Dynamics of temporal activity in multi-state neural networks

1997 J. Phys. A: Math. Gen. 30 2637

(http://iopscience.iop.org/0305-4470/30/8/012)

J. Phys. A: Math. Gen. **30** (1997) 2637–2652. Printed in the UK PII: S0305-4470(97)78149-6

Dynamics of temporal activity in multi-state neural networks

G M Shim†‖, K Y M Wong‡¶ and D Bollé§+

† Condensed Matter section, ICTP, 34014 Trieste, Italy
‡ Department of Physics, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
§ Instituut voor Theoretische Fysica, KU Leuven, B-3001 Leuven, Belgium

Received 18 September 1996

Abstract. We consider the behaviour of multi-state neural networks averaged over an extended monitoring period of their dynamics. Pattern reconstruction by clipping the activities is proposed, leading to an improvement in retrieval precision.

1. Introduction

Artificial neural networks have been widely applied to memorize and retrieve information. Recently, a study of *activity dynamics* in binary neural networks showed that information can be retrieved much more accurately when the network behaviour is monitored over an extended period of time, rather than at a particular instant [1]. Furthermore, the basins of attraction can be widened, and the storage capacity can be increased.

To understand this, we note that the instantaneous network state may be degenerated by the presence of noises (arising from the interference of concurrently stored information, or the stochasticity of the retrieval dynamics). In contrast, measuring the averaged network state, or the *temporal activities*, allows noises to be averaged out. In general, their distributions are strongly biased in the direction of the stored patterns. We may then reconstruct the information bits by quantizing the activities, which is referred to as *activity clipping*.

For example, an information bit may be $+1$ or -1 in two-state networks. During the dynamical evolution, the bit may flip between the two states. If it spends more time, on average, in the $+1$ state, then it has a positive temporal activity. Activity clipping is done by assigning the retrieved state to be the more frequently occurring one, namely $+1$ in this example. This is equivalent to making a Bayesian decision on the information bits. In extremely diluted binary networks, the improvement in retrieval quality is so drastic that, for example, the overlap undergoes a discontinuous transition from the retrieval phase to the non-retrieval phase, in contrast to the continuous one for ordinary dynamics [2]. More sophisticated sequences of activity clipping lead to further improvement.

‖ Present address: Department of Physics, Chungnam National University, Yuseong, Taejon 305-764, Korea. E-mail address: gmshim@nsphys.chungnam.ac.kr
¶ E-mail address: phkywong@usthk.ust.hk
+ Also at: Interdisciplinair Centrum voor Neurale Netwerken, KU Leuven, Belgium. E-mail address: desire.bolle@fys.kuleuven.ac.be

0305-4470/97/082637+16$19.50

Since multi-state networks and analogue networks are widely implemented to process information with a graded or grey scale, it is practically useful to generalize such techniques of pattern reconstruction from two-state to multi-state networks [3–5]. However, the generalization to multi-state networks is not so straightforward. One major problem is the determination of the quantization boundaries for clipping the temporal activities, since there is much more freedom in multi-state networks. The present paper addresses this problem.

Concretely, we study the dynamics of temporal activity in multi-state neural networks, and propose two schemes of activity clipping: the maximum likelihood clipping and Bayesian clipping. In the maximum likelihood clipping, one monitors the dynamical state of each neuron, and reconstructs the pattern bit on a node by assigning the most frequently occurring state to it. Though this clipping scheme sounds natural, it does not make full use of the information available in the activity distribution. The quantization boundaries for activity clipping turn out to be independent of the network parameters such as the storage level, and hence the scheme is inflexible. Thus, we also propose the Bayesian clipping scheme, in which one reconstructs the pattern bit by assigning the state with the maximum Bayesian probability to it, given its dynamical activity and a knowledge of the state dependence of the activity distribution. This results in quantization boundaries for activity clipping which adjust with the network parameters. Indeed, it yields excellent retrieval quality.

In section 2 we introduce the relevant dynamical variables for activity dynamics in two kinds of network architecture with exactly solvable dynamics: extremely diluted networks [6] and layered networks [7]. In section 3 we discuss freezing transitions [2], and in sections 4 and 5 we introduce the maximum likelihood clipping and Bayesian clipping, respectively. Simulation results are presented in section 6, followed by the conclusion in section 7. In the appendix we derive the evolution relations describing the dynamics of temporal activity.

2. Activity dynamics in multi-state neural networks

2.1. A review of previous results

Consider a layered neural network composed of multi-state neurons arranged in layers, each layer containing N neurons. A neuron can take values in a discrete set $\mathcal{S} \equiv \{-1 = s_1 < s_2 < \cdots < s_{Q-1} < s_Q = +1\}$. The elements in the set are equidistant in general. Each neuron in layer t is unidirectionally feeding all neurons on layer $t+1$. Given a configuration $\{\sigma_j(t),\ j = 1, 2, \ldots, N\}$, the local field $h_i(t+1)$ in neuron i on layer $t+1$ is

$$h_i(t+1) = \sum_j J_{ij}(t+1)\sigma_j(t) \tag{1}$$

where $J_{ij}(t+1)$ is the strength of the coupling from neuron j on layer t to neuron i on layer $t+1$. The state $\{\sigma_j(t+1)\}$ of layer $t+1$ is determined by the state $\{\sigma_j(t)\}$ of the previous layer t according to the zero-temperature updating rule

$$\sigma_i(t+1) = g(h_i(t+1)) \tag{2}$$

where

$$g(h) \equiv s_k \qquad \text{for } b(s_{k-1} + s_k) \leqslant h \leqslant b(s_k + s_{k+1}) \qquad k = 1, \ldots, Q \tag{3}$$

with $b > 0$ and $s_0 \equiv -\infty$, $s_{Q+1} \equiv \infty$. For finite Q, $g(h)$ is a series of step functions. The gain parameter, b^{-1}, controls the average slope of the transfer function $g(h)$.

The network receives an input configuration in the first layer $\{\sigma_j(t=1)\}$. Updating in subsequent layers then proceeds in parallel: at the next time step, the second layer is updated according to rule (2), and so on.

The stored patterns on layer t are a collection of independent and identically distributed random variables (iidrv) $\{\xi_i^\mu(t) \in \mathcal{S}\}$, $\mu \in \{1, 2, \ldots, p = \alpha N\}$. Their distribution is specified by the probabilities $p(s_k)$ that $\xi_i^\mu(t) = s_k$, $k = 1, \ldots, Q$, with the normalization $\sum_k p(s_k) = 1$. In this paper we are interested in uniform pattern distributions with zero mean. For neural networks with high connectivity, it turns out that only the variance $A = \mathrm{Var}[\xi_i^\mu(t)]$ is relevant to the network dynamics.

The task of the layered network is to retrieve, with high precision, a given pattern on each layer when a noisy version of that pattern is input in the first layer. To achieve this the pattern information has to be encoded in the synaptic couplings between adjacent layers according to some learning rule. Here we consider the Hebb rule

$$J_{ij}(t+1) = \frac{1}{NA}\sum_\mu \xi_i^\mu(t+1)\xi_j^\mu(t). \tag{4}$$

The Hebb rule is not the most efficient learning rule. Previous studies have shown that it has a low retrieval precision and storage capacity [4], and other better learning rules exist. However, as will be shown, activity clipping is able to improve the retrieval precision in the retrieval phase.

Since the patterns on different layers are chosen independently, the simple form of the Hebb rule allows the possibility of an analytic treatment of the dynamics. The following recursion relations were obtained in [4], starting from the initial data $\{\sigma_j(1)\}$ being a collection of iidrv and correlated with only one stored pattern, say $\mu = 1$,

$$m^\mu(t+1) = \delta_{\mu,1}\frac{1}{A}\left\langle\!\!\left\langle \xi^1(t+1)\int \mathrm{D}z\; g\left(\xi^1(t+1)m^1(t) + \sqrt{D(t)}z\right)\right\rangle\!\!\right\rangle \tag{5}$$

$$a(t+1) = \left\langle\!\!\left\langle \int \mathrm{D}z\; g^2\left(\xi^1(t+1)m^1(t) + \sqrt{D(t)}z\right)\right\rangle\!\!\right\rangle \tag{6}$$

$$D(t+1) = \alpha a(t+1) + \left[\left\langle\!\!\left\langle \int \mathrm{D}z\; zg\left(\xi^1(t+1)m^1(t) + \sqrt{D(t)}z\right)\right\rangle\!\!\right\rangle\right]^2. \tag{7}$$

where $m^\mu(t)$, $a(t)$ and $D(t)$ are the *overlap*, *spatial activity* and *noise*, respectively defined as

$$m^\mu(t) \equiv \lim_{N\to\infty}\frac{1}{NA}\sum_i \xi_i^\mu(t)\sigma_i(t) \tag{8}$$

$$a(t) \equiv \lim_{N\to\infty}\frac{1}{N}\sum_i \sigma_i^2(t) \tag{9}$$

$$D(t) \equiv \lim_{N\to\infty}\frac{1}{N}\sum_i [h_i(t+1) - \xi_i^1(t+1)m^1(t)]^2. \tag{10}$$

In dilute networks each neuron, σ_i, is fed by C randomly chosen neurons, σ_j, through the synaptic weights J_{ij}. The updating rule is still given by (2), but for $h_i(t) = \sum_j J_{ij}\sigma_j(t)$, the summation runs over the chosen neurons feeding neuron i. For the Hebb rule storing $p \equiv \alpha C$ patterns, $J_{ij} = \sum_\mu \xi_i^\mu\xi_j^\mu/CA$.

The evolution equations can be analytically derived for extremely diluted networks where it can be shown (see, e.g. [8]) that the probability of two sites, i and i', having disjoint clusters of ancestors approaches $\exp(-C^{2n}/N)$ for $N \gg 1$ with n being the number of time steps in the dynamics. This implies that for $C \ll \ln N$, feedback is completely suppressed

irrespective of n. In practice it turns out that correlations appear when $n \approx \mathcal{O}(\ln N / \ln C)$. So unless we are very near the region of a phase transition the number of time steps required in order to reach the steady state is mainly small. In this case the condition for strong dilution can be relaxed to $\ln C << \ln N$ for $N \to \infty$ [9]. The evolution equations governing the parallel dynamics for extremely diluted networks are given by [6, 9]

$$m^{\mu}(t+1) = \delta_{\mu,1} \frac{1}{A} \left\langle\!\!\left\langle \xi^1 \int \mathrm{D}z \; g\left(\xi^1 m^1(t) + \sqrt{\alpha a(t)}z\right) \right\rangle\!\!\right\rangle \tag{11}$$

$$a(t+1) = \left\langle\!\!\left\langle \int \mathrm{D}z \; g^2\left(\xi^1 m^1(t) + \sqrt{\alpha a(t)}z\right) \right\rangle\!\!\right\rangle. \tag{12}$$

2.2. Dynamical variables for activity dynamics

Until now we have been considering the network dynamics for a single initial condition. To understand the temporal activity of a neuron in the network, consider an *ensemble* of input configurations in layered networks [1], or an *ensemble* of initial conditions in dilute networks [10]. The ensemble averaged state of a node is called its *activity*. In layered networks we may consider presenting an input configuration at each time step, and the ensemble averaged network behaviour becomes the temporal averaged behaviour. In dilute networks, the network state converges to a single chaotic attractor for a single set of macroscopic variables [2], instead of the cloud of attractors suggested in earlier studies [6], therefore the ensemble averaged asymptotic state is the same as the temporal averaged state in the attractor. We will hereafter treat temporal averaging and ensemble averaging as being equivalent, and refer to them as *activity dynamics*.

In activity dynamics, there are three relevant dynamical variables: the overlap, $m(t)$, the dynamic noise, $u(t)$, and the frozen noise $v(t)$. The dynamic noise is the variance of the local fields about their ensemble averaged values. It describes the dynamical variations from one input configuration to another within the ensemble. The frozen noise is the variance of the temporal averaged local fields about their mean value when the nodes, i, are sampled. It describes the spatial variations independent of particular input configurations in the ensemble. As will be shown, the sum of the two noise terms, $u(t) + v(t)$, is the noise term $D(t)$ introduced previously for the ordinary dynamics. The distinction between the dynamic and frozen nature of the noise is irrelevant for ordinary dynamics, since it describes the *instantaneous* state of the network. However, this disctinction is needed to describe the activity dynamics, since they play different roles in determining the temporal activity of a node.

Consider an ensemble of input configurations that are a collection of iidrv with overlap and variance respectively given by (8) and (9) for $t = 1$, $\mu = 1$. It is natural to split the local field on site i into two terms

$$h_i(t+1) = \sum_j J_{ij}(t+1)\langle\sigma_j(t)\rangle + X_i(t+1) \tag{13}$$

with

$$X_i(t+1) = \sum_j J_{ij}(t+1)(\sigma_j - \langle\sigma_j(t)\rangle). \tag{14}$$

The brackets $\langle\ \rangle$ denote the average over this ensemble. When varying the input configurations the first term on the r.h.s. of equation (13) remains constant while $X_i(t+1)$ fluctuates around its mean zero. Inserting the learning rule (4) for the couplings into equation (14) and noting that the patterns $\{\xi_i^{\mu}(t+1)\}$ on layer $t+1$ are uncorrelated with

the patterns $\{\xi_i^\mu(t)\}$ and the state $\{\sigma_i(t)\}$ on layer t, we see that $X_i(t+1)$ is a Gaussian random variable with mean zero and variance $u(t)$ given by

$$u(t) = \frac{1}{N^2 A} \sum_\mu \sum_{j,k} \xi_j^\mu(t)\xi_k^\mu(t)[\langle\sigma_j(t)\sigma_k(t)\rangle - \langle\sigma_j(t)\rangle\langle\sigma_k(t)\rangle]. \tag{15}$$

Note that the dynamic noise, $u(t)$, is now independent of the node label i, and is determined only by the variables on the previous layer t.

The temporal averaged local fields in (13) can in turn be decomposed into a signal and a noise term

$$\sum_j J_{ij}(t+1)\langle\sigma_j(t)\rangle = \xi_i^\mu(t+1)m(t) + Y_i(t+1) \tag{16}$$

with

$$m(t) = \frac{1}{NA} \sum_j \xi_j^1(t)\langle\sigma_j(t)\rangle \tag{17}$$

$$Y_i(t+1) = \frac{1}{NA} \sum_{\mu>1} \sum_j \xi_i^\mu(t+1)\xi_j^\mu(t)\langle\sigma_j(t)\rangle. \tag{18}$$

Since $\{\xi_i^\mu(t+1)\}$ are uncorrelated with both $\{\langle\sigma_j(t)\rangle\}$ and $\{\xi_j^\mu(t)\}$, $Y_i(t+1)$ becomes a Gaussian random variable with mean zero and variance

$$v(t) = \frac{1}{N^2 A} \sum_{\mu>1} \sum_{j,k} \xi_j^\mu(t)\xi_k^\mu(t)\langle\sigma_j(t)\rangle\langle\sigma_k(t)\rangle. \tag{19}$$

Again, note that the frozen noise $v(t)$ is independent of the node label i, and is determined only by variables on the previous layer t.

Recursion relations describing the dynamics of temporal activity in the network are, as derived in the appendix,

$$m(t+1) = \frac{1}{A}\left\langle\!\!\left\langle \xi(t+1) \int \mathrm{D}z\, g\left(\xi(t+1)m(t) + \sqrt{u(t)+v(t)}z\right)\right\rangle\!\!\right\rangle \tag{20}$$

$$u(t+1) = \alpha[a(t+1) - C(t+1)] + \frac{u(t)}{u(t)+v(t)}\left[\left\langle\!\!\left\langle \int \mathrm{D}z\, zg\left(\xi(t+1)m(t) + \sqrt{u(t)+v(t)}z\right)\right\rangle\!\!\right\rangle\right]^2 \tag{21}$$

$$v(t+1) = \alpha C(t+1) + \frac{v(t)}{u(t)+v(t)}\left[\left\langle\!\!\left\langle \int \mathrm{D}z\, zg\left(\xi(t+1)m(t) + \sqrt{u(t)+v(t)}z\right)\right\rangle\!\!\right\rangle\right]^2. \tag{22}$$

We note that $u(t) + v(t) = D(t)$, the noise term (10) in ordinary dynamics (with a single input configuration).

For the extremely diluted networks we recall that the architecture is a directed tree so that the correlations among the ancestors feeding a given node are negligible. Therefore the recursion relations can easily be read off from those for the layered feedforward networks by neglecting the off-diagonal terms in the dynamic and frozen noises, arriving at

$$m(t+1) = \frac{1}{A}\left\langle\!\!\left\langle \xi \int \mathrm{D}z\, g\left(\xi m(t) + \sqrt{u(t)+v(t)}z\right)\right\rangle\!\!\right\rangle \tag{23}$$

$$u(t+1) = \alpha[a(t+1) - C(t+1)] \tag{24}$$

$$v(t+1) = \alpha C(t+1). \tag{25}$$

3. Freezing transitions

To measure the performance of the temporal averaged behaviour of the network, it is interesting to consider the distribution of the temporal averaged states for a given pattern bit ξ being one of the Q possible states:

$$P_t(f_1, \ldots, f_Q|\xi) = \lim_{N\to\infty} \frac{Q}{N} \sum_{\xi_i^1(t)=\xi} \prod_{k=1}^{Q} \delta[\langle \delta_{g(h_i),s_k} \rangle - f_k] \tag{26}$$

where f_k is the fraction of time a given node stays in state s_k. For perfect retrieval, the distribution should be a delta peak at $f_k = \delta_{\xi,s_k}$. Since the local field $h_i(t+1)$ is a Gaussian variable with mean $\xi m(t)$, temporal variance $u(t)$ and spatial variance $v(t)$, we have

$$P_{t+1}(f_1, \ldots, f_Q|\xi) = \int \mathrm{D}y \prod_k \delta \left[\frac{1}{2} \operatorname{erf}\left(\frac{b(s_k + s_{k+1}) - \xi m(t) - \sqrt{v(t)}y}{\sqrt{2u(t)}} \right) - \frac{1}{2} \operatorname{erf}\left(\frac{b(s_{k-1} + s_k) - \xi m(t) - \sqrt{v(t)}y}{\sqrt{2u(t)}} \right) - f_k \right]. \tag{27}$$

It is often convenient to merely consider the distribution function of the rth moment, η_{ri}, of the temporal activities of node i, defined by

$$\eta_{ri} \equiv \sum_{k=1}^{Q} \langle \delta_{g(h_i),s_k} \rangle s_k^r. \tag{28}$$

The distribution function is given by

$$\mathcal{D}_{t+1}^{(r)}(\eta_r|\xi) = \int \mathrm{D}y\, \delta\left(\eta_r - \int \mathrm{D}x\, g^r \left(\xi m(t) + \sqrt{u(t)}x + \sqrt{v(t)}\, y \right) \right). \tag{29}$$

Simulation results in figures 1(a) and (b) show the evolution of the activity distributions $\mathcal{D}^{(1)}(\eta_1|\xi)$ for the pattern bits $\xi = 1$, 0 up to 20 layers in the layered $Q = 3$ network with uniformly distributed stored patterns ($A = \frac{2}{3}$). It is evident that the activity distribution is highly dependent on the different pattern bits. For instance, the distribution for $\xi = 1$ is strongly biased towards activities near the value of 1, whereas the distribution is much more even for the case of $\xi = 0$. In later sections this differentiation will be utilized for improving the precision in pattern reconstruction.

We note that in the activity distributions, there is a divergence occurring at the values of $\eta_1 = \pm 1$. We say that the network is in a *partially frozen* phase, since this corresponds to nodes with overwhelming probability of aligning with (or against) the pattern bits throughout the ensemble averaging process. If no such divergence appears, we say that the network is in an *unfrozen* phase. As will be shown, there is a *freezing transition* from the unfrozen phase to the partially frozen phase as α decreases in extremely diluted networks.

We remark that while *retrieval* and *freezing transitions* are related to the non-ergodicity of the network dynamics, they are different notions. Retrieval implies that the dynamics converges to a non-ergodic attractor highly correlated with one of the stored patterns, which may be a fixed point, a cycle or a chaotic attractor. Thus, a retrieved state may either be in a partially frozen phase (which corresponds to a less chaotic attractor) or an unfrozen phase (which corresponds to a more chaotic attractor). Similarly, a non-retrieval attractor may either be unfrozen or partially frozen, though it turns out that the non-retrieval states in extremely diluted networks are unfrozen.

To consider the conditions for the occurrence of a partially frozen phase, we restrict ourselves to networks with $Q = 3$ with uniformly distributed stored patterns ($A = \frac{2}{3}$). As

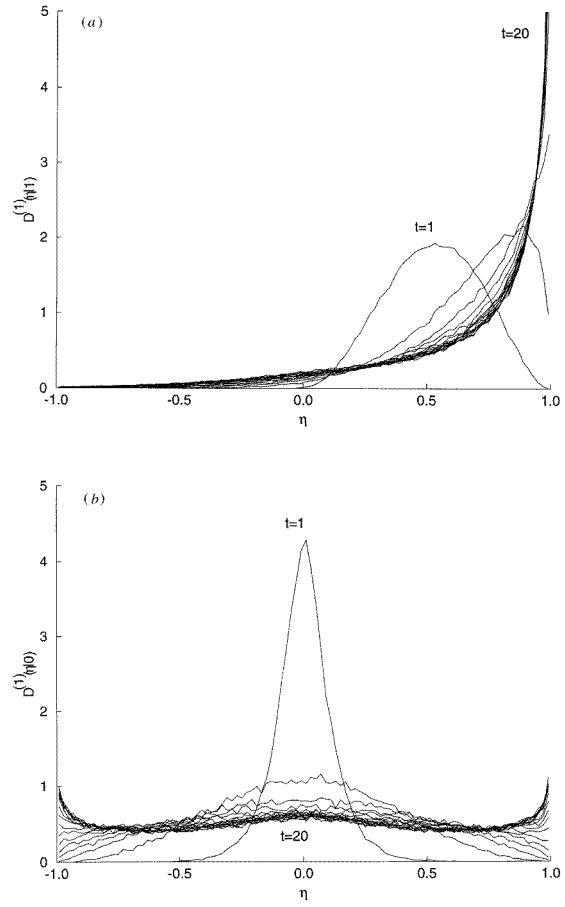

Figure 1. The activity distribution $\mathcal{D}^{(1)}(\eta_1|\xi)$ in layered networks from $t = 1$ to $t = 20$ for (*a*) $\xi = 1$ and (*b*) $\xi = 0$. We have used $N = 300$, $p = 40$, $b = 0.3$, $m(1) = 0.3$, $a(1) = 0.3$, averaged over 300 samples with an ensemble size of 500 in the simulations.

$t \to \infty$ the macroscopic variables $m(t)$, $u(t)$ and $v(t)$ converge to the fixed-point values m, u and v, respectively. The activity distribution obtains the form

$$\mathcal{D}^{(1)}(\eta_1|\xi) = \sqrt{\frac{u}{v}}\frac{\exp[-\frac{u}{v}(Y_\eta - \frac{\xi m}{\sqrt{2u}})^2]}{\exp[-(Y_\eta + \frac{b}{\sqrt{2u}})^2] + \exp[-(Y_\eta - \frac{b}{\sqrt{2u}})^2]} \tag{30}$$

where Y_η is the solution of the equation

$$\operatorname{erf}\left(\frac{b}{\sqrt{2u}} + Y\right) - \operatorname{erf}\left(\frac{b}{\sqrt{2u}} - Y\right) = 2\eta_1. \tag{31}$$

If $u < v$, the reduced distribution diverges for any ξ at $\eta_1 = \pm 1$, corresponding to a partially frozen phase in the system. For $u > v$, these divergences disappear, corresponding to an unfrozen phase. As shown in figure 2 for extremely diluted networks, a freezing transition exists for low values of b, where b is the inverse gain parameter (for $Q = 3$, b is also the zero-step size, i.e. the local field for the pattern bit $\xi = 0$ lies between $\pm b$). The entire non-retrieval phase is unfrozen. This is analogous to the freezing transition in two-state networks [2]. On the other hand, in feedforward layered networks, there is only the partially frozen

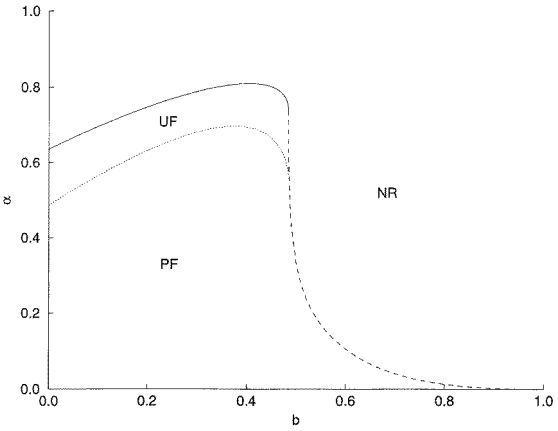

Figure 2. The freezing transition (dotted curve) in the space of the storage level α and the zero-step size b for the extremely diluted network. The full curve represents a continuous transition from the retrieval to the non-retrieval states while the broken curve represents a discontinuous one. PF, UF and NR stand for the partially frozen, unfrozen and non-retrieval phases respectively.

phase in the entire retrieval regime. It seems that in the layered feedforward networks the correlation effects among different sites enhance the partially frozen phase.

However, there is a difference between two- and three-state networks in the non-retrieval regime. In two-state networks the system is in the unfrozen phase for both the extremely diluted structure and the layered structure (in the asymptotic limit). In contrast, for three-state networks there is a partially frozen phase in the layered network for any value of b, and in the diluted network for $b > 0.48$.

4. The maximum likelihood clipping rule

As we have seen, neurons are not completely frozen in either extremely diluted or in layered feedforward three-Ising networks. In general, the neurons keep flipping. If one monitors a single neuron at the stationary state over an extended period, one obtains a sequence of neuron states whose statistics is described by the first and second moments η_1, η_2.

The problem is then how to select out of this sequence the state that is the same as the neuron state in the retrieved pattern. One of the natural choices is to take the state that appears most frequently in the hope that it is strongly correlated with the state in the pattern:

$$\sigma^{\mathrm{M}} = s_k : k = \arg\max_l(f_l). \tag{32}$$

For three-state networks the temporal activities f_k can be expressed in terms of the first and second moments η_1 and η_2. Since the fractions of time for a node to be in states $\pm 1, 0$ are $(\eta_1 \pm \eta_2)/2$ and $1 - \eta_2$ respectively, the maximum likelihood rule can be summarized as

$$\sigma^{\mathrm{M}} = \mathrm{sign}(\eta_1)\Theta(3\eta_2 + |\eta_1| - 2). \tag{33}$$

Using the activity distribution (29), one may calculate the overlap and spatial activity of the maximum likelihood clipping:

$$m^{\mathrm{M}} = \frac{1}{2}\,\mathrm{erf}\left(\frac{m}{\sqrt{2v}} + \sqrt{\frac{u}{v}}Y^{\mathrm{M}}\right) + \frac{1}{2}\,\mathrm{erf}\left(\frac{m}{\sqrt{2v}} - \sqrt{\frac{u}{v}}Y^{\mathrm{M}}\right) \tag{34}$$

$$a^{\mathrm{M}} = 1 - \frac{1}{3}\,\mathrm{erf}\left(\frac{m}{\sqrt{2v}} + \sqrt{\frac{u}{v}}Y^{\mathrm{M}}\right) + \frac{1}{3}\,\mathrm{erf}\left(\frac{m}{\sqrt{2v}} - \sqrt{\frac{u}{v}}Y^{\mathrm{M}}\right) - \frac{1}{3}\,\mathrm{erf}\left(\sqrt{\frac{u}{v}}Y^{\mathrm{M}}\right) \tag{35}$$

where Y^{M} is the positive solution of the equation

$$\mathrm{erf}\left(\frac{b}{\sqrt{2u}} + Y\right) + 2\,\mathrm{erf}\left(\frac{b}{\sqrt{2u}} - Y\right) = 1 \tag{36}$$

in its region of existence $\mathrm{erf}(b/\sqrt{2u}) > \frac{1}{3}$, and zero otherwise. Retrieval precision can be measured by the performance

$$F(t) \equiv \lim_{N\to\infty} \frac{1}{N} \sum_i \delta_{\xi_i^1(t),\sigma_i(t)} \tag{37}$$

which counts the relative number of correct bits. For the maximum likelihood clipping rule, the result is

$$F^{\mathrm{M}} = \frac{1}{3}\left[1 + \mathrm{erf}\left(\frac{m}{\sqrt{2v}} - \sqrt{\frac{u}{v}}Y^{\mathrm{M}}\right) + \mathrm{erf}\left(\sqrt{\frac{u}{v}}Y^{\mathrm{M}}\right)\right]. \tag{38}$$

For dilute networks at $b = 0.2$, figures 3(a)–(c) show that when α is small, both the ordinary and maximum likelihood curves are very close together. As α approaches the critical storage capacity, α_c, there is a significant difference: m^{M} remains finite while m goes to zero. This illustrates the advantage of the clipping procedure in precisely retrieving the non-zero pattern bits. In fact, the improvement in retrieval precision is most marked just below the storage capacity, where the maximum likelihood overlap undergoes a discontinuous transition instead of the continuous one in the ordinary overlap. This behaviour has also been found for extremely diluted binary networks [2]. However, the improvement in retrieval precision is less marked for larger values of b. Figures 4(a)–(c) show the relevant curves with $b = 0.4$, where both m^{M} and m undergo a continuous transition.

For a sufficiently low value of b and near the storage limit, maximum likelihood clipping may suffer from *overclipping*, i.e. it fails to retrieve the zero pattern bits by mistakenly clipping them to non-zero values. This can be seen from the clipping rule (33). When the lower bound of the expression $3\eta_2 + |\eta_1| - 2$ is positive, the clipped state can never be zero. This happens when $\mathrm{erf}(b/\sqrt{2u}) > \frac{1}{3}$. As shown in figure 3($b$) for $b = 0.2$, the spatial activity of the maximum likelihood clipping becomes unity for $\alpha > 0.61$, which means that almost all states are either -1 or 1. This indicates that there is room for improvement for the maximum likelihood clipping. For the case of the larger value of b in figure 4(b), the local interval for zero states is sufficiently wide, and zero clipped states are possible.

The retrieval quality of the maximum likelihood rule can also be measured by the Hamming distance between the state and the retrieved pattern, i.e. $\langle[\xi_i(t) - \sigma_i(t)]^2\rangle$. It turns out that in some region of the retrieval phase the maximum likelihood rule is not better than ordinary dynamics. Figure 5 shows that region in the case of extremely diluted networks.

For layered feedforward networks the overlap, m, remains finite and the storage capacity is rather small, implying that the maximum likelihood overlap and activity are rather similar to the ordinary ones. A numerical calculation confirms this picture. Like the dilute network, Hamming distances between the stored pattern and the retrieved pattern are often less in the case of maximum likelihood clipping. However, there exist regions in which the maximum likelihood clipping performs worse, showing that the clipping rule is not optimal.

5. The Bayesian clipping rule

The Bayesian clipping rule is better than the maximum likelihood rule. This improvement is obtained by maximizing the (conditional) probability of the observed temporal activities,

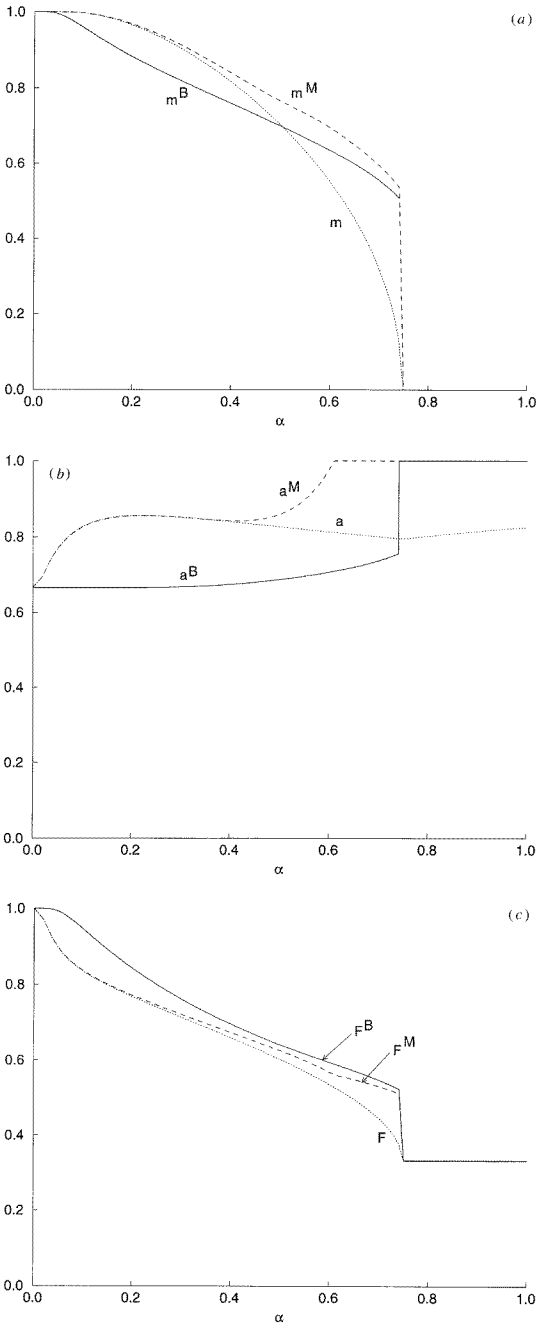

Figure 3. (*a*) The ordinary overlap m (dotted curve), the maximal likelihood overlap m^{M} (broken curve) and the Bayesian overlap m^{B} (full curve) as a function of α for $b = 0.2$ in the extremely diluted network. (*b*) The corresponding spatial activities. (*c*) The corresponding performances.

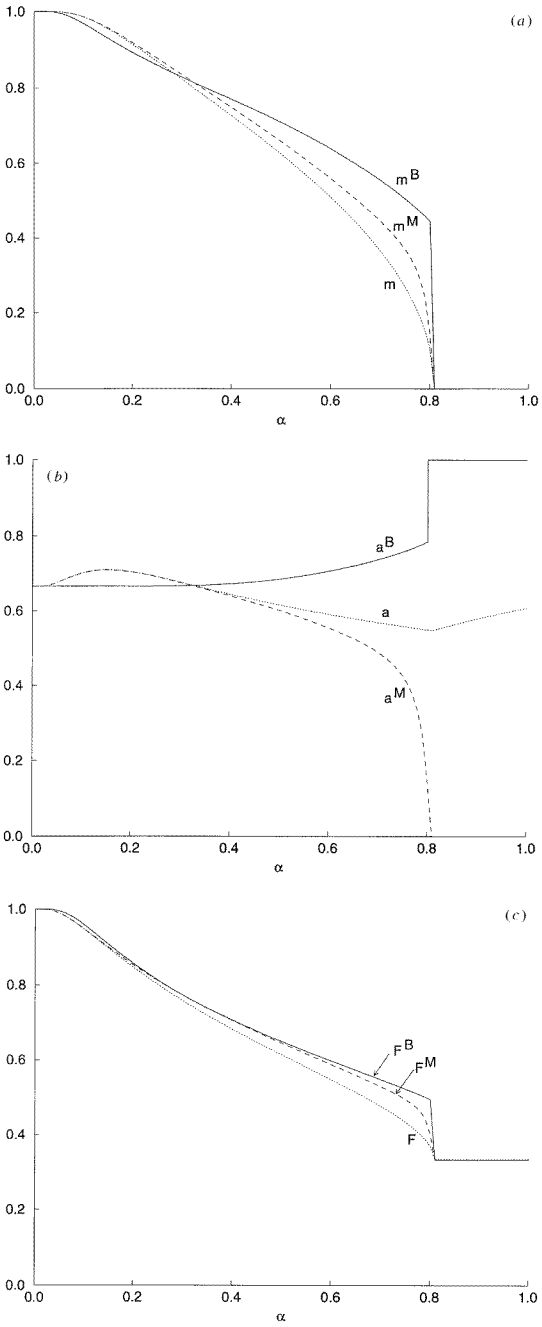

Figure 4. As in figure 3 for $b = 0.4$.

which takes into account that the patterns are uniformly distributed instead of just counting the most frequently appearing state. The posterior probability that the nominated pattern at a given node is ξ after the observation of the sequence of neuron states is given by $P(\xi|\{\eta_r\}) = \mathcal{D}(\{\eta_r\}|\xi)P(\xi)/P(\{\eta_r\})$, according to the Bayesian rule. Considering models

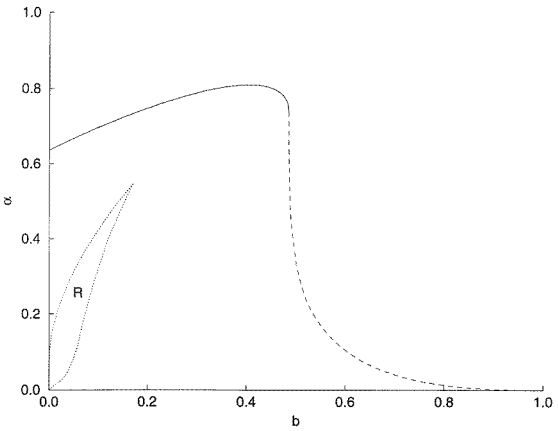

Figure 5. The region R is where the maximal likelihood Hamming distance is larger than the ordinary one in the extremely diluted network.

in which the prior probabilities of the nominated pattern at a specific neuron being $\xi = s_k$ are equal, $P(\xi|\{\eta_r\})$ is proportional to $\mathcal{D}(\{\eta_r\}|\xi)$. Therefore the most probable Bayesian state is given by taking the value of ξ having the maximum posterior probability, namely,

$$\sigma^{\mathrm{B}} = \arg\max_{\xi} P(\xi|\{\eta_r\}) = \arg\max_{\xi} \mathcal{D}(\{\eta_r\}|\xi). \tag{39}$$

According to (30) for $Q = 3$,

$$\sigma^{\mathrm{B}} = \mathrm{sign}(\eta_1)\Theta(|\eta_1| - \theta^{\mathrm{B}}) \tag{40}$$

$$\theta^{\mathrm{B}} = \frac{1}{2}\,\mathrm{erf}\left(\frac{\frac{1}{2}m + b}{\sqrt{2u}}\right) + \frac{1}{2}\,\mathrm{erf}\left(\frac{\frac{1}{2}m - b}{\sqrt{2u}}\right). \tag{41}$$

From this, one may calculate

$$m^{\mathrm{B}} = \frac{1}{2}\,\mathrm{erf}\left(\frac{3m}{2\sqrt{2v}}\right) + \frac{1}{2}\,\mathrm{erf}\left(\frac{m}{2\sqrt{2v}}\right) \tag{42}$$

$$a^{\mathrm{B}} = 1 - \frac{1}{3}\,\mathrm{erf}\left(\frac{3m}{2\sqrt{2v}}\right) \tag{43}$$

$$F^{\mathrm{B}} = \frac{1}{3}\left[1 + 2\,\mathrm{erf}\left(\frac{m}{2\sqrt{2v}}\right)\right]. \tag{44}$$

For dilute networks these quantities are displayed in figures 3(*a*)–(*c*) for $b = 0.2$ and figures 4(*a*)–(*c*) for $b = 0.4$. It is interesting to note that the Bayesian overlap undergoes a first-order transition at α_c for any b. Overclipping is not observed. This demonstrates the superior performance of Bayesian clipping over ordinary dynamics and maximum likelihood clipping. For low values of b or α, one may note that m^{B} is lower than m, indicating that the non-zero bits are retrieved rather inaccurately by Bayesian clipping. However, this sacrifice results in a higher *overall* performance.

6. Simulations

In this section we compare the analytic results with simulations. For dilute networks, the theoretical limit of $\ln C \ll \ln N$ is difficult to realize in simulations. Hence, we will focus on simulations of layered networks.

To speed up the simulations, we apply a gauge transformation for the -1 bits and pattern 1 is assigned to have $\xi_i^1(t) = 1$ for $1 \leqslant i \leqslant 2N/3$ and $\xi_i^1(t) = 0$ for $2N/3 < i \leqslant N$. The input configurations are generated according to the six probabilities $P(S_i(1)|\xi_i^1(t))$, which are determined by maximizing the entropy $-\sum_{S_i(1),\xi_i^1(1)} P(S_i(1)|\xi_i^1(1)) \ln P(S_i(1)|\xi_i^1(1))$, subject to the constraints that the overlap is $m(1)$ and the spatial activity is $a(1)$. Another technique to speed up the dynamics is to rewrite the local fields as

$$h_i(t+1) = \sum_j J_{ij}(t+1)\xi_j^1(t) + \sum_{j=1}^{2N/3} J_{ij}(t+1)(-\delta_{S_j(t),0} - 2\delta_{S_j(t),-1}) + \sum_{j=2N/3+1}^{N} J_{ij}(t+1)(\delta_{S_j(t),1} - \delta_{S_j(t),-1}). \tag{45}$$

In the expression of the local field, the first term corresponds to the value when layer t perfectly retrieves pattern 1, and can be computed once for the entire ensemble. The following terms are the corrections due to imperfect retrieval of pattern 1, and are the only terms to be computed for every input configuration, thereby saving the computational efforts when the retrieval errors are few.

As shown in figures 1(a) and (b), the difference in distribution allows a Bayesian clipping procedure to be applied efficiently. The Bayesian clipping threshold can be estimated by the crossover point of the two distributions on the same layer, which is $\theta^{\mathrm{B}} \approx 0.44$ asymptotically for the case of figures 1(a) and (b). Simulations with varying clipping thresholds show that the clipped performance, F, is maximized around this value of θ^{B}, confirming the validity of the clipping scheme.

Figure 6 shows the evolution of the performance, F, for both ordinary dynamics and activity dynamics. In activity dynamics, the clipping threshold is varied and results which maximize the performance, F, are chosen for presentation. The values of the chosen thresholds roughly agree with the theoretical values of the Bayesian threshold, but uncertainties are present due to the flatness of the performance, F, near their maxima.

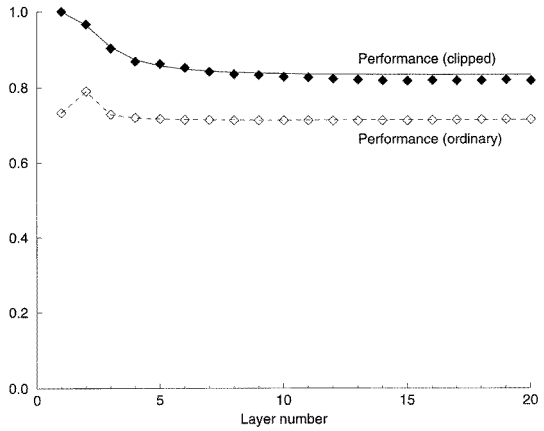

Figure 6. The performance, F, for ordinary dynamics and clipped activity in layered networks. Data points represent simulation results, full and broken curves represent analytic results for clipped activity and ordinary dynamics respectively. We have used $N = 300$, $p = 30$, $b = 0.1$, $m(1) = 0.6$, $a(1) = 0.4$, averaged over 300 samples with an ensemble size of 500 in the simulations.

Figure 6 demonstrates that activity dynamics improves the retrieval precision. Analytic results presented in the same figure agree with simulations.

7. Conclusion

We have studied two clipping schemes for activity dynamics, and found from both analysis and simulation that Bayesian clipping is most effective in improving the retrieval precision. The extent of improvement is most marked near the boundary of the retrieval phase at low values of b, where the overlap vanishes continuously for ordinary dynamics, but discontinuously for Bayesian clipped dynamics.

It is interesting to generalize the Bayesian clipping scheme to networks with weights trained by learning rules more efficient than the Hebb rule. For example, in two-state layered networks with the pseudo-inverse rule or the maximally stable rule, perfect retrieval by activity dynamics is possible. One may also consider Q-state ($Q > 3$) and analogue networks [5]. For Q-state networks with $Q > 3$, a set of thresholds have to be introduced, whereas in analogue networks the clipping scheme may consist of a mapping function from the temporal activities to the analogue states of the nodes. These clipping schemes remain to be studied.

Recently, more sophisticated clipping schemes, such as selective freezing and sequential selective freezing, have been introduced [1]. Besides improving the retrieval precision, they can also extend the basins of attraction, and even increase the storage capacity. Perfect retrieval is possible in principle, given sufficient activity statistics. The potentials of these procedures in multi-state networks remain to be explored.

The clipping techniques are useful in applications, for example the retrieval of two-dimensional patterns using line-by-line temporal codes. A layered network of L layers and N nodes on each layer can be used to process two-dimensional patterns of width N and length L, the first layer being an edge where, and only where, external information is supplied. If different edge specimens are collected and the network is run repeatedly, the clipping of temporal activities enables the entire pattern to be retrieved precisely, even when external cues are inaccessible in the interior region of the pattern.

Acknowledgments

This work has been supported in part by the Research Fund of the KU Leuven (grant OT/94/9). The authors are indebted to Reimer Kühn for clarifying discussions. One of us (DB) thanks the Belgian National Fund for Scientific Research for financial support. We thank one of the referees for the suggestion on the line-by-line temporal encoding of two-dimensional patterns.

Appendix. Recursion relations for activity dynamics

First, from definition (8) and the updating rule (2) one easily obtains the recursion relation (20) for $m(t)$.

Secondly, for the dynamic noise $u(t+1)$, we first split (15) into two terms: a diagonal term coming from the same sites, $j = k$, and an off-diagonal term arising from different sites, $j \neq k$. In the diagonal term, first and second moments of the ensemble averaged

states appear. Hence, we introduce the temporal activity $a(t)$ and the correlation $C(t)$

$$a(t) = \lim_{N\to\infty} \frac{1}{N} \sum_i \langle [\sigma_i(t)]^2 \rangle \tag{46}$$

$$C(t) = \lim_{N\to\infty} \frac{1}{N} \sum_i [\langle \sigma_i(t) \rangle]^2. \tag{47}$$

In the limit $N \to \infty$, the diagonal term becomes $\alpha(a(t) - C(t))$. The nth moment of the state on layer $t+1$ with respect to the ensemble can be written as

$$\langle [\sigma_i(t+1)]^n \rangle = \int \mathrm{D}x\, g^n \left(\sum_j J_{ij}(t+1) \langle \sigma_i(t) \rangle + \sqrt{u(t)} x \right) \tag{48}$$

leading to the results

$$a(t+1) = \left\langle\!\!\left\langle \int \mathrm{D}z\, g^2 \left(\xi(t+1) m(t) + \sqrt{u(t) + v(t)}\, z \right) \right\rangle\!\!\right\rangle \tag{49}$$

$$C(t+1) = \left\langle\!\!\left\langle \int \mathrm{D}y \left[\int \mathrm{D}x\, g \left(\xi(t+1) m(t) + \sqrt{u(t)} x + \sqrt{v(t)} y \right) \right]^2 \right\rangle\!\!\right\rangle. \tag{50}$$

In the off-diagonal term, it is important to note that contributions from different sites, j and k, do not vanish in layered networks because they receive shared information from common ancestor nodes in the previous layer. This is in contrast to networks with directed tree structures, such as in extremely diluted networks, in which contributions from different sites, j and k, can be considered independent. However, the correlations between sites j and k are weak, allowing a Taylor expansion of $\langle \sigma_l(t+1)\sigma_i(t+1) \rangle - \langle \sigma_l(t+1) \rangle \langle \sigma_i(t+1) \rangle$ in terms of the small covariance

$$\begin{aligned} \langle X_j(t+1) X_k(t+1) \rangle = {} & \frac{1}{N^2 A^2} \sum_{\mu,\nu} \sum_{l,i} \xi_j^\mu(t+1) \xi_k^\nu(t+1) \xi_l^\mu(t) \xi_i^\nu(t) \\ & \times [\langle \sigma_l(t) \sigma_i(t) \rangle - \langle \sigma_l(t) \rangle \langle \sigma_i(t) \rangle]. \end{aligned} \tag{51}$$

These results lead to the evolution equation (21).

Finally, the recursion relation for the frozen noise $v(t+1)$ can be obtained analogously. The diagonal term in (19) with $j = k$ becomes $\alpha C(t+1)$. In the off-diagonal term with $j \neq k$, again the weak correlation between $\xi_j^\mu(t+1)$ and $\langle \sigma_j(t+1) \rangle$ can be seen from writing

$$Y_j(t+1) = Y_j^{(\mu)}(t+1) + \xi_j^\mu(t+1) \frac{1}{NA} \sum_k \xi_k^\mu(t) \langle \sigma_k(t) \rangle. \tag{52}$$

Here all dependence on pattern μ are contained in the last term, and the first term is the field fluctuations if pattern μ was absent. The initial condition implies that the correlations between the network state and the non-condensed patterns $\mu > 1$ are weak. Hence the last term is $\mathrm{O}(1/\sqrt{N})$. Expanding $\langle \sigma_j(t+1) \rangle$ with respect to this term finally leads to the recursion relation (22).

References

[1] Wong K Y M 1996 *Europhys. Lett.* **36** 631
[2] Wong K Y M and Ho C 1994 *J. Phys. A: Math. Gen.* **27** 5167

[3] Bollé D, Shim G M, Vinck B and Zagrebnov V A 1994 *J. Stat. Phys.* **74** 565
[4] Bollé D, Shim G M and Vinck B 1994 *J. Stat. Phys.* **74** 583
[5] Bollé D and Vinck B 1996 *Physica* **223A** 293
[6] Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167
[7] Domany E, Kinzel W and Meir R 1989 *J. Phys. A: Math. Gen.* **22** 2081
[8] Bollé D, Vinck B and Zagrebnov V A 1993 *J. Stat. Phys.* **70** 1099
[9] Kree R and Zippelius A 1991 *Models of Neural Networks* ed E Domany, J L van Hemmen and K Schulten (Berlin: Springer)
[10] Derrida B 1989 *J. Phys. A: Math. Gen.* **22** 2069