# Attractor properties of dynamical systems: neural network models

1994 J. Phys. A: Math. Gen. 27 5167

(http://iopscience.iop.org/0305-4470/27/15/017)

# Attractor properties of dynamical systems: neural network models

K Y M Wong and C Ho†

Department of Physics, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

Received 6 January 1994, in final form 23 May 1994

**Abstract.** We propose tools to probe the nature of attractors in dynamical systems. These include the activity distribution, the evolutions of the state damage, activity damage and temporal correlation damage. When they are used to study the retrieval attractors in dilute asymmetric neural networks, a transition from a partially frozen phase to an unfrozen phase is found for networks trained with sufficiently noisy data near storage saturation, and this confirms that the retrieval attractors are more chaotic in this case. We are also able to demonstrate that the retrieval attractors in dilute asymmetric neural networks are not clouds of attractors, but consist of a single chaotic attractor for each stored pattern. Furthermore, they facilitate the device of effective freezing procedures, which significantly improve the quality of retrieval in neural networks.

## 1. Introduction

The application of statistical mechanical techniques to the study of dynamical systems is well illustrated by its contributions to the study of neural networks [1–4]. Such studies gain increasing importance with VLSI implementations of neural networks becoming increasingly available. In the retrieval phase the network dynamics drives the system to an attractor which is often described by an order parameter specified by the overlap of the network state with the pattern to be retrieved, analogous to the magnetization order parameter in spin systems. This order parameter vanishes continuously [3, 5] or discontinuously [4, 5] in the retrieval to non-retrieval transition. However, besides its being specified by the overlap, much less is known about the nature of the attractors in neural networks, and possibly in similar dynamical systems.

The description of the attractor may be simpler for dynamical spin systems with symmetric weights ($J_{ij} = J_{ji}$). In these systems the attractor at zero temperature corresponds to the ground state of an energy function, and the attractor state is a fixed point in the phase space. On the other hand, for systems with non-symmetric weights the attractor may be a fixed point, a limit cycle or becomes chaotic in the phase space, and the overlap order parameter provides no information to distinguish these cases. While the number of metastable states have been calculated in some systems [6], they only provide information on the possible fixed point attractors, but attractors of more extended structures remain unprobed.

In this paper we further extend the previously proposed notions of *damage evolution* [3] and *activity distribution* [7, 8] to probe the nature of attractors in general complex dynamical

† Present address: Department of Physics, University of Virginia, Charlottesville, VA 22905, USA.

0305-4470/94/155167+19$19.50

systems, as introduced in section 2 and detailed in sections 3 to 6. They are quantities of interest applicable to both simulation and analytical approaches to network dynamics. In this paper we illustrate their applicability to an exactly solvable case, namely the case of dilute asymmetric neural networks [3].

As a consequence, we are able to demonstrate that the retrieval attractors in dilute asymmetric neural networks are not clouds of attractors as previously proposed [3], but consist of a single chaotic attractor for each stored pattern. Furthermore, while previous studies on the activity distribution are limited to dilute asymmetric networks in which the synaptic weights are simple Hebbian [7, 8], we demonstrate in section 4 that they can be generalized to dilute asymmetric networks with *arbitrary* synaptic prescription, provided that the aligning field distribution [9–11] is known. This allows us the freedom to explore a wide class of networks, of which the simple Hebbian case [7, 8] and the maximally stable network [9–11] are only particular cases. For instance, we may tune the synaptic weights so that the basins of the retrieval attractors may be wide and interfering, or narrow and with more retrieval precision; the width of the basins being conveniently controlled by the amount of random errors present in the training data of the network [12].

It turns out that by considering the activity distribution, we will observe a transition from a partially frozen phase to an unfrozen phase when the number of stored patterns increases. We will also see that the presence of the unfrozen phase is a manifestation of wide and interfering basins of attraction, in contrast to the absence of this phase for the case of narrow and more precise basins. This observation is consistent with the increase of associativity, and the decrease of retrieval precision, storage capacity and selectivity for networks whose basins are tuned by the training process to have increasing width and interference [12].

The study of the activity distribution leads to a very useful application as described in section 7, namely that by averaging the evolving network states over an extended period in the attractor, the stored patterns are retrieved much better. Remarkably, when the pattern bits are retrieved by clipping the activity, this *clipped activity* undergoes a first-order transition to non-retrieval in dilute Hebbian networks, in contrast to the second-order transition for instantaneous overlap as previously reported [3]. This sheds new light on the importance of the dynamics during the readout process of neural networks.

## 2. Formulation

We proceed by considering a network of $N$ neurons $S_i = \pm 1$. Each neuron is fed by $C$ other neurons chosen randomly, and coupled from neuron $j$ to neuron $i$ through the synaptic weights $J_{ij}$, where $J_{ij}$ satisfy the spherical constraint $\sum_{j\in J(i)} J_{ij}^2 = C$; $J(i)$ being the set of $C$ randomly chosen neurons coupled to neuron $i$. Only parallel dynamics with Gaussian noise [13] will be considered in this paper:

$$S_i(t+1) = \operatorname{sgn}\left(\frac{1}{\sqrt{C}} \sum_{j\in J(i)} J_{ij} S_j(t) + T_n z_i(t)\right) \tag{2.1}$$

where $z_i(t)$ is a random Gaussian number of mean 0 and width 1, and $T_n$ is the retrieval noise temperature measuring the amount of stochastic noise present in the updating dynamics of the network. Alternatively, one may consider the use of discrete noise [3] instead of Gaussian noise, but the results are essentially the same.

The values of the synaptic weights $J_{ij}$ are assigned during the learning process [1] so that a set of $p$ patterns $\xi_j^\mu = \pm 1$, $j = 1 \ldots N$ and $\mu = 1 \ldots p$ can be retrieved. The actual prescription for the synaptic weights $J_{ij}$ depends on the individual learning processes. It turns out, at least for the dilute asymmetric neural networks on which we will focus our discussion, that the dynamics for the retrieval of a single pattern, say pattern $\mu$, is completely determined by a knowledge of the aligning field distribution $\rho(\Lambda)$ given by

$$\rho(\Lambda) = \lim_{N\to\infty} \frac{1}{N} \sum_i \delta(\Lambda - \Lambda_i^\mu) \tag{2.2}$$

where $\Lambda_i^\mu \equiv \xi_i^\mu \sum_j J_{ij}\xi_j^\mu/\sqrt{C}$ is the aligning field of pattern $\mu$ at neuron $i$. By varying the details of the learning process, one may adjust the aligning field distribution, and hence obtain different structures of the retrieval attractors. In appendix A we illustrate this by controlling the noise level present in the training data. In the low training noise limit, the resultant network is the maximally stable network (MSN) [9–11], whereas in the high training noise limit, one recovers the Hebbian network [2].

Retrieval of a pattern, say pattern 1, is monitored by observing the evolution of an (instantaneous) overlap order parameter $m(t)$ given by

$$m(t) = \frac{1}{N} \sum_i \xi_i^1 S_i(t) . \tag{2.3}$$

In general, $m(t)$ approaches a fixed point value $m^*$ in the asymptotic limit. A non-zero value of the attractor overlap $m^*$ corresponds to the retrieval phase. However, when the storage level $\alpha \equiv p/C$ is greater than a critical value $\alpha_c$, $m^*$ vanishes corresponding to the non-retrieval phase. $\alpha_c$ is therefore the storage capacity of the network.

Further information on the retrieval attractor can be obtained by monitoring the activity at a neuron $i$, which is defined as the time average of its projection on pattern 1 in the asymptotic limit

$$a_i = \lim_{T,T'\to\infty} \frac{1}{T'} \sum_{t=T+1}^{T+T'} \xi_i^1 S_i(t) . \tag{2.4}$$

Note that both the overlap and the activity involve the averaging of the same arguments; the difference is that in the overlap they are averaged over neurons for a fixed time, whereas in the activity they are averaged over time for a fixed neuron. As a result this provides information on the fraction of time a neuron remains firing or non-firing, and hence its most probable state, as well as the structure and size of the attractor. For instance, $a_i = \pm 1$ corresponds to a neuron always retrieving the correct/incorrect pattern bit asymptotically, i.e. it is 'frozen'. Alternatively, this implies that the attractor is restricted to the $(N-1)$-dimensional subspace with $S_i = \pm\xi_i^\mu$ in the $N$-dimensional space of network states. An attractor with many frozen spins (or nearly frozen ones) is therefore a small attractor. We note that our definition of activity is slightly different from Derrida's [7], which involves averaging over an ensemble of initial conditions instead of averaging over an extended period of time in the asymptotic limit; if the elements in the ensemble of initial network states belong to different attractors, the two definitions may lead to different values of the activity.

In the thermodynamic limit one may be more interested in the activity distribution. The probability distribution $Q(a)$ of local activities is defined by

$$Q(a) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \delta(a - a_i) . \tag{2.5}$$

Thus the activity distribution of a point attractor consists of two delta peaks at $a = \pm 1$. This means that the neurons are spending all of their time in the state of correct/incorrect pattern bits. On the other hand, the activity distribution of the attractor in the high-temperature paramagnetic phase consists of a single delta peak at $a = 0$, since thermal noise is flipping the neuronal states randomly. Furthermore, it turns out that at low temperature *it is possible to have the activity distribution diverging as the activity approaches* $\pm 1$, i.e. $\lim_{a\to\pm 1} Q(a) = \infty$ [7]. This corresponds to a phase with *partially frozen* spins. On the other hand, if the activity distribution does not diverge in these limits, the system has *unfrozen* spins.

To elucidate the structure of the retrieval attractors, the notion of *damage evolution* has been proposed [3]. Here damage refers to two slightly different network state configurations (and does not associate with any physical damage of the network). This involves monitoring the evolution of two network states $S_i(t)$ and $\tilde{S}_i(t)$, *both subject to the same stochastic noise* $z_i(t)$ in their dynamics. Damage spreads when their difference propagates on time evolution and heals when it vanishes. We define the *state overlap* as the overlap of the two instantaneous network states

$$r_s(t) = \frac{1}{N}\sum_i S_i(t)\tilde{S}_i(t) \tag{2.6}$$

and the fractional state damage is given by $d_s = (1 - r_s)/2$. $r_s = 1$ is always a fixed point since this corresponds to the dynamics of a single network state. However, this fixed point may either be stable or unstable, corresponding to the cases of damage healing or spreading, respectively.

In the case when damage spreads, one may conclude that the attractor is either a single attractor or a cloud of attractors. In this paper we propose new techniques to extract more information on these two possibilities. (We refer to *chaotic dynamics* with its unpredictability as manifested by damage spreading, which is different from *noisy dynamics*, which refers to the presence of stochastic noise $z_i(t)$ in the update of the neuronal states.) We will consider the evolution of the *activity damage*. This is the equivalence of the notion of the state damage evolution with respect to instantaneous network states, in which we monitor the overlap of the *activities* of two evolving network states in their respective attractors, and measure whether they converge to each other or not. One may define the *activity overlap* of the network states $S_i(t)$ and $\tilde{S}_i(t)$ given by

$$r_a = \frac{1}{N}\sum_i a_i\tilde{a}_i = \langle a_i\tilde{a}_i\rangle_i \tag{2.7}$$

where $\langle\cdot\rangle_i$ represents averaging over all neurons. Again, $r_a = \langle a_i^2\rangle_i$ or $\langle\tilde{a}_i^2\rangle_i$ is always a fixed point since this corresponds to the dynamics of a single network state. However, when this fixed point is unstable, arbitrarily close network states will flow towards different attractors, and the possibility of a single attractor is excluded. Note that the converse may not be valid, for it is possible to have two different attractors but with identical activities for their neurons. An example is shown in table 1, in which the neuronal states in two distinct attractors may have the same fraction of $+1$ and $-1$ states, but arranged in different sequences. Thus the stability of the fixed point $r_a = \langle a_i^2\rangle_i$ is a necessary but not sufficient condition for a single attractor.

To further differentiate single and multiple attractors by examining their sequence of neuronal states, we introduce the notion of *temporal correlation damage*. Here we monitor

**Table 1.** An example of network states having the same activity for neuron $i$, namely $a_i = \frac{1}{3}$, but may belong to different attractors. Here all attractors are limit cycles of period 6. Comparing $S_i^{(1)}(t)$ and $S_i^{(2)}(t)$, damage spreads but they belong to the same attractor. Comparing $S_i^{(1)}(t)$ and $S_i^{(3)}(t)$, damage spreads and they belong to different attractors although $a_i^{(1)} = a_i^{(3)}$.

| $t$ | $T+1$ | $T+2$ | $T+3$ | $T+4$ | $T+5$ | $T+6$ |
|---|---|---|---|---|---|---|
| $S_i^{(1)}(t)$ | + | + | − | + | − | + |
| $S_i^{(2)}(t)$ | + | + | + | − | + | − |
| $S_i^{(3)}(t)$ | + | + | + | + | − | − |

the time correlation functions of two evolving network states in their respective attractors, and measure their differences. Consider *correlation overlaps* given by

$$r_c(\tau) = \frac{1}{N} \sum_i c_i(\tau)\tilde{c}_i(\tau) \tag{2.8}$$

where $c_i(\tau)$ are temporal correlation functions given by

$$c_i(\tau) = \lim_{T,T' \to \infty} \frac{1}{T} \sum_{t=T+1}^{T+T'} S_i(t)S_i(t+\tau)\,. \tag{2.9}$$

The necessary and sufficient condition for the adjacent network states $S_i(t)$ and $\tilde{S}_i(t)$ to belong to a single attractor, rather than a cloud of attractors, is that $r_c(\tau) = \langle c_i^2(\tau)\rangle_i$ is a stable fixed point for all values of $\tau$.

## 3. Damage evolution

Consider two network states $S_i(t)$ and $\tilde{S}_i(t)$ subject to the same stochastic noise and both having macroscopic overlaps $m(t)$ and $\tilde{m}(t)$, respectively with pattern 1, and only microscopic random overlaps with the other patterns. These instantaneous overlaps at successive time-steps are related, on using (2.3) and (2.1), by

$$m(t+1) = \frac{1}{N} \sum_i \operatorname{sgn}\left[\frac{1}{\sqrt{C}} \sum_j \xi_i^1 J_{ij} S_j(t) + T_n z_i(t)\right] \tag{3.1}$$

where we have used the fact that $\xi_i^1 z_i(t)$ and $z_i(t)$ obey the same distribution. Noting that the averaged value of $S_j(t)$ is $m\xi_j^1$ and introducing the notation $\Lambda_i^1$ for the aligning field in (2.2), we obtain

$$m(t+1) = \frac{1}{N} \sum_i \operatorname{sgn}\left[m(t)\Lambda_i^1 + X_i(t) + T_n z_i(t)\right] \tag{3.2}$$

where $X_i(t) \equiv \xi_i^1 \sum_j J_{ij}(S_j(t) - m(t)\xi_j^1)/\sqrt{C}$. In the limit $C \gg 1$, $X_i(t)$ becomes a Gaussian variable of mean 0 and width

$$\langle X_i(t)^2\rangle_i = \frac{1}{C} \sum_{jk} \langle J_{ij} J_{ik}\rangle_i \left[S_j(t) - m(t)\xi_j^1\right]\left[S_k(t) - m(t)\xi_k^1\right]. \tag{3.3}$$

For input states with no further correlations than those specified by the overlap $m(t)$, non-vanishing contribution to (3.3) only comes from terms with $j = k$, and, since $\langle J_{ij}^2\rangle_i = 1$, we have $\langle X_i(t)^2\rangle_i = 1 - m^2$. Thus, on performing the averaging over neurons $i$ in (3.2), we arrive at the recursion relation

$$m(t+1) = \int \mathrm{d}\Lambda\, \rho(\Lambda) \mathrm{erf}\left(\frac{m\Lambda}{\sqrt{2(1-m^2+T_\mathrm{n}^2)}}\right) \tag{3.4}$$

and a similar expression for $\tilde{m}(t+1)$. In general, this recursion relation cannot be extended beyond one time-step, since correlation between network states of various time-steps will be involved. However, for dilute asymmetric networks with $\ln C \ll \ln N$, correlations beyond one time-step can be neglected and (3.4) can be extended to an arbitrary number of time-steps [9–11]. Henceforth this is the case we focus on.

For state overlap $r_\mathrm{s}(t)$ between the two states, the state overlap for the next time-step is derived in appendix B by the standard techniques of introducing integral representations for delta functions, factorizing and then performing pattern average. Here we present a more direct derivation when $C \gg 1$. Using the definition of state overlap (2.6) we write, in analogy with (3.2),

$$r_\mathrm{s}(t+1) = \frac{1}{N}\sum_i \mathrm{sgn}\big[m(t)\Lambda_i^1 + X_i(t) + T_\mathrm{n}z_i(t)\big]\mathrm{sgn}\big[\tilde{m}(t)\Lambda_i^1 + \tilde{X}_i(t) + T_\mathrm{n}z_i(t)\big] \tag{3.5}$$

where $X_i(t)$ and $\tilde{X}_i(t)$ are defined as in (3.2) for the network states $S_i(t)$ and $\tilde{S}_i(t)$. Hence $\langle X_i(t)\rangle_i = \langle\tilde{X}_i(t)\rangle_i = 0$ and

$$\langle X_i(t)^2\rangle_i = 1 - m^2(t) \qquad \langle\tilde{X}_i(t)^2\rangle_i = 1 - \tilde{m}^2(t)\,. \tag{3.6$a$}$$

Furthermore, correlation exists between $X_i(t)$ and $\tilde{X}_i(t)$, since

$$\langle X_i(t)\tilde{X}_i(t)\rangle_i = \frac{1}{C}\sum_{jk}\langle J_{ij}J_{ik}\rangle_i\big[S_j(t) - m(t)\xi_j^1\big]\big[\tilde{S}_k(t) - \tilde{m}(t)\xi_k^1\big]\,.$$

Again, for random input states with no further correlations than those described by $m(t)$, $\tilde{m}(t)$ and $r_\mathrm{s}(t)$, non-vanishing contribution only comes from terms with $j = k$, yielding

$$\langle X_i(t)\tilde{X}_i(t)\rangle_i = r_\mathrm{s}(t) - m(t)\tilde{m}(t)\,. \tag{3.6$b$}$$

This implies that $X_i(t)$ and $\tilde{X}_i(t)$ can be expressed as a linear combination of two independent Gaussian variables $u$ and $v$. A choice of their coefficients consistent with the correlations (3.6$a$) and (3.6$b$) is given by

$$\begin{aligned} X_i(t) &= \sqrt{1-m^2}\left(\sqrt{\frac{1+\eta_\mathrm{s}}{2}}u + \sqrt{\frac{1-\eta_\mathrm{s}}{2}}v\right) \\ \tilde{X}_i(t) &= \sqrt{1-\tilde{m}^2}\left(\sqrt{\frac{1+\eta_\mathrm{s}}{2}}u - \sqrt{\frac{1-\eta_\mathrm{s}}{2}}v\right) \end{aligned} \tag{3.7}$$

where $\eta_\mathrm{s} = (r_\mathrm{s} - m\tilde{m})/\sqrt{(1-m^2)(1-\tilde{m}^2)}$. (Note, however, that the above decomposition is not unique. For example, it is algebraically more convenient in some cases to write

$X_i(t) = \sqrt{1-m^2}u$ and $\tilde{X}_i(t) = \sqrt{1-\tilde{m}^2}(\eta_s u + \sqrt{1-\eta_s^2}v)$.) Averaging over neurons $i$ in (3.5), we arrive at the recursion relation

$$\begin{aligned} r_s(t+1) = &\int \mathrm{d}\Lambda\,\rho(\Lambda) \int \mathrm{D}z \int \mathrm{D}u \int \mathrm{D}v \\ &\times \mathrm{sgn}\left[m\Lambda + \sqrt{1-m^2}\left(\sqrt{\frac{1+\eta_s}{2}}u + \sqrt{\frac{1-\eta_s}{2}}v\right) + T_n z\right] \\ &\times \mathrm{sgn}\left[\tilde{m}\Lambda + \sqrt{1-\tilde{m}^2}\left(\sqrt{\frac{1+\eta_s}{2}}u - \sqrt{\frac{1-\eta_s}{2}}v\right) + T_n z\right] \end{aligned} \tag{3.8}$$

where D$z$ is the Gaussian measure introduced in appendix A.

In the asymptotic limit, $m(t)$ and $\tilde{m}(t)$ approach the attractor overlap $m^*$. It is easily shown that $r_s = 1$ is a fixed point, corresponding to the dynamics of a single network state. However, this fixed point is unstable, as can be shown by the expansion around $r_s(t) = 1 - \epsilon$,

$$r_s(t+1) = 1 - \frac{2}{\pi}\sqrt{\frac{2\epsilon}{1-m^2+T_n^2}} \int \mathrm{d}\Lambda\,\rho(\Lambda)\exp\left(-\frac{m^2\Lambda^2}{2(1-m^2+T_n^2)}\right). \tag{3.9}$$

This implies that damage in network states spreads for all temperatures. Note that this behaviour is different from that observed in finite-dimensional spin systems, in which the system undergoes a dynamical phase transition to a healed damage phase when the temperature increases [14]. Nevertheless, the magnitude of the damage vanishes asymptotically in the high-temperature limit, irrespective of whether the system is in retrieval or non-retrieval phase, since

$$\lim_{T_n \gg 1} r_s^* = 1 - \frac{8}{\pi^2 T_n^2}. \tag{3.10}$$

This shows that the retrieval attractor is either a chaotic attractor or a cloud of attractors. The $T_n^{-2}$ dependence of state damage is also found in the case of discrete noise, and it is interesting to note that the state damage has the same temperature dependence in the SK model [15].

Although damage spreads for all storage levels in the retrieval phase, the degree of chaoticity does increase with the storage level, as measured by the coefficient of $\sqrt{\epsilon}$ in (3.9). For example, in the Hebbian network, substitution of the aligning field distribution (A.8) reduces (3.9) to

$$\epsilon(t+1) = \frac{2}{\pi}\sqrt{\frac{2}{1+T_n^2}}\exp\left(-\frac{m^2/\alpha}{2(1+T_n^2)}\right)\epsilon(t)^{1/2} \tag{3.11}$$

and the coefficient of $\epsilon(t)^{1/2}$ increases from 0 at $\alpha = 0$ to $(2/\pi)\sqrt{2/(1+T_n^2)}$ at $\alpha = \alpha_c$.

## 4. The activity distribution

The activity distribution provides further information on the degree of chaoticity of the attractor. Consider the activity of neuron $i$, which is given by [7]

$$a_i(t+1) = \int \mathrm{D}z \sum_{\tau_j = \pm 1} \prod_j \left(\frac{1+a_j(t)\tau_j}{2}\right) \mathrm{sgn}\left(\frac{1}{\sqrt{C}} \sum_j \xi_i^1 J_{ij} \xi_j^1 \tau_j + T_n z\right) \tag{4.1}$$

where $a_i(t)$ is the activity of node $i$ over $T$ time-steps starting from the instant $t$, and the limit $T \to \infty$ is taken. The activity distribution can therefore be determined, in dilute networks, by the recursion relation

$$Q(a,t+1) = \prod_j \int \mathrm{d}a_j \, Q(a_j,t) \times \delta\left[\int \mathrm{D}z \sum_{\tau_j=\pm 1} \prod_j \left(\frac{1+a_j\tau_j}{2}\right) \mathrm{sgn}\left(\frac{1}{\sqrt{C}} \sum_j \xi_i^1 J_{ij} \xi_j^1 \tau_j + T_{\mathrm{n}} z\right) - a\right]. \quad (4.2)$$

In the asymptotic limit $Q(a,t)$ approaches a fixed point distribution $Q^*(a)$. This equation can be simplified using techniques developed in appendix B, but again, we present a more physical derivation here. Consider first the activity $a_i(t+1)$, which can be written as

$$a_i(t+1) = \left\langle \mathrm{sgn}\left[\frac{1}{\sqrt{C}} \sum_j \xi_i^1 J_{ij} \xi_j^1 a_j + Y_i(t) + T_{\mathrm{n}} z_i(t)\right]\right\rangle_{\mathrm{t}} \quad (4.3)$$

where $Y_i(t) \equiv \sum_j \xi_i^1 J_{ij}(S_j(t) - a_j\xi_j^1)/\sqrt{C}$, and $\langle\cdot\rangle_{\mathrm{t}}$ represents averaging over a period of $T$ time-steps. $Y_i(t)$ is a Gaussian variable with mean $\langle Y_i(t)\rangle_{\mathrm{t}} = 0$ and width

$$\langle Y_i(t)^2\rangle_{\mathrm{t}} = \frac{1}{C} \sum_{jk} J_{ij} J_{ik} \langle [S_j(t) - a_j\xi_j^1][S_k(t) - a_k\xi_k^1]\rangle_{\mathrm{t}} .$$

Non-vanishing contribution only comes from terms with $j = k$, yielding $\langle Y_i(t)^2\rangle_{\mathrm{t}} = 1 - \sum_j J_{ij}^2 a_j^2/C$. Note, however, that in dilute asymmetric networks $a_j$ depends on the neurons and synapses feeding neuron $j$, and is therefore independent of $J_{ij}$, which is emanating from $j$ to $i$. Hence $\langle J_{ij}^2 a_j^2\rangle_j = \langle J_{ij}^2\rangle_j \langle a_j^2\rangle_j$, leading to

$$\langle Y_i(t)^2\rangle_{\mathrm{t}} = 1 - q \quad (4.4)$$

where $q \equiv \langle a_i^2\rangle_i$. $Y_i(t)$ represents the dynamic contribution of the disorder to the activity. For example, $\langle Y_i(t)^2\rangle_{\mathrm{t}} = 0$ when all spins are frozen, and $\langle Y_i(t)^2\rangle_{\mathrm{t}} = 1$ in the paramagnetic phase. Performing the temporal averaging in (4.2) we obtain

$$a_i(t+1) = \mathrm{erf}\left(\frac{\sum_j \xi_i^1 J_{ij} \xi_j^1 a_j/\sqrt{C}}{\sqrt{2(1-q+T_{\mathrm{n}}^2)}}\right). \quad (4.5)$$

Next we consider the numerator in the expression (4.5), which can be written as

$$\frac{1}{\sqrt{C}} \sum_j \xi_i^1 J_{ij} \xi_j^1 a_j = m\Lambda_i^1 + V_i \quad (4.6)$$

where $V_i \equiv \sum_j \xi_i^1 J_{ij} \xi_j^1 (a_j - m)/\sqrt{C}$. $V_i$ is a Gaussian variable with $\langle V_i\rangle_i = 0$ and

$$\langle V_i^2\rangle_i = \frac{1}{C} \sum_{jk} \langle J_{ij} J_{ik}\rangle_i \xi_j^1 \xi_k^1 (a_j - m)(a_k - m) .$$

Again, the non-vanishing contribution only comes from terms with $j = k$, rendering

$$\langle V_i^2\rangle_i = q - m^2 \quad (4.7)$$

where $V_i$ represents frozen disorder. For example, $\langle V_i^2 \rangle_i = 0$ in the paramagnetic phase, and $\langle V_i^2 \rangle_i$ reaches its maximum value of $1 - m^2$ when all spins are frozen. In fact, for a given time-step, the disorder term $X_i(t)$ in the instantaneous overlap update (3.2) is merely the sum of the dynamic component $Y_i(t)$ and the frozen component $V_i$, since the distinction between the two is irrelevant within one time-step. Performing the averaging over neurons $i$ we arrive at an equation $Q(a)$ in the asymptotic limit

$$Q(a) = \int \mathrm{d}\Lambda\, \rho(\Lambda) \int \mathrm{D}y\, \delta\left[\operatorname{erf}\left(\frac{m\Lambda + \sqrt{q - m^2}y}{\sqrt{2(1 - q + T_\mathrm{n}^2)}}\right) - a\right]. \tag{4.8}$$

The parameters $m$ and $q$ are the first and second moments of the activity distribution, which can be determined self-consistently from

$$m = \int \mathrm{d}a\, Q(a)a \qquad q = \int \mathrm{d}a\, Q(a)a^2 . \tag{4.9}$$

Note that the first moment is merely the attractor overlap, hence the notation $m$. For the Hebbian network one may substitute the aligning field distribution (A.8) and recover the results of Derrida [7]. Here we have generalized the result to networks with an *arbitrary* aligning field distribution.

Derrida found that for the Hebbian network the activity distribution in the limits $a \to \pm 1$ behave differently when the storage level varies. For low storage levels the activity distribution at the retrieval attractor diverges at $a = \pm 1$, corresponding to a partially frozen phase. When the storage level is sufficiently high, the divergence disappears, corresponding to an unfrozen phase, although pattern retrieval in this phase is still possible. Since the presence of flipping spins may be considered as a measure of the degree of chaoticity of a dynamical system, the transition from the partially frozen to unfrozen phase shows that the system becomes increasingly chaotic with the storage level. This provides an another measure of chaoticity in addition to the damage evolution.

However, the unfrozen phase is not necessarily present in networks other than the Hebbian network. In fact, in the MSN, patterns can be stored perfectly up to the retrieval storage capacity $\alpha_\mathrm{c} = 2$ in a dilute asymmetric network architecture at $T_\mathrm{n} = 0$ [5]. This shows that the retrieval attractor is a point attractor throughout the retrieval phase of the MSN, i.e. the spins are completely frozen. One may interpolate the two extreme behaviours of the MSN and the Hebbian network by controlling the noise level present in the training data [12]. Extensive studies on the associativity, retrieval precision, storage capacity, selectivity, robustness against weight dilution and temperature, and the weight space organization of these networks show that the basin structure can be tuned by the training noise, as described by the training overlap $m_\mathrm{t}$ in appendix A [12]. When the network is optimally trained with very noisy data ($m_\mathrm{t} \to 0$), the basins of the retrieval attractors are wide and interfering, corresponding to high associativity, high robustness against weight dilution and temperature, but low precision, low storage capacity and low selectivity, as exemplified by the Hebbian network. It is therefore expected that, when the training noise level increases from the MSN limit ($m_\mathrm{t} \to 1$) to the Hebbian limit ($m_\mathrm{t} \to 0$), the unfrozen phase will be increasingly significant, signalling the increasing degree of chaoticity as the retrieval basins widen.

[t]

The condition for the existence of the partially frozen phase can be easily obtained from the activity distribution $Q(a)$ in the limit $a \to \pm 1$. Introducing $z = \operatorname{erf}^{-1} a$, equation (4.8) reduces to

$$Q(a) = \frac{1}{2}\sqrt{\frac{1 - q + T_\mathrm{n}^2}{q - m^2}} \int \mathrm{d}\Lambda\, \rho(\Lambda) \exp\left[z^2 - \frac{(\sqrt{2(1 - q + T_\mathrm{n}^2)}z - m\Lambda)^2}{2(q - m^2)}\right]. \tag{4.10}$$

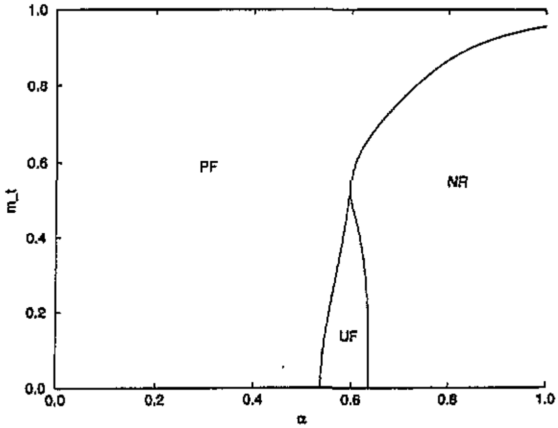

**Figure 1.** The phase diagram of the attractor structure in the space of the training overlap $m_t$ and the storage level $\alpha$ at $T_n = 0$. In figures 1 and 2, PF, UF and NR represent the partially frozen, unfrozen and non-retrieval phases, respectively.

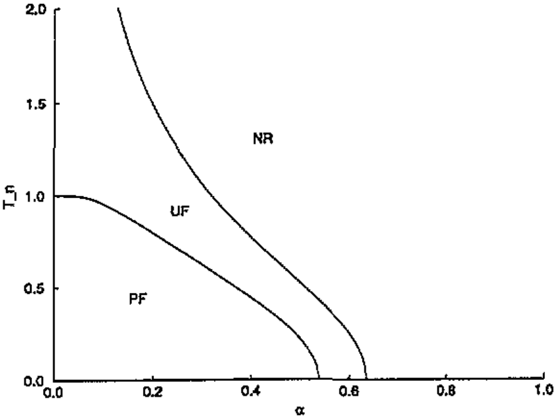

**Figure 2.** The phase diagram of the attractor structure in the space of the storage level $\alpha$ and the temperature $T_n$ for the Hebbian network.

For the neural networks satisfying the generic condition $\lim_{|t|\to\infty} \lambda(t)/t = 1$ in (A.4), this yields

$$\lim_{|a|\to 1} Q(a) = \frac{1}{2}\sqrt{\frac{1-q+T_n^2}{q}} \exp\left(\frac{2q-T_n^2-1}{q}z^2\right). \tag{4.11}$$

Thus the transition between the partially frozen and unfrozen phases takes place at

$$q = \tfrac{1}{2}(1+T_n^2). \tag{4.12}$$

Figure 1 shows the phase diagram in the space of the storage level $\alpha$ and the training overlap $m_t$ at $T_n = 0$, assuming that the network is replica symmetric in the weight space [12]. The retrieval phase is further divided into the partially frozen phase at low storage level and unfrozen phase at higher storage level. Note, however, that for sufficiently low training noise ($m_t > 0.55$), no unfrozen phase is present, showing that the retrieval attractors are less chaotic. This is consistent with the narrowing of the retrieval basins with increasing precision in the low training noise limit.

To demonstrate the effects of temperature, figure 2 shows the phase diagram in the space of the storage level $\alpha$ and temperature $T_n$ for the Hebbian network. The unfrozen phase is present at all temperatures, whereas the partially frozen phase is present only for $T_n \leqslant 1$.

## 5. The biconfigurational activity distribution and the activity damage evolution

To study further the structure of the attractors we again consider two network states $S_i(t)$ and $\tilde{S}_i(t)$ both subject to the same stochastic noise and monitor their joint activity distribution. The biconfigurational activity distribution is defined by

$$Q(a,\tilde{a}) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \delta(a-a_i)\delta(\tilde{a}-\tilde{a}_i) . \tag{5.1}$$

In analogy with (4.2), a recursion relation for this distribution can be obtained as

$$\begin{aligned} Q(a,\tilde{a},t+1) = \prod_j \int & \mathrm{d}a_j\, \mathrm{d}\tilde{a}_j\, Q(a_j,\tilde{a}_j,t) \\ & \times \delta\left[\int \mathrm{D}z \sum_{\tau_j=\pm 1} \prod_j \left(\frac{1+a_j\tau_j}{2}\right) \mathrm{sgn}\left(\frac{1}{\sqrt{C}} \sum_j \xi_i^1 J_{ij} \xi_j^1 \tau_j + T_n z\right) - a\right] \\ & \times \delta\left[\int \mathrm{D}z \sum_{\tau_j=\pm 1} \prod_j \left(\frac{1+\tilde{a}_j\tau_j}{2}\right) \mathrm{sgn}\left(\frac{1}{\sqrt{C}} \sum_j \xi_i^1 J_{ij} \xi_j^1 \tau_j + T_n z\right) - \tilde{a}\right] . \end{aligned} \tag{5.2}$$

Using techniques similar to the derivation of (4.8), we can show that the fixed-point distribution $Q(a,\tilde{a})$ at $T_n = 0$ is given by

$$\begin{aligned} Q(a,\tilde{a}) = \int & \mathrm{d}\Lambda\, \rho(\Lambda) \int \mathrm{D}u \int \mathrm{D}v \\ & \times \delta\left[\mathrm{erf}\left(\frac{m\Lambda + \sqrt{q-m^2}\left(\sqrt{(1+\eta_a)/2}u + \sqrt{(1-\eta_a)/2}v\right)}{\sqrt{2(1-q)}}\right) - a\right] \\ & \times \delta\left[\mathrm{erf}\left(\frac{\tilde{m}\Lambda + \sqrt{\tilde{q}-\tilde{m}^2}\left(\sqrt{(1+\eta_a)/2}u - \sqrt{(1-\eta_a)/2}v\right)}{\sqrt{2(1-\tilde{q})}}\right) - \tilde{a}\right] \end{aligned} \tag{5.3}$$

where $\eta_a = (r_a - m\tilde{m})/\sqrt{(q-m^2)(\tilde{q}-\tilde{m}^2)}$, and $r_a \equiv \langle a_i \tilde{a}_i \rangle_i$ is the activity overlap in (2.7). *It is determined self-consistently by*

$$r_a = \int \mathrm{d}a\, \mathrm{d}\tilde{a}\, Q(a,\tilde{a}) a\tilde{a} . \tag{5.4}$$

If the network states $S_i(t)$ and $\tilde{S}_i(t)$ belong to the same attractors, $m$, $\tilde{m}$ and $q$, $\tilde{q}$ converge to the same fixed points $m^*$ and $q^*$, respectively, as given by (4.9). The biconfigurational activity distribution is then completely determnined by $m$, $q$ and $r_{\rm a}$. The asymptotic value of $r_{\rm a}$ is therefore given by

$$r_{\rm a} = \int {\rm d}\Lambda\, \rho(\Lambda) \int {\rm D}u \int {\rm D}v\, {\rm erf}\bigg(\frac{m\Lambda + \sqrt{q-m^2}(\sqrt{(1+\eta_{\rm a})/2}u + \sqrt{(1-\eta_{\rm a})/2}v)}{\sqrt{2(1-q)}}\bigg) \times {\rm erf}\bigg(\frac{m\Lambda + \sqrt{q-m^2}(\sqrt{(1+\eta_{\rm a})/2}u - \sqrt{(1-\eta_{\rm a})/2}v)}{\sqrt{2(1-q)}}\bigg) \tag{5.5}$$

where $\eta_{\rm a} = (r_{\rm a}-m^2)/(q-m^2)$. When $r_{\rm a} = q$, $\langle a_i\tilde{a}_i\rangle_i = \langle a_i^2\rangle_i = \langle \tilde{a}_i^2\rangle_i$ and the configurations $S_i(t)$ and $\tilde{S}_i(t)$ become identical. It can easily be verified from (5.5) that $r_{\rm a} = q$ is indeed a fixed point. To study the stability of this fixed point, we consider the Taylor expansion of (5.5) around $r_{\rm a} = q$, yielding

$$r_{\rm a} = q - \bigg[\frac{2}{\pi\sqrt{(1-q)(1-2m^2+q)}}\int {\rm d}\Lambda\, \rho(\Lambda) \exp\bigg(-\frac{m^2\Lambda^2}{1-2m^2+q}\bigg)\bigg](r_{\rm a}-q) + {\rm O}(r_{\rm a}-q)^2 . \tag{5.6}$$

This may be compared with the Taylor expansion of (4.9) around the stable fixed point $q = q^*$, which reads

$$q = q^* - \bigg[\frac{2}{\pi\sqrt{(1-q)(1-2m^2+q)}}\int {\rm d}\Lambda\, \rho(\Lambda) \exp\bigg(-\frac{m^2\Lambda^2}{1-2m^2+q}\bigg)\bigg] \times (q-q^*) + {\rm O}(q-q^*)^2 . \tag{5.7}$$

It turns out that both expansions have the same first-order coefficient. Hence we deduce that $r_{\rm a} = q$ is always a stable fixed point. This shows that in the retrieval attractor, activity damage heals, although state damage spreads. This indicates that the retrieval attractor may be a single but chaotic attractor, as will be argued in the following section.

## 6. The temporal correlation damage evolution

To verify that the retrieval attractors are indeed single attractors, we have to consider the correlation overlaps (2.8) for all time intervals $\tau$. It turns out that in dilute asymmetric networks, it is more convenient to consider the temporal correlation of a neuron with its 'ancestor' $\tau$ time-steps before, instead of the local correlation (2.9), i.e.

$$c_{ij}(\tau) = \lim_{T,T'\to\infty} \frac{1}{T} \sum_{t=T+1}^{T+T'} \xi_i^1 S_i(t+\tau)\xi_j^1 S_j(t) \tag{6.1}$$

where $j \in J^\tau(i)$. For dilute asymmetric networks, $c_{ij}$ can be expressed in terms of the activities $a_i$ and $a_j$. The example of $\tau = 1$ is given in appendix C.

We can now generalize the argument to two evolving configurations $S_i(t)$ and $\tilde{S}_i(t)$. Since we have verified that $r_{\rm a} = q$ in the attractor, $a_i = \tilde{a}_i$ for all neurons. Since $c_{ij}(\tau)$ and $\tilde{c}_{ij}(\tau)$ are determined by $a_i$, $a_j$ and $\tilde{a}_i$, $\tilde{a}_j$, respectively, we conclude that $c_{ij}(\tau) = \tilde{c}_{ij}(\tau)$ for all values of $\tau$. Hence the two configurations belong to the same attractor. More precisely, the attractors of the two configurations are indistinguishable to the order of a fraction of $C^0$ neurons when a finite number of time-steps is being monitored.

Although the spreading of state damage implies that the dynamics is chaotic, the uniqueness of the attractor for each stored pattern implies that it is compact in dilute asymmetric neural networks, in contrast to the existence of strange attractors in low-dimensional systems. This can be seen by considering network states which differ by a finite number of neuronal states. In the thermodynamic limit the stability of the dynamical equation for the overlap $m$ implies that they both converge to the retrieval attractor, which is unique.

## 7. Enhanced retrieval by activity averaging

The study of the activity distribution leads to a very useful application, namely that by averaging the evolving network states over an extended period in the attractor, the stored patterns are retrieved much better than by monitoring the instantaneous network states as described by the usual overlap order parameter. This means that during the retrieval process, one first allows the system to equilibrate, and then monitors the time-averaged state of each neuron. The output state ±1 on each neuron is then determined according to the sign of the time-averaged state. The resultant overlap $m_{\mathrm{a}}$ is referred to as the *clipped activity*. Its value is given by

$$m_{\mathrm{a}} = \int \mathrm{d}a\, Q(a) \operatorname{sgn} a\,. \tag{7.1}$$

Substituting (4.8), we have

$$m_{\mathrm{a}} = \int \mathrm{d}\Lambda\, \rho(\Lambda) \operatorname{erf}\left(\frac{m\Lambda}{\sqrt{2(q-m^2)}}\right). \tag{7.2}$$

Comparing with (3.4), this corresponds to the output overlap for an effective input overlap of $\sqrt{1+T_{\mathrm{n}}^2}m/\sqrt{q}$. Since the right-hand side of (3.4) is an increasing function of $m$, this guarantees an increase of overlap from $m^*$ to $m_{\mathrm{a}}$. Furthermore, since activity damage heals, the pattern retrieved by activity clipping is unique for a sufficiently long monitoring period. Figure 3 compares the clipped activity with the instantaneous overlap for the Hebbian network at $T_{\mathrm{n}} = 0$.

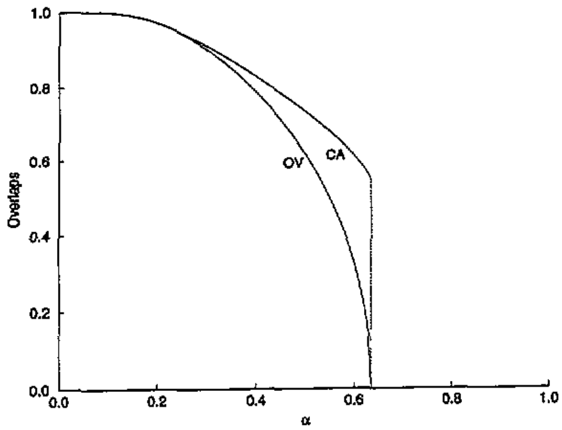

**Figure 3.** The dependence of the instantaneous overlap (OV) and the time-averaged overlap (or the clipped activity CA) on the storage level $\alpha$ for the Hebbian network at $T_{\mathrm{n}} = 0$.

This enhancement in the overlap is most dramatic immediately below the storage capacity. Consider the family of networks for which the instantaneous attractor overlap undergoes a continuous transition at the storage capacity $\alpha_c$; this can be achieved for a training overlap $m_t$ less than 0.48 [12]. By considering the fixed-point equations (4.9) near $\alpha_c$, we have

$$\frac{m}{\sqrt{q}} = \sqrt{\frac{\frac{1}{2}\pi(1+T_n^2)-1}{\langle\Lambda^2\rangle - 1}}\,. \tag{7.3}$$

This implies that both the mean and the width of the activity distribution (4.8) approaches zero with a constant ratio, so that the fraction of neurons with the correct sign of their activities is fixed. For the Hebbian network, for example, $\langle\Lambda^2\rangle = \pi/2 + 1$ by (A.8). Thus on substituting (7.3) into (7.2) for $T_n = 0$,

$$m_a = \operatorname{erf}\left(\frac{\sqrt{\pi - 2}}{2}\right) = 0.55\,. \tag{7.4}$$

This is a significant improvement in retrieval, for the clipped activity now undergoes a discontinuous transition at the storage capacity $\alpha_c$.

Encouraged by the performance of the above time-averaged procedure, we propose a *selective freezing* procedure for further improvement. We note that the activity distribution is highly asymmetric with respect to a change of sign of the activity. This means that those neurons with large magnitude of activity are much more likely to be aligned with the correct patten bits. The selective freezing procedure makes use of this fact. Here one allows the system to equilibrate and then monitors the time-averaged state on each node. Nodes with the magnitude of the time-averaged state exceeding a given threshold value are then assigned a state ±1 according to the sign of the time-averaged state. The states of these nodes become fixed in the subsequent procedure. In the next stage, one again allows the system to equilibrate with the rest of the nodes dynamic, and the output states are determined by clipping the time-averaged states as before. While further details will be reported elsewhere, here we report that this method can produce a jump in the retrieval overlap as high as 0.59 at the retrieval to non-retrieval transition for the Hebbian network at $T_n = 0$.

## 8. Discussion

In this paper we proposed a number of probes to understand the nature of attractors in dynamical systems. These probes are: (i) the activity distribution; (ii) the evolution of the state damage; (iii) the evolution of the activity damage; (iv) the evolution of the temporal correlation damage. These quantitites of interest are applicable to both simulational and analytical approaches to the dynamics of neural networks, both dilute or more extensively connected, as well as general complex dynamical systems and spin glasses. Here we illustrate their applicability to an exactly solvable case, namely the case of dilute asymmetric neural networks. These notions find a direct application in retrieval enhancement procedures making use of time-averaged states on each node. Activity clipping and selective freezing greatly improve the precision in retrieval, even in the vicinity of the storage capacity.

Using these probes we observed that the degree of chaoticity of the retrieval attractors increases with the storage level as well as the level of training noise, which is associated with a phase transition from the partially frozen to unfrozen phase. The tuning of the chaoticity of the attractors is consistent with the variation of the size of retrieval basin by adjusting the level of training noise previously studied in neural networks.

For networks with adaptable weights, Hebb learning proceeds by the synaptic modification rule $\Delta J_{ij} \sim a_i a_j$, our study on activity distribution has two further implications. First, when the network state falls into a retrieval attractor for a sufficiently long time, the synaptic modification is unique, since activity damage heals in dilute asymmetric networks. Secondly, when the storage capacity is near saturation, the network is in an unfrozen phase with lower magnitudes of activity, and Hebb learning by activity correlation becomes hard. On the other hand for lower storage levels, the network is in the partially frozen phase, and Hebb learning is effective. This provides a mechanism to guarantee the quality of patterns to be learned, for when patterns are learned, they are learned with generally high quality, whereas poorly retrieved patterns near saturation are not effectively learned.

These probes also enabled us to demonstrate that the retrieval attractors in dilute asymmetric networks are single chaotic attractors. We may compare this result with that in dilute *symmetric* neural networks [16] in which replica symmetry breaking effects are significant, corresponding to the existence of multiple attractors.

This comparison shows that the symmetry of the node couplings determines the attractor structure of the system, namely that asymmetric couplings result in more chaotic attractors, which, however, exist in a single valley around each stored pattern; on the other hand symmetric souplings result in less chaotic attractors, which, however, may be clustered in different valleys.

This observation is consistent with previous findings in spin systems of other architectures, which show that dynamical systems become increasingly chaotic with coupling asymmetry, often with larger but fewer attractors. They are observed in fully connected asymmetric spin glasses [17–19], asymmetric neural networks with soft spins for both connected [20] and highly diluted [21] architectures, layered neural networks [22] and the Kauffman model [23].

It is interesting to compare our results with the simulation of three-dimensional spin glasses [14], in which three possible phases have been observed: (i) a high-temperature $T > T_1$ regime in which state damage heals; (ii) an intermediate-temperature regime $T_2 < T < T_1$ in which damage spreads, but the asymptotic state damage is independent of the initial state overlap; (iii) a low-temperature regime $T < T_2$ in which state damage spreads, and the asymptotic state overlap depends on the initial state overlap. The transition temperature $T_2$ is generally recognized as the spin-glass temperature, whereas the transition at $T_1$ is identified as a dynamical phase transition. In dilute asymmetric neural networks we find that only the intermediate phase is present. We have already excluded the possibility of many attractors characteristic of the low-temperature spin-glass phase, by considering the evolution of activity damage and temporal correlation damage, and we attribute this to the asymmetry of the couplings. On the other hand, the high-temperature damage-healing phase is absent, probably because dilute asymmetric networks lack the strong correlations of finite-dimensional systems, which may drive them towards ordered behaviour. This is consistent with a comparative study of dynamical phase transitions in short-ranged and long-ranged neural network models [24].

Although in this paper we only considered dynamical systems with binary states, the techniques can be generalized to networks made up of spins or neurons with multi-states or continuous states, except that the formulation may be more complicated. In these cases the activity distribution is generalized to the joint distribution $Q(a, \xi)$ in neural networks, where $a$ is the time-averaged state and $\xi$ is the pattern state on the same node; damage evolution (whether state, activity or temporal correlation) can be studied by monitoring the Hamming distance in the phase space.

## Acknowledgments

We thank B Derrida, A Engel, Y Stein, E Domany, J T L Ho and D Sherrington for fruitful discussions. This work is supported by the University and Polytechnic Grant Council of Hong Kong under grant number DAG91/92.SC30.

## Appendix A. Aligning field distribution in neural networks optimally trained with noisy data

In this appendix we summarize the results derived in [4] for the aligning field distribution in neural networks optimally trained with noisy data. The noise level in the training data is specified by the training overlap $m_t$, which defines the probability distribution of the example pattern bits $\{R_j^{\mu\nu}\}$ as

$$P(R_j^{\mu\nu}) = \tfrac{1}{2}(1+m_t)\delta(R_j^{\mu\nu} - \xi_j^\mu) + \tfrac{1}{2}(1-m_t)\delta(R_j^{\mu\nu} + \xi_j^\mu)\,. \tag{A.1}$$

When this set of training data is used, learning can be considered as the optimization of the performance function $\sum_\mu g(\Lambda_i^\mu)$ for each neuron $i$, where

$$g(\Lambda) = \operatorname{erf}\left(\frac{m_t\Lambda}{\sqrt{2(1-m_t^2)}}\right)\,. \tag{A.2}$$

For an arbitrary form of the performance function $g(\Lambda)$, the aligning field distribution is given by

$$\rho(\Lambda) = \int \mathrm{D}t\,\delta(\Lambda - \lambda(t)) \tag{A.3}$$

where $\mathrm{D}t \equiv \exp(-t^2/2)\mathrm{d}t/\sqrt{2\pi}$ is the Gaussian measure and $\lambda(t)$ is the value of $\lambda$ which maximizes the expression

$$g(\lambda) - (\lambda - t)^2/2\gamma \tag{A.4}$$

and $\gamma$ is the susceptibility parameter determined by the condition

$$\int \mathrm{D}t\,(\lambda(t) - t)^2 = \alpha^{-1}\,. \tag{A.5}$$

In the low training noise limit, the resultant network is the MSN whose aligning field distribution is given by

$$\rho(\Lambda) = \frac{1}{2}\left(1 + \operatorname{erf}\frac{\kappa}{\sqrt{2}}\right)\delta(\Lambda - \kappa) + \Theta(\Lambda - \kappa)\frac{\mathrm{e}^{-\Lambda^2/2}}{\sqrt{2\pi}} \tag{A.6}$$

where $\kappa$ is the stability parameter given by

$$\int_{-\infty}^{\kappa} \mathrm{D}t(\kappa - t)^2 = \alpha^{-1} . \tag{A.7}$$

In the high training noise limit ($m_t \to 0$), one obtains the Hebbian network whose aligning field distribution is given by

$$\rho(\Lambda) = \frac{\exp\left[ -\frac{1}{2}(\Lambda - 1/\sqrt{\alpha})^2\right]}{\sqrt{2\pi}} . \tag{A.8}$$

For intermediate training noise levels, one may obtain a one-band or two-band aligning field distribution.

Using (3.4), the aligning field distribution allows us to determine the attractor overlap $m^*$ and hence the storage capacity $\alpha_c$ for dilute asymmetric networks. For example, the attractor overlap of the Hebbian network is given, on substituting (A.8) into (3.4), by

$$m^* = \mathrm{erf}\left(\frac{m^*}{\sqrt{2\alpha(1+T_n^2)}}\right) . \tag{A.9}$$

Hence the value of $\alpha_c$ is $2/(\pi(1+T_n^2))$ for the Hebbian network. Similarly, for the MSN, $\alpha_c = 2$, making use of (A.6) and (3.4). For intermediate training noise levels $\alpha_c$ is computed in [12], and reproduced in figure 1 as the phase boundary of NR.

## Appendix B. Recursion relation for the state overlap $r_s$

In this appendix we derive the recursion relation for the state overlap $r_s(t)$. The state overlap at time $t+1$ is given by

$$r_s(t+1) = \frac{1}{N}\sum_i \mathrm{sgn}\left[\frac{1}{\sqrt{C}}\sum_j J_{ij}S_j(t) + T_n z_i(t)\right]\mathrm{sgn}\left[\frac{1}{\sqrt{C}}\sum_j J_{ij}\tilde{S}_j(t) + T_n z_i(t)\right] . \tag{B.1}$$

Introducing delta functions for $\Delta = \sum_j J_{ij}S_j(t)/\sqrt{C}$ and $\tilde{\Delta} = \sum_j J_{ij}\tilde{S}_j(t)/\sqrt{C}$, and then using integral representations for the delta functions, we obtain, after factorizing over the neurons,

$$\begin{aligned} r_s(t+1) = \frac{1}{N}\sum_i \int \frac{\mathrm{d}\Delta\,\mathrm{d}x}{2\pi}\int\frac{\mathrm{d}\tilde{\Delta}\,\mathrm{d}\tilde{x}}{2\pi} &\exp(\mathrm{i}x\Delta + \mathrm{i}\tilde{x}\tilde{\Delta})\mathrm{sgn}(\Delta + T_n z_i(t))\mathrm{sgn}(\tilde{\Delta} + T_n z_i(t)) \\ &\times \prod_j \exp\left[-\frac{\mathrm{i}}{\sqrt{C}}\sum_j J_{ij}(xS_j(t) + \tilde{x}\tilde{S}_j(t))\right] . \end{aligned} \tag{B.2}$$

Averaging over $S_j(t)$ and $\tilde{S}_j(t)$, we obtain

$$\begin{aligned} r_s(t+1) = \frac{1}{N}\sum_i \int \frac{\mathrm{d}\Delta\,\mathrm{d}x}{2\pi}\int\frac{\mathrm{d}\tilde{\Delta}\,\mathrm{d}\tilde{x}}{2\pi} &\exp(\mathrm{i}x\Delta + \mathrm{i}\tilde{x}\tilde{\Delta})\mathrm{sgn}(\Delta + T_n z_i(t))\mathrm{sgn}(\tilde{\Delta} + T_n z_i(t)) \\ &\times \exp\left\{ -\frac{\mathrm{i}}{\sqrt{C}}(xm + \tilde{x}\tilde{m})\sum_j J_{ij}\xi_j^1 \right. \\ &\left. -\tfrac{1}{2}\left[(1-m^2)x^2 + 2(r_s - m\tilde{m})x\tilde{x} + (1-\tilde{m}^2)\tilde{x}^2\right]\right\} . \end{aligned} \tag{B.3}$$

Introducing the aligning field $\Lambda_i^1 = \sum_j \xi_i^1 J_{ij} \xi_j^1/\sqrt{C}$, and using the quadratic identity

$$Ax^2 + 2Bx\tilde{x} + C\tilde{x}^2 = \frac{1}{2}\left(1 + \frac{B}{\sqrt{AC}}\right)(\sqrt{A}x + \sqrt{C}\tilde{x})^2 + \frac{1}{2}\left(1 - \frac{B}{\sqrt{AC}}\right)(\sqrt{A}x - \sqrt{C}\tilde{x})^2$$

together with the Hubbard–Stratonovich identity

$$\mathrm{e}^{a^2/2} = \int \mathrm{D}z\, \mathrm{e}^{az}$$

this reduces to

$$\begin{aligned} r_s(t+1) = \frac{1}{N}\sum_i \int \mathrm{D}u \int \mathrm{D}v \\ &\times \mathrm{sgn}\left[m\xi_i^1\Lambda_i^1 + \sqrt{1-m^2}\left(\sqrt{\frac{1+\eta_s}{2}}u + \sqrt{\frac{1-\eta_s}{2}}v\right) + T_n z_i(t)\right] \\ &\times \mathrm{sgn}\left[\tilde{m}\xi_i^1\Lambda_i^1 + \sqrt{1-\tilde{m}^2}\left(\sqrt{\frac{1+\eta_s}{2}}u - \sqrt{\frac{1-\eta_s}{2}}v\right) + T_n z_i(t)\right]. \end{aligned} \tag{B.4}$$

Averaging over the thermal noise, and using the definition of the aligning field distribution (2.2), we arrive at (3.8).

## Appendix C. Temporal correlation for $\tau = 1$

In this appendix we derive the temporal correlation $c_{ij}(1)$ in terms of the activities $a_i$ and $a_j$ for $T_n = 0$. Consider the correlation function in which neuron $i$ is fed by neurons $k = 1 \ldots C$ including $j$. Then

$$c_{ij}(1) = \sum_{\tau_k=\pm1}\prod_k\left(\frac{1+a_k\tau_k}{2}\right)\mathrm{sgn}\left[\frac{1}{\sqrt{C}}\sum_{k\neq j}\xi_i^1 J_{ik}\xi_k^1\tau_k + \frac{1}{\sqrt{C}}\xi_i^1 J_{ij}\xi_j^1\tau_j\right]\tau_j. \tag{C.1}$$

Following an averaging procedure over the neurons $k \neq j$ similar to that in section 4 or appendix B, we obtain

$$c_{ij}(1) = \sum_{\tau_j=\pm1}\frac{1+a_j\tau_j}{2}\mathrm{erf}\left(\frac{\sum_{k\neq j}\xi_i^1 J_{ik}\xi_k^1 a_k/\sqrt{C} + \xi_i^1 J_{ij}\xi_j^1\tau_j/\sqrt{C}}{\sqrt{2\sum_{k\neq j}J_{ik}^2(1-a_k^2)/C}}\right)\tau_j. \tag{C.2}$$

Comparing this expression with the product $a_i a_j$ using (4.5), we obtain, on subtraction,

$$\begin{aligned} c_{ij}(1) - a_i a_j = \sum_{\tau_j=\pm1}\frac{\tau_j + a_j}{2}\Bigg[&\mathrm{erf}\left(\frac{\sum_{k\neq j}\xi_i^1 J_{ik}\xi_k^1 a_k/\sqrt{C} + \xi_i^1 J_{ij}\xi_j^1\tau_j/\sqrt{C}}{\sqrt{2\sum_{k\neq j}J_{ik}^2(1-a_k^2)/C}}\right) \\ &- \mathrm{erf}\left(\frac{\sum_{k\neq j}\xi_i^1 J_{ik}\xi_k^1 a_k/\sqrt{C} + \xi_i^1 J_{ij}\xi_j^1 a_j/\sqrt{C}}{\sqrt{2(1-q)}}\right)\Bigg]. \end{aligned} \tag{C.3}$$

To order $1/\sqrt{C}$, this yields

$$c_{ij}(1) = a_i a_j + \frac{2}{\sqrt{C}}\xi_i^1 J_{ij}\xi_j^1\frac{\exp\left[-(\mathrm{erf}^{-1}a_i)^2\right]}{\sqrt{2\pi(1-q)}}. \tag{C.4}$$

Therefore the correlation $c_{ij}(1)$ is completely determined by the single neuron activities. This argument can be generalized to correlations for other values of $\tau$.

## References

[1] Hertz J, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neurocomputation* (Redwood City, CA: Addison-Wesley)

[2] Amit D, Gutfreund H and Sompolinsky H 1987 *Ann. Phys., NY* **173** 30

[3] Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167

[4] Wong K Y M and Sherrington D 1990 *J. Phys. A: Math. Gen.* **23** 4659

[5] Amit D, Evans M, Horner H and Wong K Y M 1990 *J. Phys. A: Math. Gen.* **23** 3361

[6] Treves A and Amit D J 1988 *J. Phys. A: Math. Gen.* **21** 3155

[7] Derrida B 1989 *J. Phys. A: Math. Gen.* **22** 2069

[8] Kree R and Zippelius A 1990 *Models of Neural Networks* ed E Domany, J L van Hemmen and K Schulten (Berlin: Springer) p 193

[9] Kepler T B and Abbott L F 1988 *J. Physique* **49** 1657

[10] Krauth W, Nadal J-P and Mézard M 1988 *J. Phys. A: Math. Gen.* **21** 2995

[11] Gardner E 1989 *J. Phys. A: Math. Gen.* **22** 1969

[12] Wong K Y M and Sherrington D 1993 *Phys. Rev.* E **47** 4465

[13] Peretto P 1984 *Biol. Cybern.* **50** 51

[14] Derrida B and Weisbuch G 1987 *Europhys. Lett.* **4** 657

[15] Derrida B 1989 *Phys. Rep.* **184** 207

[16] Watkin T L H and Sherrington D 1991 *Europhys. Lett.* **14** 791

[17] Spitzner P and Kinzel W 1989 *Z. Phys.* B **77** 511

[18] Crisanti A and Sompolinsky H 1988 *Phys. Rev.* A **37** 4865

[19] Gutfreund H, Reger J D and Young A P 1988 *J. Phys. A: Math Gen.* **21** 2775

[20] Sompolinsky H, Crisanti A and Sommers H J 1988 *Phys. Rev. Lett.* **61** 259

[21] Tirozzi B and Tsodyks M 1991 *Europhys. Lett.* **14** 727

[22] Derrida B and Meir R 1988 *Phys. Rev.* A **38** 3116

[23] Derrida B and Pomeau Y 1986 *Europhys. Lett.* **1** 45

[24] Kürten K E 1989 *J. Physique* **50** 2313