Relative stability in the dynamics of a two-pattern neural net

1989 J. Phys. A: Math. Gen. 22 5117

(http://iopscience.iop.org/0305-4470/22/23/016)

J. Phys. A: Math. Gen. **22** (1989) 5117–5129. Printed in the UK

Relative stability in the dynamics of a two-pattern neural net

Ferenc Pázmándi and Tamás Geszti
Department of Atomic Physics, Eötvös University, H-1088 Budapest, Hungary

Received 2 June 1989

Abstract. We investigate how patterns of different acquisition strengths influence each other's stability. A neural network model with two strictly stable patterns stored on a noisy background is studied by a novel approximation of short-time dynamics that singles out coherent contributions systematically and uses a Gaussian approximation for incoherent sums. The basin of attraction of the weaker pattern is found to shrink depending on the two acquisition strengths. For a diluted Hopfield-type version of the model the weaker pattern may become unrecognisable.

1. Introduction

The dynamics of adaptable neural networks close to an attractor corresponding to a stored 'pattern' is dominated by an interplay between signal and noise. Signal is the modification of connection strengths storing information about the 'patterns', whereas noise is of twofold origin: partly quenched into the same connection strengths by the presence of patterns other than the one corresponding to the given attractor, partly 'thermal' and belonging to the dynamical law.

The simplest versions of the model (Hopfield 1982) are open to an analytical treatment (Amit *et al* 1986). The basic fact about a model with a number of equivalent random stored patterns is a sharp forgetting phase transition in which the attractors associated with the individual patterns become unstable at a given level of noise of both origins. For this reason, a signal-to-noise analysis has often been used to obtain rough information about the expected forgetting transition in various versions of the model. The most sophisticated quantitative elaboration along this line is the one-pattern network (OPN) model (Krauth *et al* 1988) in which randomising all information beyond a few parameters pertaining to one of the patterns is shown to cause only slight modifications of the dynamics close to the corresponding attractor.

There are indications, however, that the validity of this signal-to-noise or one-pattern philosophy is restricted to the case of equivalent patterns. This refers in particular to the learning-within-bound model, in which recently taught patterns gradually erase the older ones. Although qualitatively this effect can be accounted for (Nadal *et al* 1986, Parisi 1986) by the reducing signal level against a steady noise background, more quantitatively (Mézard *et al* 1986, Geszti and Pázmándi 1987) the dominant physics is seen to be different: old patterns become unstable with respect to the fresh ones which attract the dynamical flow from an old pattern above a critical ratio of the amplitudes of the two.

0305-4470/89/235117+13$02.50

The present paper is motivated by these observations. Here, however, we study mostly strict-stability networks like those produced by the 'Minover' algorithm (Krauth and Mézard 1987), which is easy to extend to patterns of unequal acquisition strengths. In this case finite, although different, stabilities for each stored pattern are sustained by construction, and the effect described above is transformed into a shrinking of the basin of attraction of the low-stability patterns. The behaviour of such models is then compared with that of the asymmetrically diluted Hopfield model (Derrida *et al* 1987) for two different patterns.

As a technical tool to obtain detailed insight into the dynamics of such cases, the original one-pattern network is extended to take two patterns of different stabilities into account. For this case the temporal variation of the overlaps with the two patterns and that of the distance between two initially close configurations are studied in the presence of thermal noise, by means of a novel technique utilising the trick of linearising in small but coherent contributions to a local field acting on a neuron (Peretto 1988). This, starting from the simple no-symmetry case (Krauth *et al* 1988, Crisanti and Sompolinsky 1988, Gutfreund *et al* 1988, equivalent to the solvable asymmetrically diluted model of Derrida *et al* 1987), allows one to single out dominant coherent corrections due to a specified symmetry of the connection strengths. As a result, closed formulae can be obtained for arbitrary asymmetry and temperature, containing integrals over a Gaussian distribution of a dimensionality growing with time.

2. Description of the two-pattern model

In what follows we generalise the one-pattern model (Krauth *et al* 1988) to the case of two patterns ξ_i^1 and ξ_i^2 (two-pattern model: TPN) of given stabilities Δ_1 and Δ_2 independent of the neuron i ($i = 1 \ldots N$), and a given value of the coupling symmetry parameter η:

$$J_{ii} = 0 \qquad \sum_j \xi_i^{1,2} J_{ij} \xi_j^{1,2} = \Delta_{1,2} \qquad \sum_j J_{ij} J_{ji} = \eta \qquad \forall i \tag{1}$$

where the connection strengths $J_{ij} = \pm N^{-1/2}$ satisfy the above constraints but are otherwise random. As a further simplification, we assume that the two stored patterns are orthogonal.

Two features of the above model are responsible for the relative simplicity of its dynamics to be described below. Firstly, the required strict site-independence of the stabilities Δ_1 and Δ_2 guarantees a property of restricted self-averaging (see equation (10)). Secondly, since there are only two patterns and noise here, all gauge-invariant combinations of the couplings beyond Δ_1, Δ_2 and η (e.g. the triple coupling $\sum_{j,k} J_{ij} J_{jk} J_{ki}$ that would make the calculation much more complicated although not impossible) vanish in the thermodynamical limit, which is always taken in our calculation, at least as fast as $N^{-1/2}$. It is easy to see that none of these important simplifying features is present in the Hopfield model.

In the same way as the one-pattern model (Krauth *et al* 1988) gives a faithful representation of the dynamics of strict-stability networks with patterns of equal acquisition strengths, the TPN is expected to give insight into how such networks with unequal pattern stabilities work.

3. The dynamics of the model

3.1. Parallel dynamics: definitions

At time t the configuration of the model is characterised by a vector $\boldsymbol{S}^t = \{S_i^t\}$. We start by generating initial configurations out of a distribution

$$P_1(\boldsymbol{S}^0) = \prod_{i=1}^{N} \frac{1 + S_i^0 \sum_{\mu=1,2} \xi_i^\mu m_\mu}{2} \tag{2}$$

specified by the initial values m_μ of the overlaps with the stored patterns

$$m_\mu(t) = \frac{1}{N} \sum_i \xi_i^\mu \langle S_i^t \rangle. \tag{3}$$

Let those configurations evolve by parallel dynamics at temperature T and investigate the change of some average properties, like the overlaps themselves.

Parallel dynamics of the model (Little 1974) proceeds in independent simultaneous single-spin flips after which the probability of a given value S_i^t at time t is given by

$$W(S_i^t \mid \boldsymbol{S}^{t-1}) = \frac{1 + S_i^t f(z_i^{t-1})}{2} \tag{4}$$

with $f(x) = \tanh \beta x$, $\beta = 1/T$ and

$$z_i^t = \sum_j J_{ij} S_j^t. \tag{5}$$

This defines a Markovian evolution of the probability distribution $P_1(\boldsymbol{S}^t)$ with the transition probability

$$W(\boldsymbol{S}^t \mid \boldsymbol{S}^{t-1}) = \prod_{i=1}^{N} W(S_i^t \mid \boldsymbol{S}^{t-1}). \tag{6}$$

In order to calculate the time evolution of the overlaps (3), where

$$\begin{aligned} \langle S_i^t \rangle &= \mathrm{Tr}_{\boldsymbol{S}^t} S_i^t P_1(\boldsymbol{S}^t) \\ &= \mathrm{Tr}_{S_i^t} S_i^t P_1(S_i^t) \end{aligned} \tag{7}$$

we need an approximate expression for

$$P_1(S_i^t) = \mathrm{Tr}_{\boldsymbol{S}^{t-1}, \boldsymbol{S}^{t-2}, \ldots, \boldsymbol{S}^0} W(S_i^t \mid \boldsymbol{S}^{t-1}) W(\boldsymbol{S}^{t-1} \mid \boldsymbol{S}^{t-2}) \ldots W(\boldsymbol{S}^1 \mid \boldsymbol{S}^0) P_1(\boldsymbol{S}^0). \tag{8}$$

As we shall see below, dynamics are dominated by the parameters Δ_1, Δ_2 and η defined in section 2. These parameters are site independent. Therefore we expect $\langle S_i^t \rangle$ to depend on i only through the values of ξ_i^1 and ξ_i^2 at the same site i. Since these are binary variables, $\langle S_i^t \rangle$ can be written as a linear function of them. Moreover, this function is homogeneous because it is multiplied by -1 if both of its arguments have been. Then its two coefficients are determined by the orthogonality of the two patterns, which gives the important formula

$$\langle S_i^t \rangle = \sum_\mu \xi_i^\mu m_\mu(t) \tag{9}$$

which inverts equation (3). The same reasoning holds for all gauge-invariant quantities, which are therefore self-averaging and can be calculated by averaging over the random patterns instead of over sites:

$$\frac{1}{N}\sum_i \ldots \rightarrow \overline{\langle\ldots\rangle}^{\xi}. \tag{10}$$

We emphasise that this holds only for the restricted class of strict-stability models with site-independent stabilities as discussed in section 2.

3.2. The first two time steps

For $t = 1$ a straightforward application of equations (3), (7), (8) and (4) gives

$$m_\nu(1) = \langle\langle \xi_i^\nu f(z_i^0)\rangle\rangle_{\xi,z} \tag{11}$$

where $\langle\langle\ldots\rangle\rangle_{\xi,z}$ means averaging over patterns and a distribution of the random sums z_i^t defined in equation (5). By the central limit theorem, z_i^0 is a Gaussian random variable. Its mean value x_i^0 and squared dispersion D_0 can be calculated from equation (9):

$$x_i^0 = \sum_\mu \xi_i^\mu \Delta_\mu m_\mu \tag{12}$$

and, since $\overline{(z_i^0)^2} = 1 + \sum_\mu(\Delta_\mu^2 - 1)m_\mu^2$,

$$D_0 = 1 - \sum_\mu m_\mu^2. \tag{13}$$

Then

$$m_\nu(1) = \int \frac{dy}{\sqrt{2\pi}} \exp(-y^2/2)\ \overline{[\xi_i^\nu f(x_i^0 + \sqrt{D_0}y)]}^{\xi}. \tag{14}$$

Let us turn to the next time step, $t = 2$. By equation (4), we have to average over the conditional distribution $P(z_i^1 \mid \boldsymbol{S}^0)$. Now, the conditional mean value $\langle z_i^1\rangle_{\boldsymbol{S}^0} = \sum_j J_{ij}\langle S_j^1\rangle_{\boldsymbol{S}^0}$ is determined iteratively, by the first time step as calculated above. Here, however, due to the sum weighted by J_{ij}, an important coherent contribution arises. To see this, in

$$\langle S_j^1\rangle_{\boldsymbol{S}^0} = \langle\langle f(z_j^0)P(z_j^0 \mid \boldsymbol{S}^0)\rangle\rangle \tag{15}$$

let us single out from $z_j^0 = \sum_k J_{jk}S_k^0$ both its mean and the fluctuation of the term $k = i$:

$$z_j^0 = x_j^0 + J_{ji}(S_i^0 - \langle S_i^0\rangle) + \sqrt{D_0}y \tag{16}$$

(explicitly adding one fluctuation term gives a negligible change in D_0), and linearise $f(z_j^0)$ in the J_{ji} term which is small since $J_{ji} \propto N^{-1/2}$ whereas $z_j^0 = \mathcal{O}(1)$:

$$(z_j^0) \approx f(x_j^0 + \sqrt{D_0}y) + J_{ji}f'(x_j^0 + \sqrt{D_0}y)(S_i^0 - \langle S_i^0\rangle). \tag{17}$$

Substituting into z_i^1 (see equation (5)), the j summation gives a contribution proportional to η (equation (1)), which is $\mathcal{O}(1)$, expressing the dominant influence of the initial value of a given spin S_i on its own mean value *two* time steps later, through the correlation between J_{ij} and J_{ji} expressed by $\eta \neq 0$.

The rest of the calculation is trivial and gives

$$m_\nu(2) = \left\langle\left\langle \xi_i^\nu \mathrm{Tr}_{S_i^0} \frac{1 + S_i^0 \sum_\mu \xi_i^\mu m_\mu}{2} f(x_i^1 + \sqrt{D_1}y)\right\rangle\right\rangle \tag{18}$$

where

$$x_i^1 = \sum_\mu \xi_i^\mu \Delta_\mu m_\mu(1) + (S_i^0 - \sum_\mu \xi_i^\mu m_\mu)\eta V_{01} \tag{19}$$

$$D_1 = 1 - \sum_\mu m_\mu^2(1) \tag{20}$$

(based on the estimate $\sum_{ik} J_{ij}J_{ik}J_{jk} = \mathcal{O}(N^{-1/2})$ which is not true, e.g., for the fully connected Hopfield net; see Gardner *et al* 1987), and

$$V_{01} = \langle\langle f'(x_i^0 + \sqrt{D_0}y)\rangle\rangle. \tag{21}$$

3.3. $t > 2$: *the framework*

For longer times we have to do some bookkeeping of correlations due to coherent terms mediated by the symmetry η, always connecting the same spin two time units apart. They appear in two different forms: explicitly for $t-2$, $t-4\ldots$; implicitly as Gaussian noise cross-correlations for $t-1$, $t-3\ldots$

Anticipating this structure, let us rewrite equation (8) in the form

$$P_1(S_i^t) = \mathrm{Tr}_{S_i^{t-2}, S_i^{t-4}\ldots}\{\mathrm{Tr}_{\boldsymbol{S}^{t-1}, \boldsymbol{S}^{t-3}\ldots}[W(S_i^t \mid \boldsymbol{S}^{t-1})W(S_i^{t-2} \mid \boldsymbol{S}^{t-3})\ldots] \times P(\boldsymbol{S}^{t-1}, \boldsymbol{S}^{t-3}\ldots \mid S_i^{t-2}, S_i^{t-4}\ldots)\} \tag{22}$$

where

$$P(\boldsymbol{S}^{t-1}, \boldsymbol{S}^{t-3}\ldots \mid S_i^{t-2}, S_i^{t-4}\ldots) = \mathrm{Tr}_{\boldsymbol{S}_{(i)}^{t-2}, \boldsymbol{S}_{(i)}^{t-4}\ldots} W(\boldsymbol{S}^{t-1} \mid \boldsymbol{S}^{t-2})W(\boldsymbol{S}_{(i)}^{t-2} \mid \boldsymbol{S}^{t-3})W(\boldsymbol{S}^{t-3} \mid \boldsymbol{S}^{t-4})W(\boldsymbol{S}_{(i)}^{t-4} \mid \ldots \tag{23}$$

is the conditional probability distribution of the spin variables $\boldsymbol{S}^{t-1}$, $\boldsymbol{S}^{t-3}\ldots$, and $\boldsymbol{S}_{(i)}^t$ is a vector of $N-1$ components with the i component omitted.

At this point we notice from equations (4) and (5) that $W(S_i^t \mid \boldsymbol{S}^{t-1})$ depends on $\boldsymbol{S}^{t-1}$ only through the variable z_i^t. Therefore the conditional probability of spin vectors defined in equation (23) can be contracted into that of n scalar variables $P_n(z_i^{t-1}, z_i^{t-3}\ldots \mid S_i^{t-2}, S_i^{t-4}\ldots)$, and the trace over $\boldsymbol{S}^{t-1}$, $\boldsymbol{S}^{t-3}\ldots$ becomes an average over this distribution function.

Using the restricted self-averaging property, equation (10), and defining the double average (already used in §3.2) as

$$\langle\langle\ldots\rangle\rangle = \int \mathrm{d}z_i^{t-1}\mathrm{d}z_i^{t-3}\ldots P_n(z_i^{t-1}, z_i^{t-3}\ldots \mid S_i^{t-2}, S_i^{t-4}\ldots)\overline{(\ldots)}^\xi \tag{24}$$

starting from equation (3) and carrying out the trivial trace over S_i^t, our final result will appear in the form

$$m_\nu(t) = \mathrm{Tr}_{S_i^{t-2}, S_i^{t-4}\ldots}\langle\langle \xi_i^\nu f(z_i^{t-1})W(S_i^{t-2} \mid z_i^{t-3})W(S_i^{t-4} \mid z_i^{t-5})\ldots\rangle\rangle. \tag{25}$$

Our next task is to obtain an explicit expression for the conditional probability distribution P_n.

3.4. $t > 2$: *calculating the noise distribution*

Since $J_{ii} = 0$ (see equation (1)), none of the terms in $z_i^\tau = \sum_j J_{ij} S_j^\tau$ for a given time τ is immediately correlated with S_i^t for any time. Some harder content to this intuitive statement is that the evolution of any S_j^τ is influenced by that of many spins S_k^t, among which only $k = j$ has a distinguished role, $k = i$ is just one out of many.

For this reason we expect it is reasonable to approximate the joint distribution of variables z_i^τ for different values of τ by a multidimensional Gaussian

$$P_n(z_i^{t-1}, z_i^{t-3} \ldots \mid S_i^{t-2}, S_i^{t-4} \ldots)$$
$$\approx (2\pi)^{-n/2} \mid D \mid^{-1/2} \exp\left(-\frac{1}{2} \sum_{\tau,\tau'=t-1,t-3\ldots} D^{-1}_{\tau\tau'} [(z_i^\tau - x_i^\tau)(z_i^{\tau'} - x_i^{\tau'})] \right) \tag{26}$$

where

$$x_i^\tau = \sum_j J_{ij} \langle S_j^\tau \rangle_{(i)} \tag{27}$$

(here $\langle \ldots \rangle_{(i)}$ means that, on calculating the average, S_i^{t-2}, $S_i^{t-4} \ldots$ have fixed values), and

$$D_{\tau\tau'} = \sum_{jk} J_{ij} J_{ik} (\langle S_j^\tau S_k^{\tau'} \rangle - \langle S_j^\tau \rangle \langle S_k^{\tau'} \rangle) \tag{28}$$

in which the same restriction does not have to be taken into account explicitly since its effect is negligible in the thermodynamic limit. Finally $\mid D \mid$ denotes the determinant of the matrix $D_{\tau\tau'}$.

The rest of our task is to determine the parameters x_i^τ and $D_{\tau\tau'}$ for $\tau, \tau' = t-1, t-3 \ldots$. We expect terms due to strong correlations induced by the symmetry of coupling coefficients and proportional to η arising in two kinds of terms: in x_i^τ as corrections due to the constraint that S_i^{t-2}, $S_i^{t-4} \ldots$ are fixed, and as the non-diagonal elements of $D_{\tau,\tau'}$ which would not appear at all without this effect. All such terms drop out in the case $\eta = 0$, which is then equivalent to the diluted model of Derrida *et al* (1987).

Let us start with

$$x_i^{t-1} = \sum_j J_{ij} \mathrm{Tr}_{S_j^{t-1}, S_j^{t-3} \ldots} \int \mathrm{d}z_j^{t-2} \mathrm{d}z_j^{t-4} \ldots \left[S_j^{t-1} \frac{1 + S_j^{t-1} f(z_j^{t-2})}{2} \frac{1 + S_j^{t-3} f(z_j^{t-4})}{2} \ldots \right]$$
$$\times P(z_j^{t-2}, z_j^{t-4} \ldots \mid S_j^{t-3}, S_j^{t-5} \ldots ; S_i^{t-2}, S_i^{t-4} \ldots). \tag{29}$$

As above, the multivariate distribution P can be approximated by a Gaussian and one has to calculate its mean value vector and dispersion matrix. All that would be already known iteratively from the results determined for earlier time steps, apart from the extra constraint of fixing previous values of a variable different from S_j, namely, of S_i.

The effect of this extra constraint on the dispersion matrix is easily seen to vanish in the $N \to \infty$ limit, even after all summations have been carried out. The shift it causes in the mean values, however, has to be retained because on the operation $\sum_j J_{ij} \ldots$

it gives a coherent contribution proportional to η (as in the $t = 2$ calculation after equation (14)). Indeed, to work out the first of the doubly constrained averages arising,

$$\langle z_j^{t-2}\rangle_{(ji)} = \sum_k J_{jk}\langle S_k^{t-2}\rangle_{(j)} + J_{ji}(S_i^{t-2} - \langle S_i^{t-2}\rangle_{(j)}). \tag{30}$$

The integral in (29), then, differs from those appearing in equations (24) and (25) only by shifting the centre of the Gaussian by the extra J_{ji} term in equation (30) and its homologues for $t-4$, etc. As already exploited in section 3.2, this shift is small since $J_{ij} = \mathcal{O}(N^{-1/2})$; therefore—after accordingly shifting the integration variables—one can linearise the functions $f(z_j^\tau)$ to obtain

$$x_i^{t-1} = \sum_\mu \xi_i^\mu \Delta_\mu m_\mu^{t-1} + \eta \sum_{\tau=t-2,t-4\ldots} \left(S_i^\tau - \sum_\mu \xi_i^\mu m_\mu^\tau\right) V_{\tau,t-1} \tag{31}$$

(and analogously for x_i^{t-3}, etc), where

$$V_{\tau,t-1} = \mathrm{Tr}_{S_j^{t-1},S_j^{t-3}\ldots}\left\langle\left\langle \frac{1}{2} S_j^{t-1} S_j^\tau f'(z_j^\tau) \prod_{\tau'=t-1,t-3\ldots}^{\tau'\neq\tau} \frac{1 + S_j^{\tau'} f(z_j^{\tau'-1})}{2}\right\rangle\right\rangle \tag{32}$$

and we have used equation (9) along with the fact that in the last term of equation (30) the constraint of fixed S_j on the average gives an $\mathcal{O}(N^{-1/2})$ correction for the two-pattern network (cf section 2) and can be neglected. Equation (32) can be rewritten into the elegant, but not particularly useful, form

$$V_{\tau,t-1} = \frac{\hat{c} x_j^{t-1}}{\hat{c} x_i^\tau}. \tag{33}$$

Let us turn to the determination of the dispersion matrix $D_{\tau\tau'}$. By now it should be clear, and it can be confirmed by detailed calculation, that the only non-vanishing matrix elements are those for which $|\tau - \tau'| = 0, 2\ldots$, and even for that, in the double sum of equation (28) only the $j = k$ terms contribute, the others summing to just $\mathcal{O}(N^{-1/2})$. Thus

$$D_{\tau\tau'} = q_{\tau\tau'} - \sum_\mu m_\mu(\tau) m_\mu(\tau') \tag{34}$$

where

$$q_{\tau\tau'} = \langle\langle \mathrm{Tr}_{S_j^\tau S_j^{\tau'}} S_j^\tau S_j^{\tau'} P_2(S_j^\tau, S_j^{\tau'})\rangle\rangle \tag{35}$$

and the double-time distribution function P_2 is defined analogously to equation (22), by omitting the trace over both S_j^τ and $S_j^{\tau'}$. In the Gaussian approximation, for example, for $\tau > \tau'$ it gives

$$P_2(S_j^\tau, S_j^{\tau'}) = \mathrm{Tr}_{S_j^{\tau-2},S_j^{\tau-4}\ldots S_j^{\tau'-2},S_j^{\tau'-4}\ldots} \int \mathrm{d}z_j^{\tau-1}\mathrm{d}z_j^{\tau-3}\ldots$$
$$\left[\frac{1 + S_j^\tau f(z_j^{\tau-1})}{2}\frac{1 + S_j^{\tau-2} f(z_j^{\tau-3})}{2}\ldots\right] P(z_i^{\tau-1}, z_j^{\tau-3}\ldots \mid S_j^{\tau-2} S_j^{\tau-4}\ldots) \tag{36}$$

where P is a Gaussian with parameters already iteratively determined in the earlier time steps. In particular, the diagonal elements are

$$D_{\tau\tau} = 1 - \sum_{\mu} m_{\mu}^2(\tau). \tag{37}$$

To illustrate the above formalism, we give the formulae pertinent to $t = 3$, to be added to the list of equations (18) to (21):

$$m_{\nu}(3) = \left\langle\left\langle \xi^{\nu} \mathrm{Tr}_{S^1} \frac{1 + S^1 f(z^0)}{2} f(z^2) \right\rangle\right\rangle \tag{38}$$

$$x_2 = \sum_{\mu} \xi^{\mu} \Delta_{\mu} m_{\mu}(2) + \left(S^1 - \sum_{\mu} \xi^{\mu} m_{\mu}(1) \right) \eta V_{12} \tag{39}$$

$$V_{12} = \left\langle\left\langle \mathrm{Tr}_{S^0} \frac{1 + S^0 \sum_{\mu} \xi^{\mu} m_{\mu}}{2} f'(z^1) \right\rangle\right\rangle \tag{40}$$

$$q_{02} = \left\langle\left\langle \mathrm{Tr}_{S^0} \frac{S^0 + \sum_{\mu} \xi^{\mu} m_{\mu}}{2} f(z^1) \right\rangle\right\rangle. \tag{41}$$

Treating longer times is straightforward, apart from the need to calculate averages over Gaussian distributions of more and more dimensions.

4. The $\eta = 0$ case

In the zero-symmetry case, as already noticed by various authors (Krauth *et al* 1988, Crisanti and Sompolinsky 1988, Gutfreund *et al* 1988), the dynamics will be simple and solvable. For the present model, as for the one-pattern case (Krauth *et al* 1988), this happens because not only do the noise centre shifts in equation (31) vanish, but so do all the non-diagonal elements of $D_{\tau\tau'}$ (the connected correlation functions between different time values). By the method used in the present paper this can easily be proven in two steps: one demonstrates that (i) $q_{0\tau} = \sum_{\mu} m_{\mu} m_{\mu}(\tau)$ i.e. $D_{0\tau} = 0\ \forall \tau \neq 0$; and (ii) the vanishing of non-diagonal elements propagates from $(\tau - 1, \tau' - 1)$ to (τ, τ'), which proves the assertion by induction.

In view of this simplification one now has a one-step iteration for the overlaps. Moreover, the requirement of site-independent stabilities can be relaxed to that of fluctuating stabilities Δ_{i1}, Δ_{i2} of the same distribution $P(\Delta_1, \Delta_2)$ on all sites. Then the iteration is

$$m_{\nu}(t+1) = \int \mathrm{d}\Delta_1 \mathrm{d}\Delta_2 P(\Delta_1 \Delta_2) \left\langle\left\langle \xi^{\nu} f\left(\sum_{\mu} \xi^{\mu} \Delta_{\mu} m_{\mu}(t) + \sqrt{D_0} z \right) \right\rangle\right\rangle \tag{42}$$

where z is a Gaussian variable of zero mean and unit dispersion, and

$$D_0 = 1 - \sum_{\mu} m_{\mu}^2(t) \tag{43}$$

under the normalisation

$$\sum_j J_{ij}^2 = 1. \tag{44}$$

For the TPN model defined in equation (1) Δ_1 and Δ_2 are sharply determined and $\sqrt{D_0}z$ is the only noise disturbing the pattern retrieval. Then for $\eta = 0$

$$m_\nu(t+1) = \left\langle\left\langle \xi^\nu f\left(\sum_\mu \xi^\mu \Delta_\mu m_\mu(t) + \sqrt{D}z\right)\right\rangle\right\rangle \tag{45}$$

with

$$D = D_0 \tag{46}$$

given by equation (43), which expresses an important property of such strict-stability networks: the reduction of the noise level as any of the patterns is approached by the spin configuration.

The same effect does not work in Hopfield-type asymmetrically diluted networks where the stabilities have a Gaussian distribution, which adds noise terms of amplitudes proportional to $m_1(t)$ and $m_2(t)$. This—for the case of Hebb-rule learning—just compensates the noise reduction terms in equation (43) and for the evolution of $m_\nu(t)$ we obtain an equation of the same form as equation (45), however, with D_0 replaced by

$$D = 1. \tag{47}$$

This enhanced noise level is the main reason why patterns in Hopfield-type networks have to be kept at a rather high level of stability to assure any retrieval.

5. Two-pattern dynamics

For the simpler case $\eta = 0$ we distinguish the TPN defined in equation (1) in which the pattern stabilities are site-independent, and the asymmetrically diluted Hopfield-type network model of Derrida *et al* (1987) in which they are of Gaussian distribution. The evolution of one-pattern dynamics is described in both cases by equation (45). However, the noise level D is different in the two cases: reduced close to the patterns for TPN according to equation (46); constant for diluted Hopfield according to equation (47).

This difference has a decisive influence on the dynamics of the two models, as displayed in the illustrative trajectories (figures 1 and 2), summarised in the phase diagrams of figure 3. Different 'phases' are characterised by a different pattern of fixed points of different types. Such fixed points being located from equation (45) as solutions to $m_\nu(t+1) = m_\nu(t) = m_\nu^*$, their character can be inferred from linearising the system of equations (45) for $\nu = 1, 2$ around such solutions. Then in the phase diagram on the Δ_1, Δ_2 plane a phase boundary indicates the change of stability of a fixed point

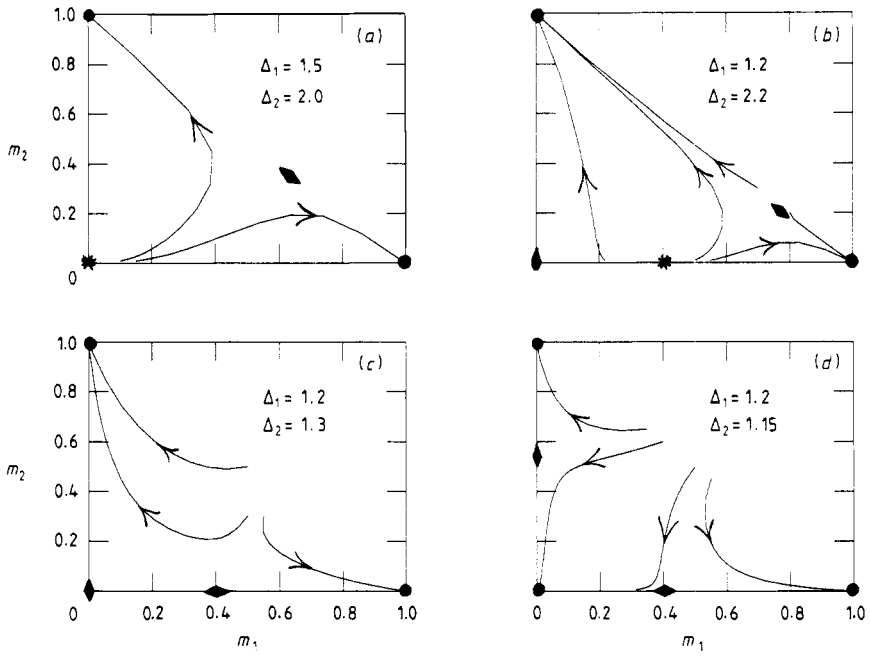

Figure 1. Trajectories of a strict-stability zero-symmetry two-pattern network on the plane of overlaps m_1 and m_2 with the two respective patterns, for different values of the two stabilities: (*a*) both patterns stable on and around their respective axes, (*b*) the strong pattern attracts most initial configurations of some overlap with it and/or weak overlap with the weak pattern, (*c*) the strong pattern attracts initial configurations of weak overlap with the weak pattern, (*d*) only configurations strongly overlapping with one of the patterns are recognised.

on some of the axes in some direction, often accompanied by the bifurcation of a fixed point or vice versa.

The main features can be summarised as follows. Taking $T = 0$ for simplicity, for $\Delta_1 < \sqrt{\pi/2}$ and $\Delta_2 < \sqrt{\pi/2}$ the origin is stable. However, for the diluted Hopfield case it is then the only stable fixed point, whereas for TPN there are also attractors for the patterns. For $\Delta_1 > \sqrt{\pi/2} > \Delta_2$ the origin loses its stability in the 1 direction. As anticipated in section 1, for the diluted Hopfield case both patterns can be retrieved only in a restricted region of the phase diagram bounded by a two-branch curve:

$$\Delta_2 = \sqrt{\frac{\pi D}{2}} \exp\left(\frac{(\Delta_1 m_1^*)^2}{2D}\right) \tag{48}$$

if

$$m_1^* = \operatorname{erf}(\Delta_1 m_1^* / \sqrt{D}) \tag{49}$$

has a non-trivial solution, and vice versa.

The regions of strong imbalance between Δ_1 and Δ_2 on both phase diagrams: phases *b* and *f* are those in which the anticipated effect appears most clearly, but in different forms in the two cases: for the Hopfield-type model the weaker pattern is

unstable there with respect to displacements towards the stronger one, whereas for the TPN the basin of attraction of the weak pattern is shrinking because of a hyperbolic fixed point (actually a couple of them from above and below) approaching from the plane.

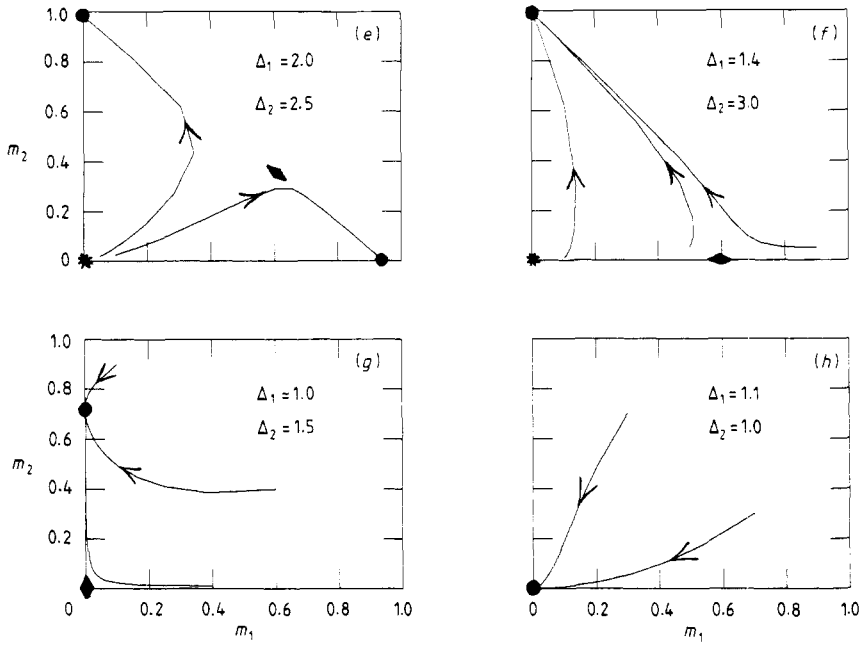

Figure 2. The same as figure 1 for a diluted Hopfield-type network: (*e*) both patterns stable, (*f*) the weaker pattern is unstable against displacements towards the strong one, (*g*) the weaker pattern is fully unstable, (*h*) both patterns unstable.

For $\eta \neq 0$ we state here only preliminary results. These seem to indicate that—slightly differing from the statement of Krauth *et al* (1988)—it is not the high stabilities but the vicinity of a decision surface which renders a trajectory very sensitive to a change of η (figure 4). There is also some indication that in such places the evolution of $m_\nu(t)$ may be an average over very different individual trajectories. The η-dependence of the evolution can be studied advantageously by calculating the distance of two configurations (Derrida *et al* 1987, Gardner *et al* 1987). Our method is applicable to this task too; the calculations are currently in progress.

As a feature pertinent to our starting problem: the loss of retrievability of a pattern because of the presence of a stronger one, $\eta \neq 0$ seems to promote the deterioration of the weaker pattern. If this is an undesirable feature, then this is another reason for looking for learning algorithms balanced close to $\eta = 0$ (Krauth *et al* 1988).

6. Discussion

The effect mentioned in section 1, namely the competition between two patterns of different stabilities, has been studied here mainly for the zero-symmetry case which is

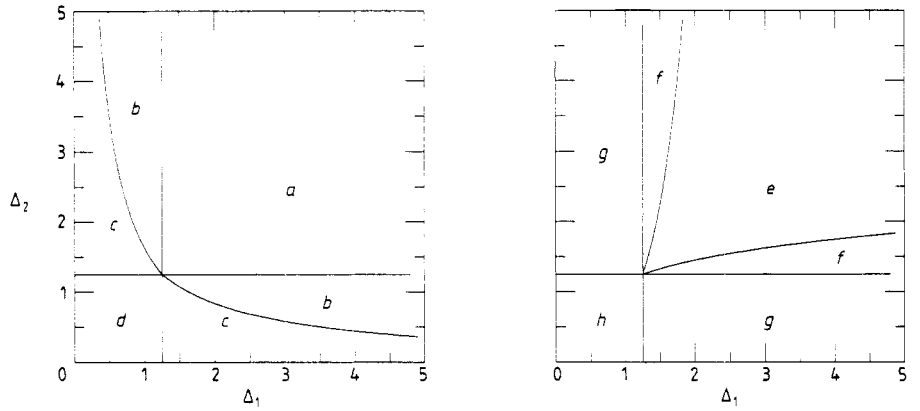

Figure 3. Phase diagrams of two zero-symmetry two-pattern networks: (*a*) strict stability, (*b*) diluted Hopfield (phases are denoted by letters referring to figures 1 and 2, indicating the different types of trajectories).

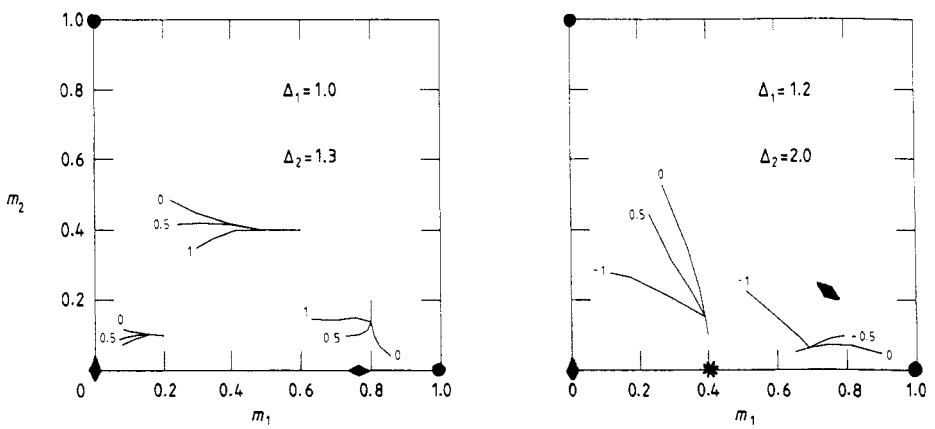

Figure 4. Some trajectories for non-vanishing symmetry of the connection strengths (numbers are the values of the symmetry parameter η defined in equation (1); fixed points pertinent to $\eta = 0$ are shown for orientation).

solvable like the diluted Hopfield-type model, and differs from the latter only by the effect of noise reduction close to the pattern configurations.

We have presented a novel method applicable to the study of the case of arbitrary symmetry up to arbitrarily long times, at the expense of calculating multidimensional integrals over a Gaussian distribution. The consequences of non-vanishing symmetry on the dynamics with patterns of different acquisition strengths will be exploited in further work.

Our method can be extended to the case when higher-order coupling loops like $\sum_{j,k} J_{ij}J_{jk}J_{ki}$ are non-vanishing. For a non-vanishing symmetry of the couplings it is more difficult to do without the beneficial self-averaging effect of the site-independence

of the gauge-invariant coupling combinations (equation (10)).

As pointed out by Virasoro (private communication), there can be a broad analogy between the loss of stability of a weak pattern in the presence of a strong one as discussed here, and the so-called *regularisation* observed in some patients suffering from *dyslexia* and modelled through a lesioned feed-forward neural network (Virasoro 1988, 1989). In these cases the network loses its ability to associate exceptional responses to exceptional patterns in the presence of a dominant rule covering most patterns: all patterns create a response obeying the strong rule. Further work is needed to decide whether the connection between the two phenomena is more than superficial.

Acknowledgments

We are indebted to Jean-Pierre Nadal for most helpful discussions in the initial stage of this work and for encouragement, and to M A Virasoro for pointing out the possible connection with the regularisation problem. TG is indebted to Professors Roger Serneels and Marc Bouten for their hospitality at Limburgs Universitair Centrum, Diepenbeek, Belgium, where the idea of this work arose. Our work was partly supported by the Hungarian Research Foundation OTKA, contract No 3150113.

References

Amit D J, Gutfreund H and Sompolinsky H 1987 *Ann. Phys., NY* **173** 30
Crisanti A and Sompolinsky H 1988 *Phys. Rev.* A **37** 4685
Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167
Gardner E, Derrida B and Mottishaw P 1987 *J. Physique* **48** 741
Geszti T and Pázmándi F 1987 *J. Phys. A: Math. Gen.* **20** L1299
Gutfreund H, Reger J D and Young A P 1988 *J. Phys. A: Math. Gen.* **21** 2775
Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554
Krauth W and Mézard M 1987 *J. Phys. A: Math. Gen.* **20** L745
Krauth W, Nadal J-P and Mézard M 1988 *J. Phys. A: Math. Gen.* **21** 2995
Little W A 1974 *Math. Biosci.* **19** 101
Mézard M, Nadal J-P and Toulouse G 1986 *J. Physique* **47** 1457
Nadal J-P, Toulouse G, Changeux J-P and Dehaene S 1986 *Europhys. Lett.* **1** 535
Parisi G 1986 *J. Phys. A: Math. Gen.* **19** L617
Peretto P 1988 *J. Physique* **49** 711
Virasoro M A 1988 *Europhys. Lett.* **7**
—— 1989 *J. Phys. A: Math. Gen.* **22** 2227