Order parameter evolution in a feedforward neural network

1995 J. Phys. A: Math. Gen. 28 1603

(http://iopscience.iop.org/0305-4470/28/6/015)

J. Phys. A: Math. Gen. 28 (1995) 1603–1614. Printed in the UK

Order parameter evolution in a feedforward neural network

K Y M Wong†, C Campbell‡ and D Sherrington§

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

Received 17 October 1994

Abstract. We consider layered neural networks in which the weights are trained with the pseudoinverse rule to store a set of random patterns. Using many-body diagrammatic techniques, the evolution in the network can be described by the overlap order parameter m and the noise parameter Δ. Looping effects are shown to be significant, in contrast to a previous conjecture. Order parameter pairs corresponding to various input conditions are found to collapse on a universal curve.

1. Introduction

In recent years, statistical mechanics has been successfully applied to the study of neural networks [1]. The advantage of this approach lies in its ability to focus on simplifying but representative models which retain the gross features of biological and technological neural networks, whilst also allowing the macroscopic description of many-body physics. By considering the model of layered networks [2] trained with the pseudoinverse rule [3, 4], this paper is an illustration of this approach, which quantifies the rich network dynamics by statistical parameters, uses many-body diagrammatic techniques, explores a new class of exactly solvable models and corrects a previous conjecture [2].

The dynamics of highly connected neural networks is complex. In fully connected networks the number of correlation parameters grows quadratically with time, restricting their study to the first few time steps [5]. Nevertheless, recent progress has been made in reducing the number of order parameters using a self-averaging and equipartitioning ansatz [6]. Up to now this applies to networks with local weight prescriptions such as the Hebb rule [7], but the non-local cases such as the pseudoinverse rule [3,4] or the maximum stability rule [8] remain unsolved. On the other hand, the dynamics of extremely dilute networks is also exactly solvable [9]; since correlations beyond one time step are negligible, the small number of parameters facilitates its extensions to networks with various weight prescriptions, local and non-local included [10].

Intermediate between the fully connected and extremely diluted architectures, the feedforward layered networks of Domany, Kinzel and Meir [2] resemble, on the one hand, the fully connected models in their full connectivity between successive layers and, on the other hand the widely applied multi-layer perceptrons [11] in their feedforward

† E-mail address: phkywong@usthk.ust.hk

‡ Permanent address: Department of Engineering Mathematics, University of Bristol, Queen's Building, Bristol BS8 1TR, UK (E-mail address: campbell@bristol.ac.uk).

§ Permanent address: Department of Physics, University of Oxford, Theoretical Physics, 1 Keble Road, Oxford OX1 3NP, UK (E-mail address: sherr@thphys.ox.ac.uk).

0305-4470/95/061603+12$19.50

nature. However, unlike the fully connected networks, in which the noise distribution is non-Gaussian because of their feedback nature [6], layered networks have Gaussian noise distributions [2], implying a more tractable treatment. Yet the presence of loops in the information path among consecutive layers produces a much more complex dynamical behaviour when compared with extremely diluted networks. Taking into account the Gaussian noise distribution and the looping effects, the evolution equations have been derived for layered networks in which the stored patterns are embedded by the Hebb rule, but layered networks with more general weight prescriptions have not been studied. The purpose of this paper is to analyse the evolution of the pseudoinverse layered networks (PIL); results will be extended to networks with more general weight prescriptions in a future paper [12]. It is hoped that the present work will contribute to solving the dynamics of networks with general architecture and general weight prescriptions.

While the replica method has been the popular approach in the theoretical study of macroscopic behaviours of neural networks and spin glasses [13], it is not as convenient as the diagrammatic approach [14] in dealing with the microscopic correlations of patterns. In contrast to the replica method, which performs the pattern averaging procedure at an early stage and subsequently obscures the microscopic correlations, the diagrammatic approach performs the pattern averaging exlicitly term by term, rendering it convenient to compute pattern correlations by series summation. Diagrammatics are discussed in this paper, while the alternative replica approach will be published elsewhere [12].

As shown in this paper, the evolution in PIL can be described by the overlap order parameter m and the noise parameter Δ, analogous to the fully connected Hopfield network in the Gaussian approximation [15] and the self-averaging and equipartitioning ansatz [6], and the layered network with Hebbian weight precription [2]. Furthermore, looping effects are shown to be significant in PIL, in contrast to a previous conjecture [2]. Supported by Monte Carlo simulations, order parameter pairs corresponding to various input conditions are found to collapse onto a universal curve.

2. Formulation

In the layered neural network there are N neurons on each layer, each may take the states $S_i(l) = \pm 1$, where l and i are the layer and neuron indices respectively. Below we employ the notation $S(l) = (S_1(l), \ldots, S_N(l))$. Each neuron $S_i(l+1)$ is fed by all neurons $S_j(l)$ on the previous layer through a set of weights $J_{ij}(l)$. The network dynamics is then given by

$$\text{Prob}(S_i(l+1) = \pm 1) = \frac{\exp(\pm \beta h_i(l+1))}{2\cosh(\beta h_i(l+1))} \tag{2.1}$$

where $h_i(l+1) = \sum_j J_{ij}(l) S_j(l)$. β is the inverse temperature and, since generalization to non-zero temperature is straightforward, we focus on the case of zero temperature in this paper:

$$S_i(l+1) = \operatorname{sgn} h_i(l+1). \tag{2.2}$$

The network dynamics is therefore feeding forward without recurrence. The layered network is assigned to retrieve $p \equiv \alpha N$ sequences of patterns, labelled by $\xi_{i\mu}(l) = \pm 1$ for node i, layer l and pattern μ. To achieve this, the pattern information has to be encoded in the weights $J_{ij}(l)$ through a learning process. The most direct (but not necessarily

the most efficient) weight prescription (or learning rule) is the Hebb rule, in which $J_{ij}(l) = \sum_\mu \xi_{i\mu}(l+1)\xi_{j\mu}(l)$, and evolutionary equations have been derived [2]. It has a low storage capacity of $\alpha_c = 0.27$, and the stored patterns cannot be perfectly retrieved even below the storage capacity. In this paper we consider another common weight prescription—the pseudoinverse rule [3, 4]. It has an explicit algebraic form convenient for analysis, as shown by studies in fully connected networks [16–19]. Its storage capacity is as high as $\alpha_c = 1$, and perfect retrieval of patterns is possible.

The pseudoinverse prescription is given by the condition that the aligning fileds are unity for all pattern bits labelled by i, μ and $l+1$:

$$\sum_j J_{ij}(l)\xi_{j\mu}(l) = \xi_{i\mu}(l+1). \tag{2.3}$$

The solution to this equation is given by

$$J_{ij}(l) = \frac{1}{N}\sum_{\mu\nu} \xi_{i\mu}(l+1)[Q(l)^{-1}]_{\mu\nu}\xi_{j\nu}(l) \tag{2.4}$$

where $Q(l)$ is the correlation matrix given by

$$Q_{\mu\nu}(l) = \frac{1}{N}\sum_j \xi_{j\mu}(l)\xi_{j\nu}(l). \tag{2.5}$$

An input state $S(1)$ is presented to the input layer $l = 1$. Assuming that $S(1)$ has a non-vanishing overlap $m(1)$ with pattern 1 only, we are interested in monitoring the evolution of the overlap $m(l)$ in subsequent layers, given by

$$m(l) = \frac{1}{N}\sum_i \xi_{i1}(l)S_i(l). \tag{2.6}$$

Separating the local field into a signal and noise term, we arrive at

$$m(l+1) = \frac{1}{N}\sum_i \operatorname{sgn}\left[\xi_{i1}(l+1)\sum_j J_{ij}(l)\xi_{j1}(l)m(l) + X_i(l+1)\right] \tag{2.7}$$

where $X_i(l+1) = \xi_{i1}(l+1)\sum_j J_{ij}(l)[S_j(l) - m(l)\xi_{j1}(l)]$ is the noise term. For randomly chosen patterns on layer $l+1$, the noise terms $X_i(l+1)$ are independently distributed. Thus the noise has a Gaussian distribution with mean $\langle X_i(l+1)\rangle_i = 0$ and variance $\langle X_i(l+1)^2\rangle_i = \Delta(l)$. Substituting (2.3) into (2.7), and averaging over the Gaussian distribution of $X_i(l+1)$, we obtain

$$m(l+1) = \operatorname{erf}\left(\frac{m(l)}{\sqrt{2\Delta(l)}}\right). \tag{2.8}$$

The next step is to derive a recursion relation for $\Delta(l+1)$ in terms of $m(l)$ and $\Delta(l)$. As demonstrated in the Hebbian case [2], a recursion relation in terms of $m(l)$ and $\Delta(l)$ can then be used to describe the evolution of the macroscopic behaviour. Consider the expression for $\Delta(l)$:

$$\Delta(l) = \sum_{jk}\langle J_{ij}(l)J_{ik}(l)\rangle_i[S_j(l) - m(l)\xi_{j1}(l)][S_k(l) - m(l)\xi_{k1}(l)]. \tag{2.9}$$

It can be divided into the $j = k$ and $j \neq k$ terms. For $j = k$, it was shown that [20]

$$\sum_j \langle J_{ij}(l)^2 \rangle_i = \frac{\alpha}{1-\alpha}. \tag{2.10}$$

In extremely dilute networks, the term with $j \neq k$ vanishes, because the probability of having a loop in the network structure is negligible, and the dynamical contributions at distinct nodes j and k are statistically independent. In the layered network, however, contributions from distinct j and k are correlated, since they receive input from all and the same nodes in the previous layer, forming multiple interlayer loops. Hence we have

$$\Delta(l) = \frac{\alpha}{1-\alpha}[1 - m(l)^2] + \sum_{j \neq k} \langle J_{ij}(l) J_{ik}(l) \rangle_i [S_j(l) - m(l)\xi_{j1}(l)][S_k(l) - m(l)\xi_{k1}(l)]. \tag{2.11}$$

In the following section we will evaluate the weight correlation $\langle J_{ij}(l) J_{ik}(l) \rangle_i$ using the diagrammatic approach, yielding a result proportional to the pattern correlation $\sum_\mu \xi_{j\mu}(l)\xi_{k\mu}(l)/Np$. This reduces the second term of (2.11) to an expression involving variables on layer l only—the same form which appeared in the analysis of the Hebbian case [2]. As demonstrated in section 4, the decomposition technique developed in the Hebbian case can then be generalized to obtain the recursion relation.

3. The diagrammatic approach

Using the explicit form (2.4) for the weights we have, for $j \neq k$,

$$\langle J_{ij}(l) J_{ik}(l) \rangle_i = \frac{1}{N^2} \sum_{\mu\nu\lambda\rho} \langle \xi_{i\mu}(l+1)\xi_{i\lambda}(l+1) \rangle_i [Q(l)^{-1}]_{\mu\nu}(l)\xi_{j\nu}(l)[Q(l)^{-1}]_{\lambda\rho}(l)\xi_{k\rho}(l). \tag{3.1}$$

For random patterns on layer $l+1$, non-vanishing contributions come from terms with $\mu = \lambda$, yielding

$$\langle J_{ij}(l) J_{ik}(l) \rangle_i = \frac{1}{N^2} \sum_{\mu\nu} \xi_{j\mu}(l)[Q(l)^{-2}]_{\mu\nu}\xi_{k\nu}(l). \tag{3.2}$$

Since dependence on layer $l+1$ has been averaged out, hereafter the layer label l will be omitted for convenience. The correlation matrix Q^{-2} can be written as

$$Q^{-2} = \lim_{\omega \to 0} \frac{\partial}{\partial \omega}(-\omega^{-1}) \sum_{r=0}^{\infty} (-\omega^{-1})^r Q^r. \tag{3.3}$$

Substituting the geometric series into (3.2), and expressing the correlation matrices in terms of their components (2.5), an infinite series involving products of the pattern bits is obtained. Non-vanishing contributions are obtained by diagrammatic contractions and the resulting series conveniently re-summed to yield a simple form for $\langle J_{ij}(l) J_{ik}(l) \rangle_i$. Similar manipulations have been encountered in the Adaline learning of the perceptron and the contraction rules can be generalized to our case [14]. We use a slanted line to represent a pattern bit, the top and bottom ends of the line corresponding to pattern label μ and node

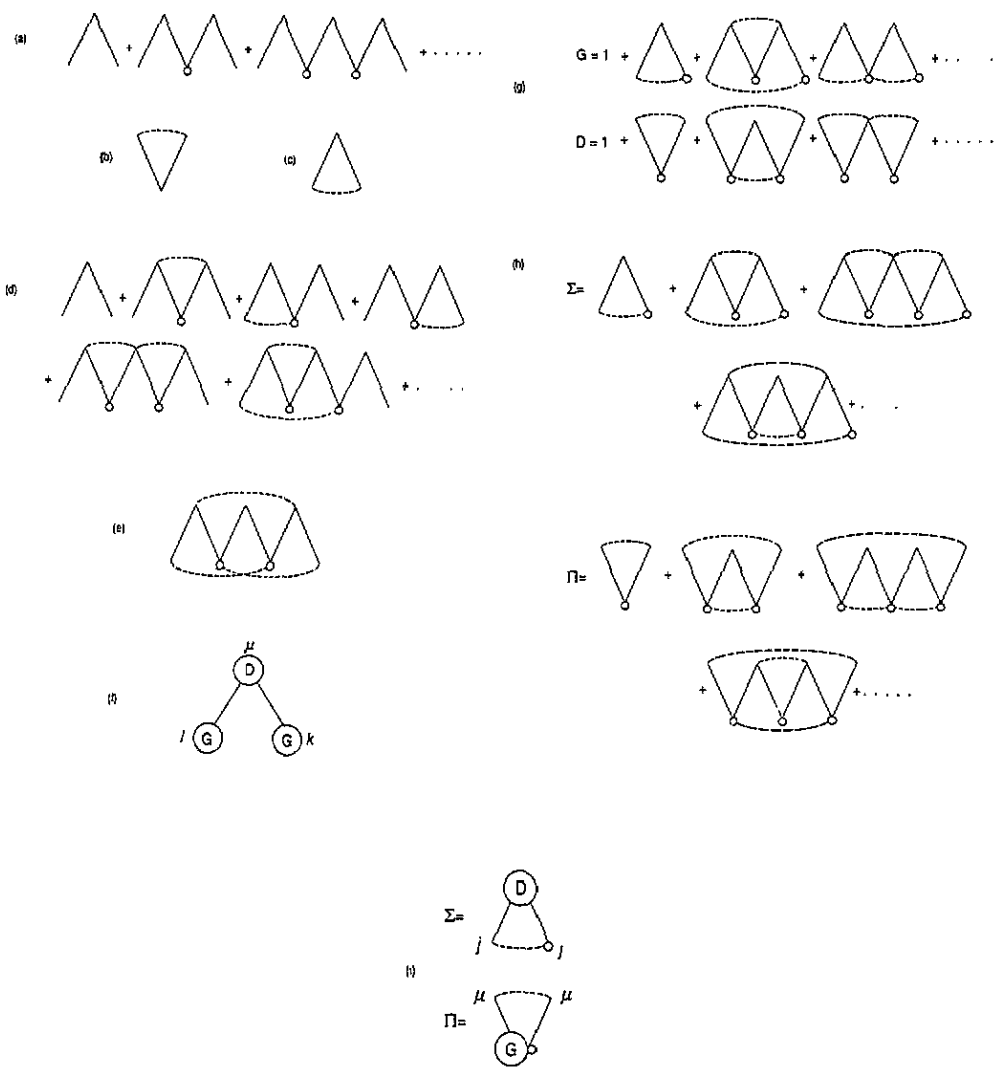

Figure 1. (a) Diagrammatic expansion of (3.2), where $\langle J_{ij}J_{ik}\rangle_i = \lim_{\omega\to 0}\partial/\partial\omega\,(-\omega N)^{-1}$(diagram); (b) pattern pairing in (3.4); (c) node pairing in (3.5); (d) diagrammatic expansion of (3.2) with pairings; (e) a crossing diagram; (f) the skeleton diagram; (g) the diagrammatic representations of G and D; (h) the self-energies Σ and Π for G and D respectively; (i) expressing Σ and Π in terms of G and D.

label j of pattern $\xi_{j\mu}$, respectively. A factor of N^{-1} is associated with the summation of a pattern or node label. Thus a matrix element $Q_{\mu\nu}$ is represented by two slanted lines connected at the bottom, sharing a common node label. $-\omega^{-1}$ is represented by a circle. Thus (3.2) is given by $\lim_{\omega\to 0}\partial/\partial\omega(-\omega N)^{-1}$ acting on the series in figure 1(a).

We note that a pairing of pattern labels across a matrix element yield a factor 1 because

$$\frac{1}{N}\sum_j \xi_{j\mu}\xi_{j\nu} = \delta_{\mu\nu} \tag{3.4}$$

whereas a pairing of node labels across adjacent matrix elements yield a factor α because

$$\frac{1}{N}\sum_{\mu}\xi_{j\mu}\xi_{k\mu}=\alpha\delta_{jk}. \tag{3.5}$$

Pattern pairings are represented by broken lines above the full lines as illustrated in figure 1(b), and node pairings below as illustrated in figure 1(c). The sum of figure 1(a) over the unrestricted node and pattern labels reduces to the restricted set shown in figure 1(d).

Noting that 'crossing' diagrams (such as illustrated in figure 1(e)) do not contribute in the thermodynamic limit ($N \gg 1$) [13], the diagrams in figure 1(d) can be factorized into the skeleton diagram in figure 1(f), corresponding to

$$\langle J_{ij}J_{ik}\rangle_i=\frac{1}{N^2}\sum_{\mu}\xi_{j\mu}\xi_{k\mu}\frac{\partial}{\partial\omega}[-\omega^{-1}G^2D] \tag{3.6}$$

where G is a 'node propagator' and D a 'pattern propagator' (in the language of many-body physics) as shown in figure 1(g). G and D can be expressed via the Dyson-like equations as

$$G=1+\Sigma+\Sigma^2+\cdots=(1-\Sigma)^{-1} \tag{3.7a}$$

$$D=1+\Pi+\Pi^2+\cdots=(1-\Pi)^{-1} \tag{3.7b}$$

where the 'self-energy' Σ contains all those diagrams with the leftmost node label paired with the rightmost node label, and with no other node labels in between (figure 1(h)). It can therefore be expressed in terms of D (see figure 1(i)):

$$\Sigma=\frac{1}{N}\sum_{j}\xi_{j\mu}\xi_{j\mu}(-\omega^{-1}D)=-\frac{\alpha}{\omega}D. \tag{3.8a}$$

Similarly,

$$\Pi=\frac{1}{N}\sum_{\mu}\xi_{j\mu}\xi_{j\mu}(-\omega^{-1}G)=-\frac{G}{\omega}. \tag{3.8b}$$

Eliminating the self-energies from (3.7*a*) and (3.7*b*), we have

$$G^{-1}=1+\frac{\alpha}{\omega}D \tag{3.9a}$$

$$D^{-1}=1+\frac{G}{\omega}. \tag{3.9b}$$

Substituting the solutions into (3.6), we finally arrive at

$$\langle J_{ij}J_{ik}\rangle_i=\frac{\alpha}{1-\alpha}(1-2\alpha)\frac{1}{Np}\sum_{\mu}\xi_{j\mu}\xi_{k\mu}. \tag{3.10}$$

It is interesting to compare this result with the Hebbian case, in which $\langle J_{ij}J_{ik}\rangle_i/\langle J_{ij}^2\rangle_i = \sum_\mu \xi_{j\mu}\xi_{k\mu}/Np$, whereas the present result has an extra factor of $1-2\alpha$. The similarity is a consequence of the assumption that the pattern bits on the different nodes i in the layer $l+1$ are random variables independent of each other and of layer l. Hence the averaging

over i reduces to terms involving patterns on layer l only. The specific weight prescription only affects the magnitude macroscopically, yielding factors of 1 and $1-2\alpha$ for the Hebbian and pseudoinverse rules respectively.

An interpretation of the factor $1-2\alpha$ can be made by squaring (2.3):

$$\sum_{jk} J_{ij}J_{ik}\xi_{j\mu}\xi_{k\mu} = 1. \tag{3.11}$$

Averaging over i and separating terms with $j = k$ and $j \neq k$, we obtain

$$\sum_{j}\langle J_{ij}^2\rangle_i + \sum_{j\neq k}\langle J_{ij}J_{ik}\rangle_i\xi_{j\mu}\xi_{k\mu} = 1. \tag{3.12}$$

Assuming the form $\langle J_{ij}J_{ik}\rangle_i = a\sum_\mu \xi_{j\mu}\xi_{k\mu}/Np$ and employing (2.10), the result (3.10) is recovered. Equation (3.12) means that, for $\alpha < 1/2$, the weight vector has a magnitude $\sum_j J_{ij}^2 = \alpha/(1-\alpha) < 1$. To stabilize the patterns with the aligning field 1, the correlation $\langle J_{ij}J_{ik}\rangle_i$ has to be positive. On the other hand, the weight magnitude is greater than 1 and $\langle J_{ij}J_{ik}\rangle_i$ becomes negative. At $\alpha = 1/2$ the correlation $\langle J_{ij}J_{ik}\rangle_i$ vanishes and, as shown in the next section, looping effects are absent reducing the dynamical equations to those of dilute asymmetric networks.

4. The recursion relations

To obtain a recursion relation for the noise Δ, we substitute (3.10) into (2.11), giving

$$\Delta = \frac{\alpha}{1-\alpha}[1 - m^2 + (1-2\alpha)I] \tag{4.1}$$

where I is the term accounting for the effects of interlayer looping given by

$$I = \frac{1}{Np}\sum_{j\neq k,\mu}\xi_{j\mu}\xi_{k\mu}[S_j - m\xi_{j1}][S_k - m\xi_{k1}]. \tag{4.2}$$

Since the recursion relation for m has been derived in (2.8), it is sufficient to consider the recursion relation for I. Hereafter primed and unprimed variables correspond to layers $l+1$ and l, respectively.

$$I' = \frac{1}{Np}\sum_{j\neq k,\mu}\xi'_{j1}\xi'_{j\mu}\xi'_{k1}\xi'_{k\mu}[\mathrm{sgn}(m + X'_j) - m'][\mathrm{sgn}(m + X'_k) - m']. \tag{4.3}$$

For pattern $\mu = 1$, the contribution to I' vanishes to order $O(N^0)$. For each pattern $\mu > 1$, the above summation is performed by the decomposition technique of [2]. Let $X_j'^{\backslash\mu}$ be the value of X'_j if $\xi'_{j\mu}$ were set to zero. Then

$$X'_j \approx X_j'^{\backslash\mu} + \frac{1}{N}\sum_{l\nu}\xi'_{j1}\xi'_{j\mu}[Q^{-1}]_{\mu\nu}\xi_{l\nu}(S_l - m\xi_{l1}) \tag{4.4}$$

and therefore

$$\mathrm{sgn}(m + X'_j) \approx \mathrm{sgn}(m + X_j'^{\backslash\mu}) + 2\delta(m + X_j'^{\backslash\mu})\frac{1}{N}\sum_{l\nu}\xi'_{j1}\xi'_{j\mu}[Q^{-1}]_{\mu\nu}\xi_{l\nu}(S_l - m\xi_{l1}). \tag{4.5}$$

Substituting into (4.3) and summing over pattern μ, the only non-vanishing term is

$$I' = \sum_{\mu>1}\left\{\frac{1}{N^2}\sum_{j\neq k} 2\delta(m + X_j^{\prime\backslash\mu})2\delta(m + X_k^{\prime\backslash\mu})\right\} \times \left\{\frac{1}{Np}\sum_{l n \nu\lambda}[Q^{-1}]_{\nu\mu}[Q^{-1}]_{\mu\lambda}\xi_{l\nu}\xi_{n\lambda}(S_l - m\xi_{l1})(S_n - m\xi_{n1})\right\}. \tag{4.6}$$

Adding the negligible contributions of $j = k$ and $\mu = 1$, and noting that $X_j^{\prime\backslash\mu} \approx X_j'$ to the lowest order, this reduces to

$$I' = \left\{\frac{1}{N}\sum_j 2\delta(m + X_j')\right\}^2\left\{\frac{1}{Np}\sum_{l n \mu\nu}[Q^{-2}]_{\mu\nu}\xi_{j\mu}\xi_{k\nu}(S_j - m\xi_{j1})(S_k - m\xi_{k1})\right\}. \tag{4.7}$$

In the first bracket, X_j' is a Gaussian variable of variance Δ. The term in the second bracket can be reduced to Δ/α using (2.9) and (3.2). This results in the looping term

$$I' = \frac{2}{\pi\alpha}\exp\left(-\frac{m^2}{\Delta}\right). \tag{4.8}$$

In summary, the recursion relations for the PIL is given by

$$m' = \operatorname{erf}\left(\frac{m}{\sqrt{2\Delta}}\right) \tag{4.9a}$$

$$\Delta' = \frac{\alpha}{1-\alpha}\left[1 - m^2 + (1-2\alpha)\frac{2}{\pi\alpha}\exp\left(-\frac{m^2}{\Delta}\right)\right]. \tag{4.9b}$$

At the input layer, no looping effects are present, and $\Delta(1) = (1 - m(1)^2)\alpha/(1-\alpha)$. Knowing $m(1)$ and $\Delta(1)$, the parameters in the subsequent layers can be obtained iteratively.

These recursion relations are different from the conjecture of [2] that the parameter evolution is identical to the case of dilute asymmetric pseudoinverse networks, in which the looping effects are neglected.

5. Results

Figure 2 shows the evolution of the overlap given by the theory for various initial conditions. For sufficiently high initial overlaps, the network converges to the perfectly retrieved state of pattern 1 after a few layers, whereas for low initial overlaps the network converges to the non-retrieval state with vanishing overlap. Note that in the case of non-retrieval the overlap increases from the input layer to the second layer before monotonically decreasing in subsequent layers, since looping effects have not developed in the second layer. This transient behaviour is also present in the Hebbian layered networks [2], but absent in dilute asymmetric networks [10]. It indicates that a single parameter, namely the overlap m, is not sufficient to describe the evolution, in contrast to the cases of dilute asymmetric networks and the conjecture of [2] for layered networks.

The theory is compared with Monte Carlo simulations. For initial states well inside the basin of attraction of the retrieval state or non-retrieval state, simulations agree with the theory very well. For initial states near the basin boundary, the agreement extends to the

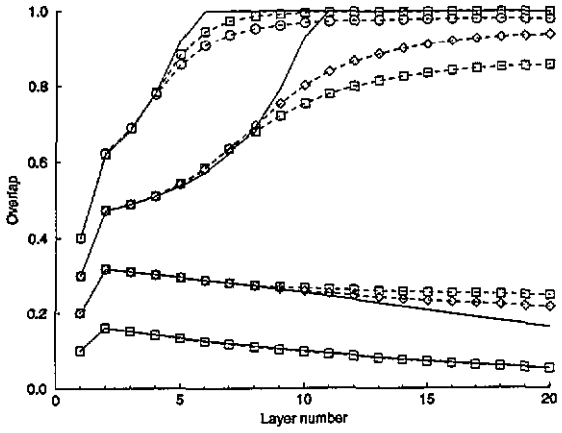

Figure 2. The evolution of the overlap for $\alpha = 0.2$ starting with various initial overlaps. The full and broken curves correspond to the theory and Monte Carlo simulations, respectively. Depending on the extent of finite-size effects, networks with 200, 400 and 800 nodes are used in simulations, represented by circles, squares and diamonds, respectively. 600, 150 and 40 sets of random patterns are generated respectively for each value of initial overlap, and 500 initial configurations are used for each set. Error bars are smaller than the size of the symbols.

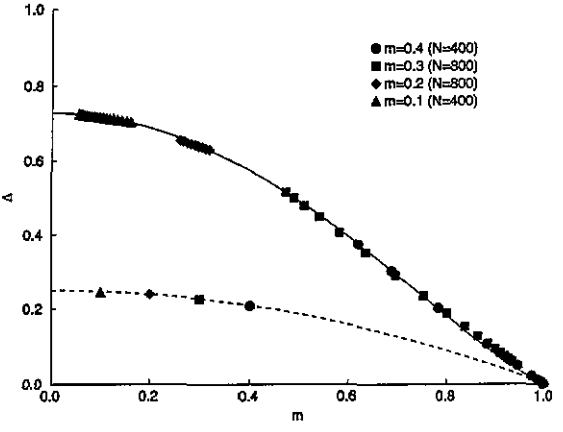

Figure 3. The universal curves in the space of m and Δ for $\alpha = 0.2$. In figures 3 and 4, the broken and full curves correspond to the first and subsequent layers, respectively.

first few layers, and finite-size effects become significant thereafter, but results extrapolated to infinite size agree with the theory.

Figure 3 shows the trajectory of the parameter pairs (m, Δ) for a given value of α. Eliminating parameters of the previous layer from the recursion relation (4.9*a*,*b*), we see that starting from the second layer, the parameters lie on the universal curve

$$\Delta = \frac{\alpha}{1-\alpha}\left\{1 - m^2 + (1-2\alpha)\frac{2}{\pi\alpha}\exp[-2(\mathrm{erf}^{-1}\, m)^2]\right\}. \tag{5.1}$$

On the other hand, the parameters for the input layer lie on the universal curve

$$\Delta = \frac{\alpha}{1-\alpha}(1-m^2). \tag{5.2}$$

Results from Monte Carlo simulations for various layers and initial overlaps are also presented in figure 3. Data for the input layer lie on the universal curve (5.2), and those for subsequent layers lie on the universal curve (5.1). This further supports the theory that looping effects are important in PIL, and invalidates the conjecture of [2], which predicts that data for all layers should fall on the universal curve (5.2).

Figure 4 shows the universal curves and simulation data for a value of $\alpha > 1/2$. Note that since the looping term is negative for $\alpha > 1/2$, the universal curve (5.1) lies inside (5.2).

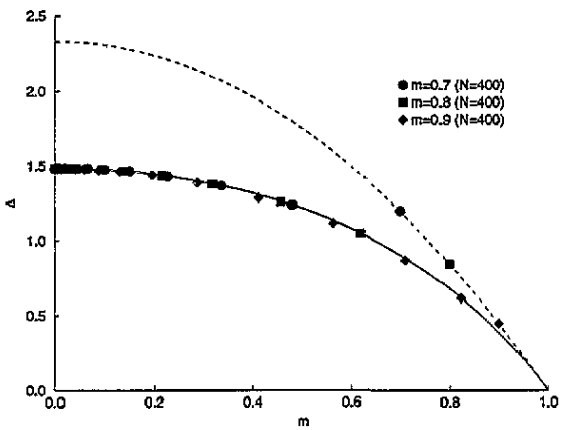

Figure 4. The universal curves for $\alpha = 0.7$.

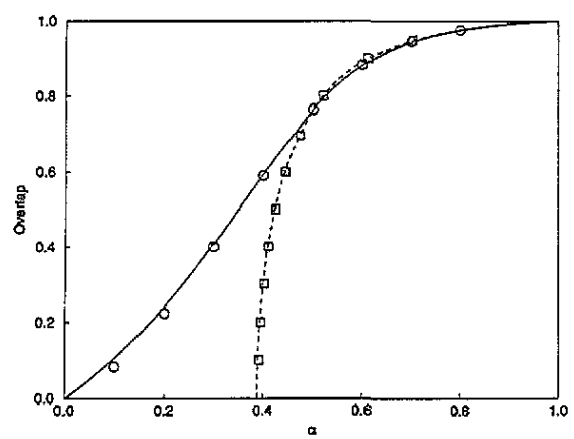

Figure 5. The basin and transient boundaries in the space of the storage level α and initial overlap $m(1)$, represented by full and broken curves respectively. Simulation results are shown in circles and squares for the basin and transient boundaries respectively.

Figure 5 shows the phase diagram in the space of α and initial overlap $m(1)$. The basin boundary separates the retrieval phases on the high $m(1)$ side from the non-retrieval phases on the low side. When α approaches zero, the initial overlap on the basin boundary vanishes linearly with α:

$$m(1) = \sqrt{\frac{3}{2}\left(\frac{\pi}{2} - 1\right)}\alpha. \tag{5.3}$$

When α approaches the storage capacity $\alpha_c = 1$, $m(1)$ at the basin boundary approaches 1. The fully connected pseudoinverse network has a similar phase diagram [16, 17].

We have also shown in figure 5 the basin boundary line conjectured in [2]. Since it still describes the correct evolution from the input to second layer, we rename this line the transient boundary. The basin and transient boundaries divide the phase space into four regions: (i) for high $m(1)$ and low α, the overlap monotonically increases; (ii) for low $m(1)$ and low α, the overlap increases for one layer and then monotonically decreases; (iii) for low $m(1)$ and high α, the overlap monotonically decreases; (iv) for a narrow region of high $m(1)$ and high α, the overlap decreases for one layer and then monotonically increases. All four behaviours are observed in simulations, although in the last case the result is obscured by the small area of this phase and the masking finite-size effects. Simulation results of the basin and transient boundaries agree with the theory.

6. Conclusion

We have studied the parameter evolution in layered networks storing patterns with the pseudoinverse rule. The explicit algebraic form of this rule facilitates the derivation of the evolution equation using the diagrammatic method, which involves an overlap m and a noise Δ. The parameters lie on a universal curve for various initial conditions and a given storage level, and correct a previous conjecture which neglects looping effects [2]. Despite the non-local nature of the pseudoinverse rule, we note that this is the same set of parameters used to describe the dynamics in the case of the local Hebb rule in layered networks [2] and the fully connected networks in the Gaussian approximation [15] and the self-averaging and equipartitioning ansatz [6]. They also form a subset of the parameters describing the dynamics of fully connected networks with the pseudoinverse rule [18, 19].

It is possible to generalize the study to the cases of correlated patterns [14], multi-state and continuous patterns [21], non-zero temperature, static synaptic noise, random dilution and nonlinear synapses [2] and to investigate pattern selectivity [22]. It is also possible to consider the activity distribution of the network, in which the neuronal states are averaged over an ensemble of input patterns [23]. The procedure of 'activity clipping' is shown to give perfect retrieval for all layers over a wide range of storage levels in the PIL.

Generalization to other non-local learning rules such as the maximally stable network [8] is more difficult, since no explicit algebraic form of such rules are available. However, recent progress in the cavity analysis of the class of optimal learning rules shows that the weights can be expressed in terms of the cavity fields [24], and the parameter evolution equations can be derived accordingly [11]. The present step is a significant step towards understanding the dynamics of networks with non-local learning rules, having layered as well as general structures, including the important example of the widely applied 'backprop' networks. We have demonstrated that many-body physics have the necessary tools to understand them.

Acknowledgment

This work is supported by the British Council UK/HK Joint Research Scheme.

References

[1] Hertz J, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Redwood City: Addison Wesley)
[2] Domany E, Kinzel W and Meir R 1989 *J. Phys. A: Math. Gen.* **22** 2081
[3] Kohonen T 1984 *Self Organization and Associative Memory* (Berlin: Springer)
[4] Personnaz L, Guyon I and Dreyfus G 1985 *J. Physique Lett.* **16** 359
[5] Gardner E, Derrida B and Mottishaw P 1987 *J. Physique* **48** 741
[6] Coolen A C C and Sherrington D 1993 *Phys. Rev. Lett.* **71** 3886
[7] Hopfield J J 1982 *Proc. Natl Acad. USA* **79** 2554
[8] Gardner E J 1988 *J. Phys. A: Math. Gen.* **21** 257
[9] Derrida B, Gardner E J and Zippelius A 1987 *Europhys. Lett.* **4** 167
[10] Wong K Y M and Sherrington D 1993 *Phys. Rev.* E **47** 4465
[11] Rumelhart D E, Hinton G E and McClelland J L 1986 *Parallel Distributed Processing* vol 1 (Cambridge: MIT)
[12] Wong K Y M 1994 *Dynamics in a class of optimally trained feedforward neural networks* in preparation
[13] *Mézard M, Parisi G and Virasoro M 1987 Spin Glass Theory and Beyond* (Singapore: World Scientific)
[14] Hertz J, Krogh A and Thorborgssen G I 1989 *J. Phys. A: Math. Gen.* **22** 2133
[15] Amari A and Maginu K 1988 *Neural Networks* **1** 63
[16] Kanter I and Sompolinsky H 1987 *Phys. Rev.* A **35** 380
[17] Opper M, Kleinz J, Köhler H and Kinzel W 1989 *J. Phys. A: Math. Gen.* **22** L407
[18] Henkel R D and Opper M 1990 *Europhys. Lett.* **11** 403
[19] Henkel R D and Opper M 1991 *J. Phys. A: Math. Gen.* **24** 2201
[20] Krauth W, Mézard M and Nadal J-P 1988 *Complex Systems* **2** 387
[21] Bollé D, Shim G M and Vinck B 1994 *J. Stat. Phys.* at press
[22] Rau A, Wong K Y M and Sherrington D 1993 *J. Phys. A: Math. Gen.* **26** 2901
[23] Wong K Y M 1994 *J. Korean Phys. Soc.* **26** S387
[24] Wong K Y M 1994 Microscopic optimal equations and their stability in neural networks *Phys. Rev. Lett.* submitted