Layered neural networks

1989 J. Phys. A: Math. Gen. 22 2081

(http://iopscience.iop.org/0305-4470/22/12/013)

J. Phys. A: Math. Gen. 22 (1989) 2081-2102. Printed in the UK

Layered neural networks

Eytan Domany†, Wolfgang Kinzel‡ and Ronny Meir§

† Department of Electronics, Weizmann Institute of Science, Rehovot 76100, Israel
‡ Institut fur Theoretische Physik, Justus-Liebig-Universitat Giessen, Heinrich-Buff-Ring 16, D-6300 Giessen, Federal Republic of Germany
§ Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125, USA

Received 17 January 1989

Abstract. Exact solutions for the dynamics of layered feedforward neural networks are presented. These networks are expected to respond to an input by going through a sequence of preassigned states on the various layers. The family of networks considered has a variety of interlayer couplings: linear and non-linear Hebbian, Hebbian with Gaussian synaptic noise and with various kinds of dilution. In addition, we also solve the problem of layered networks with the pseudoinverse (projector) matrix of couplings. In all cases our solutions take the form of layer-to-layer recursions for the mean overlap with a (random) key pattern and for the width of the embedding field distribution. The dynamics is governed by the fixed points of these recursions. For all cases, non-trivial domains of attraction of the memory states are found and graphically displayed.

1. Introduction

In this paper we extend results previously obtained for layered feedforward neural networks. The family of networks considered was introduced by us two years ago [1]. Subsequently, for the case of linear Hebbian couplings an exact solution of the dynamics was found, domains of attraction were calculated analytically [2], various extensions of the basic model were discussed [3], and chaotic behaviour of the dynamics was established and studied [4]. Here we present results for such networks with couplings that are not linear Hebbian; we treat the case of non-linearity [5, 6], dilution [6] and also that of couplings obtained by the pseudoinverse method [7, 8].

The main purpose for introducing these networks was to bridge a gap between what we can loosely term physicists' and non-physicists' models.

Contributions of the theoretical physics community to the field of neural networks concentrated primarily (for recent reviews of physicists' contributions see [9]) on the Hopfield model [10, 11] and extensions thereof [6, 9, 12]. The typical physicist's network is characterised by three general features (see, e.g., [13]). First, it is *homogeneous* and uniform. By this we mean that all elements of the network are functionally similar. A second characteristic of these networks is the kind of task they attempt to perform; in most cases the networks studied by physicists deal with *random patterns*. That is, the network is to be designed in such a way that some preset, randomly selected states play a special role in its dynamics. Physicists are interested primarily in typical or average properties of an ensemble of such networks; these properties are usually calculated in the thermodynamic limit, i.e. when the number of basic units

$N \to \infty$. The third attribute of physicists' networks is that in general a 'cognitive event', such as recall of a memory, is associated with a *stable state* of the network's dynamics†.

On the other hand, neural networks were studied by non-physicists (in particular computer scientists) for many years [8, 14, 15]. The networks considered by this community differ, by and large, from physicists' networks in all the above-mentioned respects. These networks are usually not functionally uniform: for example, some units receive input only from the external world, and not from other units of the network. Whereas these elements constitute the *input*, a different set of units passes information out of the network, and serves to generate its *output*. There may also be processing units, which have no direct contact with the external world: all their inputs come from, and all their outputs go to, other units of the network itself.

More importantly, most computer scientists are interested in *finite* networks, that perform *specific*, well defined tasks, such as determining whether a given geometrical figure is singly or multiply connected [16]. The answer of the network to a posed problem is read off its output units at some preset time. Hence stable attractors do not necessarily play a special role in the dynamics of such networks. In many of the most popular and widely studied examples the network architecture is purely feedforward [15].

The family of networks studied by us belongs to the non-physicists' class with respect to the first and third of the attributes listed above. It has an input layer and can be viewed as producing an output on some remote layer; hence it is inhomogenous. Being feedforward, it does not have stable attractors of its dynamics. Instead, a wave of activity that passes the layers sequentially may be interpreted [17] as a 'cognitive event'. On the other hand, with regard to the second attribute, ours is a physicist's network in that random patterns are assigned to *each* layer (including the internal or 'hidden' layers), and average properties are calculated in the thermodynamic limit. We do not address here various learning schemes which generate internal representations that help to realise desired input-output associations [18-20].

Nearly all the analytic results for our model were derived [2, 3] by saddle point integration. In this paper we present solutions to a number of problems that were not previously treated in the context of layered networks. We solve these problems here in a simple manner, based on the observation that the fields generated by cells of layer l on the units of layer $l+1$ are Gaussian distributed independent random variables.

The paper is organised as follows. In § 2 the basic model, with Hebbian couplings, is briefly defined, and the Gaussian independent nature of the fields is established. Next, the basic recursions of the mean overlap and of the width of the field distribution are rederived; these layer-to-layer recursions constitute the solution of the dynamics of our network. One of the most interesting observations that resulted from our studies was that the layered network behaves in a manner which is surprisingly similar to that of recurrent networks. Therefore we devote § 3 to a discussion of the similarities and differences between the layered network and two recurrent networks; the fully connected Hopfield model and the strongly diluted model introduced and solved by Derrida, Gardner and Zippelius [12]. In § 4 we solve the layered model for a number of particular cases; first, the solution of static synaptic noise [5] is rederived [3], and the problem of diluted bonds [6] is solved by mapping it to the previous one. Similarly we solve the model with non-linear synapses (i.e. the bonds are non-linear functions of the Hebbian couplings [5]). Finally we present a solution of a layered network in

† For some applications a stable *cycle* may be embedded in a network.

which the interlayer couplings were obtained by the pseudoinverse method [7, 8, 21]. All the above-mentioned solutions appear in the form of layer-to-layer recursions. These recursions are analysed and the corresponding domains of attraction are calculated and presented in § 5. Our results are discussed and summarised in § 6.

2. The model: its definition and solution by Gaussian transforms

The network is composed of binary valued units (cells, spins) arranged in layers:

$$S_i^l = \pm 1$$

where $l = 1, 2, \ldots, L$ is a layer index, and each layer contains N cells. The state of unit S_i^{l+1} is determined by the state of the units on the previous layer l, according to the stochastic law

$$\begin{aligned} &P(S_i^{l+1} | S_1^l, S_2^l, \ldots, S_N^l) = \exp(\beta S_i^{l+1} h_i^{l+1}) / 2 \cosh(\beta h_i^{l+1}) \\ &h_i^{l+1} = \sum_j J_{ij}^l S_j^l. \end{aligned} \tag{2.1}$$

Here J_{ij}^l is the strength of the connection from cell j on layer l to cell i on layer $l+1$. The quantity h_i^{l+1} is the field produced by the entire layer l on site i of the next layer. The parameter $\beta = 1/T$ controls stochasticity; the $T \to 0$ limit reduces to the deterministic form

$$S_i^{l+1} = \mathrm{sgn}(h_i^{l+1}).$$

The dynamics of such a network can be defined as follows. Initially the first (input) layer is set in some fixed state externally. In response to that all units of the second layer are set synchronously at the next time step, according to the rule (2.1), the next layer follows at the next time step, and so on. Thus the response of the network to an initial state is an 'avalanche' of coherent activity that produces the appearance of a sequence of states, on layer l at *time* l.

An alternative interpretation of this dynamics is as follows. Imagine a single-layer recurrent network with couplings K_{ij}, in which units update their states synchronously (such as is the case of the Little model [22]). If we set in the layered network all $J_{ij}^l = K_{ij}$, i.e. independent of the layer index, the resulting layered network dynamics will be precisely identical to the dynamics of the recurrent network, with the layer index of the former playing the role of time for the latter. Hence *every* recurrent network is equivalent to a properly defined layered feedforward network [18]. Letting the bonds depend on and vary with the layer index is, therefore, completely equivalent to allowing the couplings of a recurrent network to vary with time. Therefore one can interpret solutions of these layered networks as an 'annealed approximation' [23] to the dynamics of networks with feedback.

Returning to our feedforward network, we have to specify the bonds J_{ij}^l. These are chosen so that the network performs a desired task. A reasonable task for a layered network is to require that in response to a particular input, a preset sequence of states develops on subsequent layers. We consider the problem of embedding in the network different random sequences of patterns, associating pattern $\xi_{i\mu}^l$ with layer l; the pattern index μ takes one of $\mu = 1, 2, \ldots, \alpha N$ possible values. These states, $\xi_{i\mu}^l$, are the key patterns of the network. The interlayer couplings can be chosen according to any one of a variety of standard learning procedures. The simplest choice is that of outer product, of Hebbian couplings [10, 22]

$$J^l_{ij} = \frac{1}{N} \sum_{\nu=1}^{\alpha N} \xi^{l+1}_{i\nu} \xi^l_{i\nu}. \tag{2.2}$$

Other choices will be presented in § 4.

By 'solving the model' we mean that for a given initial state, i.e. the state of the first layer, we can predict the state on layer l that results from the network's dynamics. Of course we predict the state in the sense of *averages* over the thermal noise associated with the dynamics, as well as over the choice of key patterns. An initial state is characterised by its overlap with the key patterns on the first ($l=0$) layer:

$$M^0_\mu = \frac{1}{N} \sum_i S^0_i \xi^0_{i\mu}. \tag{2.3}$$

In this paper we consider initial states with finite overlap with *one* key pattern, i.e. $M^0_1 = \mathrm{O}(1)$, whereas for $\mu > 1$ we have $M^0_\mu = \mathrm{O}(1/\sqrt{N})$. With this initial condition we let the network develop in time according to the stochastic dynamic rule (2.1). Denote by M^l_μ the overlap of any particular state $\{S^l_i\}$, obtained in the course of this dynamic process, for some particular choice of the key patterns $\{\xi^l_{i\mu}\}$, by

$$M^l_\mu = \frac{1}{N} \sum_{i=1}^{N} S^l_i \xi^l_{i\mu}. \tag{2.4}$$

Our aim is to calculate the average overlap for $l > 1$:

$$\begin{aligned} m^l_\mu &= \overline{\langle M^l_\mu \rangle} = \frac{1}{N} \sum_i \overline{\langle S^l_i \xi^l_{i\mu} \rangle} \\ &= \overline{\langle S^l_i \xi^l_{i\mu} \rangle}. \end{aligned} \tag{2.5}$$

In this expression the brackets $\langle \cdot \rangle$ denote thermal average over the stochastic dynamic process; the overbar denotes average over the key pattern assignments. With the initial conditions specified above, we expect $M^l_\mu = \mathrm{O}(1/\sqrt{N})$ for all $\mu > 1$, whereas M^l_1 may be of order unity. A network that corrects errors of the input is expected to start with a low but finite initial overlap with one key pattern, and yield increasingly larger overlaps on subsequent layers. In order to compress notation we suppress the layer index, and prime all variables associated with layer $l+1$ (i.e. use S'_i and $\xi'_{i\mu}$). Unprimed variables refer to layer l. Another simplifying notation is the following: all brackets and overbars will refer to averages taken over *primed* variables. The fact that unprimed variables have also to be averaged over is implicitly assumed everywhere; as we will demonstrate, all unprimed averages are taken care of by the law of large numbers, and only primed variables have to be averaged explicitly. Consider therefore

$$m'_1 = \overline{\langle \xi'_{i1} S'_i \rangle}.$$

First we perform the thermal average over the dynamics of the last (primed) layer. From (2.1) we immediately get

$$m'_1 = \overline{\tanh \beta \xi'_{i1} h'_i} = \overline{\tanh \beta H'_i}$$

where H'_i is the 'embedding field' associated with pattern $\nu = 1$. This can be rewritten, using (2.1) and (2.2), as

$$m'_1 = \overline{\tanh \beta \left(\frac{1}{N} \sum_j \xi_{j1} S_j + \frac{1}{N} \sum_{\nu>1} \xi'_{i1} \xi'_{i\nu} \sum_j \xi_{j\nu} S_j \right)}. \tag{2.6}$$

With respect to averaging over the patterns, we explicitly perform averages over ξ', keeping in mind that thermal and configurational averages are to be taken (if needed; see below) for previous layers as well. We can rewrite (2.6) as

$$\overline{m'_1 = \tanh \beta\left(M_1 + \sum_{\nu>1} \xi'_{i1}\xi'_{i\nu}M_\nu \right)}. \tag{2.7}$$

The stochastic variable M_1 is the average of N stochastic variables $\xi_{j1}S_j$. Therefore, using the law of large numbers we have

$$M_1 = m_1 + \mathrm{O}(1/\sqrt{N}). \tag{2.8}$$

Since we assume that $M_1 = \mathrm{O}(1)$, deviations of M_1 from m_1 can be neglected. Hence with respect to the first term in the brackets in equation (2.7), all thermal and configurational averaging has been taken into account. Turning now to the second term,

$$x = \sum_{\nu>1} \xi'_{i1}\xi'_{i\nu}M_\nu \tag{2.9}$$

we will show that it is a Gaussian distributed random variable. The essence of the method used throughout this paper is the treatment of the embedding field (i.e. the argument of tanh in (2.7)) as the sum of a 'signal', m_1, and a Gaussian 'noise' x. Once the noise distribution is known, averages such as (2.7) can be performed by integration.

Note that in (2.9) we also have

$$M_\nu = m_\nu + \mathrm{O}(1/\sqrt{N}).$$

But since $m_\nu = 0$ for $\nu > 1$, we cannot replace M_ν by its average value and neglect its fluctuations. Keeping in mind, however, the fact that for $\nu > 1$ all $M_\nu = \mathrm{O}(1/\sqrt{N})$, we note that the Lindeberg condition (see, e.g., [24]) is satisfied for x.

Therefore we can use the central limit theorem, according to which the stochastic variable x is Gaussian distributed, with mean

$$\bar{x} = 0$$

and variance

$$\Delta^2 = \sum_{\nu,\mu>1} \overline{\xi'_{i\nu}\xi'_{i\mu}}M_\nu M_\mu.$$

Using the fact that $\overline{\xi'_{i\nu}\xi'_{i\mu}} = \delta_{\nu\mu}$, we obtain

$$\Delta^2 = \sum_{\nu>1} M_\nu^2. \tag{2.10}$$

Recall now that M_ν are fluctuating (thermal and configurational fluctuation on layers below $l+1$ were not yet averaged). However, we are summing in (2.10) over αN such independent variables: invoking the law of large numbers, we can write

$$\Delta^2 = \alpha N \overline{\langle M_\nu^2 \rangle} + \mathrm{O}(1/\sqrt{N}). \tag{2.11}$$

In this equation the explicitly displayed averaging refers to the *unprimed* variables.

Thus we can express (2.7) in the form

$$m'_1 = \int \mathrm{d}x \tanh[\beta(m_1 + x)] \frac{\exp[-\frac{1}{2}(x/\Delta)^2]}{\sqrt{2\pi\Delta^2}}. \tag{2.12}$$

This relation constitutes a recursion that determines the average overlap on layer $l+1$ in terms of the average overlap m_1 and the width Δ^2, both characteristic of layer l. To complete the solution, a recursion for the width is also needed, in order to express

$$(\Delta')^2 = \sum_{\mu>1}^{\alpha N} \overline{\langle (M'_\mu)^2 \rangle} \tag{2.13}$$

in terms of m_1 and Δ^2. We must evaluate, for $\mu > 1$,

$$\overline{\langle (M'_\mu)^2 \rangle} = \frac{1}{N} + \frac{1}{N^2} \sum_{i \neq j} \overline{\tanh(\beta \xi'_{i\mu} h'_i) \tanh(\beta \xi'_{j\mu} h'_j)}. \tag{2.14}$$

Here, as before, the thermal average over the state of the last (primed) layer has explicitly been carried out. We still have to average over all ξ, and take the thermal average over previous layers. We first rewrite (2.14) as

$$\begin{aligned} \overline{\langle (M'_\mu)^2 \rangle} &= \frac{1}{N} + \frac{1}{N^2} \sum_{i \neq j} \overline{\tanh(\beta H'_i) \tanh(\beta H'_j)} \\ H'_i &= \xi'_{i\mu} \xi'_{i1} m_1 + M_\mu + \sum_{\nu \neq 1,\mu} \xi'_{i\mu} \xi'_{i\nu} M_\nu . \end{aligned} \tag{2.15}$$

Again, we replaced M_1 by m_1, neglecting fluctuations, but kept M_ν for $\nu > 1$. Averages over ξ' are carried out, as before, by noting that the variables x_i and x_j, defined† as

$$x_i = \sum_{\nu \neq 1,\mu} \xi'_{i\mu} \xi'_{i\nu} M_\nu$$

are independent (for $i \neq j$) and Gaussian distributed, with mean $\bar{x}_i = 0$ and width Δ. Furthermore, denote

$$\xi'_{i\mu} \xi'_{i1} = \eta \qquad \xi'_{j\mu} \xi'_{j1} = \eta' .$$

Hence (2.15) takes the form

$$\begin{aligned} \overline{\langle (M'_\mu)^2 \rangle} = \frac{1}{N} + \frac{1}{4} \sum_{\eta,\eta'=\pm 1} \int \frac{\mathrm{d}x\,\mathrm{d}y}{2\pi\Delta^2} \exp[-\tfrac{1}{2}(x^2+y^2)/2\Delta^2] \\ \times \tanh \beta(\eta m_1 + M_\mu + x) \tanh \beta(\eta' m_1 + M_\mu + y). \end{aligned} \tag{2.16}$$

Noting that $M_\mu = \mathrm{O}(1/\sqrt{N})$ we expand the integrand, and find to leading order in M_μ

$$\overline{\langle (M'_\mu)^2 \rangle} = \frac{1}{N} + I \beta^2 M_\mu^2 \tag{2.17}$$

with the constant I given by

$$I = \left(\int \frac{\mathrm{d}x}{\sqrt{2\pi\Delta^2}} \frac{\exp[-\frac{1}{2}(x/\Delta)^2]}{\cosh^2 \beta(m_1+x)} \right)^2 = (1-q)^2 \tag{2.18}$$

where

$$q = \overline{\langle S_i \rangle^2} = \int \frac{\mathrm{d}x}{\sqrt{2\pi\Delta^2}} \exp[-\tfrac{1}{2}(x/\Delta)^2] \tanh^2[\beta(m_1+x)].$$

† It is very important to realise that the embedding fields H_i and H_j are *not* independent. Their correlation is $M_\mu^2 \sim 1/N$. This correlation gives rise to the layer-to-layer recursive variation of the width parameter Δ', which in turn causes the appearance of non-trivial domains of attraction.

Finally, substituting (2.17) into (2.13) we find that the new width, $(\Delta')^2$, is given by

$$(\Delta')^2 = \alpha + I\beta^2\Delta^2 \qquad I = (1-q)^2. \tag{2.19}$$

Again we used the fact that summing over αN fluctuating variables $(M_\mu)^2$ yields the average of this quantity (with respect to thermal and configurational fluctuations associated with layers below $l+1$).

In summary, the solution of a layered network with random key patterns on each layer and Hebbian interlayer couplings is given by recursions of the form

$$m^{l+1} = \frac{1}{\sqrt{2\pi}} \int \mathrm{d}y \exp(-\tfrac{1}{2}y^2) \tanh[\beta(m^l + \Delta^l y)] \tag{2.20a}$$

$$(\Delta^{l+1})^2 = \alpha + \beta^2 I^l (\Delta^l)^2 \tag{2.20b}$$

with

$$I^l = \left(\int \frac{\mathrm{d}y}{\sqrt{2\pi}} \frac{\exp(-\frac{1}{2}y^2)}{\cosh^2 \beta(m^l + \Delta^l y)} \right)^2 = (1-q^l)^2.$$

In the deterministic limit, $\beta \to \infty$, these recursions become

$$m^{l+1} = \mathrm{erf}(m^l/\sqrt{2}\Delta^l) \tag{2.21a}$$

$$(\Delta^{l+1})^2 = \alpha + (2/\pi)\exp[-(m^l/\Delta^l)^2]. \tag{2.21b}$$

In order to find the overlap on layer $l+1$ we have to iterate these recursions. The initial state determines m_1^0, the overlap on the first layer, and

$$\Delta^0 = \alpha.$$

These recursions were derived previously by a lengthier method. They were analysed for general α and T. We found two 'phases' in the (α, T, m_1^0) space; a memory phase, in which $m_1^l \to m_1^* = 1$ for large l, and a phase of no recall, in which $m_1^l \to 0$. When α and T are such that the memory phase exists, the network flows to $m_1^* = 1$ provided the initial overlap is large enough:

$$m_1^0 > m_c^0(\alpha, T).$$

Phase diagrams (i.e. domains of attraction) in the full (α, T, m_1^0) space can be found elsewhere [2]; here we display results graphically for $T=0$ only (see § 5).

3. Comparison with the Hopfield model

The essential mechanism leading to (2.20) and (2.21) is the Gaussian distribution of the internal fields h_i. For this the layered feedforward structure of the network is important since

$$h_i^{l+1} = \sum_\nu^p \xi_{i\nu}^{l+1} M_\nu^l \tag{3.1}$$

and since the $\xi_{i\nu}^{l+1}$ are independent random variables uncorrelated to the overlaps M_ν^l; h_i^{l+1} is a large sum of uncorrelated random variables. This does not hold for the Hopfield model, i.e. the completely connected feedback network. In that case the internal fields are given by [11]

$$h_i = \sum_\nu \xi_{i\nu} M_\nu \tag{3.2}$$

but M_ν is correlated to $\xi_{i\nu}$. Therefore, the distribution of h_i has a more complicated structure; in particular, in thermal equilibrium the field h_i acting on a site depends on and is correlated with the value of the spin S_i on that site. Nevertheless, the equations describing the stationary states are very similar for both models; the purpose of this section is to present their similarities and differences.

Consider the Hopfield model with couplings

$$J_{ij} = \frac{c_{ij}}{cN} \sum_\nu \xi_{i\nu}\xi_{j\nu}. \tag{3.3}$$

If all bonds are occupied ($c = 1$), the thermal equilibrium states of the model can be calculated by the methods of statistical mechanics [11]. If only a small fraction c of bonds per site is occupied asymmetrically, one can also solve exactly the complete dynamics [12]. In this case we consider $c \ll (\log N)/N$ only; i.e. $K = cN$, the number of bonds connected to a site, is large, but with $K \ll \log N$.

In both versions of the Hopfield model, i.e. fully connected and diluted, as well as in the layered system the average overlap m of the stationary state with one pattern is given by

$$m = \int \frac{\mathrm{d}z}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}z^2) \tanh \beta(m + \Delta z). \tag{3.4}$$

As in (2.20), the overlap is a Gaussian average of an embedding strength $H = \xi_1 h$, with mean $\overline{\langle H \rangle} = m$, and variance $\overline{\langle H^2 \rangle - \langle H \rangle^2} = \Delta^2$. As in § 2, we denote thermal and configurational averaging by brackets and an overbar, respectively. However, the equations for the width Δ of the H distribution differ. One has, for the three models,

$$\begin{aligned} &\Delta^2 = \alpha && \text{extremely diluted} \\ &\Delta^2 = \frac{\alpha}{1 - \beta^2(1-q)^2} && \text{layered} \\ &\Delta^2 = \frac{\alpha q}{[1 - \beta(1-q)]^2} && \text{full feedback} \end{aligned} \tag{3.5}$$

with

$$q = \overline{\langle S_i \rangle^2} = \int \frac{\mathrm{d}z}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}z^2) \tanh^2 \beta(m + \Delta z) \tag{3.6}$$

where $\alpha = p/(cN)$ is the storage capacity.

For the layered and completely connected network Δ depends on the state itself (through q); hence the properties of these models are different from the extremely diluted model. In these two models (3.4)-(3.6) give a discontinuous transition from a high retrieval overlap $m \simeq 1$ to $m = 0$. At zero temperature ($\beta \to \infty$) this transition occurs when the storage capacity α exceeds a critical value $\alpha_c = 0.14$ (Hopfield) and $\alpha_c = 0.27$ (layered), respectively. The diluted model, however, has a continuous transition at $\alpha_c = 2/\pi$.

For the diluted and layered models the dynamics can be solved exactly (in the latter case if the layer index l is interpreted as time step t). Hence for these two models basins of attraction can be calculated. Both cases may be considered as an approximation to the completely connected Hopfield network. The diluted approach is equivalent to neglecting correlations in the full network [25], while the layered structure is an annealed approximation [23], i.e. at each time step the patterns are changed. Solving the dynamics yields a non-trivial basin of attraction only for the layered model; in the diluted case all initial overlaps, except $m = 0$, flow towards the attractor given by (3.4) and (3.5).

We now sketch the solution of the three different models, to see why the stationary state is given by rather similar equations (3.4)-(3.6). Let us assume that the state is condensed into the pattern $\nu = 1$, i.e. $\overline{\langle M_\nu\rangle} = m\delta_{\nu,1}$. For all three models the dynamics (2.1) gives

$$\overline{\langle S_i\rangle} = \overline{\langle \tanh \beta h_i\rangle} \tag{3.7}$$

with $h_i = \Sigma_j J_{ij}S_j$, where we have dropped the time or layer indices.

In the diluted as well as in the layered model, h_i is Gaussian distributed; in the first case because spins at different sites are uncorrelated [12] and in the second case because $\xi_{i\nu}^{l+1}$ are independent random variables, uncorrelated to M_ν^l. Hence, to obtain $m = \overline{\langle S_i\xi_{i1}\rangle}$ it is sufficient to find the first and second moments of the distribution of $H_i = h_i\xi_{i1}$. One obtains

$$\overline{\langle h_i\xi_{i1}\rangle} = \overline{\left\langle \sum_\nu \xi_{i\nu}\xi_{i1}M_\nu \right\rangle} = m \tag{3.8}$$

and

$$\Delta^2 = \overline{\langle (h_1\xi_{i1})^2\rangle} - m^2 = \sum_j \overline{\langle J_{ij}^2\rangle} + \sum_{j\neq k} \overline{\langle J_{ij}S_jJ_{ik}S_k\rangle} - m^2 \tag{3.9}$$

with

$$\sum_j \overline{\langle J_{ij}^2\rangle} = \frac{cN}{c^2N^2}\sum_{\nu\mu} \overline{\langle \xi_{i\nu}\xi_{j\nu}\xi_{i\mu}\xi_{j\mu}\rangle} = \frac{p}{cN} = \alpha. \tag{3.10}$$

In the diluted case the sites j and k are uncorrelated

$$\overline{\langle J_{ij}S_jJ_{ik}S_k\rangle} = \overline{\langle J_{ij}S_j\rangle}\,\overline{\langle J_{ik}S_k\rangle} = m^2/(cN)^2$$

hence one has $\Delta^2 = \alpha$. In the layered model, however, the correlations (of order $1/N$) between different sites yield (see the calculation of § 2)

$$\sum_{j\neq k} \overline{\langle J_{ij}S_jJ_{ik}S_k\rangle} - m^2 = \beta^2(1-q)^2\Delta^2 \tag{3.11}$$

and hence (3.5) results. Now, with the mean m and variance Δ^2, one can perform the average of (3.7) and obtain (3.4).

In the densely connected Hopfield network the field distribution is *not* Gaussian. The results (3.3)-(3.6) have been derived using the replica method with a saddle point which is symmetric in the replicas [11]. One can, however, derive the same results without replicas, by using the cavity approach [26]. This method reveals the physical origin of the Gaussian average of (3.4) and (3.6). Here we want to present the cavity method in a somewhat simpler form.

The thermal average of a spin in a system of N spins is given by [26]

$$\langle S_i\rangle = \langle \tanh \beta h_i\rangle = \tanh(\beta\langle h_i\rangle_{N-1}) \tag{3.12}$$

where $\langle \ldots \rangle_{N-1}$ is the thermal average of the system of $N-1$ spins *without* spin S_i. With (3.2) one has

$$\langle S_i\rangle = \tanh\left(\beta \sum_\nu \xi_{i\nu}\langle M_\nu\rangle_{N-1}\right). \tag{3.13}$$

Note that $\xi_{i\nu}$ is a random number which is now *uncorrelated* to $\langle M_\nu\rangle_{N-1}$; hence the average over $\xi_{i\nu}$ yields a Gaussian distribution of the variable $\langle h_i\rangle_{N-1}$ (instead of h_i in (3.7)). Therefore, the average of (3.13) is again (3.4) with $m = \overline{\langle h_i\rangle_{N-1}}$ and $\Delta^2 = \overline{\langle h_i\rangle^2_{N-1}} - m^2$. One finds

$$\Delta^2 = \overline{\left(\sum_{\nu>1} \xi_{i\nu}\langle M_\nu\rangle_{N-1}\right)^2} = \alpha N \overline{\langle M_\nu\rangle^2_{N-1}}. \tag{3.14}$$

By definition one has

$$\langle M_\nu\rangle_{N-1} = \frac{1}{N}\sum_i \xi_{i\nu}\langle S_i\rangle_{N-1} = \frac{1}{N}\sum_i \xi_{i\nu} \tanh\left(\beta\sum_\mu \xi_{i\mu}\langle M_\mu\rangle_{N-2}\right). \tag{3.15}$$

For $\nu > 1$, $\langle M_\nu\rangle_{N-2}$ is a small quantity of order $1/\sqrt{N}$; hence the tanh can be expanded:

$$\begin{aligned}\tanh\left(\beta\sum_\mu \xi_{i\mu}\langle M_\mu\rangle_{N-2}\right) &= \tanh\left(\beta\sum_{\mu\neq\nu} \xi_{i\mu}\langle M_\mu\rangle_{N-2}\right)\\ &\quad + \left[1-\tanh^2\left(\beta\sum_{\mu\neq\nu}\xi_{i\mu}\langle M_\mu\rangle_{N-2}\right)\right]\beta\xi_{i\nu}\langle M_\nu\rangle_{N-2}\\ &= \langle S_i\rangle_{N-1,p-1} + (1-\langle S_i\rangle^2_{N-1,p-1})\beta\xi_{i\nu}\langle M_\nu\rangle_{N-2}\end{aligned} \tag{3.16}$$

where $\langle\ldots\rangle_{N-1,p-1}$ is the thermal average of $N-1$ spins in a model without pattern ν. With the self-averaging quantity

$$q_{N-1,p-1} = \frac{1}{N}\sum_i \langle S_i\rangle^2_{N-1,p-1} \tag{3.17}$$

one obtains

$$\langle M_\nu\rangle_{N-1} = \frac{1}{N}\sum_i \xi_{i\nu}\langle S_i\rangle_{N-1,p-1} + \beta(1-q_{N-1,p-1})\langle M_\nu\rangle_{N-2}. \tag{3.18}$$

In the thermodynamic limit we expect that adding a spin or a pattern gives rise to a small correlation of order $1/N$; hence we can approximate

$$\langle M_\nu\rangle_{N-1} \simeq \langle M_\nu\rangle_{N-2} \qquad q_{N-1,p-1} \simeq q \tag{3.19}$$

and we obtain

$$\langle M_\nu\rangle_{N-1} = \frac{\sum_i \xi_{i\nu}\langle S_i\rangle_{N-1,p-1}}{N[1-\beta(1-q)]}. \tag{3.20}$$

Note that $\langle S_i\rangle_{N-1,p-1}$ is uncorrelated to $\xi_{i\nu}$; hence we can easily average the square of this equation, which gives

$$\overline{\langle M_\nu\rangle^2_{N-1}} = \frac{Nq}{N^2[1-\beta(1-q)]^2}. \tag{3.21}$$

Using this in (3.14), one obtains the result (3.5). As mentioned above, these results are not exact for the fully connected model. Apparently the approximation involved in choosing the replica symmetric solution of the saddle point equations [11] is equivalent to the approximation used here in (3.19).

4. Applications for particular cases

In this section we use the Gaussian method to evaluate layer-to-layer recursions for the overlap for various choices of the couplings that were not treated previously for layered networks. For completeness' sake we also present the recursions for static synaptic noise, even though these were given elsewhere. Again, we use the same initial conditions on the first layer as before; i.e. significant overlap with key pattern 1, whereas for all $\nu > 1$ we have $M_\nu = \mathrm{O}(1/\sqrt{N})$.

4.1. Static synaptic noise

We consider the same problem as in § 2, but with a noise term added [5] to the Hebbian couplings

$$J_{ij}^{l} = \frac{1}{N} \sum_{\nu}^{\alpha N} \xi_{i\nu}^{l+1} \xi_{j\nu}^{l} + \frac{1}{\sqrt{N}} z_{ij}^{l}. \tag{4.1}$$

Here z_{ij}^{l} are independent, Gaussian distributed random variables, with mean and variance given by

$$[z_{ij}^{l}] = 0 \qquad [z_{ij}^{l} z_{km}^{l}] = \alpha \Delta_0^2 \delta_{ik} \delta_{jm} \delta_{ll'} \tag{4.2}$$

where $[\,\cdot\,]$ denotes average over the synaptic noise. Following the notation of § 2, we have

$$\begin{aligned} m_1^{l} &= [\overline{\langle \xi_{i1}^{l} S_i^{l} \rangle}] \\ &= \overline{\left[\tanh \beta \left(M_1 + \frac{1}{\sqrt{N}} \sum_j z_{ij} S_j \xi_{i1}^{l} + \sum_{\nu > 1} \xi_{i1}^{l} \xi_{i\nu}^{l} M_\nu \right) \right]}. \end{aligned} \tag{4.3}$$

The argument of the tanh has, as before, a 'signal' term, m_1, since

$$M_1 = m_1 + \mathrm{O}(1/\sqrt{N})$$

but now there are two distinct noise terms:

$$x = \sum_{\nu > 1} \xi_{i1}^{l} \xi_{i\nu}^{l} M_\nu \qquad y = \frac{1}{\sqrt{N}} \sum_j z_{ij} S_j \xi_{i1}^{l}.$$

We have shown that x is Gaussian with

$$\bar{x} = 0 \qquad \overline{x^2} = \sum_{\nu > 1} M_\nu^2 = \Delta^2. \tag{4.4}$$

As to y, we have

$$[\bar{y}] = 0$$

whereas

$$[\overline{y^2}] = \frac{1}{N} \sum_{jk} [z_{ij} z_{ik}] S_j S_k = \alpha \Delta_0^2. \tag{4.5}$$

Hence, using the central limit theorem, we conclude that the total 'noise' in the argument of tanh β in (4.3) is Gaussian distributed, with mean zero and variance

$$\delta^2 = \Delta^2 + \alpha \Delta_0^2. \tag{4.6}$$

Therefore we get

$$m_1' = \int \frac{\mathrm{d}u}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}u^2) \tanh \beta(m_1 + \delta u). \tag{4.7}$$

As in § 2, in order to complete the recursion, we must express

$$(\Delta')^2 = \sum_{\nu>1} [\overline{\langle (M_\nu')^2 \rangle}] \tag{4.8}$$

in terms of m_1 and Δ. We have for $\mu > 1$

$$[\overline{\langle (M_\mu')^2 \rangle}] = \frac{1}{N} + \frac{1}{N^2} \sum_{i \neq j} \overline{[\tanh(\beta H_i') \tanh(\beta H_j')]} \tag{4.9}$$

where

$$H_i' = \xi_{i\mu}' \xi_{i1}' m_1 + M_\mu + \sum_{\nu \neq 1,\mu} \xi_{i\mu}' \xi_{i\nu}' M_\nu + \frac{1}{\sqrt{N}} \sum_j z_{ij} S_j \xi_{i\mu}'. \tag{4.10}$$

Here we encounter again two Gaussian noise terms, with the same mean and variance as above. Therefore we obtain, using the same steps and arguments that led to (2.19), the relation

$$(\Delta')^2 = \alpha + I\beta^2\Delta^2 \tag{4.11}$$

$$I = \left(\int \frac{\mathrm{d}u}{\sqrt{2\pi}} \frac{\exp(-\frac{1}{2}u^2)}{\cosh^2 \beta(m + u\delta)} \right)^2. \tag{4.12}$$

The deterministic ($\beta \to \infty$) limit of the layer-to-layer recursions is [3]

$$\begin{aligned} m_1^{l+1} &= \mathrm{erf}(m^l/\sqrt{2}\delta^l) \\ (\Delta^{l+1})^2 &= \alpha + \frac{2}{\pi}\left(\frac{\Delta^l}{\delta^l}\right)^2 \exp[-(m^l/\delta^l)^2] \\ (\delta^l)^2 &= (\Delta^l)^2 + \alpha\Delta_0^2. \end{aligned} \tag{4.13}$$

4.2. Dilution

Consider the layered network with randomly diluted [6] Hebbian couplings:

$$J_{ij}^l = c_{ij}^l \frac{1}{cN} \sum_\nu \xi_{i\nu}^{l+1} \xi_{j\nu}^l \tag{4.14}$$

where c_{ij}^l are independent random variables, chosen from the distribution

$$P(c_{ij}) = c\delta(c_{ij}, 1) + (1 - c)\delta(c_{ij}, 0). \tag{4.15}$$

The concentration of non-zero bonds is c. Using again the notation of § 2, and denoting by [·] averages over the variables c_{ij}^l, we want to calculate

$$\begin{aligned} m_1' &= [\overline{\langle \xi_{i1}' S_i' \rangle}] \\ &= \left[\overline{\tanh \beta \left(\tilde{M}_1 + \sum_{\nu>1} \xi_{i1}' \xi_{i\nu}' \tilde{M}_\nu \right)} \right] \end{aligned} \tag{4.16}$$

where

$$\tilde{M}_\mu = \frac{1}{cN} \sum_{j=1}^{N} c_{ij} \xi_{j\mu} S_j. \tag{4.17}$$

It is important to note that $\tilde{M}_\mu$, as defined here, should also carry a site index i, since it depends on c_{ij}. Fluctuations of $\tilde{M}_\mu$ are caused by three sources; thermal, configurational (ξ) and due to dilution. Since S_j does not depend on the c_{ij}, we have

$$[c_{ij}\xi_{j\mu}S_j] = [c_{ij}]\xi_{j\mu}S_j. \tag{4.18}$$

Therefore, using the law of large numbers for the first term in the argument of the tanh in (4.16), we get again the 'signal'

$$\tilde{M}_1 = m_1 + \mathrm{O}(1/\sqrt{N}). \tag{4.19}$$

As to the noise term

$$x = \sum_{\nu>1} \xi'_{i1}\xi'_{i\nu}\tilde{M}_\nu \tag{4.20}$$

its average and variance are given by

$$[\bar{x}] = 0 \qquad [\overline{x^2}] = \left[\sum_{\nu>1} \tilde{M}_\nu^2\right] \tag{4.21}$$

where again only averages (over patterns and dilution) associated with the *last* layer are explicitly displayed; all averages with respect to earlier layers are implied. The variance of the noise term can be easily rewritten as

$$[\overline{x^2}] = \alpha\left[\frac{1-c}{c}\right] + \sum_{\nu>1} M_\nu^2. \tag{4.22}$$

This variance is reminiscent of (4.6) for the case of static synaptic noise. Indeed, denoting

$$\sum_{\nu>1} M_\nu^2 = \Delta^2 \qquad \frac{1-c}{c} = \Delta_0^2 \tag{4.23}$$

allows us to rewrite (4.22) as

$$[\overline{x^2}] = \delta^2 = \Delta^2 + \alpha\Delta_0^2. \tag{4.24}$$

The average overlap m'_1 is given by the same expression, (4.7), as in the case of static noise; the width of the effective noise distribution is Δ, given by (4.23) and (4.24). Again we obtain

$$[\overline{\langle (M'_\mu)^2\rangle}] = \frac{1}{N} + \frac{1}{4}\sum_{\eta,\eta'=\pm1}\int \frac{\mathrm{d}x\,\mathrm{d}y}{2\pi\delta^2}\exp[-(x^2+y^2)/2\delta^2]$$
$$\times \tanh\beta(\eta m_1 + \tilde{M}_{i\mu} + x)\tanh(\eta' m_1 + \tilde{M}_{j\mu} + y). \tag{4.25}$$

Here we emphasised that $\tilde{M}_\mu$ depends on i and j, two *distinct* site indices. Using the law of large numbers we get

$$\sum_{\mu>1} \tilde{M}_{i\mu}\tilde{M}_{j\mu} = \sum_{\mu>1}[\tilde{M}_{i\mu}\tilde{M}_{j\mu}] = \sum_{\mu>1}[\hat{M}_{i\mu}][\tilde{M}_{j\mu}]$$
$$= \sum_{\mu>1} M_\mu^2 = \Delta^2. \tag{4.26}$$

As before, expanding (4.25) in $\tilde{M}_\mu$, summing over μ, and using (4.26) we obtain precisely the recursion (4.11) for the variable $(\Delta')^2$. Therefore, the problem of dilution maps exactly onto that of static synaptic noise, with the effective variance of the noise given by (4.23).

4.3. *Non-linear synapses*

We turn now to the case of couplings of the form [5, 6]

$$J_{ij}^{l}=\frac{\sqrt{\alpha N}}{N}F\left(\frac{1}{\sqrt{\alpha N}}\sum_{i\nu}^{\alpha N}\xi_{i\nu}^{l+1}\xi_{j\nu}^{l}\right) \tag{4.27}$$

where $F(x)$ is a (generally non-linear) function of x. This general class of models includes some interesting cases, such as clipped synapses (i.e. $J_{ij}=\mathrm{sgn}[\Sigma\xi_{i\nu}^{l+1}\xi_{j\nu}^{l}]$), and selective dilution (see below). Derivation of the recursion for this case is along the same lines as that of the previous sections, and of the similar problem for fully connected Hopfield networks [5, 6]. The restrictions on F are also similar to those given in that case [6].

As before, we wish to calculate the thermal and configurational average of the overlap

$$m_1'=\overline{\langle\xi_{i1}'S_i\rangle}=\overline{\tanh[\beta H_{i1}']}. \tag{4.28}$$

The embedding field is now given by

$$H_{i1}'=\xi_{i1}'\frac{\sqrt{\alpha N}}{N}\sum_j F\left(\frac{1}{\sqrt{\alpha N}}\sum_\nu \xi_{i\nu}'\xi_{j\nu}\right)S_j. \tag{4.29}$$

To perform the average over ξ', we found that H_{i1}' is,to a good approximation, a Gaussian distributed random variable, with mean value and variance given by

$$\overline{H_{i1}'}=m_1\overline{F'} \tag{4.30}$$

where

$$\overline{F'}=\overline{\frac{\mathrm{d}F}{\mathrm{d}x}}$$

and

$$\overline{(H_{i1}'-\overline{H_{i1}'})^2}=(\overline{F'})^2\sum_{\nu>1}M_\nu^2+\alpha[\overline{F^2}-(\overline{F'})^2]. \tag{4.31}$$

Using the notation

$$\sum_{\nu>1}M_\nu^2=\Delta^2 \qquad \overline{F^2}/(\overline{F'})^2-1=\Delta_0^2 \tag{4.32}$$

it is convenient to define a width parameter

$$\delta^2=\Delta^2+\alpha\Delta_0^2 \tag{4.33}$$

and an effective inverse temperature

$$\bar\beta=\beta\overline{F'}. \tag{4.34}$$

In terms of these variables, we obtain for m_1' the same recursion (4.7) as we had for the case of static synaptic noise, but with β replaced by $\bar\beta$:

$$m_1'=\int \mathrm{d}x\frac{\exp(-x^2/2\delta^2)}{\sqrt{2\pi}\delta}\tanh\bar\beta[m_1+x]. \tag{4.35}$$

The recursion for the width is derived as before, and we find for $(\Delta')^2$ the recursion (4.11), again with $\bar\beta$ replacing β.

4.4. *Pseudoinverse*

The network with Hebbian couplings, (2.2), had the disadvantage that for $\alpha > 0$ the internal fields $h_{i\nu}^{l+1} = \Sigma_j J_{ij}^l \xi_{j\nu}^l$ of the patterns $\xi_{i\nu}^l$ had a broad distribution. Hence the patterns are not stable fixed points of the dynamics (2.1) due to the negative Gaussian tail of the distribution of $\xi_{i\nu}^{l+1} h_{i\nu}^{l+1}$. But there is a matrix J_{ij}^l which gives a sharp distribution

$$\xi_{i\nu}^{l+1} h_{i\nu}^{l+1} = 1 \tag{4.36}$$

for all ν, l and i. This does not only hold for random patterns, but for any set $\{\xi_{i\nu}^l\}$ of linearly independent patterns too. Since there are at most N such patterns, the network has a maximal capacity of $\alpha_c = 1$.

This matrix can be calculated from the pseudoinverse [7, 8] of (4.36); for the layered structure one obtains

$$J_{ij}^l = \frac{1}{N} \sum_{\mu\nu} \xi_{i\nu}^{l+1} [C_l^{-1}]_{\nu,\mu} \xi_{j\mu}^l \tag{4.37}$$

with the correlation matrix

$$[C_l]_{\nu,\mu} = \frac{1}{N} \sum_i \xi_{i\nu}^l \xi_{i\mu}^l . \tag{4.38}$$

J_{ij}^l has two properties which are important for an associative memory: (i) it is a projector onto the linear space spanned by the p patterns $\xi_{j\nu}^l$, i.e. the orthogonal space is projected out; (ii) of all matrices for which (4.36) holds, J_{ij}^l has the minimal norm $\Sigma_j (J_{ij}^l)^2$; hence one expects a maximal basin of attraction [27]. The completely connected feedback network has been studied previously [7]. A feedforward network with one layer of couplings (4.37) has recently been considered [21], and the extremely diluted anisotropic network [12] with these couplings was solved exactly [28, 29].

Here we want to solve the layered network with the pseudoinverse couplings (4.37). For simplicity we consider an initial state S_i^0 of layer $l = 0$ which has a non-vanishing overlap m_1^0 with pattern ξ_{i1}^0 only, i.e. one has

$$\langle S_i^0 \rangle = \xi_{i1}^0 m_1^0 . \tag{4.39}$$

The brackets denote here an average over initial states. The internal fields

$$h_i^{l+1} = \sum_\nu \xi_{i\nu}^{l+1} \sum_\mu [C_l^{-1}]_{\nu\mu} m_\mu^l \tag{4.40}$$

are Gaussian distributed, since $\xi_{i\nu}^{l+1}$ is uncorrelated to C_l and m_μ^l. Let us consider the average over the initial state first; one obtains for the second ($l = 1$) layer

$$\langle h_i^1 \rangle = \sum_j J_{ij}^0 \langle S_j^0 \rangle = \xi_{i1}^1 m_1^0 \tag{4.41}$$

and

$$\begin{aligned}
(\Delta^0)^2 &= \langle (h_i^1)^2 \rangle - (m_1^0)^2 \\
&= \left\langle \sum_{jk} J_{ij}^0 S_j^0 J_{ik}^0 S_k^0 \right\rangle - (m_1^0)^2 \\
&= \sum_j (J_{ij}^0)^2 + \left(\sum_j J_{ij}^0 \langle S_j^0 \rangle \right)^2 - \sum_j (J_{ij}^0)^2 \langle S_j^0 \rangle^2 - (m_1^0)^2 \\
&= \sum_j (J_{ij}^0)^2 (1 - (m_1^0)^2).
\end{aligned} \tag{4.42}$$

The norm of the matrix J^0_{ij} is self-averaging with the result [21]

$$\sum_j (J^0_{ij})^2 = \frac{\alpha}{1-\alpha}. \tag{4.43}$$

Hence one obtains

$$m^1_1 = \xi^1_{i1}\langle S^1_i\rangle = \int \frac{dz}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}z^2) \tanh[\beta(m^0_1 + \Delta^0 z)]. \tag{4.44}$$

Note, that $\langle S^1_i\rangle$ does not depend on i, hence one does not have to average over the patterns $\xi^l_{i\nu}$; the average over the initial state S^0_i is sufficient to obtain the equation for the overlap m^1_1.

The result (4.44) is of the same form as the equation for the initial state (4.39). Furthermore, one has $\langle h^1_j h^1_k\rangle = \langle h^1_j\rangle\langle h^1_k\rangle$, and hence $\langle S^1_j S^1_k\rangle = \langle S^1_j\rangle\langle S^1_k\rangle$. Therefore one can apply the calculation that led to (4.41) and (4.42) to any layer l, with the result

$$m^{l+1} = \int \frac{dz}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}z^2) \tanh[\beta(m^l + \Delta^l z)]. \tag{4.45}$$

At zero temperature this reduces to

$$m^{l+1} = \operatorname{erf}[m^l/\sqrt{2}\Delta^l]. \tag{4.46}$$

These recursions are supplemented by that of the width

$$(\Delta^l)^2 = \frac{\alpha}{1-\alpha}(1-(m^l)^2). \tag{4.47}$$

The width of the field distribution is zero at $m^0 = 1$, hence at $T = 0$, $m^l = 1$ is a fixed point as it should be. For $m^0 \neq 1$ the width diverges for $\alpha_c = 1$.

Note that one has

$$q^l = \langle S^l_i\rangle^2 = (m^l)^2 \tag{4.48}$$

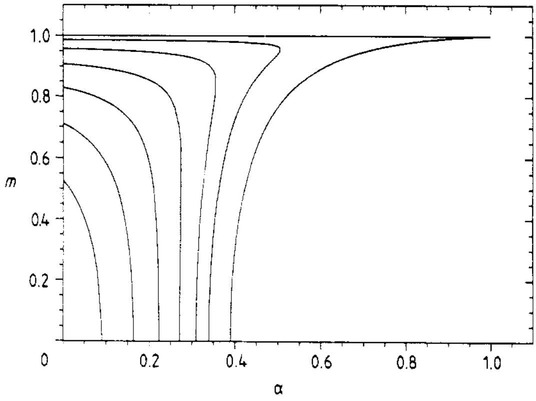

Figure 1. Fixed points of equations (4.46) and (4.47) as functions of storage capacity α for different temperatures T (from left to right: $T = 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0$). For $T \leqslant 0.6$ there exists a lower branch of unstable fixed points. The basin of attraction is the range of overlaps between the upper and lower branches.

which is the same result as in the full feedback network [7]. In both models this holds only if the initial state or the equilibrium, respectively, are condensed into a single pattern. Mixture states add an additional width to the field distribution which is much harder to calculate than (4.46); this has not yet been done for the layered network.

For the corresponding extremely asymmetrically diluted network one obtains [28] the same equations as in the layered model, (4.45)-(4.47).

Figure 1 shows the attractor and the basins of attraction as functions of α for different temperatures (taken from [29]). At zero temperature and $\alpha < 1$, $m = 1$ is a stable attractor. For small m, (4.47) gives an unstable fixed point at

$$\alpha_u = 2/(2+\pi) \simeq 0.39. \tag{4.49}$$

Hence for $\alpha < \alpha_u$ the network has a maximal possible basin of attraction; an initial state with any $m^0 = O(1) > 0$ will flow to $m^* = 1$.

5. Basins of attraction of the various models

In this section we display graphically the basins of attraction, as obtained from the various analytically obtained recursions. We display only results of the deterministic ($T = 0$) limit of the recursion relations.

Consider first the results for the basic model with Hebbian couplings, (2.2), to be referred to as outer product or OP. The problem we solved here concerns the case of an initial state with sizeable overlap, m^0, with a *single* key pattern on the first layer; overlaps with all other ($\nu > 1$) patterns are of order $1/\sqrt{N}$. In this case we had the following layer-to-layer recursions for the overlap and the width of the embedding-field distribution:

$$m^{l+1} = \mathrm{erf}(m^l/\sqrt{2}\Delta^l) \tag{5.1a}$$

$$(\Delta^{l+1})^2 = \alpha + (2/\pi)\exp[-(m^l/\Delta^l)^2]. \tag{5.1b}$$

In order to find the overlap on layer $l+1$ we have to iterate these recursions. The initial state determines m^0, and the overlaps with patterns $\nu > 1$ determine Δ^0 via

$$(\Delta^0)^2 = \sum_{\nu>1}^{\alpha N} (M_\nu^0)^2 = \alpha.$$

These recursions were analysed by first locating their fixed points. For $\alpha < \alpha_c \simeq 0.27$ a stable branch, with $m^* \simeq 1$ coexists with an unstable branch; the two merge at α_c. In addition to these two branches one has a stable fixed point at $m^* = 0$, for all values of α. For $\alpha > \alpha_c$ all initial states flow to the $m^* = 0$ fixed point; however, for $\alpha < \alpha_c$ the overlap develops in a manner that depends on its initial value. If the initial overlap is large enough, i.e.

$$m^0 > m_c^0(\alpha)$$

we obtain $m^l \to m^* \simeq 1$ for large l, whereas $m_1^l \to 0$ for $m_1^0 < m_c^0(\alpha)$. A convenient measure for the size of the domain of attraction of the memory state is its radius R, defined by

$$R = 1 - m_c^0. \tag{5.2}$$

R measures how close the initial configuration must be to one of the key patterns in order to guarantee convergence to that key pattern on subsequent layers.

We present below results for various choices of the couplings. For all cases (with the exception of the pseudoinverse) we have demonstrated that the recursions are the same as those derived for OP with static synaptic noise with width Δ_0:

$$\begin{aligned} m_1^{l+1} &= \operatorname{erf}(m^l/\sqrt{2}\delta^l) \\ (\Delta^{l+1})^2 &= \alpha + \frac{2}{\pi}\left(\frac{\Delta^l}{\delta^l}\right)^2 \exp[-(m^l/\delta^l)^2] \qquad (5.3) \\ (\delta^l)^2 &= (\Delta^l)^2 + \alpha\Delta_0^2. \end{aligned}$$

By iterating these recursions for different Δ_0 we obtain the critical values $\alpha_c(\Delta_0)$, presented in figure 2(*a*). As expected, the storage capacity decreases with increasing static synaptic noise. An additional degrading effect of the noise is shown in figure 2(*b*): the limiting overlap with the recalled pattern, m^*, also decreases. We present in the figure only the critical overlap, i.e. $m^*(\alpha_c)$ as a function of Δ_0.

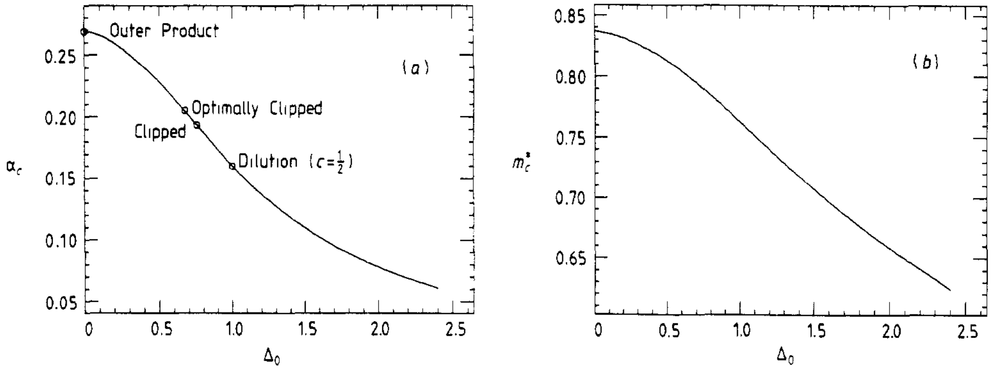

Figure 2. (*a*) Critical storage capacity α_c as a function of the width of the static synaptic noise distribution Δ_0. Various models are mapped onto this problem, with different values of Δ_0, as indicated. (*b*) The limiting overlap with the recalled pattern, $m^*(\alpha_c)$ as a function of Δ_0.

Next we demonstrate the effect of the static noise Δ_0 on the basin size. In figure 3 we plot R as a function of Δ_0 for various values of α. As can be expected, the basin size decreases with increasing α and static noise Δ_0.

The different models studied above (see § 4) map onto different initial values for the width of the static noise. We had, for dilution,

$$\Delta_0^2 = (1-c)/c \qquad (5.4)$$

where c is the concentration of bonds present, whereas for non-linear synapses

$$\Delta_0^2 = \overline{F^2}/(\overline{F'})^2 - 1 \qquad (5.5)$$

where $F(x)$ is the non-linear function that gives the coupling J_{ij}^l in terms of the outer product $x = (1/\sqrt{\alpha N})\sum \xi_{i\nu}^{l+1}\xi_{j\nu}^l$ (see (4.27)). We give below results for three cases; clipped (C), optimally clipped (OC) and linear diluted. The function $F(x)$ takes for the three cases the following forms. For the clipped case

$$F(x) = \operatorname{sgn}(x).$$

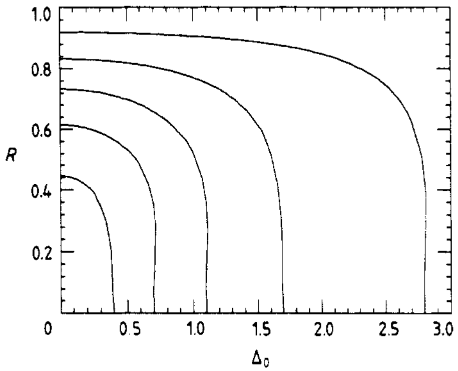

Figure 3. Size of basin of attraction R plotted against static noise width Δ_0 for various values of the storage capacity α (from left to right: $\alpha=0.25$, 0.20, 0.15, 0.10, 0.05).

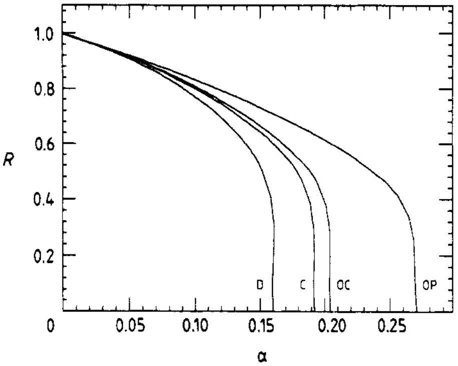

Figure 4. Size of basin of attraction R plotted against storage capacity α, evaluated at values of Δ_0 (the effective static noise width) that correspond to models discussed; fully connected outer product (OP); randomly diluted outer product with $c=0.5$ bond concentration (D); clipped (C) and optimally clipped (OC)

Optimal clipping is obtained by using

$$F(x)=\begin{cases}\operatorname{sgn}(x) & \text{if } |x|>x_0\\ 0 & \text{otherwise}\end{cases}$$

and choosing x_0 so that the resulting Δ_0 is minimal. The last case, of optimal dilution, is given by

$$F(x)=\begin{cases}x & \text{if } |x|>x_0\\ 0 & \text{otherwise.}\end{cases}$$

This choice corresponds to deleting bonds with small value of the coupling, and keeping the outer product for strong bonds:

$$J_{ij}=\begin{cases}J_{ij}^{\mathrm{OP}} & \text{if } |J_{ij}^{\mathrm{OP}}|>J_0\\ 0 & \text{otherwise}\end{cases}$$

where $J_0=\sqrt{\alpha/N}\,x_0$, and J^{OP} is given by (2.2). Incidentally, for the last two cases the fraction of existing bonds is given by

$$c_0=1-\operatorname{erf}(x_0/\sqrt{2}). \qquad (5.6)$$

In figure 2(a) we indicated also the Δ_0, α_c values that correspond to clipped, diluted (with concentration $c=\frac{1}{2}$) and optimally clipped bonds. For each of these, as well as for the model with outer product couplings, we calculated the basin size R, plotted against α in figure 4. As can be expected, the best performance (largest basin) is achieved for the original outer product couplings studied previously.

Turning to the case of optimal dilution, we plot in figure 5(a) α_c, the critical α, as a function of c_0, the fraction of existing bonds, while in figure 5(b) we plot α_c/c_0, i.e. the storage capacity per bond, against c_0 (full curve). Whereas the storage capacity decreases with the dilution, we find that the capacity per bond *increases* with the dilution. A similar effect has been observed by Sompolinsky [6] in the context of the Hopfield model. Finally, to compare the effect of random dilution with that of the

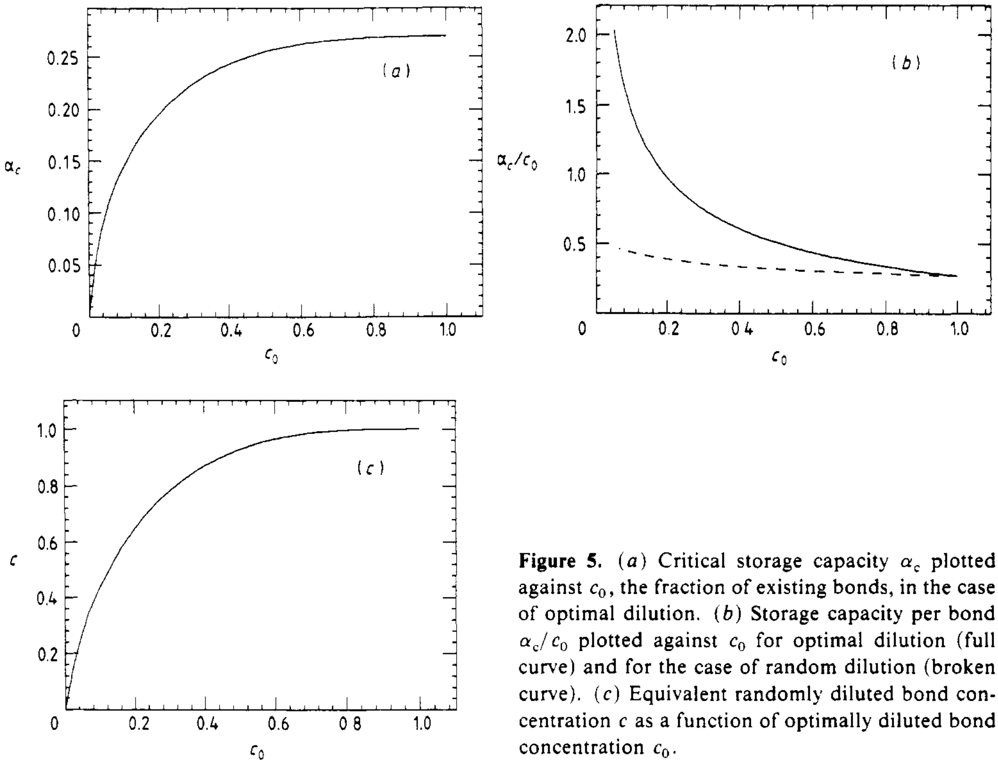

Figure 5. (*a*) Critical storage capacity α_c plotted against c_0, the fraction of existing bonds, in the case of optimal dilution. (*b*) Storage capacity per bond α_c/c_0 plotted against c_0 for optimal dilution (full curve) and for the case of random dilution (broken curve). (*c*) Equivalent randomly diluted bond concentration c as a function of optimally diluted bond concentration c_0.

optimal dilution, we plot in figure 5(*b*) α_c/c_0 for random dilution as well (broken curve), and present, in figure 5(*c*), the equivalent *random* concentration c against the actual *optimally diluted* concentration c_0. By equivalent random concentration we mean that the randomly diluted model with N_c bonds has the same storage capacity as the optimally diluted model with N_{c_0} bonds. As can be seen from the figure, the system functions much better in the case of the non-random dilution.

Finally we mention the last model considered in § 4, with pseudoinverse couplings. For this case figure 1 presents the limiting overlap $m_c(\alpha)$ for a range of temperatures; the radius R of the domain of attraction is given by $R = 1 - m_c$. For completeness'

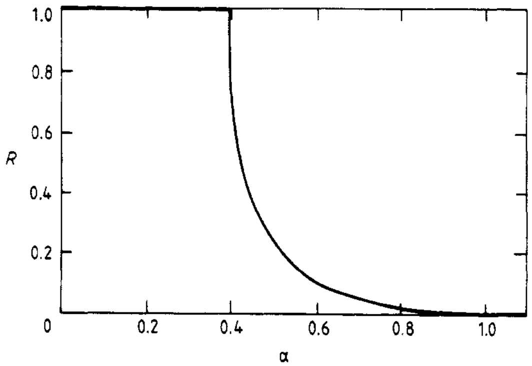

Figure 6. Size of basin of attraction R plotted against the storage capacity α for pseudo-inverse interlayer couplings.

sake we present R as a function of α in figure 6 for the case of $T = 0$ and pseudoinverse couplings between the layers. This situation is intermediate between layered Hebbian and strongly diluted [12] networks. For $\alpha < \alpha_u \simeq 0.39$ all initial states with non-vanishing overlap flow to the correct pattern (as is the case for the diluted model), whereas for $\alpha_u < \alpha < 1$ the memory state has a non-trivial domain of attraction (as in the layered network with Hebbian couplings).

6. Summary and discussion

We have presented exact results for layered feedforward neural networks with a variety of interlayer couplings. In addition to linear Hebbian (outer product) connections, we studied the effect of adding static noise, of diluting the bonds in various ways, and of introducing couplings whose strength is a non-linear function of the outer product. In addition we also solved the case of pseudoinverse couplings between neighbouring layers, that take the network with no error through a sequence of random patterns assigned to the sequence of layers.

This family of neural networks is unique in that its dynamics is exactly soluble and is controlled by attractors with non-trivial domains of attraction. The only other model with exactly soluble dynamics is the extremely diluted asymmetric model of [12]. We found that many of our results exhibit qualitative agreement with those obtained for the dilute model and also with the fully connected Hopfield model.

We analysed these three models on similar footing. For all three one finds that the stationary states (and their layered analogues) satisfy similar equations for the overlap with the recalled key pattern (3.4). This equation is easily interpreted as the Gaussian average of the embedding field. The only difference between the three models is in the expressions for the width of this distribution. One should keep in mind that the embedding field is Gaussian distributed for the layered and dilute networks, whereas in the fully connected case this result is incorrect, and is based on approximations (such as using a replica-symmetric solution).

The class of models studied here bridges a gap between physicists' and non-physicists' models, in that it has input and output units, does not utilise stable states of the dynamics but does use random key patterns, over which averages are calculated in the thermodynamic limit. More work in the 'no man's land' between physics and computer science will, hopefully, advance diffusion of ideas, problems and new results between these disciplines.

Acknowledgments

WK thanks Manfred Opper for many very useful discussions and the Einstein Center and the Deutsche Forschungsgemeinschaft for support. The research of ED and RM was partially supported by Minerva and the US-Israel Binational Science Foundation.

References

[1] Domany E, Meir R and Kinzel W 1986 *Europhys. Lett.* **2** 175

[2] Meir R and Domany E 1987 *Phys. Rev. Lett.* **59** 359; 1988 *Europhys. Lett.* **4** 645; 1988 *Phys. Rev.* A **37** 608

[3] Meir R 1988 *J. Physique* **49** 201

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

[5] Sompolinsky H 1986 *Phys. Rev.* A **34** 2571
van Hemmen J L and Kuhn R 1986 *Phys. Rev. Lett.* **57** 913
van Hemmen J L 1987 *Phys. Rev.* A **36** 1959

[6] Sompolinsky H 1987 *Heidelberg Colloquium on Glassy Dynamics* (*Lecture Notes in Physics* **275**) ed J L van Hemmen and I Morgenstern (Berlin: Springer)

[7] Kanter I and Sompolinsky H 1987 *Phys. Rev.* A **35** 380
Personnaz L, Guyon I and Dreyfus G 1986 *Phys. Rev.* A **34** 4217

[8] Kohonen T 1984 *Self Organization and Associative Memory* (Berlin: Springer)

[9] Kinzel W 1988 Statistical Mechanics of Neural Networks *Preprint* Justus-Liebig-Universitat Giessen
Hertz J A 1988 Statistical Mechanics of Neural Computation *Preprint* NORDITA 88/25 S

[10] Hopfield J J 1982 *Proc. Natl. Acad. USA* **79** 2554

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

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

[13] Domany E 1988 *J. Stat. Phys.* **51** 743

[14] Cowan J D and Sharp D H 1988 *Proc. Am. Acad. Arts Sci.* **117** 85
Lippmann R P 1987 *IEEE ASSP Mag.* **4** 4

[15] Rumelhart D E and mcClelland J L 1986 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* vols 1 and 2 (Cambridge, MA: MIT Press)

[16] Minsky M and Papert S 1988 *Perceptrons* (Cambridge, MA: MIT Press) (Expanded edition)

[17] Abeles M 1982 *Local Cortical Circuits* (Berlin: Springer)

[18] Rumelhart D E, Hinton G E and Williams R J 1986 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* vol 1, ed D E Rumelhart and J L McClelland (Cambridge, MA: MIT Press) p 318

[19] Le Cun Y 1985 *Proc. Cognitiva* **85** 593
Werbos P J 1974 *PhD thesis* Harvard

[20] Grossman T, Meir R and Domany E 1988 *Complex Systems* **2** 555

[21] Krauth W, Mezard M and Nadal J 1988 *Complex Systems* **2** 387

[22] Little W A 1975 *Math. Biosci.* **19** 101

[23] Amari S and Maginu K 1988 *Neural Networks* **1** 63

[24] Feller W 1966 *An Introduction to Probability Theory and its Applications* vol 2 (New York: Wiley) p 256

[25] Kinzel W 1985 *Z. Phys.* B **60** 205

[26] Mezard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore: World Sceintific) chap 5

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

[28] Opper M, Kleinz J, Kohler H and Kinzel W 1989 *J. Phys. A: Math. Gen.* **22** L407

[29] Kleinz J 1988 *Diplomthesis* Justus-Liebig-Universitat Giessen