, and order for $T < |J|$. This difference between the two types of dynamics for $J < 0$ is explained by studying dynamics. As we will see in a subsequent chapter, for the present (toy) model in the limit $N \to \infty$ the average activity evolves in time according to the deterministic laws

$$\frac{\mathrm{d}}{\mathrm{d}t} m = \tanh[\beta J m] - m, \qquad m(t+1) = \tanh[\beta J m(t)]$$

for sequential and parallel dynamics, respectively. For $J < 0$ the sequential system always decays towards the trivial state $m = 0$, whereas for sufficiently large β the

parallel system enters the stable limit-cycle $m(t) = M_\beta(-1)^t$ (where M_β is the nonzero solution of (39)). The concepts of 'distance' and 'local minima' are quite different for the two dynamics types; in contrast to the sequential case, parallel dynamics allows the system to make the transition $m \to -m$ in equilibrium.

3.2. Phenomenology of Hopfield models

3.2.1. The ideas behind the Hopfield model

The Hopfield model [9] is a network of binary neurons of the type (1) and (2), with threshold noise $w(\eta) = \frac{1}{2}[1 - \tanh^2(\eta)]$, and with a specific recipe for the synapses J_{ij} aimed at storing patterns, motivated by suggestions made in the late 1940s [10]. The original model was in fact defined more narrowly, as the zero noise limit of the system (2), but the term has since then been accepted to cover a larger network class. Let us first consider the simplest case and try to store a single pattern $\boldsymbol{\xi} \in \{-1, 1\}^N$ in noise-less infinite-range binary networks. Appealing candidates for interactions and thresholds would be $J_{ij} = \xi_i \xi_j$ and $\theta_i = 0$ (for sequential dynamics we put $J_{ii} = 0$ for all i). With this choice the Lyapunov function (17) becomes:

$$L_{\text{seq}}[\boldsymbol{\sigma}] = \frac{1}{2}N - \frac{1}{2}\left[\sum_i \xi_i \sigma_i\right]^2.$$

It will have to decrease monotonically during the dynamics, from which we immediately deduce

$$\sum_i \xi_i \sigma_i(0) > 0 : \ \boldsymbol{\sigma}(\infty) = \boldsymbol{\xi}, \quad \sum_i \xi_i \sigma_i(0) < 0 : \ \boldsymbol{\sigma}(\infty) = -\boldsymbol{\xi}.$$

This system indeed reconstructs dynamically the original pattern $\boldsymbol{\xi}$ from an input vector $\boldsymbol{\sigma}(0)$, at least for sequential dynamics. However, *en passant* we have created an additional attractor: the state $-\boldsymbol{\xi}$. This property is shared by all binary models in which the external fields are zero, where the Hamiltonians $H(\boldsymbol{\sigma})$ (30) and $\tilde{H}(\boldsymbol{\sigma})$ (32) are invariant under an overall sign change $\boldsymbol{\sigma} \to -\boldsymbol{\sigma}$. A second feature common to several (but not all) attractor neural networks is that *each* initial state will lead to pattern reconstruction, even nonsensical (random) ones.

The Hopfield model is obtained by generalizing the previous simple one-pattern recipe to the case of an arbitrary number p of binary patterns $\boldsymbol{\xi}^\mu = (\xi_1^\mu, \ldots, \xi_N^\mu) \in \{-1, 1\}^N$:

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu, \quad \theta_i = 0 \quad \text{for all } i \quad (\text{sequential dynamics: } J_{ii} \to 0 \text{ for all } i). \quad (40)$$

The prefactor N^{-1} has been inserted to ensure that the limit $N \to \infty$ will exist in future expressions. The process of interest is that where, triggered by correlation between the initial state and a stored pattern $\boldsymbol{\xi}^\lambda$, the state vector $\boldsymbol{\sigma}$ evolves towards $\boldsymbol{\xi}^\lambda$. If this happens, pattern $\boldsymbol{\xi}^\lambda$ is said to be recalled. The similarity between a state vector and the stored patterns is measured by so-called overlaps

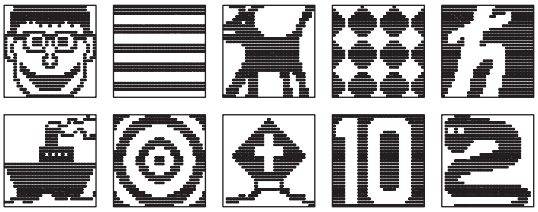

Fig. 3. Information represented as specific microscopic neuronal firing patterns $\boldsymbol{\xi}$ in an $N = 841$ Hopfield network and drawn as images in the plane (black pixels: $\xi_i = 1$, white pixels: $\xi_i = -1$).

$$m_\mu(\boldsymbol{\sigma}) = \frac{1}{N}\sum_i \xi_i^\mu \sigma_i. \tag{41}$$

Numerical simulations illustrate the functioning of the Hopfield model as an associative memory, and the description of the recall process in terms of overlaps. Our simulated system is an $N = 841$ Hopfield model, in which $p = 10$ patterns have been stored (see Fig. 3) according to prescription (40). The two-dimensional arrangement of the neurons in this example is just a guide to the eye; since the network is fully connected the physical location of the neurons is irrelevant. The dynamics is as given by (2), with $T = 0.1$. In Fig. 4 we first show (left column) the result of letting the system evolve in time from an initial state, which is a noisy version of one of the stored patterns (here 40% of the neuronal states σ_i where corrupted, according to $\sigma_i \to -\sigma_i$). The top left row of graphs shows snapshots of the microscopic state as the system evolves in time. The bottom left row shows the values of the $p = 10$ overlaps m_μ, as defined in (41), as functions of time; the one which evolves towards the value 1 corresponds to the pattern being reconstructed. The right column of Fig. 4 shows a similar experiment, but here the initial state is drawn at random. The system subsequently evolves towards a mixture of the stored patterns, which is found to be very stable, due to the fact that the patterns involved (see Fig. 3) are significantly correlated. It will be clear that, although the idea of information storage via the creation of attractors does work, the choice (40) for the synapses is still too simple to be optimal; in addition to the desired states $\boldsymbol{\xi}^\mu$ and their mirror images $-\boldsymbol{\xi}^\mu$, even more unwanted spurious attractors are created. Yet this model will already push the analysis to the limits, as soon as we allow for the storage of an extensive number of patterns $\boldsymbol{\xi}^\mu$.

3.2.2. Issues related to saturation: storage capacity and non-trivial dynamics

In our previous simulation example the loading of the network was modest; a total of $\frac{1}{2}N(N-1) = 353,220$ synapses were used to store just $pN = 8410$ bits of information. Let us now investigate the behavior of the network when the number of patterns scales with the system size as $p = \alpha N$ ($\alpha > 0$); now for large N the number of bits stored per synapse will be $pN/\frac{1}{2}N(N-1) \approx 2\alpha$. This is called the saturation regime. Again numerical simulations, but now with finite α, illustrate the main features and complications of recall dynamics in the saturation regime. In our ex-

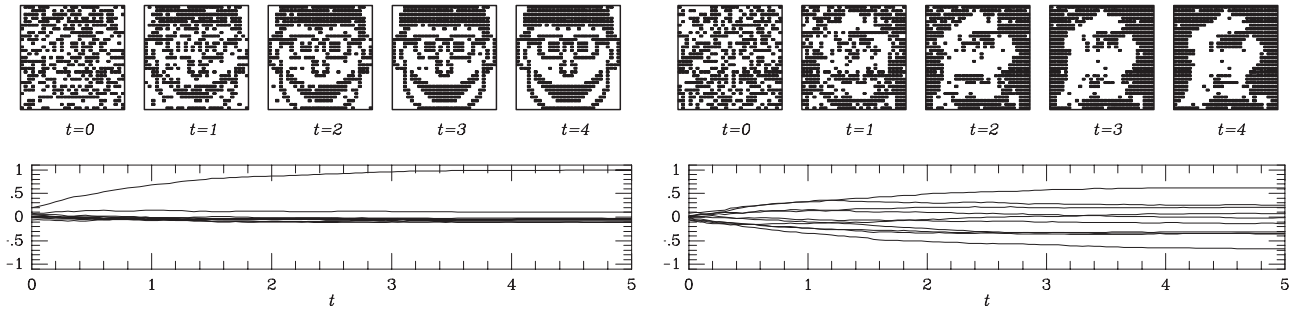

Fig. 4. Information processing in a sequential dynamics Hopfield model with $N = 841$, $p = 10$ and $T = 0.1$, and with the $p = 10$ stored patterns shown in Fig. 3. Left pictures: dynamic reconstruction of a stored pattern from an initial state which is a corrupted version thereof. Top left: snapshots of the system state at times $t = 0, 1, 2, 3, 4$ iterations/neuron. Bottom left: values of the overlap order parameters as functions of time. Right pictures: evolution towards a spurious state from a randomly drawn initial state. Top right: snapshots of the microscopic system state at times $t = 0, 1, 2, 3, 4$ iterations/neuron. Bottom right: values of the overlap order parameters as functions of time.

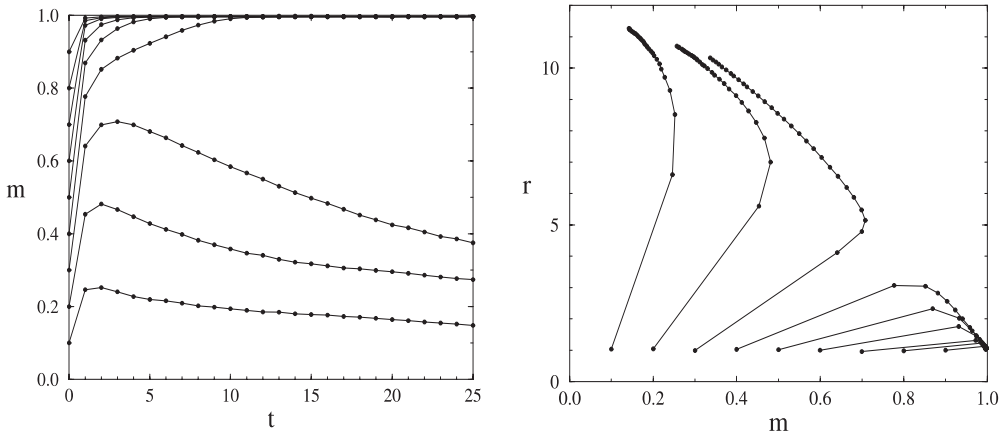

Fig. 5. Simulations of a parallel dynamics Hopfield model with $N = 30,000$ and $\alpha = T = 0.1$, and with random patterns. Left: overlaps $m = m_1(\boldsymbol{\sigma})$ with pattern one as functions of time, following initial states correlated with pattern one only, with $m_1(\boldsymbol{\sigma}(0)) \in \{0.1, \ldots, 0.9\}$. Right: corresponding flow in the (m, r) plane, with $r = \alpha^{-1} \sum_{\mu>1} m_\mu^2(\boldsymbol{\sigma})$ measuring the overlaps with nonnominated patterns.

ample the dynamics is given by (1) (parallel updates), with $T = 0.1$ and threshold noise distribution $w(\eta) = \frac{1}{2}[1 - \tanh^2(\eta)]$; the patterns are chosen randomly. Figure 5 shows the result of measuring in such simulations the two quantities

$$m = m_1(\boldsymbol{\sigma}), \quad r = \alpha^{-1} \sum_{\mu>1} m_\mu^2(\boldsymbol{\sigma}) \tag{42}$$

following initial states which are correlated with pattern $\boldsymbol{\xi}^1$ only. For large N we can distinguish structural overlaps, where $m_\mu(\boldsymbol{\sigma}) = \mathcal{O}(1)$, from accidental ones, where $m_\mu(\boldsymbol{\sigma}) = \mathcal{O}(N^{-\frac{1}{2}})$ (as for a randomly drawn $\boldsymbol{\sigma}$). Overlaps with nonnominated patterns are seen to remain $\mathcal{O}(N^{-\frac{1}{2}})$, i.e. $r(t) = \mathcal{O}(1)$. We observe competition between pattern recall ($m \to 1$) and interference of nonnominated patterns ($m \to 0$, with r increasing), and a profound slowing down of the process for nonrecall trajectories. The initial overlap (the 'cue') needed to trigger recall is found to increase with increasing α (the loading) and increasing T (the noise). Further numerical experimentation, with random patterns, reveals that at any noise level T there is a critical storage level $\alpha_c(T)$ above which recall is impossible, with an absolute upper limit of $\alpha_c = \max_T \alpha_c(T) = \alpha_c(0) \approx 0.139$. The competing forces at work are easily recognized when working out the local fields (1), using (40):

$$h_i(\boldsymbol{\sigma}) = \xi_i^1 m_1(\boldsymbol{\sigma}) + \frac{1}{N} \sum_{\mu>1} \xi_i^\mu \sum_{j \neq i} \xi_j^\mu \sigma_j + \mathcal{O}(N^{-1}). \tag{43}$$

The first term in (43) drives $\boldsymbol{\sigma}$ towards pattern $\boldsymbol{\xi}^1$ as soon as $m_1(\boldsymbol{\sigma}) > 0$. The second terms represent interference, caused by correlations between $\boldsymbol{\sigma}$ and nonnominated patterns. One easily shows (to be demonstrated later) that for $N \to \infty$ the fluctua-

tions in the values of the recall overlap m will vanish, and that for the present types of initial states and threshold noise the overlap m will obey

$$m(t+1) = \int \mathrm{d}z\, P_t(z) \tanh[\beta(m(t)+z)],$$
$$P_t(z) = \lim_{N\to\infty} \frac{1}{N} \sum_i \left\langle \delta\left[z - \frac{1}{N} \sum_{\mu>1} \xi_i^1 \xi_i^\mu \sum_{j\neq i} \xi_j^\mu \sigma_j(t) \right] \right\rangle. \tag{44}$$

If all $\sigma_i(0)$ are drawn independently, $\mathrm{Prob}[\sigma_i(0) = \pm\xi_i^1] = \frac{1}{2}[1 \pm m(0)]$, the central limit theorem states that $P_0(z)$ is Gaussian. One easily derives $\langle z\rangle_0 = 0$ and $\langle z^2\rangle_0 = \alpha$, so at $t = 0$ Eq. (44) gives

$$m(1) = \int \frac{\mathrm{d}z}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}z^2} \tanh[\beta(m(0) + z\sqrt{\alpha})]. \tag{45}$$

The above ideas, and Eq. (45) in particular, go back to [11]. For times $t > 0$, however, the independence of the states σ_i need no longer hold. As a simple approximation one could just assume that the σ_i remain uncorrelated at all times, i.e. $\mathrm{Prob}[\sigma_i(t) = \pm\xi_i^1] = \frac{1}{2}[1 \pm m(t)]$ for all $t \geqslant 0$, such that the argument given for $t = 0$ would hold generally, and where (for randomly drawn patterns) the mapping (45) would describe the overlap evolution at all times:

$$m(t+1) = \int \frac{\mathrm{d}z}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}z^2} \tanh[\beta(m(t) + z\sqrt{\alpha})]. \tag{46}$$

This equation, however, must be generally incorrect. Firstly, Fig. 5 already shows that knowledge of $m(t)$ *only* does not yet permit prediction of $m(t+1)$. Secondly, upon working out its bifurcation properties one finds that Eq. (46) predicts a storage capacity of $\alpha_c = 2/\pi \approx 0.637$, which is no way near to what is actually being observed. We will see in the paper on dynamics that only for certain types of extremely diluted networks (where most of the synapses are cut) Eq. (46) is indeed correct on finite times; in these networks the time it takes for correlations between neuron states to build up diverges with N, so that correlations are simply not yet noticeable on finite times.

For fully connected Hopfield networks storing random patterns near saturation, i.e. with $\alpha > 0$, the complicated correlations building up between the microscopic variables in the course of the dynamics generate an interference noise distribution which is intrinsically non-Gaussian, see e.g. Fig. 6. This leads to a highly nontrivial dynamics which is fundamentally different from that in the $\lim_{N\to\infty} p/N = 0$ regime. Solving models of recurrent neural networks in the saturation regime boils down to calculating this non-Gaussian noise distribution, which requires advanced mathematical techniques (in statics and dynamics), and constitutes the main challenge to the theorist. The simplest way to evade this challenge is to study situations where the interference noise is either trivial (as with asymmetric extremely diluted models) or where it vanishes, which happens in fully connected networks when $\alpha = \lim_{N\to\infty} p/N = 0$ (as with finite p). The latter $\alpha = 0$ regime is the one we will explore first.

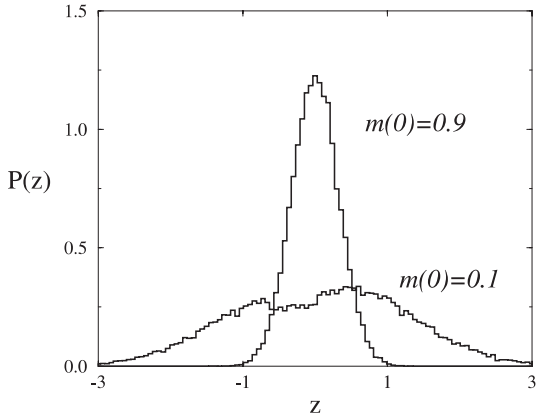

Fig. 6. Distributions of interference noise variables $z_i = \frac{1}{N}\sum_{\mu>1}\xi_i^1\xi_i^\mu\sum_{j\neq i}\xi_j^\mu\sigma_j$, as measured in the simulations of Fig. 5, at $t = 10$. Uni-modal histogram: noise distribution following $m(0) = 0.9$ (leading to recall). Bi-model histogram: noise distribution following $m(0) = 0.1$ (not leading to recall).

3.3. Analysis of Hopfield models away from saturation

3.3.1. Equilibrium order parameter equations

A binary Hopfield network with parameters given by (40) obeys detailed balance, and the Hamiltonian $H(\boldsymbol{\sigma})$ (30) (corresponding to sequential dynamics) and the pseudo-Hamiltonian $\tilde{H}(\boldsymbol{\sigma})$ (32) (corresponding to parallel dynamics) become

$$H(\boldsymbol{\sigma}) = -\frac{1}{2}N\sum_{\mu=1}^{p} m_\mu^2(\boldsymbol{\sigma}) + \frac{1}{2}p, \quad \tilde{H}(\boldsymbol{\sigma}) = -\frac{1}{\beta}\sum_i \log\, 2\cosh\left[\beta\sum_{\mu=1}^{p}\xi_i^\mu m_\mu(\boldsymbol{\sigma})\right] \tag{47}$$

with the overlaps (41). Solving the statics implies calculating the free energies F and $\tilde{F}$:

$$F = -\frac{1}{\beta}\log\sum_{\boldsymbol{\sigma}} \mathrm{e}^{-\beta H(\boldsymbol{\sigma})}, \quad \tilde{F} = -\frac{1}{\beta}\log\sum_{\boldsymbol{\sigma}} \mathrm{e}^{-\beta\tilde{H}(\boldsymbol{\sigma})}.$$

Upon introducing the shorthand notation $\mathbf{m} = (m_1, \ldots, m_p)$ and $\boldsymbol{\xi}_i = (\xi_i^1, \ldots, \xi_i^p)$, both free energies can be expressed in terms of the density of states $\mathscr{D}(\mathbf{m}) = 2^{-N}\sum_{\boldsymbol{\sigma}}\delta[\mathbf{m} - \mathbf{m}(\boldsymbol{\sigma})]$:

$$F/N = -\frac{1}{\beta}\log 2 - \frac{1}{\beta N}\log\int \mathrm{d}\mathbf{m}\,\mathscr{D}(\mathbf{m})\mathrm{e}^{-\frac{1}{2}\beta N\mathbf{m}^2} + \frac{p}{2N} \tag{48}$$

$$\tilde{F}/N = -\frac{1}{\beta}\log 2 - \frac{1}{\beta N}\log\int \mathrm{d}\mathbf{m}\,\mathscr{D}(\mathbf{m})\mathrm{e}^{\sum_{i=1}^{N}\log 2\cosh[\beta\boldsymbol{\xi}_i\cdot\mathbf{m}]} \tag{49}$$

(note: $\int \mathrm{d}\mathbf{m}\,\delta[\mathbf{m} - \mathbf{m}(\boldsymbol{\sigma})] = 1$). In order to proceed we need to specify how the number of patterns p scales with the system size N. In this section we will follow [12]

(equilibrium analysis following sequential dynamics) and [13] (equilibrium analysis following parallel dynamics), and assume p to be finite. One can now easily calculate the leading contribution to the density of states, using the integral representation of the δ-function and keeping in mind that according to (48) and (49) only terms exponential in N will retain statistical relevance for $N \to \infty$:

$$\lim_{N\to\infty} \frac{1}{N} \log \mathscr{D}(\mathbf{m}) = \lim_{N\to\infty} \frac{1}{N} \log \int d\mathbf{x}\, e^{iN\mathbf{x}\cdot\mathbf{m}} \left\langle e^{-i\sum_{i=1}^{N} \sigma_i \boldsymbol{\xi}_i \cdot \mathbf{x}} \right\rangle_{\boldsymbol{\sigma}}$$
$$= \lim_{N\to\infty} \frac{1}{N} \log \int d\mathbf{x}\, e^{N[i\mathbf{x}\cdot\mathbf{m} + \langle \log \cos[\boldsymbol{\xi}\cdot\mathbf{x}]\rangle_{\boldsymbol{\xi}}]}$$

with the abbreviation $\langle \Phi(\boldsymbol{\xi})\rangle_{\boldsymbol{\xi}} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \Phi(\boldsymbol{\xi}_i)$. The leading contribution to both free energies can be expressed as a finite-dimensional integral, for large N dominated by that saddle-point (extremum) for which the extensive exponent is real and maximal:

$$\lim_{N\to\infty} F/N = -\frac{1}{\beta N} \log \int d\mathbf{m}\, d\mathbf{x}\, e^{-N\beta f(\mathbf{m},\mathbf{x})} = \mathrm{extr}_{\mathbf{x},\mathbf{m}}\, f(\mathbf{m},\mathbf{x})$$
$$\lim_{N\to\infty} \tilde{F}/N = -\frac{1}{\beta N} \log \int d\mathbf{m}\, d\mathbf{x}\, e^{-N\beta \tilde{f}(\mathbf{m},\mathbf{x})} = \mathrm{extr}_{\mathbf{x},\mathbf{m}}\, \tilde{f}(\mathbf{m},\mathbf{x})$$

with

$$f(\mathbf{m},\mathbf{x}) = -\frac{1}{2}\mathbf{m}^2 - i\mathbf{x}\cdot\mathbf{m} - \beta^{-1}\langle \log 2\cos[\beta\boldsymbol{\xi}\cdot\mathbf{x}]\rangle_{\boldsymbol{\xi}}$$
$$\tilde{f}(\mathbf{m},\mathbf{x}) = -\beta^{-1}\langle \log 2\cosh[\beta\boldsymbol{\xi}\cdot\mathbf{m}]\rangle_{\boldsymbol{\xi}} - i\mathbf{x}\cdot\mathbf{m} - \beta^{-1}\langle \log 2\cos[\beta\boldsymbol{\xi}\cdot\mathbf{x}]\rangle_{\boldsymbol{\xi}}.$$

The saddle-point equations for f and $\tilde{f}$ are given by:

$$f:\ \mathbf{x} = i\mathbf{m}, \qquad i\mathbf{m} = \langle \boldsymbol{\xi} \tan[\beta\boldsymbol{\xi}\cdot\mathbf{x}]\rangle_{\boldsymbol{\xi}},$$
$$\tilde{f}:\ \mathbf{x} = i\langle \boldsymbol{\xi} \tanh[\beta\boldsymbol{\xi}\cdot\mathbf{m}]\rangle_{\boldsymbol{\xi}}, \quad i\mathbf{m} = \langle \boldsymbol{\xi} \tan[\beta\boldsymbol{\xi}\cdot\mathbf{x}]\rangle_{\boldsymbol{\xi}}.$$

In saddle-points $\mathbf{x}$ turns out to be purely imaginary. However, after a shift of the integration contours, putting $\mathbf{x} = i\mathbf{x}^\star(\mathbf{m}) + \mathbf{y}$ (where $i\mathbf{x}^\star(\mathbf{m})$ is the imaginary saddle-point, and where $\mathbf{y} \in \Re^p$) we can eliminate $\mathbf{x}$ in favor of $\mathbf{y} \in \Re^p$ which does have a real saddle-point, by construction.[1] We then obtain[2]

---

1 Our functions to be integrated have no poles, but strictly speaking we still have to verify that the integration segments linking the original integration regime to the shifted one will not contribute to the integrals. This is generally a tedious and distracting task, which is often skipped. For simple models, however (e.g. networks with uniform synapses), the verification can be carried out properly, and all is found to be safe.

2 Here we used the equation $\partial f(\mathbf{m},\mathbf{x})/\partial\mathbf{m} = \mathbf{0}$ to express $\mathbf{x}$ in terms of $\mathbf{m}$, because this is simpler. Strictly speaking we should have used $\partial f(\mathbf{m},\mathbf{x})/\partial\mathbf{x} = \mathbf{0}$ for this purpose; our short-cut could in principle generate additional solutions. In the present model, however, we can check explicitly that this is not the case. Also, in view of the imaginary saddle-point $\mathbf{x}$, we cannot be certain that, upon elimination of $\mathbf{x}$, the relevant saddle-point of the remaining function $f(\mathbf{m})$ must be a minimum. This will have to be checked, for instance by inspection of the $T \to \infty$ limit.

Sequential dynamics:

$$\mathbf{m} = \langle \boldsymbol{\xi} \tanh[\beta \boldsymbol{\xi} \cdot \mathbf{m}] \rangle_{\boldsymbol{\xi}}$$

Parallel Dynamics:

$$\mathbf{m} = \langle \boldsymbol{\xi} \tanh[\beta \boldsymbol{\xi} \cdot [\langle \boldsymbol{\xi}' \tanh[\beta \boldsymbol{\xi}' \cdot \mathbf{m}] \rangle_{\boldsymbol{\xi}'}]] \rangle_{\boldsymbol{\xi}}$$

(compare to e.g. (38) and (39)). The solutions of the above two equations will in general be identical. To see this, let us denote $\hat{\mathbf{m}} = \langle \boldsymbol{\xi} \tanh[\beta \boldsymbol{\xi} \cdot \mathbf{m}] \rangle_{\boldsymbol{\xi}}$, with which the saddle point equation for $\tilde{f}$ decouples into:

$$\mathbf{m} = \langle \boldsymbol{\xi} \tanh[\beta \boldsymbol{\xi} \cdot \hat{\mathbf{m}}] \rangle_{\boldsymbol{\xi}}, \quad \hat{\mathbf{m}} = \langle \boldsymbol{\xi} \tanh[\beta \boldsymbol{\xi} \cdot \mathbf{m}] \rangle_{\boldsymbol{\xi}}$$

so

$$[\mathbf{m} - \hat{\mathbf{m}}]^2 = \langle [(\boldsymbol{\xi} \cdot \mathbf{m}) - (\boldsymbol{\xi} \cdot \hat{\mathbf{m}})][\tanh(\beta \boldsymbol{\xi} \cdot \hat{\mathbf{m}}) - \tanh(\beta \boldsymbol{\xi} \cdot \mathbf{m})] \rangle_{\boldsymbol{\xi}}.$$

Since *tanh* is a monotonicaly increasing function, we must have $[\mathbf{m} - \hat{\mathbf{m}}] \cdot \boldsymbol{\xi} = 0$ for each $\boldsymbol{\xi}$ that contributes to the averages $\langle \cdots \rangle_{\boldsymbol{\xi}}$. For all choices of patterns, where the covariance matrix $C_{\mu\nu} = \langle \xi_\mu \xi_\nu \rangle_{\boldsymbol{\xi}}$ is positive definite, we thus obtain $\mathbf{m} = \hat{\mathbf{m}}$. The final result is: for both types of dynamics (sequential and parallel) the overlap order parameters in equilibrium are given by the solution $\mathbf{m}^*$ of

$$\mathbf{m} = \langle \boldsymbol{\xi} \tanh[\beta \boldsymbol{\xi} \cdot \mathbf{m}] \rangle_{\boldsymbol{\xi}}, \tag{50}$$

which minimizes[3]

$$f(\mathbf{m}) = \frac{1}{2}\mathbf{m}^2 - \frac{1}{\beta} \langle \log 2 \cosh[\beta \boldsymbol{\xi} \cdot \mathbf{m}] \rangle_{\boldsymbol{\xi}}. \tag{51}$$

The free energies of the ergodic components are $\lim_{N\to\infty} F/N = f(\mathbf{m}^*)$ and $\lim_{N\to\infty} \tilde{F}/N = 2f(\mathbf{m}^*)$. Adding generating terms of the form $H \to H + \lambda g[\mathbf{m}(\boldsymbol{\sigma})]$ to the Hamiltonians allows us to identify $\langle g[\mathbf{m}(\boldsymbol{\sigma})] \rangle_{\text{eq}} = \lim_{\lambda\to 0} \partial F/\partial\lambda = g[\mathbf{m}^*]$. Thus, in equilibrium the fluctuations in the overlap order parameters $\mathbf{m}(\boldsymbol{\sigma})$ (41) vanish for $N \to \infty$. Their deterministic values are simply given by $\mathbf{m}^*$. Note that in the case of sequential dynamics we could also have used linearization with Gaussian integrals (which we will use for coupled oscillators with uniform synapses) to arrive at this solution, with p auxiliary integrations, but that for parallel dynamics this would not have been possible.

3.3.2. Analysis of order parameter equations: pure states and mixture states

We will restrict our further discussion to the case of randomly drawn patterns, so

$$\langle \Phi(\boldsymbol{\xi}) \rangle_{\boldsymbol{\xi}} = 2^{-p} \sum_{\boldsymbol{\xi} \in \{-1,1\}^p} \Phi(\boldsymbol{\xi}), \quad \langle \xi_\mu \rangle_{\boldsymbol{\xi}} = 0, \quad \langle \xi_\mu \xi_\nu \rangle_{\boldsymbol{\xi}} = \delta_{\mu\nu}$$

3 We here indeed know the relevant saddle-point to be a minimum: the only solution of the saddle-point equations at high temperatures, $\mathbf{m} = \mathbf{0}$, is seen to minimize $f(\mathbf{m})$, since $f(\mathbf{m}) + \beta^{-1} \log 2 = \frac{1}{2}\mathbf{m}^2(1 - \beta) + \mathcal{O}(\beta^3)$.

(generalization to correlated patterns is in principle straightforward). We first establish an upper bound for the temperature for where nontrivial solutions $\mathbf{m}^*$ could exist, by writing (50) in integral form:

$$m_\mu = \beta\left\langle \xi_\mu(\boldsymbol{\xi}\cdot\mathbf{m})\int_0^1 \mathrm{d}\lambda\big[1-\tanh^2[\beta\lambda\boldsymbol{\xi}\cdot\mathbf{m}]\big]\right\rangle_\xi$$

from which we deduce

$$\begin{aligned} 0 &= \mathbf{m}^2 - \beta\left\langle (\boldsymbol{\xi}\cdot\mathbf{m})^2\int_0^1 \mathrm{d}\lambda\big[1-\tanh^2[\beta\lambda\boldsymbol{\xi}\cdot\mathbf{m}]\big]\right\rangle_\xi \geqslant \mathbf{m}^2 - \beta\left\langle(\boldsymbol{\xi}\cdot\mathbf{m})^2\right\rangle_\xi \\ &= \mathbf{m}^2(1-\beta). \end{aligned}$$

For $T>1$ the only solution of (50) is the paramagnetic state $\mathbf{m}=0$, which gives for the free energy per neuron $-T\log 2$ and $-2T\log 2$ (for sequential and parallel dynamics, respectively). At $T=1$ a phase transition occurs, which follows from expanding (50) for small $|\mathbf{m}|$ in powers of $\tau=\beta-1$:

$$\begin{aligned} m_\mu &= (1+\tau)m_\mu - \frac{1}{3}\sum_{\nu\rho\lambda} m_\nu m_\rho m_\lambda \langle \xi_\mu\xi_\nu\xi_\rho\xi_\lambda\rangle_\xi + \mathcal{O}(\mathbf{m}^5,\tau\mathbf{m}^3) \\ &= m_\mu\left[1+\tau-\mathbf{m}^2+\frac{2}{3}m_\mu^2\right]+\mathcal{O}(\mathbf{m}^5,\tau\mathbf{m}^3) \end{aligned}$$

The new saddle-point scales as $m_\mu=\tilde{m}_\mu\tau^{1/2}+\mathcal{O}(\tau^{3/2})$, with for each μ: $\tilde{m}_\mu=0$ or $0=1-\tilde{\mathbf{m}}^2+\frac{2}{3}\tilde{m}_\mu^2$. The solutions are of the form $\tilde{m}_\mu\in\{-\tilde{m},0,\tilde{m}\}$. If we denote with n the number of nonzero components in the vector $\tilde{\mathbf{m}}$, we derive from the above identities: $\tilde{m}_\mu=0$ or $\tilde{m}_\mu=\pm\sqrt{3}/\sqrt{3n-2}$. These saddle-points are called *mixture states*, since they correspond to microscopic configurations correlated equally with a finite number n of the stored patterns (or their negatives). Without loss of generality we can always perform gauge transformations on the set of stored patterns (permutations and reflections), such that the mixture states acquire the form

$$\mathbf{m} = m_n(\overbrace{1,\ldots,1}^{n\text{ times}},\overbrace{0,\ldots,0}^{p-n\text{ times}}),\quad m_n=\left[\frac{3}{3n-2}\right]^{1/2}(\beta-1)^{1/2}+\cdots \tag{52}$$

These states are in fact saddle-points of the surface $f(\boldsymbol{m})$ (51) for any finite temperature, as can be verified by substituting (52) as an *ansatz* into (50):

$$\begin{aligned} &\mu\leqslant n:\ m_n = \left\langle \xi_\mu\tanh\left[\beta m_n\sum_{\nu\leqslant n}\xi_\nu\right]\right\rangle_\xi, \\ &\mu>n:\ 0 = \left\langle \xi_\mu\tanh\left[\beta m_n\sum_{\nu\leqslant n}\xi_\nu\right]\right\rangle_\xi. \end{aligned}$$

The second equation is automatically satisfied since the average factorises. The first equation leads to a condition determining the amplitude m_n of the mixture states:

$$m_n = \left\langle \left[\frac{1}{n} \sum_{\mu \leqslant n} \xi_\mu \right] \tanh \left[\beta m_n \sum_{\nu \leqslant n} \xi_\nu \right] \right\rangle_\xi . \tag{53}$$

The corresponding values of $f(\mathbf{m})$, to be denoted by f_n, are

$$f_n = \frac{1}{2} n m_n^2 - \frac{1}{\beta} \left\langle \log 2 \cosh \left[\beta m_n \sum_{\nu \leqslant n} \xi_\nu \right] \right\rangle_\xi . \tag{54}$$

The relevant question at this stage is whether or not these saddle-points correspond to local minima of the surface $f(\mathbf{m})$ (51). The second derivative of $f(\mathbf{m})$ is given by

$$\frac{\partial^2 f(\mathbf{m})}{\partial m_\mu \partial m_\nu} = \delta_{\mu\nu} - \beta \langle \xi_\mu \xi_\nu [1 - \tanh^2[\beta \boldsymbol{\xi} \cdot \mathbf{m}]] \rangle_\xi \tag{55}$$

(a local minimum corresponds to a positive definite second derivative). In the trivial saddle-point $\mathbf{m} = 0$ this gives simply $\delta_{\mu\nu}(1 - \beta)$, so at $T = 1$ this state destabilizes. In a mixture state of the type (52) the second derivative becomes:

$$D_{\mu\nu}^{(n)} = \delta_{\mu\nu} - \beta \left\langle \xi_\mu \xi_\nu \left[1 - \tanh^2 \left[\beta m_n \sum_{\rho \leqslant n} \xi_\rho \right] \right] \right\rangle_\xi .$$

Due to the symmetries in the problem the spectrum of the matrix $D^{(n)}$ can be calculated. One finds the following eigenspaces, with $Q = \langle \tanh^2[\beta m_n \sum_{\rho \leqslant n} \xi_\rho] \rangle_\xi$ and $R = \langle \xi_1 \xi_2 \tanh^2[\beta m_n \sum_{\rho \leqslant n} \xi_\rho] \rangle_\xi$:

	Eigenspace	Eigenvalue
I	$: \mathbf{x} = (0, \ldots, 0, x_{n+1}, \ldots, x_p)$	$1 - \beta[1 - Q]$
II	$: \mathbf{x} = (1, \ldots, 1, 0, \ldots, 0)$	$1 - \beta[1 - Q + (1 - n)R]$
III	$: \mathbf{x} = (x_1, \ldots, x_n, 0, \ldots, 0), \quad \sum_\mu x_\mu = 0$	$1 - \beta[1 - Q + R]$

Eigenspace III and the quantity R only come into play for $n > 1$. To find the smallest eigenvalue we need to know the sign of R. With the abbreviation $M_\xi = \sum_{\rho \leqslant n} \xi_\rho$ we find:

$$\begin{aligned} n(n-1)R &= \langle M_\xi^2 \tanh^2[\beta m_n M_\xi] \rangle_\xi - n \langle \tanh^2[\beta m_n M_\xi] \rangle_\xi \\ &= \langle [M_\xi^2 - \langle M_{\xi'}^2 \rangle_{\xi'}] \tanh^2[\beta m_n |M_\xi|] \rangle_\xi \\ &= \left\langle \left[M_\xi^2 - \langle M_{\xi'}^2 \rangle_{\xi'} \right] \left\{ \tanh^2 \left[\beta m_n \sqrt{M_\xi^2} \right] - \tanh^2 \left[\beta m_n \sqrt{\langle M_{\xi'}^2 \rangle_{\xi'}} \right] \right\} \right\rangle_\xi \geqslant 0. \end{aligned}$$

We may now identify the conditions for an n-mixture state to be a local minimum of $f(\mathbf{m})$. For $n = 1$ the relevant eigenvalue is I, now the quantity Q simplifies consid-

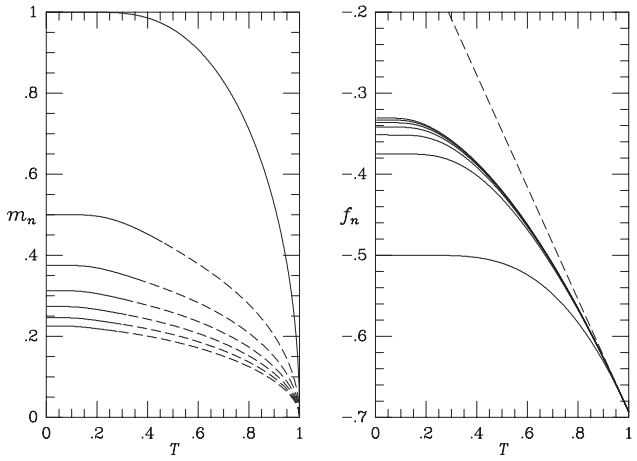

Fig. 7. Left picture: Amplitudes m_n of the mixture states as functions of temperature. From top to bottom: $n = 1, 3, 5, 7, 9, 11, 13$. Solid: region where they are stable (local minima of f). Dashed: region where they are unstable. Right picture: corresponding 'free energies' f_n. From bottom to top: $n = 1, 3, 5, 7, 9, 11, 13$. Dashed line: 'free energy' of the paramagnetic state $m = \mathbf{0}$ (for comparison).

erably. For $n > 1$ the relevant eigenvalue is *III*, here we can combine Q and R into one single average:

$$n = 1: \ 1 - \beta[1 - \tanh^2[\beta m_1]] > 0,$$
$$n = 2: \ 1 - \beta > 0,$$
$$n \geqslant 3: \ 1 - \beta\left[1 - \left\langle \tanh^2\left[\beta m_n \textstyle\sum_{\rho=3}^{n} \xi_\rho\right]\right\rangle_\xi\right] > 0.$$

The $n = 1$ states, correlated with one pattern only, are the desired solutions. They are stable for all $T < 1$, since partial differentiation with respect to β of the $n = 1$ amplitude Eq. (53) gives

$$m_1 = \tanh[\beta m_1] \to 1 - \beta[1 - \tanh^2[\beta m_1]] = m_1[1 - \tanh^2[\beta m_1]](\partial m_1/\partial\beta)^{-1}$$

(clearly $\mathrm{sgn}[m_1] = \mathrm{sgn}[\partial m_1/\partial\beta]$). The $n = 2$ mixtures are always unstable. For $n \geqslant 3$ we have to solve the amplitude Eq. (53) numerically to evaluate their stability. The result is shown in Fig. 7, together with the corresponding 'free energies' f_n (54). It turns out that only for odd n will there be a critical temperature below which the n-mixture states are local minima of $f(\mathbf{m})$. From Fig. 7 we can also conclude that, in terms of the network functioning as an associative memory, noise is actually beneficial in the sense that it can be used to eliminate the unwanted $n > 1$ ergodic components (while retaining the relevant ones: the pure $n = 1$ states). In fact the overlap equations (50) do also allow for stable solutions different from the n-mixture states discussed here. They are in turn found to be continuously bifurcating mixtures of the mixture states. However, for random (or uncorrelated) patterns they come

into existence only near $T = 0$ and play a marginal role; phase space is dominated by the odd n-mixture states.

We have now solved the model in equilibrium for finite p and $N \to \infty$. Most of the relevant information on when and to what extent stored random patterns will be recalled is summarized in Fig. 7. For nonrandom patterns one simply has to study the bifurcation properties of Eq. (50) for the new pattern statistics at hand; this is only qualitatively different from the random pattern analysis explained above. The occurrence of multiple saddle-points corresponding to local minima of the free energy signals ergodicity breaking. Although among these only the *global* minimum will correspond to the thermodynamic equilibrium state, the nonglobal minima correspond to true ergodic components, i.e. on finite time-scales they will be just as relevant as the global minimum.

4. Simple recurrent networks of coupled oscillators

4.1. Coupled oscillators with uniform synapses

Models with continuous variables involve integration over states, rather than summation. For a coupled oscillator network (13) with uniform synapses $J_{ij} = J/N$ and zero frequencies $\omega_i = 0$ (which is a simple version of the model in [14]) we obtain for the free energy per oscillator:

$$\lim_{N\to\infty} F/N = -\lim_{N\to\infty} \frac{1}{\beta N} \log \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} \mathrm{d}\boldsymbol{\phi} \times \exp\left((\beta J/2N) \left[\left[\sum_i \cos(\phi_i) \right]^2 + \left[\sum_i \sin(\phi_i) \right]^2 \right] \right).$$

We would now have to 'count' microscopic states with prescribed average cosines and sines. A faster route exploits auxiliary Gaussian integrals, via the identity

$$\mathrm{e}^{\frac{1}{2}y^2} = \int \mathrm{D}z\, \mathrm{e}^{yz} \tag{56}$$

with the shorthand $\mathrm{D}x = (2\pi)^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}x^2} \mathrm{d}x$ (this alternative would also have been open to us in the binary case; my aim in this section is to explain both methods):

$$\begin{aligned}
\lim_{N\to\infty} F/N &= -\lim_{N\to\infty} \frac{1}{\beta N} \log \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} \mathrm{d}\boldsymbol{\phi} \\
&\quad \times \int \mathrm{D}x\, \mathrm{D}y\, \exp \sqrt{\beta J/N} \left[x \sum_i \cos(\phi_i) + y \sum_i \sin(\phi_i) \right] \\
&= -\lim_{N\to\infty} \frac{1}{\beta N} \log \int \mathrm{D}x\, \mathrm{D}y \left[\int_{-\pi}^{\pi} \mathrm{d}\phi\, \mathrm{e}^{\cos(\phi)\sqrt{\beta J(x^2+y^2)/N}} \right]^N \\
&= -\lim_{N\to\infty} \frac{1}{\beta N} \log \int_0^{\infty} \mathrm{d}q\, q \mathrm{e}^{-\frac{1}{2}N\beta|J|q^2} \left[\int_{-\pi}^{\pi} \mathrm{d}\phi\, \mathrm{e}^{\beta|J|q\cos(\phi)\sqrt{\mathrm{sgn}(J)}} \right]^N,
\end{aligned}$$

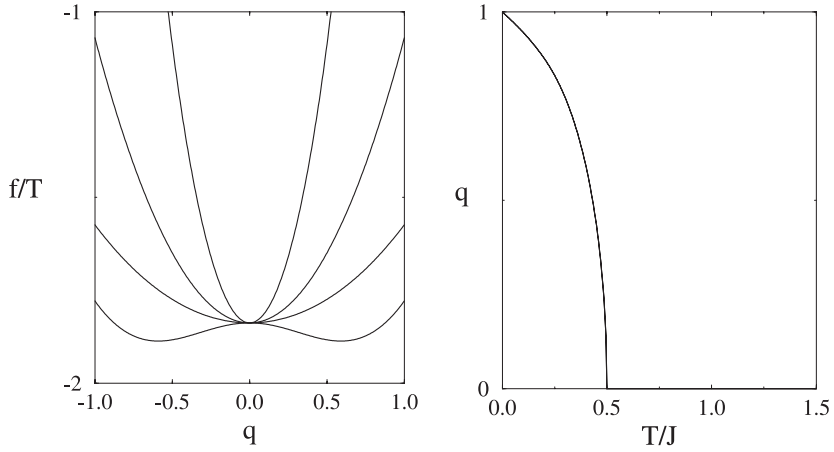

Fig. 8. The function $f(q)/T$ (left) for networks of coupled oscillators with uniform synapses $J_{ij} = J/N$, and for different choices of the re-scaled interaction strength J/T ($T = \beta^{-1}$): $J/T = -\frac{5}{2}, -1, 1, \frac{5}{2}$ (from top to bottom). The right picture gives, for $J > 0$, the location of the nonnegative minimum of $f(q)$ (which measures the overall degree of global synchronisation in thermal equilibrium) as a function of T/J. A transition to a synchronised state occurs at $T/J = \frac{1}{2}$.

where we have transformed to polar coordinates, $(x, y) = q\sqrt{\beta|J|N}(\cos\theta, \sin\theta)$, and where we have already eliminated (constant) terms which will not survive the limit $N \to \infty$. Thus, saddle-point integration gives us, quite similar to the previous cases (36) and (37):

$$\lim_{N\to\infty} F/N = \min_{q\geqslant 0} f(q) \qquad \begin{array}{l} J > 0: \ \beta f(q) = \frac{1}{2}\beta|J|q^2 - \log[2\pi I_0(\beta|J|q)] \\ J < 0: \ \beta f(q) = \frac{1}{2}\beta|J|q^2 - \log[2\pi I_0(i\beta|J|q)] \end{array} \tag{57}$$

in which the $I_n(z)$ are the modified Bessel functions (see e.g. [15]). The function $f(q)$ is shown in Fig. 8. The equations from which to solve the minima are obtained by differentiation, using $\frac{\mathrm{d}}{\mathrm{d}z} I_0(z) = I_1(z)$:

$$J > 0: \ q = \frac{I_1(\beta|J|q)}{I_0(\beta|J|q)}, \quad J < 0: \ q = i\frac{I_1(i\beta|J|q)}{I_0(i\beta|J|q)}. \tag{58}$$

Again, in both cases the problem has been reduced to studying a single nonlinear equation. The physical meaning of the solution follows from the identity $-2\partial F/\partial J = \langle N^{-1}\sum_{i\neq j}\cos(\phi_i - \phi_j)\rangle$:

$$\lim_{N\to\infty}\left\langle\left[\frac{1}{N}\sum_i \cos(\phi_i)\right]^2\right\rangle + \lim_{N\to\infty}\left\langle\left[\frac{1}{N}\sum_i \sin(\phi_i)\right]^2\right\rangle = \mathrm{sgn}(J)q^2.$$

From this equation it also follows that $q \leqslant 1$. Note: since $\partial f(q)/\partial q = 0$ at the minimum, one only needs to consider the explicit derivative of $f(q)$ with respect to J. If the synapses induce antisynchronization, $J < 0$, the only solution of (58) (and

the minimum in (57)) is the trivial state $q = 0$. This also follows immediately from the equation which gave the physical meaning of q. For synchronizing forces, $J > 0$, on the other hand, we again find the trivial solution at high noise levels, but a globally synchronized state with $q > 0$ at low noise levels. Here a phase transition occurs at $T = \frac{1}{2}J$ (a bifurcation of nontrivial solutions of (58)), and for $T < \frac{1}{2}J$ the minimum of (57) is found at two nonzero values for q. The critical noise level is again found upon expanding the saddle-point equation, using $I_0(z) = 1 + \mathcal{O}(z^2)$ and $I_1(z) = \frac{1}{2}z + \mathcal{O}(z^3)$: $q = \frac{1}{2}\beta J q + \mathcal{O}(q^3)$. Precisely at $\beta J = 2$ one finds a de-stabilization of the trivial solution $q = 0$, together with the creation of (two) stable nontrivial ones (see Fig. 8). Note that, in view of (57), we are only interested in nonnegative values of q. One can prove, using the properties of the Bessel functions, that there are no other (discontinuous) bifurcations of nontrivial solutions of the saddle-point equation. Note, finally, that the absence of a state with global antisynchronization for $J < 0$ has the same origin as the absence of an antiferromagnetic state for $J < 0$ in the previous models with binary neurons. Due to the long-range nature of the synapses $J_{ij} = J/N$ such states simply cannot exist: whereas any set of oscillators can be in a fully synchronized state, if two oscillators are in anti-synchrony it is already impossible for a third to be simultaneously in antisynchrony with the first two (since antisynchrony with one implies synchrony with the other).

4.2. Coupled oscillator attractor networks

4.2.1. Intuition and definitions

Let us now turn to an alternative realization of information storage in a recurrent network based upon the creation of attractors. We will solve models of coupled neural oscillators of the type (13), with zero natural frequencies (since we wish to use equilibrium techniques), in which real-valued patterns are stored as stable configurations of oscillator phases, following [16]. Let us, however, first find out how to store a single pattern $\boldsymbol{\xi} \in [-\pi, \pi]^N$ in a noise-less infinite-range oscillator network. For simplicity we will draw each component ξ_i independently at random from $[-\pi, \pi]$, with uniform probability density. This allows us to use asymptotic properties such as $|N^{-1}\sum_j e^{i\ell\xi_j}| = \mathcal{O}(N^{-\frac{1}{2}})$ for any integer ℓ. A sensible choice for the synapses would be $J_{ij} = \cos[\xi_i - \xi_j]$. To see this we work out the corresponding Lyapunov function (20):

$$L[\boldsymbol{\phi}] = -\frac{1}{2N^2}\sum_{ij}\cos[\xi_i - \xi_j]\cos[\phi_i - \phi_j],$$

$$L[\boldsymbol{\xi}] = -\frac{1}{2N^2}\sum_{ij}\cos^2[\xi_i - \xi_j] = -\frac{1}{4} + \mathcal{O}\left(N^{-\frac{1}{2}}\right)$$

(the factors of N have been inserted to achieve appropriate scaling in the $N \to \infty$ limit). The function $L[\boldsymbol{\phi}]$, which is obviously bounded from below, must decrease monotonically during the dynamics. To find out whether the state $\boldsymbol{\xi}$ is a stable fixed-point of the dynamics we have to calculate L and derivatives of L at $\boldsymbol{\phi} = \boldsymbol{\xi}$:

$$\left.\frac{\partial L}{\partial \phi_i}\right|_{\xi} = \frac{1}{2N^2}\sum_j \sin[2(\xi_i - \xi_j)],$$

$$\left.\frac{\partial^2 L}{\partial \phi_i^2}\right|_{\xi} = \frac{1}{N^2}\sum_j \cos^2[\xi_i - \xi_j], \qquad i \neq j : \left.\frac{\partial^2 L}{\partial \phi_i \partial \phi_j}\right|_{\xi} = -\frac{1}{N^2}\cos^2[\xi_i - \xi_j].$$

Clearly $\lim_{N\to\infty} L[\boldsymbol{\xi}] = -\frac{1}{4}$. Putting $\boldsymbol{\phi} = \boldsymbol{\xi} + \Delta\boldsymbol{\phi}$, with $\Delta\phi_i = \mathcal{O}(N^0)$, we find

$$\begin{aligned} L[\boldsymbol{\xi} + \Delta\boldsymbol{\phi}] - L[\boldsymbol{\xi}] &= \sum_i \Delta\phi_i \left.\frac{\partial L}{\partial \phi_i}\right|_{\xi} + \frac{1}{2}\sum_{ij} \Delta\phi_i \Delta\phi_j \left.\frac{\partial^2 L}{\partial \phi_i \partial \phi_j}\right|_{\xi} + \mathcal{O}(\Delta\boldsymbol{\phi}^3) \\ &= \frac{1}{4N}\sum_i \Delta\phi_i^2 - \frac{1}{2N^2}\sum_{ij} \Delta\phi_i \Delta\phi_j \cos^2[\xi_i - \xi_j] + \mathcal{O}(N^{-\frac{1}{2}}, \Delta\boldsymbol{\phi}^3) \\ &= \frac{1}{4}\left\{ \frac{1}{N}\sum_i \Delta\phi_i^2 - \left[\frac{1}{N}\sum_i \Delta\phi_i\right]^2 - \left[\frac{1}{N}\sum_i \Delta\phi_i \cos(2\xi_i)\right]^2 \right. \\ &\qquad \left. - \left[\frac{1}{N}\sum_i \Delta\phi_i \sin(2\xi_i)\right]^2 \right\} + \mathcal{O}(N^{-\frac{1}{2}}, \Delta\boldsymbol{\phi}^3). \end{aligned} \tag{59}$$

In leading order in N the following three vectors in $\Re^N$ are normalized and orthogonal:

$$\mathbf{e}_1 = \frac{1}{\sqrt{N}}(1, 1, \ldots, 1),$$

$$\mathbf{e}_2 = \frac{\sqrt{2}}{\sqrt{N}}(\cos(2\xi_1), \ldots, \cos(2\xi_N)),$$

$$\mathbf{e}_2 = \frac{\sqrt{2}}{\sqrt{N}}(\sin(2\xi_1), \ldots, \sin(2\xi_N)).$$

We may therefore use $\Delta\boldsymbol{\phi}^2 \geqslant (\Delta\boldsymbol{\phi}\cdot\mathbf{e}_1)^2 + (\Delta\boldsymbol{\phi}\cdot\mathbf{e}_2)^2 + (\Delta\boldsymbol{\phi}\cdot\mathbf{e}_3)^2$, insertion of which into (59) leads to

$$\begin{aligned} L[\boldsymbol{\xi} + \Delta\boldsymbol{\phi}] - L[\boldsymbol{\xi}] \;\geqslant\; & \left[\frac{1}{2N}\sum_i \Delta\phi_i \cos(2\xi_i)\right]^2 \\ & + \left[\frac{1}{2N}\sum_i \Delta\phi_i \sin(2\xi_i)\right]^2 + \mathcal{O}(N^{-\frac{1}{2}}, \Delta\boldsymbol{\phi}^3). \end{aligned}$$

Thus for large N the second derivative of L is nonnegative at $\boldsymbol{\phi} = \boldsymbol{\xi}$, and the phase pattern $\boldsymbol{\xi}$ has indeed become a fixed-point attractor of the dynamics of the noise-free coupled oscillator network. The same is found to be true for the states $\boldsymbol{\phi} = \pm\boldsymbol{\xi} + \alpha(1, \ldots, 1)$ (for any α).

4.2.2. Storing p phase patterns: equilibrium order parameter equations

We next follow the strategy of the Hopfield model and attempt to simply extend the above recipe for the synapses to the case of having a finite number p of phase patterns $\boldsymbol{\xi}^\mu = (\xi_1^\mu, \ldots, \xi_N^\mu) \in [-\pi, \pi]^N$, giving

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \cos[\xi_i^\mu - \xi_j^\mu] \tag{60}$$

(the factor N, as before, ensures a proper limit $N \to \infty$ later). In analogy with our solution of the Hopfield model we define the following averages over pattern variables:

$$\langle g[\boldsymbol{\xi}] \rangle_\xi = \lim_{N\to\infty} \frac{1}{N} \sum_i g[\boldsymbol{\xi}_i], \quad \boldsymbol{\xi}_i = (\xi_i^1, \ldots, \xi_i^p) \in [-\pi, \pi]^p.$$

We can write the Hamiltonian $H(\boldsymbol{\phi})$ of (34) in the form

$$\begin{aligned} H(\boldsymbol{\phi}) &= -\frac{1}{2N} \sum_{\mu=1}^{p} \sum_{ij} \cos[\xi_i^\mu - \xi_j^\mu] \cos[\phi_i - \phi_j] \\ &= -\frac{N}{2} \sum_{\mu=1}^{p} \left\{ m_{cc}^\mu(\boldsymbol{\phi})^2 + m_{cs}^\mu(\boldsymbol{\phi})^2 + m_{sc}^\mu(\boldsymbol{\phi})^2 + m_{ss}^\mu(\boldsymbol{\phi})^2 \right\} \end{aligned},$$

in which

$$m_{cc}^\mu(\boldsymbol{\phi}) = \frac{1}{N} \sum_i \cos(\xi_i^\mu) \cos(\phi_i), \quad m_{cs}^\mu(\boldsymbol{\phi}) = \frac{1}{N} \sum_i \cos(\xi_i^\mu) \sin(\phi_i), \tag{61}$$

$$m_{sc}^\mu(\boldsymbol{\phi}) = \frac{1}{N} \sum_i \sin(\xi_i^\mu) \cos(\phi_i), \quad m_{ss}^\mu(\boldsymbol{\phi}) = \frac{1}{N} \sum_i \sin(\xi_i^\mu) \sin(\phi_i). \tag{62}$$

The free energy per oscillator can now be written as

$$F/N = -\frac{1}{\beta N} \log \int \cdots \int \mathrm{d}\boldsymbol{\phi}\, \mathrm{e}^{-\beta H(\boldsymbol{\phi})} = -\frac{1}{\beta N} \log \int \cdots \int \mathrm{d}\boldsymbol{\phi}\, \mathrm{e}^{\frac{1}{2}\beta N \sum_\mu \sum_{\star\star} m_{\star\star}^\mu(\boldsymbol{\phi})^2}$$

with $\star\star \in \{cc, ss, cs, sc\}$. Upon introducing the notation $\mathbf{m}_{\star\star} = (m_{\star\star}^1, \ldots, m_{\star\star}^p)$ we can again express the free energy in terms of the density of states $\mathscr{D}(\{\mathbf{m}_{\star\star}\}) = (2\pi)^{-N} \int \cdots \int \mathrm{d}\boldsymbol{\phi} \prod_{\star\star} \delta[\mathbf{m}_{\star\star} - \mathbf{m}_{\star\star}(\boldsymbol{\phi})]$:

$$F/N = -\frac{1}{\beta} \log(2\pi) - \frac{1}{\beta N} \log \int \prod_{\star\star} \mathrm{d}\mathbf{m}_{\star\star}\, \mathscr{D}(\{\mathbf{m}_{\star\star}\}) \mathrm{e}^{\frac{1}{2}\beta N \sum_{\star\star} \mathbf{m}_{\star\star}^2}. \tag{63}$$

Since p is finite, the leading contribution to the density of states (as $N \to \infty$), which will give us the entropy, can be calculated by writing the δ-functions in integral representation:

$$\lim_{N\to\infty}\frac{1}{N}\log\mathscr{D}(\{\mathbf{m}_{\star\star}\})$$
$$=\lim_{N\to\infty}\frac{1}{N}\log\int\prod_{\star\star}\Big[\mathrm{d}\mathbf{x}_{\star\star}\mathrm{e}^{\mathrm{i}N\boldsymbol{x}_{\star\star}\cdot\mathbf{m}_{\star\star}}\Big]\times\int\ldots\int\frac{\mathrm{d}\boldsymbol{\phi}}{(2\pi)^N}$$
$$\times\exp\Bigg(-i\sum_i\sum_\mu\big[x^\mu_{cc}\cos(\xi^\mu_i)\cos(\phi_i)+x^\mu_{cs}\cos(\xi^\mu_i)\sin(\phi_i)$$
$$+x^\mu_{sc}\sin(\xi^\mu_i)\cos(\phi_i)+x^\mu_{ss}\sin(\xi^\mu_i)\sin(\phi_i)\big]\Bigg)$$
$$=\mathrm{extr}_{\{\mathbf{x}_{\star\star}\}}\Bigg\{\mathrm{i}\sum_{\star\star}\mathbf{x}_{\star\star}\cdot\mathbf{m}_{\star\star}+\Bigg\langle\log\int\frac{\mathrm{d}\phi}{2\pi}\exp\Bigg(-\mathrm{i}\sum_\mu[x^\mu_{cc}\cos(\xi_\mu)\cos(\phi)$$
$$+x^\mu_{cs}\cos(\xi_\mu)\sin(\phi)+x^\mu_{sc}\sin(\xi_\mu)\cos(\phi)$$
$$+x^\mu_{ss}\sin(\xi_\mu)\sin(\phi)]\Bigg)\Bigg\rangle_\xi\Bigg\}$$

The relevant extremum is purely imaginary so we put $\boldsymbol{x}_{\star\star}=\mathrm{i}\beta\mathbf{y}_{\star\star}$ (see also our previous discussion for the Hopfield model) and, upon inserting the density of states into our original expression for the free energy per oscillator, arrive at

$$\lim_{N\to\infty}F/N=\mathrm{extr}_{\{\mathbf{m}_{\star\star},\boldsymbol{y}_{\star\star}\}}f(\{\mathbf{m}_{\star\star},\boldsymbol{y}_{\star\star}\})$$
$$f(\{\mathbf{m}_{\star\star},\mathbf{y}_{\star\star}\})=-\frac{1}{\beta}\log(2\pi)-\frac{1}{2}\sum_{\star\star}\mathbf{m}^2_{\star\star}+\sum_{\star\star}\mathbf{y}_{\star\star}\cdot\mathbf{m}_{\star\star}$$
$$-\frac{1}{\beta}\Bigg\langle\log\int\frac{\mathrm{d}\phi}{2\pi}\exp\Bigg(\beta\sum_\mu[y^\mu_{cc}\cos(\xi_\mu)\cos(\phi)+y^\mu_{cs}\cos(\xi_\mu)\sin(\phi)$$
$$+y^\mu_{sc}\sin(\xi_\mu)\cos(\phi)+y^\mu_{ss}\sin(\xi_\mu)\sin(\phi)]\Bigg)\Bigg\rangle_\xi.$$

Taking derivatives with respect to the order parameters $\mathbf{m}_{\star\star}$ gives us $\mathbf{y}_{\star\star}=\mathbf{m}_{\star\star}$, with which we can eliminate the $\mathbf{y}_{\star\star}$. Derivation with respect to the $\mathbf{m}_{\star\star}$ subsequently gives the saddle-point equations

$$m^\mu_{cc}=\Big\langle\cos[\xi_\mu]\frac{\int\mathrm{d}\phi\cos[\phi]\exp(\beta\cos[\phi]\sum_\nu[m^\nu_{cc}\cos[\xi_\nu]+m^\nu_{sc}\sin[\xi_\nu]]+\beta\sin[\phi]\sum_\nu[m^\nu_{cs}\cos[\xi_\nu]+m^\nu_{ss}\sin[\xi_\nu]])}{\int\mathrm{d}\phi\exp(\beta\cos[\phi]\sum_\nu[m^\nu_{cc}\cos[\xi_\nu]+m^\nu_{sc}\sin[\xi_\nu]]+\beta\sin[\phi]\sum_\nu[m^\nu_{cs}\cos[\xi_\nu]+m^\nu_{ss}\sin[\xi_\nu]])}\Big\rangle_\xi,\tag{64}$$

$$m^\mu_{cs}=\Big\langle\cos[\xi_\mu]\frac{\int\mathrm{d}\phi\sin[\phi]\exp(\beta\cos[\phi]\sum_\nu[m^\nu_{cc}\cos[\xi_\nu]+m^\nu_{sc}\sin[\xi_\nu]]+\beta\sin[\phi]\sum_\nu[m^\nu_{cs}\cos[\xi_\nu]+m^\nu_{ss}\sin[\xi_\nu]])}{\int\mathrm{d}\phi\exp(\beta\cos[\phi]\sum_\nu[m^\nu_{cc}\cos[\xi_\nu]+m^\nu_{sc}\sin[\xi_\nu]]+\beta\sin[\phi]\sum_\nu[m^\nu_{cs}\cos[\xi_\nu]+m^\nu_{ss}\sin[\xi_\nu]])}\Big\rangle_\xi,\tag{65}$$

$$m_{sc}^{\mu} = \langle \sin[\xi_\mu] \frac{\int \mathrm{d}\phi \cos[\phi] \exp(\beta \cos[\phi] \sum_\nu [m_{cc}^\nu \cos[\xi_\nu] + m_{sc}^\nu \sin[\xi_\nu]] + \beta \sin[\phi] \sum_\nu [m_{cs}^\nu \cos[\xi_\nu] + m_{ss}^\nu \sin[\xi_\nu]])}{\int \mathrm{d}\phi \exp(\beta \cos[\phi] \sum_\nu [m_{cc}^\nu \cos[\xi_\nu] + m_{sc}^\nu \sin[\xi_\nu]] + \beta \sin[\phi] \sum_\nu [m_{cs}^\nu \cos[\xi_\nu] + m_{ss}^\nu \sin[\xi_\nu]])} \rangle_\xi, \tag{66}$$

$$m_{ss}^{\mu} = \langle \sin[\xi_\mu] \frac{\int \mathrm{d}\phi \sin[\phi] \exp(\beta \cos[\phi] \sum_\nu [m_{cc}^\nu \cos[\xi_\nu] + m_{sc}^\nu \sin[\xi_\nu]] + \beta \sin[\phi] \sum_\nu [m_{cs}^\nu \cos[\xi_\nu] + m_{ss}^\nu \sin[\xi_\nu]])}{\int \mathrm{d}\phi \exp(\beta \cos[\phi] \sum_\nu [m_{cc}^\nu \cos[\xi_\nu] + m_{sc}^\nu \sin[\xi_\nu]] + \beta \sin[\phi] \sum_\nu [m_{cs}^\nu \cos[\xi_\nu] + m_{ss}^\nu \sin[\xi_\nu]])} \rangle_\xi. \tag{67}$$

The equilibrium values of the observables $\mathbf{m}_{\star\star}$, as defined in (61) and (62), are now given by the solution of the coupled equations (64)–(67)which minimizes

$$f(\{\mathbf{m}_{\star\star}\}) = \frac{1}{2} \sum_{\star\star} \mathbf{m}_{\star\star}^2 - \frac{1}{\beta} \left\langle \log \int \mathrm{d}\phi \exp \left(\beta \cos[\phi] \sum_\nu \left[m_{cc}^\nu \cos[\xi_\nu] + m_{sc}^\nu \sin[\xi_\nu] \right] + \beta \sin[\phi] \sum_\nu [m_{cs}^\nu \cos[\xi_\nu] + m_{ss}^\nu \sin[\xi_\nu]] \right) \right\rangle_\xi . \tag{68}$$

We can confirm that the relevant saddle-point must be a minimum by inspecting the $\beta = 0$ limit (infinite noise levels): $\lim_{\beta \to 0} f(\{\mathbf{m}_{\star\star}\}) = \frac{1}{2} \sum_{\star\star} \mathbf{m}_{\star\star}^2 - \frac{1}{\beta} \log(2\pi)$.

4.2.3. Analysis of order parameter equations: pure states

From now on we will restrict our analysis to phase pattern components ξ_i^μ which have all been drawn independently at random from $[-\pi, \pi]$, with uniform probability density, so that $\langle g[\boldsymbol{\xi}] \rangle_\xi = (2\pi)^{-p} \int_{-\pi}^{\pi} \ldots \int_{-\pi}^{\pi} \mathrm{d}\boldsymbol{\xi}\ \ g[\boldsymbol{\xi}]$. At $\beta = 0$ ($T = \infty$) one finds only the trivial state $m_{\star\star}^\mu = 0$. It can be shown that there will be no discontinuous transitions to a nontrivial state as the noise level (temperature) is reduced. The continuous ones follow upon expansion of the equations (64)–(67) for small $\{\mathbf{m}_{\star\star}\}$,which is found to give (for each μ and each combination $\star\star$):

$$m_{\star\star}^\mu = \frac{1}{4} \beta m_{\star\star}^\mu + \mathcal{O}(\{\mathbf{m}_{\star\star}^2\}).$$

Thus a continuous transition to recall states occurs at $T = \frac{1}{4}$. Full classification of all solutions of (64)–(67) is ruled out. Here we will restrict ourselves to the most relevant ones, such as the pure states, where $m_{\star\star}^\mu = m_{\star\star} \delta_{\mu\lambda}$ (for some pattern label λ). Here the oscillator phases are correlated with only one of the stored phase patterns (if at all). Insertion into the above expression for $f(\{\mathbf{m}_{\star\star}\})$ shows that for such solutions we have to minimize

$$f(\{m_{\star\star}\}) = \frac{1}{2} \sum_{\star\star} m_{\star\star}^2 - \frac{1}{\beta} \int \frac{\mathrm{d}\xi}{2\pi} \log \int \mathrm{d}\phi \exp(\beta \cos[\phi][m_{cc} \cos[\xi] + m_{sc} \sin[\xi]] + \beta \sin[\phi][m_{cs} \cos[\xi] + m_{ss} \sin[\xi]]). \tag{69}$$

We anticipate solutions corresponding to the (partial) recall of the stored phase pattern $\boldsymbol{\xi}^\lambda$ or its mirror image (modulo overall phase shifts $\xi_i \to \xi_i + \delta$, under which the synapses are obviously invariant). Insertion into (64)–(67) of the state $\phi_i = \xi_i^\lambda + \delta$ gives $(m_{cc}, m_{sc}, m_{cs}, m_{ss}) = \frac{1}{2}(\cos\delta, -\sin\delta, \sin\delta, \cos\delta)$. Similarly, insertion into (64)–(67) of $\phi_i = -\xi_i^\lambda + \delta$ gives $(m_{cc}, m_{sc}, m_{cs}, m_{ss}) = \frac{1}{2}(\cos\delta, \sin\delta, \sin\delta, -\cos\delta)$. Thus we can identify retrieval states as those solutions which are of the form

$$\begin{aligned} &(i)\ \text{retrieval of } \boldsymbol{\xi}^\lambda: && (m_{cc}, m_{sc}, m_{cs}, m_{ss}) = m(\cos\delta, -\sin\delta, \sin\delta, \cos\delta) \\ &(ii)\ \text{retrieval of } -\boldsymbol{\xi}^\lambda: && (m_{cc}, m_{sc}, m_{cs}, m_{ss}) = m(\cos\delta, \sin\delta, \sin\delta, -\cos\delta) \end{aligned}$$

with full recall corresponding to $m = \frac{1}{2}$. Insertion into the saddle-point equations and into (69), followed by an appropriate shift of the integration variable ϕ, shows that the free energy is independent of δ (so the above two ansätze solve the saddle-point equations for any δ) and that

$$m = \frac{1}{2}\frac{\int \mathrm{d}\phi \cos[\phi] \mathrm{e}^{\beta m \cos[\phi]}}{\int \mathrm{d}\phi\, \mathrm{e}^{\beta m \cos[\phi]}}, \quad f(m) = m^2 - \frac{1}{\beta}\log \int \mathrm{d}\phi\, \mathrm{e}^{\beta m \cos[\phi]}.$$

Expansion in powers of m, using $\log(1+z) = z - \frac{1}{2}z^2 + \mathcal{O}(z^3)$, reveals that nonzero minima m indeed bifurcate continuously at $T = \beta^{-1} = \frac{1}{4}$:

$$f(m) + \frac{1}{\beta}\log[2\pi] = \left(1 - \frac{1}{4}\beta\right)m^2 + \frac{1}{64}\beta^3 m^4 + \mathcal{O}(m^6). \tag{70}$$

Retrieval states are obviously not the only pure states that solve the saddle-point equations. The function (69) is invariant under the following discrete (noncommuting) transformations:

$$\begin{aligned} &\text{I:}\ (m_{cc}, m_{sc}, m_{cs}, m_{ss}) \to (m_{cc}, m_{sc}, -m_{cs}, -m_{ss}), \\ &\text{II:}\ (m_{cc}, m_{sc}, m_{cs}, m_{ss}) \to (m_{cs}, m_{ss}, m_{cc}, m_{sc}). \end{aligned}$$

We expect these to induce solutions with specific symmetries. In particular we anticipate the following symmetric and antisymmetric states:

$$\begin{aligned} &(iii)\ \text{symmetric under I:} && (m_{cc}, m_{sc}, m_{cs}, m_{ss}) = \sqrt{2}\, m(\cos\delta, \sin\delta, 0, 0), \\ &(iv)\ \text{antisymmetric under I:} && (m_{cc}, m_{sc}, m_{cs}, m_{ss}) = \sqrt{2}\, m(0, 0, \cos\delta, \sin\delta), \\ &(v)\ \text{symmetric under II:} && (m_{cc}, m_{sc}, m_{cs}, m_{ss}) = m(\cos\delta, \sin\delta, \cos\delta, \sin\delta), \\ &(vi)\ \text{antisymmetric under II:} && (m_{cc}, m_{sc}, m_{cs}, m_{ss}) = m(\cos\delta, \sin\delta, -\cos\delta, -\sin\delta). \end{aligned}$$

Insertion into the saddle-point equations and into (69) shows in all four cases the parameter δ is arbitrary and that always

$$m = \frac{1}{\sqrt{2}}\int \frac{\mathrm{d}\xi}{2\pi}\cos[\xi]\frac{\int \mathrm{d}\phi \cos[\phi] \mathrm{e}^{\beta m\sqrt{2}\cos[\phi]\cos[\xi]}}{\int \mathrm{d}\phi\ \ \mathrm{e}^{\beta m\sqrt{2}\cos[\phi]\cos[\xi]}},$$

$$f(m) = m^2 - \frac{1}{\beta}\int \frac{\mathrm{d}\xi}{2\pi}\log \int \mathrm{d}\phi\ \ \mathrm{e}^{\beta m\sqrt{2}\cos[\phi]\cos[\xi]}.$$

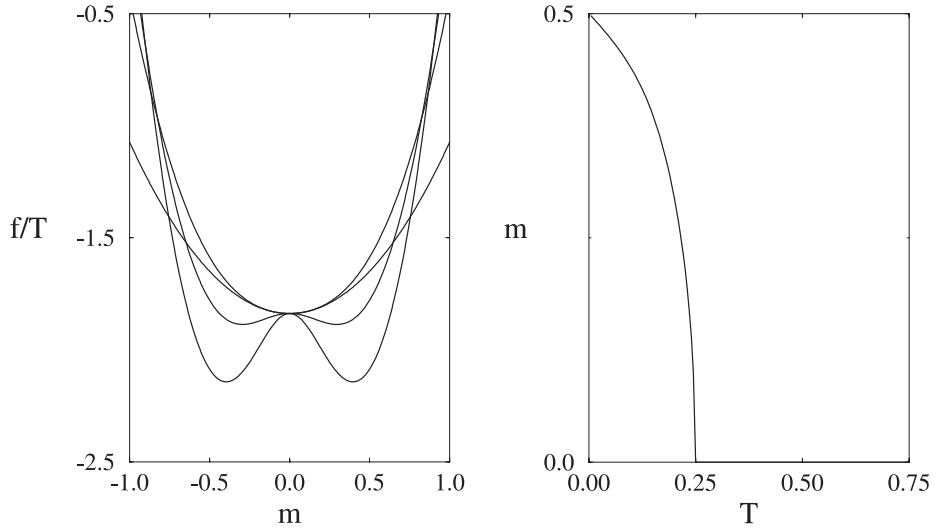

Fig. 9. The function $f(m)/T$ (left) for networks of coupled oscillators with phase patterns stored via the synapses $J_{ij} = N^{-1} \sum_{\mu} \cos[\xi_i^{\mu} - \xi_j^{\mu}]$, and for different choices of $\beta = T^{-1}$: $\beta = 1, 3, 5, 7$ (from top to bottom). The right picture gives the location of the nonnegative minimum of $f(m)$ (which measures the overall degree of global synchronization with one recalled phase pattern in thermal equilibrium) as a function of T. A transition to a recall state occurs at $T = \frac{1}{4}$.

Expansion in powers of m reveals that nonzero solutions m here again bifurcate continuously at $T = \frac{1}{4}$:

$$f(m) + \frac{1}{\beta} \log[2\pi] = \left(1 - \frac{1}{4}\beta\right) m^2 + \frac{3}{2} \frac{1}{64} \beta^3 m^4 + \mathrm{O}(m^6). \tag{71}$$

However, comparison with (70) shows that the free energy of the pure recall states is lower. Thus the system will prefer the recall states over the above solutions with specific symmetries.

Note, finally, that the free energy and the order parameter equation for the pure recall states can be written in terms of modified Bessel functions as follows:

$$m = \frac{1}{2} \frac{I_1(\beta m)}{I_0(\beta m)}, \quad f(m) = m^2 - \frac{1}{\beta} \log[2\pi I_0(\beta m)].$$

The behavior of these equations and the observable m for different noise levels are shown in Fig. 9. One easily proves that $|m| \leqslant \frac{1}{2}$, and that $\lim_{\beta \to \infty} m = \frac{1}{2}$. Following the transition to a state with partial recall of a stored phase pattern at $T = \frac{1}{4}$, further reduction of the noise level T gives a monotonic increase of retrieval quality until retrieval is perfect at $T = 0$.

5. Networks with Gaussian distributed synapses

The type of analysis presented so far to deal with attractor networks breaks down if the number of patterns stored p no longer remains finite for $N \to \infty$, but scales as $p = \alpha N$ ($\alpha > 0$). Expressions such as (48) and (49) can no longer be evaluated by saddle-point methods, since the dimension of the integral diverges at the same time as the exponent of the integrand. The number of local minima (ergodic components) of Hamiltonians such as (30) and (32) will diverge and we will encounter phenomena reminiscent of complex disordered magnetic systems, i.e. spin glasses. As a consequence we will need corresponding methods of analysis, in the present case: replica theory.

5.1. Replica analysis

5.1.1. Replica calculation of the disorder-averaged free energy

As an introduction to the replica technique we will first discuss the equilibrium solution of a recurrent neural network model with binary neurons $\sigma_i \in \{-1, 1\}$ in which a single pattern $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N) \in \{-1, 1\}^N$ has been stored (via a Hebbian-type recipe) on a background of zero-average Gaussian synapses (equivalent to the SK model, [17]):

$$J_{ij} = \frac{J_0}{N}\xi_i\xi_j + \frac{J}{\sqrt{N}}z_{ij}, \quad \bar{z}_{ij} = 0, \ \overline{z_{ij}^2} = 1, \tag{72}$$

in which $J_0 > 0$ measures the embedding strength of the pattern, and the z_{ij} ($i < j$) are independent Gaussian random variables. We denote averaging over their distribution by $\overline{\cdots}$ (the factors in (72) involving N ensure appropriate scaling and statistical relevance of the two terms, and as always $J_{ii} = 0$). Here the Hamiltonian H (30), corresponding to sequential dynamics (2), becomes

$$H(\boldsymbol{\sigma}) = -\frac{1}{2}NJ_0m^2(\boldsymbol{\sigma}) + \frac{1}{2}J_0 - \frac{J}{\sqrt{N}}\sum_{i<j}\sigma_i\sigma_j z_{ij} \tag{73}$$

with the overlap $m(\boldsymbol{\sigma}) = \frac{1}{N}\sum_k \sigma_k\xi_k$ which measures pattern recall quality. We clearly cannot calculate the free energy for every given realization of the synapses, furthermore it is to be expected that for $N \to \infty$ macroscopic observables like the free energy per neuron and the overlap m only depend on the statistics of the synapses, not on their specific values. We therefore average the free energy over the disorder distribution and concentrate on

$$\bar{F} = -\frac{1}{\beta}\lim_{N\to\infty}\overline{\log Z}, \quad Z = \sum_{\boldsymbol{\sigma}} e^{-\beta H(\boldsymbol{\sigma})}. \tag{74}$$

The disorder average is transformed into an average of powers of Z, with the identity

$$\overline{\log Z} = \lim_{n\to 0} \frac{1}{n}\left[\overline{Z^n} - 1\right] \quad \text{or, equivalently,} \quad \overline{\log Z} = \lim_{n\to 0} \frac{1}{n} \log \overline{Z^n}. \tag{75}$$

The so-called 'replica trick' consists in evaluating the averages $\overline{Z^n}$ for integer values of n, and taking the limit $n \to 0$ afterwards, under the assumption that the resulting expression is correct for noninteger values of n as well. The integer powers of Z are written as a product of terms, each of which can be interpreted as an equivalent copy, or 'replica' of the original system. The disorder-averaged free energy now becomes

$$\bar{F} = -\lim_{n\to 0} \frac{1}{\beta n} \log \overline{Z^n} = -\lim_{n\to 0} \frac{1}{\beta n} \log \sum_{\boldsymbol{\sigma}^1 \ldots \boldsymbol{\sigma}^n} \overline{\mathrm{e}^{-\beta \sum_{\alpha=1}^{n} H(\boldsymbol{\sigma}^\alpha)}}.$$

From now Roman indices will refer to sites, i.e. $i = 1 \ldots N$, whereas Greek indices will refer to replicas, i.e. $\alpha = 1 \ldots n$. We introduce a shorthand for the Gaussian measure, $\mathrm{D}z = (2\pi)^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}z^2} \mathrm{d}z$, and we will repeatedly use the identity $\int \mathrm{D}z \;\; \mathrm{e}^{xz} = \mathrm{e}^{\frac{1}{2}x^2}$. Upon insertion of the Hamiltonian (73) we obtain

$$\begin{aligned}
\bar{F} &= -\frac{1}{\beta} N \log 2 - \lim_{n\to 0} (\beta n)^{-1} \log \left\langle \mathrm{e}^{\frac{\beta J_0}{N} \sum_{i<j} \xi_i \xi_j \sum_\alpha \sigma_i^\alpha \sigma_j^\alpha} \prod_{i<j} \left[\int \mathrm{D}z \;\; \mathrm{e}^{\frac{\beta J z}{\sqrt{N}} \sum_\alpha \sigma_i^\alpha \sigma_j^\alpha} \right] \right\rangle_{\{\boldsymbol{\sigma}^\alpha\}} \\
&= -\frac{1}{\beta} N \log 2 - \lim_{n\to 0} (\beta n)^{-1} \\
&\quad \times \log \left\langle \exp\left(\frac{\beta J_0}{2N} \sum_\alpha \sum_{i\neq j} \xi_i \xi_j \sigma_i^\alpha \sigma_j^\alpha + \frac{\beta^2 J^2}{4N} \sum_{\alpha\gamma} \sum_{i\neq j} \sigma_i^\alpha \sigma_j^\alpha \sigma_i^\gamma \sigma_j^\gamma \right) \right\rangle_{\{\boldsymbol{\sigma}^\alpha\}}.
\end{aligned}$$

We now complete the sums over sites in this expression,

$$\sum_{i\neq j} \sigma_i^\alpha \sigma_j^\alpha = \left[\sum_i \sigma_i^\alpha \right]^2 - N, \qquad \sum_{i\neq j} \sigma_i^\alpha \sigma_j^\alpha \sigma_i^\gamma \sigma_j^\gamma = \left[\sum_i \sigma_i^\alpha \sigma_i^\gamma \right]^2 - N.$$

The averaging over the neuron states $\{\boldsymbol{\sigma}^\alpha\}$ in our expression for $\bar{F}$ will now factorize nicely if we insert appropriate δ-functions (in their integral representations) to isolate the relevant terms, using

$$1 = \int \mathrm{d}\mathbf{q} \prod_{\alpha\beta} \delta\left[q_{\alpha\beta} - \frac{1}{N} \sum_i \sigma_i^\alpha \sigma_i^\beta \right] = \left[\frac{N}{2\pi} \right]^{n^2} \int \mathrm{d}\mathbf{q} \, \mathrm{d}\hat{\mathbf{q}} \;\; \mathrm{e}^{\mathrm{i}N \sum_{\alpha\beta} \hat{q}_{\alpha\beta} \left[q_{\alpha\beta} - \frac{1}{N} \sum_i \sigma_i^\alpha \sigma_i^\beta \right]},$$

$$1 = \int \mathrm{d}\mathbf{m} \prod_{\alpha} \delta\left[m_\alpha - \frac{1}{N} \sum_i \xi_i \sigma_i^\alpha \right] = \left[\frac{N}{2\pi} \right]^{n} \int \mathrm{d}\mathbf{m} \, \mathrm{d}\hat{\mathbf{m}} \;\; \mathrm{e}^{\mathrm{i}N \sum_\alpha \hat{m}_\alpha \left[m_\alpha - \frac{1}{N} \sum_i \xi_i \sigma_i^\alpha \right]}.$$

The integrations are over the $n \times n$ matrices $\mathbf{q}$ and $\hat{\mathbf{q}}$ and over the n-vectors $\boldsymbol{m}$ and $\hat{\mathbf{m}}$. After inserting these integrals we obtain

$$\lim_{N\to\infty}\bar{F}/N=-\frac{1}{\beta}\log 2-\lim_{N\to\infty}\lim_{n\to 0}\frac{1}{\beta Nn}$$
$$\times\log\left\{\left[\frac{N}{2\pi}\right]^{n^2+n}\int \mathrm{d}\mathbf{q}\,\mathrm{d}\hat{\mathbf{q}}\,\mathrm{d}\mathbf{m}\,\mathrm{d}\hat{\mathbf{m}}\,\mathrm{e}^{-\frac{1}{2}n\beta J_0-\frac{1}{4}n^2\beta^2J^2}\right.$$
$$\times\exp\left(N\left[i\sum_{\alpha\gamma}\hat{q}_{\alpha\gamma}q_{\alpha\gamma}+i\sum_{\alpha}\hat{m}_\alpha m_\alpha+\frac{1}{2}\beta J_0\sum_\alpha m_\alpha^2+\frac{1}{4}\beta^2J^2\sum_{\alpha\gamma}q_{\alpha\gamma}^2\right]\right)$$
$$\left.\times\left\langle\exp\left(-i\sum_i\left[\sum_{\alpha\gamma}\hat{q}_{\alpha\gamma}\sigma_i^\alpha\sigma_i^\gamma+\sum_\alpha\hat{m}_\alpha\xi_i\sigma_i^\alpha\right]\right)\right\rangle_{\{\boldsymbol{\sigma}^\alpha\}}\right\}.$$

The neuronal averages factorize and are therefore reduced to single-site ones. A simple transformation $\sigma_i\to\xi_i\sigma_i$ for all i eliminates the pattern components ξ_i from our equations, and the remaining averages involve only one n-replicated neuron $(\sigma_1,\ldots,\sigma_n)$. Finally one assumes that the two limits $n\to 0$ and $N\to\infty$ commute. This allows us to evaluate the integral with the steepest-descent method:

$$\lim_{N\to\infty}\lim_{n\to 0}\frac{1}{Nn}\log\int \mathrm{d}\mathbf{x}\,\mathrm{e}^{N\Phi(\mathbf{x})}=\lim_{n\to 0}\lim_{N\to\infty}\frac{1}{Nn}\log\mathrm{e}^{N\,\mathrm{extr}\,\Phi+\cdots}=\lim_{n\to 0}\frac{1}{n}\,\mathrm{extr}\,\Phi. \tag{76}$$

The result of these manipulations is

$$\lim_{N\to\infty}\bar{F}/N=\lim_{n\to 0}\mathrm{extr}\,f(\mathbf{q},\mathbf{m};\hat{\mathbf{q}},\hat{\mathbf{m}}) \tag{77}$$

$$f(\mathbf{q},\mathbf{m};\hat{\mathbf{q}},\hat{\mathbf{m}})=-\frac{1}{\beta}\log 2-\frac{1}{\beta n}\left[\log\langle\mathrm{e}^{-\mathrm{i}\sum_{\alpha\gamma}\hat{q}_{\alpha\gamma}\sigma_\alpha\sigma_\gamma-i\sum_\alpha\hat{m}_\alpha\sigma_\alpha}\rangle_{\boldsymbol{\sigma}}\right.$$
$$\left.+i\sum_{\alpha\gamma}\hat{q}_{\alpha\gamma}q_{\alpha\gamma}+i\sum_\alpha\hat{m}_\alpha m_\alpha+\frac{1}{2}\beta J_0\sum_\alpha m_\alpha^2+\frac{1}{4}\beta^2J^2\sum_{\alpha\gamma}q_{\alpha\gamma}^2\right]. \tag{78}$$

Variation of the parameters $\{q_{\alpha\beta}\}$ and $\{m_\alpha\}$ allows us to eliminate immediately the conjugate parameters $\{\hat{q}_{\alpha\beta}\}$ and $\{\hat{m}_\alpha\}$, since it leads to the saddle-point requirements

$$\hat{q}_{\alpha\beta}=\frac{1}{2}i\beta^2J^2q_{\alpha\beta},\qquad \hat{m}_\alpha=i\beta J_0m_\alpha. \tag{79}$$

Upon elimination of $\{\hat{q}_{\alpha\beta},\hat{m}_\alpha\}$ according to (79) the result (77) and (78) is simplified to

$$\lim_{N\to\infty}\bar{F}/N=\lim_{n\to 0}\mathrm{extr}\,f(\mathbf{q},\mathbf{m}), \tag{80}$$

$$f(\mathbf{q},\mathbf{m})=-\frac{1}{\beta}\log 2+\frac{\beta J^2}{4n}\sum_{\alpha\gamma}q_{\alpha\gamma}^2+\frac{J_0}{2n}\sum_\alpha m_\alpha^2$$
$$-\frac{1}{\beta n}\log\left\langle\mathrm{e}^{\frac{1}{2}\beta^2J^2\sum_{\alpha\gamma}q_{\alpha\gamma}\sigma_\alpha\sigma_\gamma+\beta J_0\sum_\alpha m_\alpha\sigma_\alpha}\right\rangle_{\boldsymbol{\sigma}}. \tag{81}$$

Variation of the remaining parameters $\{q_{\alpha\beta}\}$ and $\{m_\alpha\}$ gives the final saddle-point equations

$$q_{\lambda\rho} = \frac{\left\langle \sigma_\lambda \sigma_\rho \exp\left(\frac{1}{2}\beta^2 J^2 \sum_{\alpha\gamma} q_{\alpha\gamma}\sigma_\alpha\sigma_\gamma + \beta J_0 \sum_\alpha m_\alpha \sigma_\alpha\right)\right\rangle_{\boldsymbol{\sigma}}}{\left\langle \exp\left(\frac{1}{2}\beta^2 J^2 \sum_{\alpha\gamma} q_{\alpha\gamma}\sigma_\alpha\sigma_\gamma + \beta J_0 \sum_\alpha m_\alpha \sigma_\alpha\right)\right\rangle_{\boldsymbol{\sigma}}}, \tag{82}$$

$$m_\lambda = \frac{\left\langle \sigma_\lambda \exp\left(\frac{1}{2}\beta^2 J^2 \sum_{\alpha\gamma} q_{\alpha\gamma}\sigma_\alpha\sigma_\gamma + \beta J_0 \sum_\alpha m_\alpha \sigma_\alpha\right)\right\rangle_{\boldsymbol{\sigma}}}{\left\langle \exp\left(\frac{1}{2}\beta^2 J^2 \sum_{\alpha\gamma} q_{\alpha\gamma}\sigma_\alpha\sigma_\gamma + \beta J_0 \sum_\alpha m_\alpha \sigma_\alpha\right)\right\rangle_{\boldsymbol{\sigma}}}. \tag{83}$$

The diagonal elements are always $q_{\alpha\alpha} = 1$. For high noise levels, $\beta \to 0$, we obtain the trivial result

$$q_{\alpha\gamma} = \delta_{\alpha\gamma}, \quad m_\alpha = 0.$$

Assuming a continuous transition to a non-trivial state as the noise level is lowered, we can expand the saddle-point equations (82) and (83) in powers of $\mathbf{q}$ and $\mathbf{m}$ and look for bifurcations, which gives ($\lambda \neq \rho$):

$$q_{\lambda\rho} = \beta^2 J^2 q_{\lambda\rho} + \mathcal{O}(\mathbf{q},\mathbf{m})^2, \quad m_\lambda = \beta J_0 m_\lambda + \mathcal{O}(\mathbf{q},\mathbf{m})^2.$$

Therefore we expect transitions either at $T = J_0$ (if $J_0 > J$) or at $T = J$ (if $J > J_0$). The remaining program is: find the saddle-point $(\mathbf{q},\mathbf{m})$ for $T < \max\{J_0, J\}$ which for integer n minimizes f, determine the corresponding minimum as a function of n, and finally take the limit $n \to 0$. This is in fact the most complicated part of the procedure.

5.2. Replica-symmetric solution and AT-instability

5.2.1. Physical interpretation of saddle points

To obtain a guide in how to select saddle-points we now turn to a different (but equivalent) version of the replica trick (75), which allows us to attach a physical meaning to the saddle-points $(\mathbf{m},\mathbf{q})$. This version transforms averages over a given measure W:

$$\begin{aligned}\frac{\sum_{\boldsymbol{\sigma}} \Phi(\boldsymbol{\sigma}) W(\boldsymbol{\sigma})}{\sum_{\boldsymbol{\sigma}} W(\boldsymbol{\sigma})} &= \lim_{n\to 0} \sum_{\boldsymbol{\sigma}} \Phi(\boldsymbol{\sigma}) W(\boldsymbol{\sigma}) \left[\sum_{\boldsymbol{\sigma}} W(\boldsymbol{\sigma})\right]^{n-1} \\ &= \lim_{n\to 0} \sum_{\boldsymbol{\sigma}^1 \ldots \boldsymbol{\sigma}^n} \Phi(\boldsymbol{\sigma}^1) \prod_{\alpha=1}^{n} W(\boldsymbol{\sigma}^\alpha) \\ &= \lim_{n\to 0} \frac{1}{n} \sum_{\gamma=1}^{n} \sum_{\boldsymbol{\sigma}^1 \ldots \boldsymbol{\sigma}^n} \Phi(\boldsymbol{\sigma}^\gamma) \prod_{\alpha=1}^{n} W(\boldsymbol{\sigma}^\alpha)\end{aligned} \tag{84}$$

The trick again consists in evaluating this quantity for *integer* n, whereas the limit refers to noninteger n. We use (84) to write the distribution $P(m)$ of overlaps in equilibrium as

$$P(m) = \frac{\sum_{\boldsymbol{\sigma}} \delta[m - \frac{1}{N}\sum_i \xi_i\sigma_i] \mathrm{e}^{-\beta H(\boldsymbol{\sigma})}}{\sum_{\boldsymbol{\sigma}} \mathrm{e}^{-\beta H(\boldsymbol{\sigma})}}$$
$$= \lim_{n\to 0} \frac{1}{n} \sum_{\gamma} \sum_{\boldsymbol{\sigma}^1 \ldots \boldsymbol{\sigma}^n} \delta\left[m - \frac{1}{N}\sum_i \xi_i \sigma_i^\gamma\right] \prod_\alpha \mathrm{e}^{-\beta H(\boldsymbol{\sigma}^\alpha)}.$$

If we average this distribution over the disorder, we find identical expressions to those encountered in evaluating the disorder averaged free energy. By inserting the same delta-functions we arrive at the steepest descend integration (77) and find

$$\overline{P(m)} = \lim_{n\to 0} \frac{1}{n} \sum_{\gamma} \delta\big[m - m_\gamma\big], \tag{85}$$

where $\{m_\gamma\}$ refers to the relevant solution of (82) and (83). Similarly we can imagine *two* systems $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$ with identical synapses $\{J_{ij}\}$, both in thermal equilibrium. We now use (84) to rewrite the distribution $P(q)$ for the mutual overlap between the microstates of the two systems

$$P(q) = \frac{\sum_{\boldsymbol{\sigma},\boldsymbol{\sigma}'} \delta[q - \frac{1}{N}\sum_i \sigma_i\sigma_i'] \mathrm{e}^{-\beta H(\boldsymbol{\sigma}) - \beta H(\boldsymbol{\sigma}')}}{\sum_{\boldsymbol{\sigma},\boldsymbol{\sigma}'} \mathrm{e}^{-\beta H(\boldsymbol{\sigma}) - \beta H(\boldsymbol{\sigma}')}}$$
$$= \lim_{n\to 0} \frac{1}{n(n-1)} \sum_{\lambda\neq\gamma} \sum_{\boldsymbol{\sigma}^1 \ldots \boldsymbol{\sigma}^n} \delta\left[q - \frac{1}{N}\sum_i \sigma_i^\lambda \sigma_i^\gamma\right] \prod_\alpha \mathrm{e}^{-\beta H(\boldsymbol{\sigma}^\alpha)}.$$

Averaging over the disorder again leads to the steepest descend integration (77) and we find

$$\overline{P(q)} = \lim_{n\to 0} \frac{1}{n(n-1)} \sum_{\lambda\neq\gamma} \delta\big[q - q_{\lambda\gamma}\big], \tag{86}$$

where $\{q_{\lambda\gamma}\}$ refers to the relevant solution of (82) and (83). We can now partly interpret the saddle-points $(\mathbf{m}, \mathbf{q})$, since the shape of $\overline{P(q)}$ and $\overline{P(m)}$ gives direct information on the structure of phase space with respect to ergodicity. The crucial observation is that for an ergodic system one always has

$$P(m) = \delta\left[m - \frac{1}{N}\sum_i \xi_i \langle\sigma_i\rangle_{\mathrm{eq}}\right], \quad P(q) = \delta\left[q - \frac{1}{N}\sum_i \langle\sigma_i\rangle_{\mathrm{eq}}^2\right]. \tag{87}$$

If, on the other hand, there are L ergodic components in our system, each of which corresponding to a pure Gibbs state with microstate probabilities proportional to $\exp(-\beta H)$ and thermal averages $\langle\ldots\rangle_\ell$, and if we denote the probability of finding the system in component ℓ by W_ℓ, we find

$$P(m) = \sum_{\ell=1}^{L} W_\ell\, \delta\left[m - \frac{1}{N}\sum_i \xi_i \langle\sigma_i\rangle_\ell\right], \quad P(q) = \sum_{\ell,\ell'=1}^{L} W_\ell W_{\ell'}\, \delta\left[q - \frac{1}{N}\sum_i \langle\sigma_i\rangle_\ell \langle\sigma_i\rangle_{\ell'}\right].$$

For ergodic systems both $P(m)$ and $P(q)$ are δ-functions, for systems with a finite number of ergodic components they are finite sums of δ-functions. A *diverging* number of ergodic components generally leads to distributions with continuous pieces. If we combine this interpretation with our results (85) and (86) we find that ergodicity is equivalent to the relevant saddle-point being of the form:

$$q_{\alpha\beta} = \delta_{\alpha\beta} + q\big[1 - \delta_{\alpha\beta}\big], \quad m_\alpha = m, \tag{88}$$

which is called the 'replica symmetry' (RS) ansatz. The meaning of m and q is deduced from (87) (taking into account the transformation $\sigma_i \to \xi_i\sigma_i$ we performed along the way):

$$m = \frac{1}{N}\sum_i \overline{\xi_i\langle\sigma_i\rangle_{\rm eq}}, \quad q = \frac{1}{N}\sum_i \overline{\langle\sigma_i\rangle_{\rm eq}^2}.$$

5.2.2. Replica symmetric solution

Having saddle-points of the simple form (88) leads to an enormous simplification in our calculations. Insertion of (88) as an ansatz into Eqs. (81)–(83) gives

$$f(\mathbf{q},\mathbf{m}) = -\frac{1}{\beta}\log 2 - \frac{1}{4}\beta J^2(1-q)^2 + \frac{1}{2}J_0m^2 - \frac{1}{\beta n}\log\left\langle \exp\left(\frac{1}{2}q\beta^2J^2\left[\sum_\alpha \sigma_\alpha\right]^2 + \beta J_0 m\sum_\alpha \sigma_\alpha\right)\right\rangle_{\boldsymbol{\sigma}} + \mathcal{O}(n),$$

$$q = \frac{\langle\sigma_1\sigma_2\exp(\frac{1}{2}q\beta^2J^2\left[\sum_\alpha\sigma_\alpha\right]^2+\beta J_0m\sum_\alpha\sigma_\alpha)\rangle_{\boldsymbol{\sigma}}}{\langle\exp(\frac{1}{2}q\beta^2J^2\left[\sum_\alpha\sigma_\alpha\right]^2+\beta J_0m\sum_\alpha\sigma_\alpha)\rangle_{\boldsymbol{\sigma}}},$$

$$m = \frac{\langle\sigma_1\exp(\frac{1}{2}q\beta^2J^2\left[\sum_\alpha\sigma_\alpha\right]^2+\beta J_0m\sum_\alpha\sigma_\alpha)\rangle_{\boldsymbol{\sigma}}}{\langle\exp(\frac{1}{2}q\beta^2J^2\left[\sum_\alpha\sigma_\alpha\right]^2+\beta J_0m\sum_\alpha\sigma_\alpha)\rangle_{\boldsymbol{\sigma}}}.$$

We linearize the terms $\left[\sum_\alpha\sigma_\alpha\right]^2$ by introducing a Gaussian integral, and perform the average over the remaining neurons. The solutions m and q turn out to be well-defined for $n \to 0$ so we can take the limit:

$$\lim_{n\to 0} f(\mathbf{q},\mathbf{m}) = -\frac{1}{\beta}\log 2 - \frac{1}{4}\beta J^2(1-q)^2 + \frac{1}{2}J_0m^2 - \frac{1}{\beta}\int \mathrm{D}z\,\log\cosh\big[\beta J_0m + \beta Jz\sqrt{q}\big], \tag{89}$$

$$q = \int \mathrm{D}z\,\tanh^2\big[\beta J_0m + \beta Jz\sqrt{q}\big], \quad m = \int \mathrm{D}z\,\tanh\big[\beta J_0m + \beta Jz\sqrt{q}\big]. \tag{90}$$

Writing the equation for m in integral form gives

$$m = \beta J_0 m\int_0^1 \mathrm{d}\lambda\left[1 - \int \mathrm{D}z\,\tanh^2\big[\lambda\beta J_0m + \beta Jz\sqrt{q}\big]\right].$$

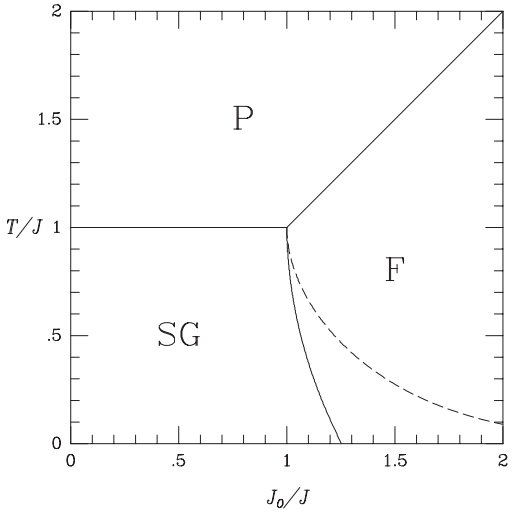

Fig. 10. Phase diagram of the model (72) with Gaussian synapses, obtained from the replica-symmetric solution. P: paramagnetic phase, $m = q = 0$ (more or less random evolution). SG: spin-glass phase, $m = 0$, $q \neq 0$ ('frozen' equilibrium states without pattern recall). F: recall ('ferro-magnetic') phase, $m \neq 0$, $q \neq 0$. Solid lines: second-order transitions. Dashed: the AT instability.

From this expression, in combination with (90), we conclude:

$$T > J_0 : \; m = 0 \qquad T > J_0 \text{ and } T > J : \; m = q = 0.$$

Linearization of (90) for small q and m shows the following continuous bifurcations:

	at	from	to
$J_0 > J$:	$T = J_0$	$m = q = 0$	$m \neq 0,\ q > 0$
$J_0 < J$:	$T = J$	$m = q = 0$	$m = 0,\ q > 0$
$T < \max\{J_0, J\}$:	$T = J_0[1 - q]$	$m = 0,\ q > 0$	$m \neq 0,\ q > 0$

Solving numerically equations $T = J_0[1 - q]$ and (90) leads to the phase diagram shown in Fig. 10.

5.2.3. Breaking of RS: the AT instability

If for the replica symmetric solution we calculate the entropy $S = \beta^2 \partial F / \partial \beta$ numerically, we find that for small temperatures it becomes negative. This is not possible. Firstly, straightforward differentiation shows $\partial S / \partial \beta = \beta[\langle H \rangle_{\text{eq}}^2 - \langle H^2 \rangle_{\text{eq}}] \leqslant 0$, so S increases with the noise level T. Let us now write $H(\boldsymbol{\sigma}) = H_0 + \hat{H}(\boldsymbol{\sigma})$, where H_0 is the ground- state energy and $\hat{H}(\boldsymbol{\sigma}) \geqslant 0$ (zero only for ground-state configurations, the number of which we denote by $N_0 \geqslant 1$). We now find

$$\lim_{T\to 0} S = \lim_{\beta\to\infty}\left\{\log\sum_{\boldsymbol{\sigma}} \mathrm{e}^{-\beta H(\boldsymbol{\sigma})} + \beta\langle H\rangle_{\mathrm{eq}}\right\}$$
$$= \lim_{\beta\to\infty}\left[\log\sum_{\boldsymbol{\sigma}} \mathrm{e}^{-\beta \hat{H}(\boldsymbol{\sigma})} + \beta\langle \hat{H}\rangle_{\mathrm{eq}}\right] \geqslant \log N_0.$$

We conclude that $S\geqslant 0$ for all T. At small temperatures the RS ansatz (88) is apparently incorrect in that it no longer corresponds to the minimum of $f(\mathbf{q},\mathbf{m})$ (81). If saddle-points without RS bifurcate continuously from the RS one, we can locate the occurrence of this 'replica symmetry breaking' (RSB) by studying the effect on $f(\mathbf{q},\mathbf{m})$ of small fluctuations around the RS solution. It was shown [19] that the 'dangerous' fluctuations are of the form

$$q_{\alpha\beta} \to \delta_{\alpha\beta} + q\big[1-\delta_{\alpha\beta}\big] + \eta_{\alpha\beta}, \quad \sum_{\beta}\eta_{\alpha\beta} = 0 \quad \forall\alpha. \tag{91}$$

in which q is the solution of (90) and $\eta_{\alpha\beta} = \eta_{\beta\alpha}$. We now calculate the resulting change in $f(\mathbf{q},\mathbf{m})$, away from the RS value $f(\mathbf{q}_{\mathrm{RS}},\mathbf{m}_{\mathrm{RS}})$, the leading order of which is quadratic in the fluctuations $\{\eta_{\alpha\beta}\}$ since the RS solution of (90) is a saddle-point:

$$f(\mathbf{q},\mathbf{m}) - f(\mathbf{q}_{\mathrm{RS}},\mathbf{m}_{\mathrm{RS}}) = \frac{\beta J^2}{4n}\sum_{\alpha\neq\gamma}\eta_{\alpha\gamma}^2 - \frac{\beta^3 J^4}{8n}\sum_{\alpha\neq\gamma}\sum_{\rho\neq\lambda}\eta_{\alpha\gamma}\eta_{\rho\lambda}G_{\alpha\gamma\rho\lambda}$$

with

$$G_{\alpha\gamma\rho\lambda} = \frac{\left\langle \sigma_\alpha\sigma_\gamma\sigma_\rho\sigma_\lambda \exp\left(\frac{1}{2}q\beta^2J^2\big[\sum_\alpha\sigma_\alpha\big]^2 + \beta m J_0\sum_\alpha\sigma_\alpha\right)\right\rangle_{\boldsymbol{\sigma}}}{\left\langle \exp\left(\frac{1}{2}q\beta^2J^2\big[\sum_\alpha\sigma_\alpha\big]^2 + \beta m J_0\sum_\alpha\sigma_\alpha\right)\right\rangle_{\boldsymbol{\sigma}}}.$$

Because of the index permutation symmetry in the above average we can write for $\alpha\neq\gamma$ and $\rho\neq\lambda$:

$$G_{\alpha\gamma\rho\lambda} = \delta_{\alpha\rho}\delta_{\gamma\lambda} + \delta_{\alpha\lambda}\delta_{\gamma\rho} + G_4[1-\delta_{\alpha\rho}][1-\delta_{\gamma\lambda}][1-\delta_{\alpha\lambda}][1-\delta_{\gamma\rho}]$$
$$+ G_2\big\{\delta_{\alpha\rho}[1-\delta_{\gamma\lambda}] + \delta_{\gamma\lambda}[1-\delta_{\alpha\rho}] + \delta_{\alpha\lambda}[1-\delta_{\gamma\rho}] + \delta_{\gamma\rho}[1-\delta_{\alpha\lambda}]\big\}$$

with

$$G_\ell = \frac{\int \mathrm{D}z\,\tanh^\ell\big[\beta J_0 m + \beta J z\sqrt{q}\big]\cosh^n\big[\beta J_0 m + \beta J z\sqrt{q}\big]}{\int \mathrm{D}z\,\cosh^n\big[\beta J_0 m + \beta J z\sqrt{q}\big]}.$$

Only terms which involve precisely two δ-functions can contribute, because of the requirements $\alpha\neq\gamma$, $\rho\neq\lambda$ and $\sum_\beta\eta_{\alpha\beta}=0$. As a result:

$$f(\mathbf{q},\mathbf{m}) - f(\mathbf{q}_{\mathrm{RS}},\mathbf{m}_{\mathrm{RS}}) = \frac{\beta J^2}{4n}\left[1-\beta^2J^2(1-2G_2+G_4)\right]\sum_{\alpha\neq\gamma}\eta_{\alpha\gamma}^2.$$

The condition for the RS solution to minimize $f(\mathbf{q},\mathbf{m})$, if compared to the so-called 'replicon' fluctuations (91), is therefore

$$1 > \beta^2 J^2 \lim_{n\to 0}(1 - 2\,G_2 + G_4).$$

After taking the limit in the expressions G_ℓ this condition can be written as

$$1 > \beta^2 J^2 \int \mathrm{D}z\, \cosh^{-4}\big[\beta J_0 m + \beta J z\sqrt{q}\big]. \tag{92}$$

The so-called AT line in the phase diagram where this condition ceases to be met, indicates a continuous transition to a complex 'spin-glass' state where ergodicity is broken (i.e. the distribution $\overline{P(q)}$ (86) is no longer a δ-function). It is shown in figure 10 as a dashed line for $J_0/J > 1$, and coincides with the line $T/J = 1$ for $J_0 < 1$.

6. The Hopfield model near saturation

6.1. Replica analysis

We now turn to the Hopfield model with an extensive number of stored patterns, i.e. $p = \alpha N$ in (40). We can still write the free energy in the form (48), but this will not be of help since here it involves integrals over an extensive number of variables, so that steepest descent integration does not apply. Instead, following the approach of the previous model (72), we assume [18] that we can average the free energy over the distribution of the patterns, with help of the replica-trick (75):

$$\bar{F} = -\lim_{n\to 0}\frac{1}{\beta n}\log \sum_{\boldsymbol{\sigma}^1\ldots\boldsymbol{\sigma}^n} \overline{\mathrm{e}^{-\beta\sum_{\alpha=1}^n H(\boldsymbol{\sigma}^\alpha)}}.$$

Greek indices will denote either replica labels or pattern labels (it will be clear from the context), i.e. $\alpha, \beta = 1, \ldots, n$ and $\mu, \nu = 1, \ldots, p$. The $p \times N$ pattern components $\{\xi_i^\mu\}$ are assumed to be drawn independently at random from $\{-1, 1\}$.

6.1.1. Replica calculation of the disorder-averaged free energy

We first add to the Hamiltonian of (30) a finite number ℓ of generating terms, that will allow us to obtain expectation values of the overlap order parameters m_μ (41) by differentiation of the free energy (since all patterns are equivalent in the calculation we may choose these ℓ nominated patterns arbitrarily):

$$H \to H + \sum_{\mu=1}^{\ell}\lambda_\mu \sum_i \sigma_i \xi_i^\mu, \quad \langle m_\mu(\boldsymbol{\sigma})\rangle_{\mathrm{eq}} = \lim_{\boldsymbol{\lambda}\to\mathbf{0}}\frac{\partial}{\partial\lambda_\mu}F/N. \tag{93}$$

We know how to deal with a finite number of overlaps and corresponding patterns, therefore we average only over the disorder that is responsible for the complications: the patterns $\{\boldsymbol{\xi}^{\ell+1}, \ldots, \boldsymbol{\xi}^p\}$ (as in the previous section we denote this disorder-averaging by $\overline{\cdots}$). Upon inserting the extended Hamiltonian into the replica-expression for the free energy, and assuming that the order of the limits $N \to \infty$ and $n \to 0$ can be interchanged, we obtain for large N:

$$\bar{F}/N = \frac{1}{2}\alpha - \frac{1}{\beta}\log 2 - \lim_{n\to 0}\frac{1}{\beta N n} \times \log\left\langle \exp\left(-\beta \sum_{\mu\leqslant\ell}\sum_{\alpha}\left[\lambda_\mu \sum_i \sigma_i^\alpha \xi_i^\mu - \frac{1}{2N}\left[\sum_i \sigma_i^\alpha \xi_i^\mu\right]^2\right]\right) \times \overline{e^{\frac{\beta}{2N}\sum_\alpha \sum_{\mu>\ell}\left[\sum_i \sigma_i^\alpha \xi_i^\mu\right]^2}}\right\rangle_{\{\boldsymbol{\sigma}^\alpha\}}.$$

We linearize the $\mu\leqslant\ell$ quadratic term using the identity (56), leading to $n\times\ell$ Gaussian integrals with $\mathrm{D}\mathbf{m} = (\mathrm{D}m_1^1,\ldots,Dm_n^\ell)$:

$$\bar{F}/N = \frac{1}{2}\alpha - \frac{1}{\beta}\log 2 - \lim_{n\to 0}\frac{1}{\beta N n} \times \log\int \mathrm{D}\mathbf{m}\left\langle \exp\left(\sum_{\mu\leqslant\ell}\sum_\alpha\sum_i \sigma_i^\alpha\xi_i^\mu\left[\sqrt{\frac{\beta}{N}}m_\alpha^\mu - \beta\lambda_\mu\right]\right)\overline{e^{\frac{\beta}{2N}\sum_\alpha\sum_{\mu>\ell}\left[\sum_i\sigma_i^\alpha\xi_i^\mu\right]^2}}\right\rangle_{\{\boldsymbol{\sigma}^\alpha\}}.$$

Anticipating that only terms exponential in the system size N will retain statistical relevance in the limit $N\to\infty$, we rescale the $n\times\ell$ integration variables $\mathbf{m}$ according to $\mathbf{m}\to\mathbf{m}\sqrt{\beta N}$:

$$\bar{F}/N = \frac{1}{2}\alpha - \frac{1}{\beta}\log 2 - \lim_{n\to 0}\frac{1}{\beta N n} \times \log\left\{\left[\frac{\beta N}{2\pi}\right]^{\frac{n\ell}{2}}\int \mathrm{d}\mathbf{m}\, e^{-\frac{1}{2}\beta N\mathbf{m}^2}\left\langle e^{\beta\sum_{\mu\leqslant\ell}\sum_\alpha\sum_i\sigma_i^\alpha\xi_i^\mu\left[m_\alpha^\mu-\lambda_\mu\right]}\overline{e^{\frac{\beta}{2N}\sum_\alpha\sum_{\mu>\ell}\left[\sum_i\sigma_i^\alpha\xi_i^\mu\right]^2}}\right\rangle_{\{\boldsymbol{\sigma}^\alpha\}}\right\}. \tag{94}$$

Next we turn to the disorder average, where we again linearize the exponent containing the pattern components using the identity (56), with $\mathrm{D}\mathbf{z} = (\mathrm{D}z_1,\ldots,\mathrm{D}z_n)$:

$$\begin{aligned}
\overline{e^{\frac{\beta}{2N}\sum_\alpha\sum_{\mu>\ell}\left[\sum_i\sigma_i^\alpha\xi_i^\mu\right]^2}} &= \left\{\overline{e^{\frac{1}{2}\sum_\alpha\left[\left(\frac{\beta}{N}\right)^{\frac{1}{2}}\sum_i\sigma_i^\alpha\xi_i\right]^2}}\right\}^{p-\ell} \\
&= \left\{\int \mathrm{D}\mathbf{z}\ \overline{e^{\left(\frac{\beta}{N}\right)^{\frac{1}{2}}\sum_\alpha z_\alpha\sum_i\sigma_i^\alpha\xi_i}}\right\}^{p-\ell} \\
&= \left\{\int \mathrm{D}\mathbf{z}\prod_i\cosh\left[\left(\frac{\beta}{N}\right)^{\frac{1}{2}}\sum_\alpha z_\alpha\sigma_i^\alpha\right]\right\}^{p-\ell} \\
&= \left\{\int \mathrm{D}\mathbf{z}\ e^{\frac{\beta}{2N}\sum_{\alpha\beta}z_\alpha z_\beta\sum_i\sigma_i^\alpha\sigma_i^\beta+\mathcal{O}\left(\frac{1}{N}\right)}\right\}^{p}.
\end{aligned} \tag{95}$$

We are now as in the previous case led to introducing the replica order parameters $q_{\alpha\beta}$:

$$1 = \int \mathrm{d}\mathbf{q} \prod_{\alpha\beta} \delta\left[q_{\alpha\beta} - \frac{1}{N}\sum_i \sigma_i^\alpha \sigma_i^\beta\right]$$
$$= \left[\frac{N}{2\pi}\right]^{n^2} \int \mathrm{d}\mathbf{q}\,\mathrm{d}\hat{\mathbf{q}}\; \mathrm{e}^{\mathrm{i}N\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\left[q_{\alpha\beta}-\frac{1}{N}\sum_i \sigma_i^\alpha\sigma_i^\beta\right]}.$$

Inserting (95) and the above identities into (94) and assuming that the limits $N \to \infty$ and $n \to 0$ commute gives:

$$\lim_{N\to\infty} \bar{F}/N = \frac{1}{2}\alpha - \frac{1}{\beta}\log 2 - \lim_{N\to\infty}\lim_{n\to 0}\frac{1}{\beta N n}\log \int \mathrm{d}\mathbf{m}\,\mathrm{d}\mathbf{q}\,\mathrm{d}\hat{\mathbf{q}}$$
$$\times \exp\left(N\left[i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}q_{\alpha\beta} - \frac{1}{2}\beta\mathbf{m}^2 + \alpha\log\int \mathrm{D}\mathbf{z}\,\mathrm{e}^{\frac{\beta}{2}\sum_{\alpha\beta} z_\alpha z_\beta q_{\alpha\beta}}\right]\right)$$
$$\times \left\langle \exp\left(\beta\sum_{\mu\leqslant\ell}\sum_\alpha\sum_i \sigma_i^\alpha \xi_i^\mu\left[m_\alpha^\mu - \lambda_\mu\right] - i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sum_i \sigma_i^\alpha\sigma_i^\beta\right)\right\rangle_{\{\boldsymbol{\sigma}^\alpha\}}.$$

The n-dimensional Gaussian integral over $\mathbf{z}$ factorizes in the standard way after appropriate rotation of the integration variables $\mathbf{z}$, with the result:

$$\log\int \mathrm{D}\mathbf{z}\,\mathrm{e}^{\frac{\beta}{2}\sum_{\alpha\beta} z_\alpha z_\beta q_{\alpha\beta}} = -\frac{1}{2}\log\det[\mathbf{I} - \beta\mathbf{q}],$$

in which $\mathbf{I}$ denotes the $n \times n$ identity matrix. The neuron averages factorize and are reduced to single-site ones over the n-replicated neuron $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$:

$$\lim_{N\to\infty} \bar{F}/N = \frac{1}{2}\alpha - \frac{1}{\beta}\log 2 - \lim_{N\to\infty}\lim_{n\to 0}\frac{1}{\beta N n}\log \int \mathrm{d}\mathbf{m}\,\mathrm{d}\mathbf{q}\,\mathrm{d}\hat{\mathbf{q}}$$
$$\times \exp\left(N\left[i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}q_{\alpha\beta} - \frac{1}{2}\beta\mathbf{m}^2 - \frac{1}{2}\alpha\log\det[\mathbf{I} - \beta\mathbf{q}]\right]\right)$$
$$\times \prod_i \left\langle \exp\left(\beta\sum_{\mu\leqslant\ell}\sum_\alpha \sigma_\alpha \xi_i^\mu\left[m_\alpha^\mu - \lambda_\mu\right] - i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sigma_\alpha\sigma_\beta\right)\right\rangle_{\boldsymbol{\sigma}}$$

and we arrive at integrals that can be evaluated by steepest descent, following the manipulations (76). If we denote averages over the remaining ℓ patterns in the familiar way

$$\boldsymbol{\xi} = (\xi_1, \ldots, \xi_\ell), \quad \langle\Phi(\boldsymbol{\xi})\rangle_{\boldsymbol{\xi}} = 2^{-\ell}\sum_{\boldsymbol{\xi}\in\{-1,1\}^\ell}\Phi(\boldsymbol{\xi})$$

we can write the final result in the form

$$\lim_{N\to\infty}\bar{F}/N = \lim_{n\to 0}\operatorname{extr} f(\mathbf{m}, \mathbf{q}, \hat{\mathbf{q}}), \tag{96}$$

$$f(\mathbf{m},\mathbf{q},\hat{\mathbf{q}}) = \frac{1}{2}\alpha - \frac{1}{\beta}\log 2 - \frac{1}{\beta n} \times \left[\left\langle \log \left\langle \exp\left(\beta \sum_{\mu\leqslant\ell} \sum_{\alpha} \sigma_\alpha \xi_\mu \left[m_\alpha^\mu - \lambda_\mu \right] - i \sum_{\alpha\beta} \hat{q}_{\alpha\beta}\sigma_\alpha\sigma_\beta \right) \right\rangle_{\boldsymbol{\sigma}} \right\rangle_{\boldsymbol{\xi}} + i \sum_{\alpha\beta} \hat{q}_{\alpha\beta} q_{\alpha\beta} - \frac{1}{2}\beta\mathbf{m}^2 - \frac{1}{2}\alpha \log\det[\mathbf{I} - \beta\mathbf{q}] \right].$$

Having arrived at a saddle-point problem we now first identify the expectation values of the overlaps with (93) (note: extremization with respect to the saddle-point variables and differentiation with respect to $\boldsymbol{\lambda}$ commute):

$$\overline{\langle m_\mu(\boldsymbol{\sigma})\rangle_{\rm eq}} = \lim_{n\to 0} \lim_{\boldsymbol{\lambda}\to 0} \frac{\partial}{\partial \lambda_\mu} \operatorname{extr} f(\mathbf{m},\mathbf{q},\hat{\mathbf{q}}) = \lim_{n\to 0} \left\langle \frac{\xi_\mu \left\langle \frac{1}{n}\sum_\alpha \sigma_\alpha \exp\left(\beta \sum_{\mu\leqslant\ell}\sum_\alpha \sigma_\alpha \xi_\mu m_\alpha^\mu - i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sigma_\alpha\sigma_\beta \right) \right\rangle_{\boldsymbol{\sigma}}}{\left\langle \exp\left(\beta \sum_{\mu\leqslant\ell}\sum_\alpha \sigma_\alpha \xi_\mu m_\alpha^\mu - i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sigma_\alpha\sigma_\beta \right) \right\rangle_{\boldsymbol{\sigma}}} \right\rangle_{\boldsymbol{\xi}} \tag{97}$$

which is to be evaluated in the $\boldsymbol{\lambda} = \mathbf{0}$ saddle-point. Having served their purpose, the generating fields $\boldsymbol{\lambda}_\mu$ can be set to zero and we can restrict ourselves to the $\boldsymbol{\lambda} = \mathbf{0}$ saddle-point problem:

$$f(\mathbf{m},\mathbf{q},\hat{\mathbf{q}}) = \frac{1}{2}\alpha - \frac{1}{\beta}\log 2 - \frac{1}{\beta n}\left[i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}q_{\alpha\beta} - \frac{1}{2}\beta\mathbf{m}^2 - \frac{1}{2}\alpha\log\det[\mathbf{I}-\beta\mathbf{q}] + \left\langle \log\left\langle \exp\left(\beta\sum_{\mu\leqslant\ell}\sum_\alpha \sigma_\alpha\xi_\mu m_\alpha^\mu - i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sigma_\alpha\sigma_\beta \right)\right\rangle_{\boldsymbol{\sigma}}\right\rangle_{\boldsymbol{\xi}} \right]. \tag{98}$$

Variation of the parameters $\{m_\alpha^\mu, \hat{q}_{\alpha\beta}, q_{\alpha\beta}\}$ gives the saddle-point equations:

$$m_\alpha^\mu = \left\langle \xi_\mu \frac{\langle \sigma_\alpha \exp(\beta\sum_{\mu\leqslant\ell}\sum_\alpha \sigma_\alpha\xi_\mu m_\alpha^\mu - i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sigma_\alpha\sigma_\beta)\rangle_{\boldsymbol{\sigma}}}{\langle \exp(\beta\sum_{\mu\leqslant\ell}\sum_\alpha \sigma_\alpha\xi_\mu m_\alpha^\mu - i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sigma_\alpha\sigma_\beta)\rangle_{\boldsymbol{\sigma}}} \right\rangle_{\boldsymbol{\xi}} \tag{99}$$

$$q_{\lambda\rho} = \left\langle \frac{\langle \sigma_\lambda\sigma_\rho \exp(\beta\sum_{\mu\leqslant\ell}\sum_\alpha \sigma_\alpha\xi_\mu m_\alpha^\mu - i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sigma_\alpha\sigma_\beta)\rangle_{\boldsymbol{\sigma}}}{\langle \exp(\beta\sum_{\mu\leqslant\ell}\sum_\alpha \sigma_\alpha\xi_\mu m_\alpha^\mu - i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sigma_\alpha\sigma_\beta)\rangle_{\boldsymbol{\sigma}}} \right\rangle_{\boldsymbol{\xi}} \tag{100}$$

$$\hat{q}_{\lambda\rho} = \frac{1}{2} i\alpha\beta \frac{\int \mathrm{d}\mathbf{z}\, z_\lambda z_\rho \mathrm{e}^{-\frac{1}{2}\mathbf{z}\cdot[\mathbf{I}-\beta\mathbf{q}]\mathbf{z}}}{\int \mathrm{d}\mathbf{z}\ \ \mathrm{e}^{-\frac{1}{2}\mathbf{z}\cdot[\mathbf{I}-\beta\mathbf{q}]\mathbf{z}}} \tag{101}$$

furthermore,

$$\overline{\langle m_\mu(\boldsymbol{\sigma})\rangle_{\rm eq}} = \lim_{n\to 0}\frac{1}{n}\sum_\alpha m_\alpha^\mu \tag{102}$$

replaces the identification (97). As expected, one always has $q_{\alpha\alpha}=1$. The diagonal elements $\hat{q}_{\alpha\alpha}$ drop out of (99) and (100), their values are simply given as functions of the remaining parameters by (101).

6.1.2. Physical interpretation of saddle points

We proceed along the lines of the Gaussian model (72). If we apply the alternative version (84) of the replica trick to the Hopfield model, we can write the distribution of the ℓ overlaps $\mathbf{m}=(m_1,\ldots,m_\ell)$ in equilibrium as

$$P(\boldsymbol{m}) = \lim_{n\to 0}\frac{1}{n}\sum_\gamma \sum_{\boldsymbol{\sigma}^1\ldots\boldsymbol{\sigma}^n}\delta\left[\mathbf{m}-\frac{1}{N}\sum_i \sigma_i^\gamma\boldsymbol{\xi}_i\right]\prod_\alpha {\rm e}^{-\beta H(\boldsymbol{\sigma}^\alpha)}$$

with $\boldsymbol{\xi}_i=(\xi_i^1,\ldots,\xi_i^\ell)$. Averaging this distribution over the disorder leads to expressions identical to those encountered in evaluating the disorder averaged free energy. By inserting the same delta-functions we arrive at the saddle-point integration (96) and (98) and find

$$\overline{P(\boldsymbol{m})} = \lim_{n\to 0}\frac{1}{n}\sum_\gamma \delta[\mathbf{m}-\mathbf{m}_\gamma], \tag{103}$$

where $\mathbf{m}_\gamma=(m_\gamma^1,\ldots,m_\gamma^\ell)$ refers to the relevant solution of (99)–(101).

Similarly we imagine two systems $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$ with identical realization of the interactions $\{J_{ij}\}$, both in thermal equilibrium, and use (84) to rewrite the distribution $P(q)$ for the mutual overlap between the microstates of the two systems

$$P(q) = \lim_{n\to 0}\frac{1}{n(n-1)}\sum_{\lambda\neq\gamma}\sum_{\boldsymbol{\sigma}^1\ldots\boldsymbol{\sigma}^n}\delta\left[q-\frac{1}{N}\sum_i \sigma_i^\lambda\sigma_i^\gamma\right]\prod_\alpha {\rm e}^{-\beta H(\boldsymbol{\sigma}^\alpha)}$$

Averaging over the disorder again leads to the steepest descend integration (96) and (98) and we find

$$\overline{P(q)} = \lim_{n\to 0}\frac{1}{n(n-1)}\sum_{\lambda\neq\gamma}\delta\big[q-q_{\lambda\gamma}\big], \tag{104}$$

where $\{q_{\lambda\gamma}\}$ refers to the relevant solution of (99)–(101).

Finally we analyze the physical meaning of the conjugate parameters $\{\hat{q}_{\alpha\beta}\}$ for $\alpha\neq\beta$. We will do this in more detail, the analysis being rather specific for the Hopfield model and slightly different from the derivations above. Again we imagine two systems $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$ with identical interactions $\{J_{ij}\}$, both in thermal equilibrium. We now use (84) to evaluate the covariance of the overlaps corresponding to non-nominated patterns:

$$r = \frac{1}{\alpha}\sum_{\mu=\ell+1}^{p} \overline{\left\langle \frac{1}{N}\sum_i \sigma_i \xi_i^\mu \right\rangle_{\rm eq} \left\langle \frac{1}{N}\sum_i \sigma_i' \xi_i^\mu \right\rangle_{\rm eq}}$$
$$= \lim_{n\to 0} \frac{N-\ell/\alpha}{n(n-1)} \sum_{\lambda\neq\gamma} \sum_{\boldsymbol{\sigma}^1\ldots\boldsymbol{\sigma}^n} \overline{\left[\frac{1}{N}\sum_i \sigma_i^\lambda \xi_i^p\right]\left[\frac{1}{N}\sum_i \sigma_i^\gamma \xi_i^p\right]\prod_\alpha {\rm e}^{-\beta H(\boldsymbol{\sigma}^\alpha)}} \tag{105}$$

(using the equivalence of all such patterns). We next perform the same manipulations as in calculating the free energy. Here the disorder average involves

$$\overline{\left[\frac{1}{\sqrt{N}}\sum_i \sigma_i^\lambda \xi_i^p\right]\left[\frac{1}{\sqrt{N}}\sum_i \sigma_i^\gamma \xi_i^p\right] {\rm e}^{\frac{\beta}{2N}\sum_\alpha \sum_{\mu>\ell}\left[\sum_i \sigma_i^\alpha \xi_i^\mu\right]^2}}$$
$$= \left\{\int {\rm D}\mathbf{z}\, \overline{{\rm e}^{\left(\frac{\beta}{N}\right)^{\frac{1}{2}}\sum_\alpha z_\alpha \sum_i \sigma_i^\alpha \xi_i}}\right\}^{p-\ell-1} \int \frac{{\rm D}\mathbf{z}}{\beta} \frac{\partial^2}{\partial z_\lambda \partial z_\gamma} \overline{{\rm e}^{\left(\frac{\beta}{N}\right)^{\frac{1}{2}}\sum_\alpha z_\alpha \sum_i \sigma_i^\alpha \xi_i}}$$
$$= \left\{\int {\rm D}\mathbf{z}\, \overline{{\rm e}^{\left(\frac{\beta}{N}\right)^{\frac{1}{2}}\sum_\alpha z_\alpha \sum_i \sigma_i^\alpha \xi_i}}\right\}^{p-\ell-1} \int {\rm D}\mathbf{z} \frac{z_\lambda z_\gamma}{\beta} \overline{{\rm e}^{\left(\frac{\beta}{N}\right)^{\frac{1}{2}}\sum_\alpha z_\alpha \sum_i \sigma_i^\alpha \xi_i}}$$

(after partial integration). We finally obtain an expression which involves the surface (98):

$$r = \frac{1}{\beta}\lim_{n\to 0}\frac{1}{n(n-1)}\sum_{\lambda\neq\rho}\lim_{N\to\infty} \frac{\int {\rm d}\mathbf{m}\,{\rm d}\mathbf{q}\,{\rm d}\hat{\mathbf{q}} \left[\frac{\int {\rm d}\mathbf{z}\, z_\lambda z_\rho\, {\rm e}^{-\frac{1}{2}\mathbf{z}\cdot[1-\beta\mathbf{q}]\mathbf{z}}}{\int {\rm d}\mathbf{z}\, {\rm e}^{-\frac{1}{2}\mathbf{z}\cdot[1-\beta\mathbf{q}]\mathbf{z}}}\right] {\rm e}^{-\beta n N f(\mathbf{m},\mathbf{q},\hat{\mathbf{q}})}}{\int {\rm d}\mathbf{m}\,{\rm d}\mathbf{q}\,{\rm d}\hat{\mathbf{q}}\, {\rm e}^{-\beta n N f(\mathbf{m},\mathbf{q},\hat{\mathbf{q}})}}.$$

The normalization of the above integral over $\{\mathbf{m},\mathbf{q},\hat{\mathbf{q}}\}$ follows from using the replica procedure to rewrite unity. The integration being dominated by the minima of f, we can use the saddle-point Eq. (101) to arrive at

$$\lim_{n\to 0}\frac{1}{n(n-1)}\sum_{\lambda\neq\rho}\hat{q}_{\lambda\rho} = \frac{1}{2}i\alpha\beta^2 r. \tag{106}$$

The result (105) and (106) provides a physical interpretation of the order parameters $\{\hat{q}_{\alpha\beta}\}$.

Ergodicity implies that the distributions $\overline{P(q)}$ and $\overline{P(\mathbf{m})}$ are δ-functions, this is equivalent to the relevant saddle-point being of the form:

$$m_\gamma^\mu = m_\mu, \quad q_{\gamma\rho} = \delta_{\gamma\rho} + q\left[1-\delta_{\gamma\rho}\right], \quad \hat{q}_{\gamma\rho} = \frac{1}{2}i\alpha\beta^2\left[R\delta_{\gamma\rho} + r\left[1-\delta_{\gamma\rho}\right]\right], \tag{107}$$

which is the RS ansatz for the Hopfield model. The RS form for $\{q_{\alpha\beta}\}$ and $\{m_\alpha^\mu\}$ is a direct consequence of the corresponding distributions being δ-functions, whereas the RS form for $\{\hat{q}_{\alpha\beta}\}$ subsequently follows from (101). The physical meaning of m_μ and q is

$$m_\mu = \overline{\langle m_\mu(\boldsymbol{\sigma})\rangle_{\rm eq}}, \quad q = \frac{1}{N}\sum_i \overline{\langle \sigma_i\rangle_{\rm eq}^2}.$$

Before proceeding with a full analysis of the RS saddle-point equations, we finally make a few tentative statements on the phase diagram. For $\beta = 0$ we obtain the trivial result $q_{\lambda\rho} = \delta_{\lambda\rho}$, $\hat{q}_{\lambda\rho} = 0$, $m_\alpha^\mu = 0$. We can identify continuous bifurcations to a nontrivial state by expanding the saddle-point equations in first-order in the relevant parameters:

$$m_\alpha^\mu = \beta m_\alpha^\mu + \cdots, \quad q_{\lambda\rho} = -2\,i\hat{q}_{\lambda\rho} + \cdots (\lambda \neq \rho),$$
$$\hat{q}_{\lambda\rho} = \frac{1}{2}\frac{i\alpha\beta}{1-\beta}\left[\delta_{\lambda\rho} + \frac{\beta}{1-\beta} q_{\lambda\rho}\left[1-\delta_{\lambda\rho}\right]\right] + \cdots$$

Combining the equations for $\mathbf{q}$ and $\hat{\mathbf{q}}$ gives $q_{\lambda\rho} = \alpha\left[\frac{\beta}{1-\beta}\right]^2 q_{\lambda\rho} + \cdots$ Thus we expect a continuous transition at $T = 1 + \sqrt{\alpha}$ from the trivial state to an ordered state where $q_{\lambda\rho} \neq 0$, but still $\langle m_\mu\rangle_{\rm eq} = 0$ (a spin-glass state).

6.2. Replica symmetric solution and AT-instability

The symmetry of the ansatz (107) for the saddle-point allows us to diagonalize the matrix $\boldsymbol{\Lambda} = \mathbf{I} - \beta\mathbf{q}$ which we encountered in the saddle-point problem, $\Lambda_{\alpha\beta} = [1-\beta(1-q)]\delta_{\alpha\beta} - \beta q$:

eigenspace	eigenvalue	multiplicity
$\mathbf{x} = (1,\ldots,1)$	$1-\beta(1-q)-\beta qn$	1
$\sum_\alpha x_\alpha = 0$	$1-\beta(1-q)$	$n-1$

so that

$$\log\det\boldsymbol{\Lambda} = \log[1-\beta(1-q)-\beta qn] + (n-1)\log[1-\beta(1-q)]$$
$$= n\left[\log[1-\beta(1-q)] - \frac{\beta q}{1-\beta(1-q)}\right] + \mathcal{O}(n^2).$$

Inserting the RS ansatz (107) for the saddle-point into (98), utilizing the above expression for the determinant and the shorthand $\mathbf{m} = (m_1,\ldots,m_\ell)$, gives

$$f(\mathbf{m}_{\rm RS}, \mathbf{q}_{\rm RS}, \hat{\mathbf{q}}_{\rm RS}) = -\frac{1}{\beta}\log 2 + \frac{1}{2}\alpha[1+\beta r(1-q)]$$
$$+\frac{1}{2}\mathbf{m}^2 + \frac{\alpha}{2\beta}\left[\log[1-\beta(1-q)] - \frac{\beta q}{1-\beta(1-q)}\right]$$
$$-\frac{1}{\beta n}\left\langle \log\left\langle e^{\beta\mathbf{m}\cdot\boldsymbol{\xi}\sum_\alpha\sigma_\alpha + \frac{1}{2}\alpha r\beta^2\left[\sum_\alpha\sigma_\alpha\right]^2}\right\rangle_{\boldsymbol{\sigma}}\right\rangle_{\boldsymbol{\xi}} + \mathcal{O}(n).$$

We now linearize the squares in the neuron averages with (56), subsequently average over the replicated neuron $\boldsymbol{\sigma}$, use $\cosh^n[x] = 1 + n\log\cosh[x] + \mathcal{O}(n^2)$, and take the limit $n \to 0$:

$$\lim_{N\to\infty} \bar{F}_{\mathrm{RS}}/N = \lim_{n\to 0} f(\mathbf{m}_{\mathrm{RS}}, \mathbf{q}_{\mathrm{RS}}, \hat{\mathbf{q}}_{\mathrm{RS}})$$
$$= \frac{1}{2}\mathbf{m}^2 + \frac{1}{2}\alpha\left[1 + \beta r(1-q) + \frac{1}{\beta}\log[1-\beta(1-q)] - \frac{q}{1-\beta(1-q)}\right]$$
$$- \frac{1}{\beta}\left\langle \int \mathrm{D}z\, \log 2\, \cosh\beta\left[\mathbf{m}\cdot\boldsymbol{\xi} + z\sqrt{\alpha r}\right]\right\rangle_{\xi}. \tag{108}$$

The saddle-point equations for $\mathbf{m}$, q and r can be obtained either by insertion of the RS ansatz (107) into (99)–(101) and subsequently taking the $n \to 0$ limit, or by variation of the RS expression (108). The latter route is the fastest one. After performing partial integrations where appropriate we obtain the final result:

$$\mathbf{m} = \left\langle \boldsymbol{\xi} \int \mathrm{D}z\, \tanh\beta\left[\mathbf{m}\cdot\boldsymbol{\xi} + z\sqrt{\alpha r}\right]\right\rangle_{\xi}, \tag{109}$$

$$q = \left\langle \int \mathrm{D}z\ \tanh^2\beta\left[\mathbf{m}\cdot\boldsymbol{\xi} + z\sqrt{\alpha r}\right]\right\rangle_{\xi}, \qquad r = q[1-\beta(1-q)]^{-2}. \tag{110}$$

By substitution of the equation for r into the remaining equations this set can easily be further reduced, should the need arise. In case of multiple solutions of (109) and (110) the relevant saddle-point is the one that minimizes (108). Clearly for $\alpha = 0$ we recover our previous results (50) and (51).

6.2.1. Analysis of RS order parameter equations and phase diagram

We first establish an upper bound for the temperature $T = 1/\beta$ for nontrivial solutions of the set (109) and (110) to exist, by writing (109) in integral form:

$$m_{\mu} = \beta\left\langle \xi_{\mu}(\boldsymbol{\xi}\cdot\mathbf{m}) \int_0^1 \mathrm{d}\lambda \int \mathrm{D}z\left[1 - \tanh^2\beta\left(\lambda\boldsymbol{\xi}\cdot\mathbf{m} + z\sqrt{\alpha r}\right)\right]\right\rangle_{\xi}$$

from which we deduce

$$0 = \mathbf{m}^2 - \beta\left\langle (\boldsymbol{\xi}\cdot\mathbf{m})^2 \int_0^1 \mathrm{d}\lambda \int \mathrm{D}z\left[1 - \tanh^2\beta\left(\lambda\boldsymbol{\xi}\cdot\mathbf{m} + z\sqrt{\alpha r}\right)\right]\right\rangle_{\xi}$$
$$\geqslant \mathbf{m}^2 - \beta\left\langle (\boldsymbol{\xi}\cdot\mathbf{m})^2\right\rangle_{\xi} = \mathbf{m}^2[1-\beta].$$

Therefore $\mathbf{m} = 0$ for $T > 1$. If $T > 1$ we obtain in turn from (110), using $\tanh^2(x) \leq x^2$ and $0 \leq q \leq 1$: $q = 0$ or $q \leq 1 + \sqrt{\alpha} - T$. We conclude that $q = 0$ for $T > 1 + \sqrt{\alpha}$. Secondly, for the free energy (108) to be well defined we must require $q > 1 - T$. Linearization of (109) and (110) for small q and $\mathbf{m}$ shows the continuous bifurcations:

	at	from	to
$\alpha > 0$:	$T = 1 + \sqrt{\alpha}$	$\mathbf{m} = \mathbf{0}$, $q = 0$	$\mathbf{m} = \mathbf{0}$, $q > 0$
$\alpha = 0$:	$T = 1$	$\mathbf{m} = \mathbf{0}$, $q = 0$	$\mathbf{m} \neq \mathbf{0}$, $q > 0$

The upper bound $T = 1 + \sqrt{\alpha}$ turns out to be the critical noise level indicating (for $\alpha > 0$) a continuous transition to a spin-glass state, where there is no significant alignment of the neurons in the direction of one particular pattern, but still a certain degree of local freezing. Since $\mathbf{m} = \mathbf{0}$ for $T > 1$ this spin-glass state persists at least down to $T = 1$. The quantitative details of the spin-glass state are obtained by inserting $\mathbf{m} = \mathbf{0}$ into (110) (since (109) is fulfilled automatically).

The impact on the saddle-point Eqs. (109) and (110) of having $\alpha > 0$, a smoothening of the hyperbolic tangent by convolution with a Gaussian kernel, can be viewed as noise caused by interference between the attractors. The natural strategy for solving (109) and (110) is therefore to make an ansatz for the nominated overlaps $\mathbf{m}$ of the type (52) (the mixture states). Insertion of this ansatz into the saddle-point equations indeed leads to self-consistent solutions. One can solve numerically the remaining equations for the amplitudes of the mixture states and evaluate their stability by calculating the eigenvalues of the second derivative of $f(\mathbf{m}, \mathbf{q}, \hat{\mathbf{q}})$, in the same way as for $\alpha = 0$. The calculations are just more involved. It then turns out that even mixtures are again unstable for any T and α, whereas odd mixtures can become locally stable for sufficiently small T and α. Among the mixture states, the pure states, where the vector $\mathbf{m}$ has only one nonzero component, are the first to stabilize as the temperature is lowered. These pure states, together with the spin-glass state ($\mathbf{m} = 0, q > 0$), we will study in more detail.

Let us first calculate the second derivatives of (108) and evaluate them in the spin-glass saddle-point. One finds, after elimination of r with (110):

$$\partial^2 f/\partial m_\mu \partial m_\nu = \delta_{\mu\nu}[1 - \beta(1-q)], \qquad \partial^2 f/\partial m_\mu \partial q = 0.$$

The $(\ell+1) \times (\ell+1)$ matrix of second derivatives with respect to variation of $(\mathbf{m}, q)$, evaluated in the spin-glass saddle-point, thereby acquires a diagonal form

$$\partial^2 f = \begin{pmatrix} 1-\beta(1-q) & & & \\ & \ddots & & \\ & & 1-\beta(1-q) & \\ & & & \partial^2 f/\partial q^2 \end{pmatrix}$$

and the eigenvalues can simply be read off. The ℓ-fold degenerate eigenvalue $1 - \beta(1-q)$ is always positive (otherwise (108) would not even exist), implying stability of the spin-glass state in the direction of the nominated patterns. The remaining eigenvalue measures the stability of the spin-glass state with respect to variation in the amplitude q. Below the critical noise level $T = 1 + \sqrt{\alpha}$ it turns out to be positive for the spin-glass solution of (110) with nonzero q. One important difference between the previously studied case $\alpha = 0$ and the present case $\alpha > 0$ is that there is now an $\mathbf{m} = \mathbf{0}$ spin-glass solution which is *stable* for all $T < 1 + \sqrt{\alpha}$. In terms of information processing this implies that for $\alpha > 0$ an initial state must have a certain nonzero overlap with a pattern to evoke a final state with $\mathbf{m} \neq \mathbf{0}$, in order to avoid ending up in the $\mathbf{m} = \mathbf{0}$ spin-glass state. This is clearly consistent with the observations in Fig. 5. In contrast, for $\alpha = 0$, the state with $\mathbf{m} = \mathbf{0}$ is unstable, so *any* initial state will eventually lead to a final state with $\mathbf{m} \neq \mathbf{0}$.

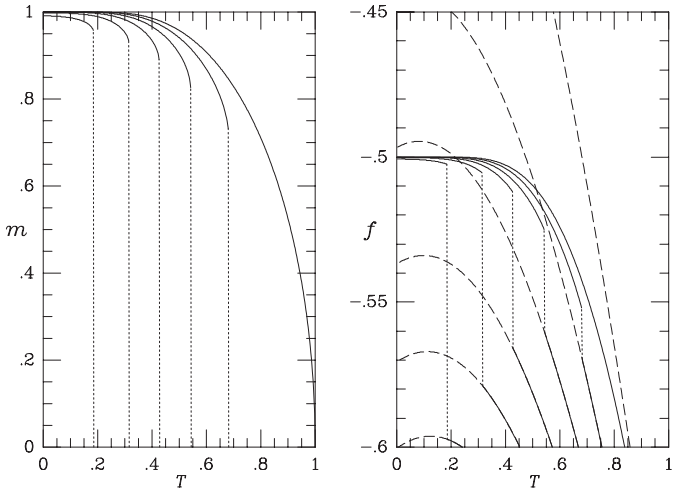

Fig. 11. Left: RS amplitudes m of the pure states of the Hopfield model versus temperature. From top to bottom: $\alpha = 0.000 - 0.125$ ($\Delta\alpha = 0.025$). Right, solid lines: 'free energies' f of the pure states. From bottom to top: $\alpha = 0.000 - 0.125$ ($\Delta\alpha = 0.025$). Right, dashed lines: 'free energies' of the spin-glass state $\mathbf{m} = 0$ (for comparison). From top to bottom: $\alpha = 0.000 - 0.125$ ($\Delta\alpha = 0.025$).

Inserting the pure state ansatz $\mathbf{m} = m(1, 0, \ldots, 0)$ into our RS equations gives

$$m = \int \mathrm{D}z \, \tanh\left[\beta m + \frac{z\beta\sqrt{\alpha q}}{1-\beta(1-q)}\right], \quad q = \int \mathrm{D}z \, \tanh^2\left[\beta m + \frac{z\beta\sqrt{\alpha q}}{1-\beta(1-q)}\right], \tag{111}$$

$$f = \frac{1}{2}m^2 + \frac{1}{2}\alpha\left[(1-q)\frac{1+\beta(1-q)(\beta-2)}{[1-\beta(1-q)]^2} + \frac{1}{\beta}\log[1-\beta(1-q)]\right] - \frac{1}{\beta}\int \mathrm{D}z \log 2\cosh\left[\beta m + \frac{z\beta\sqrt{\alpha q}}{1-\beta(1-q)}\right]. \tag{112}$$

If we solve Eq. (111) numerically for different values of α, and calculate the corresponding 'free energies' f (112) for the pure states and the spin-glass state $\mathbf{m} = 0$, we obtain Fig. 11. For $\alpha > 0$ the nontrivial solution m for the amplitude of the pure state appears *discontinously* as the temperature is lowered, defining a critical temperature $T_{\mathrm{M}}(\alpha)$. Once the pure state appears, it turns out to be locally stable (within the RS ansatz). Its 'free energy' f, however, remains larger than the one corresponding to the spin-glass state, until the temperature is further reduced to below a second critical temperature $T_{\mathrm{c}}(\alpha)$. For $T < T_{\mathrm{c}}(\alpha)$ the pure states are therefore the equilibrium states in the thermodynamics sense.

By drawing these critical lines in the (α, T) plane, together with the line $T_{\mathrm{g}}(\alpha) = 1 + \sqrt{\alpha}$ which signals the second-order transition from the paramagnetic to

the spin-glass state, we obtain the RS phase diagram of the Hopfield model, depicted in Fig. 12. Strictly speaking the line T_M would appear meaningless in the thermodynamic picture, only the saddle-point that minimizes f being relevant. However, we have to keep in mind the physics behind the formalism. The occurrence of multiple locally stable saddle-points is the manifestation of ergodicity breaking in the limit $N \to \infty$. The thermodynamic analysis, based on ergodicity, therefore applies only within a single ergodic component. Each locally stable saddle-point is indeed relevant for appropriate initial conditions and time-scales.

6.2.2. Zero temperature, storage capacity

The storage capacity α_c of the Hopfield model is defined as the largest α for which locally stable pure states exist. If for the moment we neglect the low temperature reentrance peculiarities in the phase diagram (12) to which we will come back later, the critical temperature $T_M(\alpha)$, where the pure states appear decreases monotonically with α, and the storage capacity is reached for $T = 0$. Before we can put $T \to 0$ in (111), however, we will have to rewrite these equations in terms of quantities with well-defined $T \to 0$ limits, since $q \to 1$. A suitable quantity is $C = \beta(1 - q)$, which obeys $0 \leq C \leq 1$ for the free energy (108) to exist. The saddle-point equations can now be written in the form

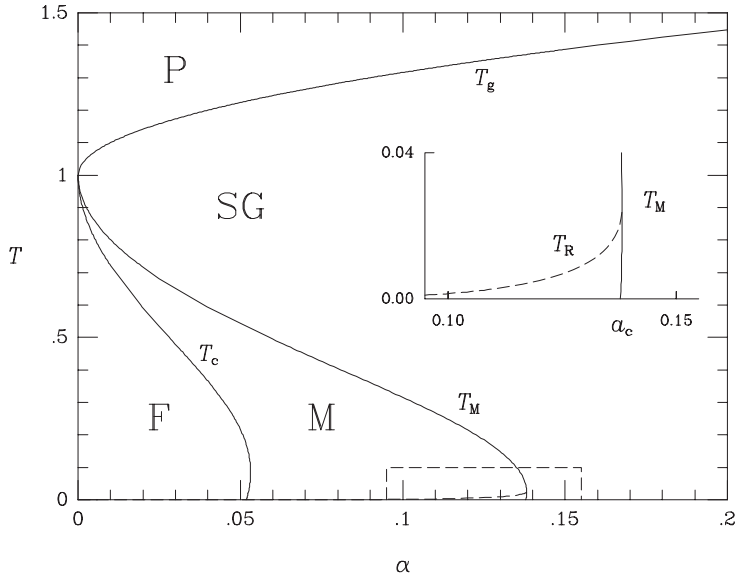

Fig. 12. Phase diagram of the Hopfield model. P: paramagnetic phase, $m = q = 0$ (no recall). SG: spin-glass phase, $m = 0$, $q \neq 0$ (no recall). F: pattern recall phase (recall states minimise f), $m \neq 0$, $q \neq 0$. M: mixed phase (recall states are local but not global minima of f). Solid lines: separations of the above phases (T_g: second-order, T_M and T_c: first-order). Dashed: the AT instability for the recall solutions (T_R). Inset: close-up of the low temperature region.

$$m = \int \mathrm{D}z \tanh\left[\beta m + \frac{z\beta\sqrt{\alpha q}}{1-C}\right], \quad C = \frac{\partial}{\partial m}\int \mathrm{D}z \tanh\left[\beta m + \frac{z\beta\sqrt{\alpha q}}{1-C}\right],$$

in which the limit $T \to 0$ simply corresponds to $\tanh(\beta x) \to \mathrm{sgn}(x)$ and $q \to 1$. After having taken the limit we perform the Gaussian integral:

$$m = \mathrm{erf}\left[\frac{m(1-C)}{\sqrt{2\,\alpha}}\right], \quad C = (1-C)\sqrt{\frac{2}{\alpha\pi}}\,\mathrm{e}^{-m^2(1-C)^2/2\alpha}.$$

This set can be reduced to a single transcendental equation by introducing $x = m(1-C)/\sqrt{2\alpha}$:

$$x\sqrt{2\,\alpha} = F(x), \qquad F(x) = \mathrm{erf}(x) - \frac{2\,x}{\sqrt{\pi}}\mathrm{e}^{-x^2}. \tag{113}$$

Eq. (113) is solved numerically (see Fig. 13). Since $F(x)$ is antisymmetric, solutions come in pairs $(x, -x)$ (reflecting the symmetry of the Hamiltonian of the system with respect to an overall state-flip $\boldsymbol{\sigma} \to -\boldsymbol{\sigma}$). For $\alpha < \alpha_c \sim 0.138$ there indeed exist pure state solutions $x \neq 0$. For $\alpha > \alpha_c$ there is only the spin-glass solution $x = 0$. Given a solution x of (113), the zero temperature values for the order parameters follow from

$$\lim_{T\to 0} m = \mathrm{erf}[x], \quad \lim_{T\to 0} C = \left[1 + \sqrt{\frac{\alpha\pi}{2}}\mathrm{e}^{x^2}\right]^{-1}$$

with which in turn we can take the zero temperature limit in our expression (112) for the free energy:

$$\lim_{T\to 0} f = \frac{1}{2}\mathrm{erf}^2[x] + \frac{1}{\pi}\mathrm{e}^{-x^2} - \frac{2}{\pi}\left[\mathrm{e}^{-x^2} + \sqrt{\frac{\alpha\pi}{2}}\right]\left[x\sqrt{\pi}\,\mathrm{erf}[x] + \mathrm{e}^{-x^2}\right].$$

Comparison of the values for $\lim_{T\to 0} f$ thus obtained, for the pure state $m > 0$ and the spin-glass state $m = 0$ leads to Fig. 13, which clearly shows that for sufficiently small α the pure states are the true ground states of the system.

6.2.3. The AT-instability

As in the case of the Gaussian model (72), the above RS solution again generates negative entropies at sufficiently low temperatures, indicating that RS must be broken. We can locate continuous RS breaking by studying the effect on $f(\mathbf{m}, \mathbf{q}, \hat{\mathbf{q}})$ (98) of small replicon [19] fluctuations around the RS solution:

$$q_{\alpha\beta} \to \delta_{\alpha\beta} + q\left[1 - \delta_{\alpha\beta}\right] + \eta_{\alpha\beta}, \quad \eta_{\alpha\beta} = \eta_{\beta\alpha}, \quad \eta_{\alpha\alpha} = 0, \quad \sum_{\alpha}\eta_{\alpha\beta} = 0. \tag{114}$$

The variation of $\mathbf{q}$ induces a similar variation in the conjugate parameters $\hat{\mathbf{q}}$ through Eq. (101):

$$\hat{q}_{\alpha\beta} \to \frac{1}{2}i\alpha\beta^2\left[R\delta_{\alpha\beta} + r\left[1 - \delta_{\alpha\beta}\right] + \hat{\eta}_{\alpha\beta}\right], \quad \hat{\eta}_{\alpha\beta} = \frac{1}{2}\sum_{\gamma\delta}\eta_{\gamma\delta}\left[g_{\alpha\beta\gamma\delta} - g_{\alpha\beta}g_{\gamma\delta}\right]$$

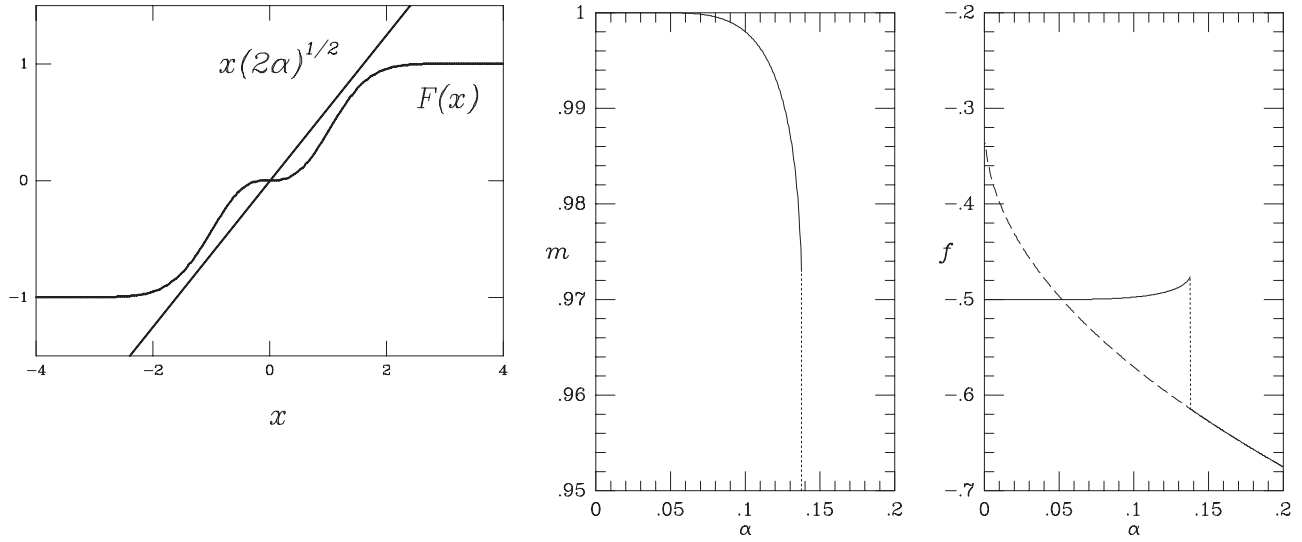

Fig. 13. Left: solution of the transcendental equation $F(x) = x\sqrt{2\alpha}$, where $x = \mathrm{erf}^{\mathrm{inv}}(m)$. The storage capacity $\alpha_c \sim 0.138$ of the Hopfield model is the largest α for which solutions $x \neq 0$ exist. Middle picture: RS amplitudes m of the pure states of the Hopfield model for $T = 0$ as a function of $\alpha = p/N$. The location of the discontinuity, where m vanishes, defines the storage capacity $\alpha_c \sim 0.138$. Right picture, solid line: $T = 0$ 'free energy' f of the pure states. Dashed lines: $T = 0$ 'free energy' of the spin-glass state $\mathbf{m} = \mathbf{0}$ (for comparison).

with

$$g_{\alpha\beta\gamma\delta} = \frac{\int \mathrm{d}\mathbf{z}\, z_\alpha z_\beta z_\gamma z_\delta \mathrm{e}^{-\frac{1}{2}\mathbf{z}\cdot[\mathbf{1}-\beta\mathbf{q}_{\mathrm{RS}}]\mathbf{z}}}{\int \mathrm{d}\mathbf{z}\, \mathrm{e}^{-\frac{1}{2}\mathbf{z}\cdot[\mathbf{1}-\beta\mathbf{q}_{\mathrm{RS}}]\mathbf{z}}}, \quad g_{\alpha\beta} = \frac{\int \mathrm{d}\mathbf{z}\, z_\alpha z_\beta \mathrm{e}^{-\frac{1}{2}\mathbf{z}\cdot[\mathbf{1}-\beta\mathbf{q}_{\mathrm{RS}}]\mathbf{z}}}{\int \mathrm{d}z\, \mathrm{e}^{-\frac{1}{2}\mathbf{z}\cdot[\mathbf{1}-\beta\mathbf{q}_{\mathrm{RS}}]\mathbf{z}}}.$$

Wick's theorem (see e.g. [4]) can now be used to write everything in terms of second moments of the Gaussian integrals only:

$$g_{\alpha\beta\gamma\delta} = g_{\alpha\beta}g_{\gamma\delta} + g_{\alpha\gamma}g_{\beta\delta} + g_{\alpha\delta}g_{\beta\gamma}$$

with which we can express the replicon variation in $\hat{\mathbf{q}}$, using the symmetry of $\{\eta_{\alpha\beta}\}$ and the saddle-point Eq. (101), as

$$\begin{aligned}\hat{\eta}_{\alpha\beta} &= \sum_{\gamma\delta} g_{\alpha\gamma}\eta_{\gamma\delta}g_{\delta\beta} \\ &= \beta^2 \sum_{\gamma\neq\delta} \big[R\delta_{\alpha\gamma} + r\big[1-\delta_{\alpha\gamma}\big]\big]\eta_{\gamma\delta}\big[R\delta_{\delta\beta} + r\big[1-\delta_{\delta\beta}\big]\big] \\ &= \beta^2(R-r)^2\eta_{\alpha\beta} \end{aligned} \tag{115}$$

since only those terms can contribute which involve precisely two δ-symbols, due to $\sum_\alpha \eta_{\alpha\beta} = 0$. We can now calculate the change in $f(\mathbf{m},\mathbf{q},\hat{\mathbf{q}})$, away from the RS value $f(\mathbf{m}_{\mathrm{RS}},\mathbf{q}_{\mathrm{RS}},\hat{\boldsymbol{q}}_{\mathrm{RS}})$, the leading order of which must be quadratic in the fluctuations $\{\eta_{\alpha\beta}\}$ since the RS solution is a saddle-point:

$$\begin{aligned} & f(\mathbf{m}_{\mathrm{RS}},\mathbf{q},\hat{\mathbf{q}}) - f(\mathbf{m}_{\mathrm{RS}},\mathbf{q}_{\mathrm{RS}},\hat{\mathbf{q}}_{\mathrm{RS}}) \\ &\quad = \frac{1}{\beta n}\left[\frac{1}{2}\alpha\log\frac{\det[\mathbf{1}-\beta(\mathbf{q}_{\mathrm{RS}}+\boldsymbol{\eta})]}{\det[\mathbf{1}-\beta\mathbf{q}_{\mathrm{RS}}]} - i\mathrm{Tr}[\hat{\mathbf{q}}_{\mathrm{RS}}\cdot\boldsymbol{\eta}]\right. \\ &\qquad \left. + \frac{1}{2}\alpha\beta^2\mathrm{Tr}[\hat{\boldsymbol{\eta}}\cdot\boldsymbol{\eta} + \hat{\boldsymbol{\eta}}\cdot\mathbf{q}_{\mathrm{RS}}] - \left\langle \log\frac{\langle \mathrm{e}^{\beta\xi\cdot\mathbf{m}_{\mathrm{RS}}\sum_\alpha\sigma_\alpha - i\boldsymbol{\sigma}\cdot[\hat{\mathbf{q}}_{\mathrm{RS}}+\frac{1}{2}i\alpha\beta^2\hat{\boldsymbol{\eta}}]\boldsymbol{\sigma}}\rangle_{\boldsymbol{\sigma}}}{\langle \mathrm{e}^{\beta\xi\cdot\mathbf{m}_{\mathrm{RS}}\sum_\alpha\sigma_\alpha - i\boldsymbol{\sigma}\cdot\hat{\mathbf{q}}_{\mathrm{RS}}\boldsymbol{\sigma}}\rangle_{\boldsymbol{\sigma}}}\right\rangle_\xi\right]. \end{aligned} \tag{116}$$

Evaluating (116) is simplified by the fact that the matrices $\mathbf{q}_{\mathrm{RS}}$ and $\boldsymbol{\eta}$ commute, which is a direct consequence of the properties (114) of the replicon fluctuations and the form of the replica-symmetric saddle-point. If we define the $n \times n$ matrix $\mathbf{P}$ as the projection onto the vector $(1,\ldots,1)$, we have

$$\begin{aligned} & P_{\alpha\beta} = n^{-1}, \quad \mathbf{P}\cdot\boldsymbol{\eta} = \boldsymbol{\eta}\cdot\mathbf{P} = 0, \quad \mathbf{q}_{\mathrm{RS}} = (1-q)\mathbf{1} + nq\mathbf{P}, \\ & \mathbf{q}_{\mathrm{RS}}\cdot\boldsymbol{\eta} = \boldsymbol{\eta}\cdot\mathbf{q}_{\mathrm{RS}} = (1-q)\boldsymbol{\eta} \\ & [\mathbf{1}-\beta\mathbf{q}_{\mathrm{RS}}]^{-1} = \frac{1}{1-\beta(1-q)}\mathbf{1} + \frac{\beta nq}{[1-\beta(1-q)-\beta nq][1-\beta(1-q)]}\mathbf{P}. \end{aligned} \tag{117}$$

We can now simply expand the relevant terms, using the identity $\log\det M = \mathrm{Tr}\,\log M$:

$$\log\frac{\det[\mathbf{1}-\beta(\mathbf{q}_{\rm RS}+\boldsymbol{\eta})]}{\det[\mathbf{1}-\beta\mathbf{q}_{\rm RS}]}=\mathrm{Tr}\,\log\Big[\mathbf{1}-\beta\boldsymbol{\eta}[\mathbf{1}-\beta\mathbf{q}_{\rm RS}]^{-1}\Big]$$
$$=\mathrm{Tr}\Big\{-\beta\boldsymbol{\eta}[\mathbf{1}-\beta\mathbf{q}_{\rm RS}]^{-1}-\frac{1}{2}\beta^2\Big[\boldsymbol{\eta}[\mathbf{1}-\beta\mathbf{q}_{\rm RS}]^{-1}\Big]^2\Big\}+\mathcal{O}(\boldsymbol{\eta}^3)$$
$$=-\frac{1}{2}\frac{\beta^2}{[1-\beta(1-q)]^2}\mathrm{Tr}\,\boldsymbol{\eta}^2+\mathcal{O}(\boldsymbol{\eta}^3) \tag{118}$$

Finally we address the remaining term in (116), again using the RS saddle-point Eqs. (109) and (110) where appropriate:

$$\left\langle\log\frac{\langle e^{\beta\xi\cdot\mathbf{m}_{\rm RS}\sum_\alpha\sigma_\alpha-i\boldsymbol{\sigma}\cdot\hat{\mathbf{q}}_{\rm RS}\boldsymbol{\sigma}}\big[1+\frac{1}{2}\alpha\beta^2\boldsymbol{\sigma}\cdot\hat{\boldsymbol{\eta}}\boldsymbol{\sigma}+\frac{1}{8}\alpha^2\beta^4(\boldsymbol{\sigma}\cdot\hat{\boldsymbol{\eta}}\boldsymbol{\sigma})^2+\cdots\big]\rangle_{\boldsymbol{\sigma}}}{\langle e^{\beta\xi\cdot\mathbf{m}_{\rm RS}\sum_\alpha\sigma_\alpha-i\boldsymbol{\sigma}\cdot\hat{\mathbf{q}}_{\rm RS}\boldsymbol{\sigma}}\rangle_{\boldsymbol{\sigma}}}\right\rangle_\xi$$
$$=\frac{1}{2}\alpha\beta^2\mathrm{Tr}[\hat{\boldsymbol{\eta}}\cdot\mathbf{q}_{\rm RS}]+\frac{1}{8}\alpha^2\beta^4\sum_{\alpha\beta\gamma\delta}\hat{\eta}_{\alpha\beta}\hat{\eta}_{\gamma\delta}[G_{\alpha\beta\gamma\delta}-H_{\alpha\beta\gamma\delta}]+\cdots \tag{119}$$

with

$$G_{\alpha\beta\gamma\delta}=\left\langle\frac{\langle\sigma_\alpha\sigma_\beta\sigma_\gamma\sigma_\delta\,e^{\beta\xi\cdot\mathbf{m}_{\rm RS}\sum_\alpha\sigma_\alpha-i\boldsymbol{\sigma}\cdot\hat{\mathbf{q}}_{\rm RS}\boldsymbol{\sigma}}\rangle_{\boldsymbol{\sigma}}}{\langle e^{\beta\xi\cdot\mathbf{m}_{\rm RS}\sum_\alpha\sigma_\alpha-i\boldsymbol{\sigma}\cdot\hat{\mathbf{q}}_{\rm RS}\boldsymbol{\sigma}}\rangle_{\boldsymbol{\sigma}}}\right\rangle_\xi,$$

$$H_{\alpha\beta\gamma\delta}=\left\langle\frac{\langle\sigma_\alpha\sigma_\beta\,e^{\beta\xi\cdot\mathbf{m}_{\rm RS}\sum_\alpha\sigma_\alpha-i\boldsymbol{\sigma}\cdot\hat{\mathbf{q}}_{\rm RS}\boldsymbol{\sigma}}\rangle_{\boldsymbol{\sigma}}}{\langle e^{\beta\xi\cdot\mathbf{m}_{\rm RS}\sum_\alpha\sigma_\alpha-i\boldsymbol{\sigma}\cdot\hat{\mathbf{q}}_{\rm RS}\boldsymbol{\sigma}}\rangle_{\boldsymbol{\sigma}}}\frac{\langle\sigma_\gamma\sigma_\delta\,e^{\beta\xi\cdot\mathbf{m}_{\rm RS}\sum_\alpha\sigma_\alpha-i\boldsymbol{\sigma}\cdot\hat{\mathbf{q}}_{\rm RS}\boldsymbol{\sigma}}\rangle_{\boldsymbol{\sigma}}}{\langle e^{\beta\xi\cdot\mathbf{m}_{\rm RS}\sum_\alpha\sigma_\alpha-i\boldsymbol{\sigma}\cdot\hat{\mathbf{q}}_{\rm RS}\boldsymbol{\sigma}}\rangle_{\boldsymbol{\sigma}}}\right\rangle_\xi.$$

Inserting the ingredients ((115), (117)–(119)) into expression (116) and rearranging terms shows that the linear terms indeed cancel, and that the term involving $H_{\alpha\beta\gamma\delta}$ does not contribute (since the elements $H_{\alpha\beta\gamma\delta}$ do not depend on the indices for $\alpha\neq\beta$ and $\gamma\neq\delta$), and we are left with:

$$f(\mathbf{m}_{\rm RS},\mathbf{q},\hat{\mathbf{q}})-f(\mathbf{m}_{\rm RS},\mathbf{q}_{\rm RS},\hat{\mathbf{q}}_{\rm RS})$$
$$=\frac{1}{\beta n}\Bigg[-\frac{1}{4}\frac{\alpha\beta^2}{[1-\beta(1-q)]^2}\mathrm{Tr}\,\boldsymbol{\eta}^2+\frac{1}{2}\alpha\beta^4(R-r)^2\mathrm{Tr}\,\boldsymbol{\eta}^2$$
$$-\frac{1}{8}\alpha^2\beta^8(R-r)^4\sum_{\alpha\beta\gamma\delta}\eta_{\alpha\beta}\eta_{\gamma\delta}G_{\alpha\beta\gamma\delta}\Bigg]+\cdots$$

Because of the index permutation symmetry in the neuron average we can write for $\alpha\neq\gamma$ and $\rho\neq\lambda$:

$$G_{\alpha\gamma\rho\lambda}=\delta_{\alpha\rho}\delta_{\gamma\lambda}+\delta_{\alpha\lambda}\delta_{\gamma\rho}+G_4\big[1-\delta_{\alpha\rho}\big]\big[1-\delta_{\gamma\lambda}\big]\big[1-\delta_{\alpha\lambda}\big]\big[1-\delta_{\gamma\rho}\big]$$
$$+G_2\big\{\delta_{\alpha\rho}\big[1-\delta_{\gamma\lambda}\big]+\delta_{\gamma\lambda}\big[1-\delta_{\alpha\rho}\big]+\delta_{\alpha\lambda}\big[1-\delta_{\gamma\rho}\big]+\delta_{\gamma\rho}[1-\delta_{\alpha\lambda}]\big\}$$

with

$$G_\ell = \left\langle \frac{\int \mathrm{D}z \tanh^\ell \beta[\mathbf{m}\cdot\boldsymbol{\xi} + z\sqrt{\alpha r}] \cosh^n \beta[\mathbf{m}\cdot\boldsymbol{\xi} + z\sqrt{\alpha r}]}{\int \mathrm{D}z \cosh^n \beta[\mathbf{m}\cdot\boldsymbol{\xi} + z\sqrt{\alpha r}]} \right\rangle_\xi .$$

Only terms which involve precisely two δ-functions can contribute, because of the replicon properties (114). As a result:

$$\begin{aligned} f(\mathbf{m}_{\mathrm{RS}}, \mathbf{q}, \hat{\mathbf{q}}) - f(\mathbf{m}_{\mathrm{RS}}, \mathbf{q}_{\mathrm{RS}}, \hat{\mathbf{q}}_{\mathrm{RS}}) \\ = \frac{1}{\beta n} \mathrm{Tr}\, \boldsymbol{\eta}^2 \Big[-\frac{1}{4} \frac{\alpha\beta^2}{[1-\beta(1-q)]^2} + \frac{1}{2}\alpha\beta^4 (R-r)^2 \\ - \frac{1}{4}\alpha^2\beta^8 (R-r)^4 [1 - 2\,G_2 + G_4] \Big] + \cdots \end{aligned}$$

Since $\mathrm{Tr}\,\boldsymbol{\eta}^2 = \sum_{\alpha\beta} \eta_{\alpha\beta}^2$, the condition for the RS solution to minimize $f(\mathbf{m}, \boldsymbol{q}, \hat{\mathbf{q}})$, if compared to the 'replicon' fluctuations, is therefore

$$-\frac{1}{[1-\beta(1-q)]^2} + 2\,\beta^2 (R-r)^2 - \alpha\beta^6 (R-r)^4 [1 - 2\,G_2 + G_4] > 0. \tag{120}$$

After taking the limit in the expressions G_ℓ and after evaluating

$$\begin{aligned} \lim_{n\to 0} R = \frac{1}{\beta} \lim_{n\to 0} g_{\alpha\alpha} = \lim_{n\to 0} \frac{1}{n\beta} \frac{\int \mathrm{d}\mathbf{z}\, \mathbf{z}^2 \mathrm{e}^{-\frac{1}{2}\mathbf{z}\cdot[\mathbf{I} - \beta\mathbf{q}_{\mathrm{RS}}]\mathbf{z}}}{\int \mathrm{d}\mathbf{z}\, \mathrm{e}^{-\frac{1}{2}\mathbf{z}\cdot[\mathbf{I} - \beta\mathbf{q}_{\mathrm{RS}}]\mathbf{z}}} \\ = \lim_{n\to 0} \frac{1}{n\beta} \left[\frac{n-1}{1-\beta(1-q)} + \frac{1}{1-\beta(1-q+nq)} \right] = \frac{1}{\beta} \frac{1-\beta+2\,\beta q}{[1-\beta(1-q)]^2} \end{aligned}$$

and using (110), condition (120) can be written as

$$[1-\beta(1-q)]^2 > \alpha\beta^2 \left\langle \int \mathrm{D}z\, \cosh^{-4} \beta\big[\mathbf{m}\cdot\boldsymbol{\xi} + z\sqrt{\alpha r}\big] \right\rangle_\xi \tag{121}$$

The AT line in the phase diagram, where this condition ceases to be met, indicates a second-order transition to a spin-glass state, where ergodicity is broken in the sense that the distribution $\overline{P(q)}$ (104) is no longer a δ-function. In the paramagnetic regime of the phase diagram, $\mathbf{m} = \mathbf{0}$ and $q = 0$, the AT condition reduces precisely to $T > T_{\mathrm{g}} = 1 + \sqrt{\alpha}$. Therefore the paramagnetic solution is stable. The AT line coincides with the boundary between the paramagnetic and spin-glass phase. Numerical evaluation of (121) shows that the RS spin-glass solution remains unstable for all $T < T_{\mathrm{g}}$, but that the retrieval solution $\mathbf{m} \neq \mathbf{0}$ is unstable only for very low temperatures $T < T_{\mathrm{R}}$ (see Fig. 12).

7. Epilogue

In this paper I have tried to give a self-contained exposé of the main issues, models and mathematical techniques relating to the equilibrium statistical mechanical

analysis of recurrent neural networks. I have included networks of binary neurons and networks of coupled (neural) oscillators, with various degrees of synaptic complexity (albeit always fully connected), ranging from uniform synapses, via synapses storing a small number of patterns, to Gaussian synapses and synapses encoding an extensive number of stored patterns. The latter (complex) cases I only worked out for binary neurons; similar calculations can be done for coupled oscillators (see [16]). Networks of graded response neurons could not be included, because these are found never to go to (detailed balance) equilibrium, ruling out equilibrium statistical mechanical analysis. All analytical results and predictions have later also been confirmed comprehensively by numerical simulations. Over the years we have learned an impressive amount about the operation of recurrent networks by thinking in terms of free energies and phase transitions, and by having been able to derive explicit analytical solutions (since a good theory always supersedes an infinite number of simulation experiments ...). I have given a number of key references along the way; many could have been added but were left out for practical reasons. Instead I will just mention a number of textbooks in which more science as well as more references to research papers can be found. Any such selection is obviously highly subjective, and I wish to apologize beforehand to the authors which I regret to have omitted. Several relevant review papers dealing with the statistical mechanics of neural networks can be found scattered over the three volumes [20–22]. Textbooks which attempt to take the interested but nonexpert reader towards the expert level are [8,23]. Finally, a good introduction to the methods and backgrounds of replica theory, together with a good collection of reprints of original papers, can be found in [24].

What should we expect for the next decades, in the equilibrium statistical mechanics of recurrent neural networks? Within the confined area of large symmetric and fully connected recurrent networks with simple neuron types we can now deal with fairly complicated choices for the synapses, inducing complicated energy landscapes with many stable states, but this involves nontrivial and cutting-edge mathematical techniques. If our basic driving force next is the aim to bring our models closer to biological reality, balancing the need to retain mathematical solvability with the desire to bring in more details of the various electro-chemical processes known to occur in neurons and synapses and spatio-temporal characteristics of dendrites, the boundaries of what can be done with equilibrium statistical mechanics are, roughly speaking, set by the three key issues of (presence or absence of) detailed balance, system size, and synaptic interaction range. The first issue is vital: no detailed balance immediately implies no equilibrium statistical mechanics. This generally rules out networks with nonsymmetric synapses and all networks of graded response neurons (even when the latter are equipped with symmetric synapses). The issue of system size is slightly less severe; models of networks with $N < \infty$ neurons can often be solved in leading order in $N^{-\frac{1}{2}}$, but a price will have to be paid in the form of a reduction of our ambition elsewhere (e.g. we might have to restrict ourselves to simpler choices of synaptic interactions). Finally, we know how to deal with fully connected models (such as those discussed in this paper), and also with models having dendritic structures which cover a long

(but not infinite) range, provided they vary smoothly with distance. We can also deal with short-range dendrites in one-dimensional (and to a lesser extent two-dimensional) networks; however, since even the relatively simple Ising model (mathematically equivalent to a network of binary neurons with uniform synapses connecting only nearest-neighbor neurons) has so far not yet been solved in three dimensions, it is not realistic to assume that analytical solution will be possible soon of general recurrent neural network models with short range interactions. On balance, although there are still many interesting puzzles to keep theorists happy for years to come, and although many of the model types discussed in this text will continue to be useful building blocks in explaining at a basic and qualitative level the operation of specific recurrent brain regions (such as the CA3 region of the hippocampus), one is therefore led to the conclusion that equilibrium statistical mechanics has by now brought us as far as can be expected with regard to increasing our understanding of biological neural networks. Dale's law already rules out synaptic symmetry, and thereby equilibrium statistical mechanics altogether, so we are forced to turn to dynamical techniques if we wish to improve biological realism.

Acknowledgements

It is my pleasure to thank David Sherrington and Nikos Skantzos for their direct and indirect contributions to this review.

References

1. Van Kampen, N.G. (1992) Stochastic Processes in Physics and Chemistry. North-Holland, Amsterdam.
2. Gardiner, C.W. (1994) Handbook of Stochastic Methods. Springer, Berlin.
3. Khalil, H.K. (1992) Nonlinear Systems. MacMillan, New York.
4. Zinn-Justin, J. (1993) Quantum Field Theory and Critical Phenomena. U.P., Oxford.
5. Yeomans, J.M. (1992) Statistical Mechanics of Phase Transitions. U.P., Oxford.
6. Plischke, M. and Bergersen, B. (1994) Equilibrium Statistical Mechanics. World Scientific, Singapore.
7. Peretto, P. (1984) Biol. Cybern. **50**, 51.
8. Peretto, P. (1992) An Introduction to the Theory of Neural Computation. U.P., Cambridge.
9. Hopfield, J.J. (1982) Proc. Natl. Acad. Sci. USA. **79**, 2554.
10. Hebb, D.O. (1949) The Organization of Behaviour. Wiley, New York.
11. Amari, S.-I. (1977) Biol. Cybern. **26**, 175.
12. Amit, D.J., Gutfreund, H. and Sompolinsky, H. (1985) Phys. Rev. A **32**, 1007.
13. Fontanari, J.F. and Köberle, R. (1988) J. Physique **49**, 13.
14. Kuramoto, Y. (1984) Chemical Oscillations, Waves and Turbulence. Springer, Berlin.
15. Abramowitz, M. and Stegun, I.A. (1972) Handbook of Mathematical Functions. Dover, New York.
16. Cook, J. (1989) J. Phys. A **22**, 2057.
17. Sherrington, D. and Kirkpatrick, S. (1975) Phys. Rev. Lett. **35**, 1972.
18. Amit, D.J., Gutfreund, H. and Sompolinsky, H. (1985) Phys. Rev. Lett. **55**.
19. de Almeida, J.R.L. and Thouless, D.J. (1978) J. Phys. A **11**, 983.
20. Domany, E., van Hemmen, J.L. and Schulten, K. eds (1991) Models of Neural Networks I. Springer, Berlin.
21. Domany, E., van Hemmen, J.L. and Schulten, K. eds (1994) Models of Neural Networks II. Springer, Berlin.

22. Domany, E., van Hemmen, J.L. and Schulten, K. eds (1995) Models of Neural Networks III. Springer, Berlin.
23. Coolen, A.C.C. and Sherrington, D. (2000) Statistical Physics of Neural Networks. U.P., Cambridge.
24. Mézard, M., Parisi, G. and Virasoro, M.A. (1987) Spin-Glass Theory and Beyond. World Scientific, Singapore.