Storage of hierarchically correlated patterns

1990 J. Phys. A: Math. Gen. 23 2587

(http://iopscience.iop.org/0305-4470/23/12/034)

J. Phys. A: Math. Gen. 23 (1990) 2587-2603. Printed in the UK

Storage of hierarchically correlated patterns

A Engel

Sektion Physik der Humboldt-Universität, Bereich 04, Invalidenstrasse 42, Berlin 1040, German Democratic Republic

Received 21 November 1989

Abstract. The maximal storage capacity for hierarchically correlated patterns is calculated without using an explicit learning rule. If the lowest-level branching ratio tends to infinity strong correlations increase the storage capacity, otherwise they decrease it significantly. There is probably no enlargement of the typical basins of attraction due to the hierarchical organisation. We consider both two-level hierarchies and those with infinitely many levels.

1. Introduction

One of the most interesting questions in the rapidly developing field of statistical mechanics of attractor neural networks concerns the storage of hierarchically correlated patterns. There are at least two reasons why hierarchies of patterns merit special consideration. Firstly it is known that the human brain prefers to store hierarchically ordered information; we even subconsciously modify inputs to make them fit into an already existing hierarchy (Parga and Virasoro 1986, Toulouse *et al* 1986). Secondly in neural networks of N formal neurons with *random* synaptic couplings there are $O(\exp(aN))$ attractors in phase space which are hierarchically ordered with respect to their mutual overlap (Mézard *et al* 1984, 1987). It is tempting to try to meet this spontaneously arising structure of low-energy states by a hierarchical organisation of the patterns in order to improve the so far rather modest storage capacity of maximally $p = \alpha_c N$ patterns, where $\alpha_c = O(1)$.

Several approaches to this problem have been discussed. One possibility is to use *hierarchically structured networks*, where different levels of the network are used to store the ancestors of the patterns (Dotsenko 1985, 1986, Gutfreund 1988, Sourlas 1988, Sutton *et al* 1988). Here we will be concerned with *homogeneous* networks that store hierarchically correlated patterns, a situation more similar to the spin-glass problem where the above-mentioned hierarchical organisation of attractors occurs. Moreover, in this case the ancestor patterns, which are only instrumental in defining the statistics of the hierarchy, do not have to be stored. Hitherto the storage abilities of such networks have only been studied by using a special learning rule which allows us to determine the synaptic couplings appropriate for a given set of patterns (Parga and Virasoro 1986, Feigelman and Ioffe 1987, Cortes *et al* 1987, Bös *et al* 1988, Gutfreund 1988, Ioffe *et al* 1989). It is therefore not clear whether the results produced are characteristic for the statistics of the patterns or for the learning rule implemented. Cortes *et al*, for example, use the pseudoinverse rule and find, after neglecting terms of order $N^{-1/2}$ in the overlap matrix, the same storage capacity as for the Hopfield model. But the Hebb rule used in the Hopfield model is also equivalent to the

0305-4470/90/122587+17$03.50

pseudoinverse rule for uncorrelated patterns if one neglects $O(N^{-1/2})$ terms in the overlap matrix. It seems likely, therefore, that by retaining these terms the learning rule of Cortes *et al* would yield $\alpha_c = 1$ as in the case of uncorrelated patterns (Kanter and Sompolinsky 1987). The learning rules of Parga and Virasoro, Feigelman and Ioffe, and Bös *et al* differ only slightly from the rule of Cortes *et al.* Hence it remains unclear whether there are more appropriate learning rules which allow for a more effective use of the correlations present in a pattern hierarchy. Actually very recently Ioffe *et al* proposed a new learning rule with much higher values of α_c than in the Hopfield model (Ioffe *et al* 1988).

The most transparent way to study the optimal storage abilities for hierarchically correlated patterns consists in applying the methods developed by Elizabeth Gardner to analyse the phase space of interactions (Gardner 1988). Using these methods it is not necessary to implement a learning rule at all so that the results for the storage capacity α_c are solely determined by the statistical properties of the pattern hierarchy. The actual value of the synaptic couplings that produce a given capacity can then be determined by iterative procedures (Gardner 1988). In the present paper we calculate in this way the maximal storage capacity α_c for a regular two-level hierarchy. The results should be typical also for hierarchies with more than two but finitely many levels. In addition we investigate for the first time pattern hierarchies with a number of levels diverging with $N \to \infty$. This allows us to store infinitely many patterns although the branching ratio at all levels remains finite.

The paper is organised as follows. Section 2 defines the pattern hierarchy and fixes the notation. Section 3 deals with the two-level hierarchy and introduces a renormalisation group technique for the determination of α_c. These methods are used in section 4 for the analysis of an infinite hierarchy. Section 5 is devoted to a study of the information content of a pattern hierarchy and finally section 6 contains a summary and discussion of the results.

2. Pattern hierarchies

We consider a network N of formal neurons $S_i = \pm 1$, $i = 1, \ldots, N$, which are connected by synaptic couplings J_{ij}. The dynamics is given by $S_i(t+1) = \operatorname{sgn}(\Sigma_{j \neq i} J_{ij} S_j(t))$. A special configuration $\{\xi_i\}$ is called a fixpoint with stability κ of the dynamics if for all $i = 1, \ldots, N$

$$\xi_i \frac{1}{N^{1/2}} \sum_{j \neq i} J_{ij} \xi_j \geqslant \kappa. \tag{2.1}$$

The normalisation

$$\sum_{j \neq i} J_{ij}^2 = N \tag{2.2}$$

for all i makes $\kappa = O(1)$ for $N \to \infty$. We are interested in the maximal number p of hierarchically correlated patterns $\{\xi_i^\mu\}$, $\mu = 1, \ldots, p$, which can be stored with stability κ in the network if $N \to \infty$.

We first define the statistics of a two-level pattern hierarchy, the generalisation to more levels being straightforward. To this end we choose the branching ratios z_0 and z_1 for the zeroth and first level respectively and set

$$\begin{aligned} &\xi_i^\mu = \xi_i^{(1),\mu'} \xi_i^{(0),\mu} \\ &i = 1, \ldots, N \qquad \mu = 1, \ldots, p \qquad \mu' = [\mu / z_0] = 1, \ldots, z_1 \end{aligned} \tag{2.3}$$

where $[x]$ denotes the largest integer smaller than x and the $\xi_i^{(k),\mu}$ are independent, identically distributed variables with distribution

$$P(\xi_i^{(k),\mu}) = \frac{1+m_k}{2}\,\delta(\xi_i^{(k),\mu}-1) + \frac{1-m_k}{2}\,\delta(\xi_i^{(k),\mu}+1). \tag{2.4}$$

Equations (2.3) and (2.4) just mean that one first generates z_1 ancestors with mutual overlap m_1^2. Then every ancestor pattern splits into z_0 descendents with mutual overlap m_0^2 (figure 1), hence there are $p = z_0 z_1$ patterns $\{\xi_i^\mu\}$.

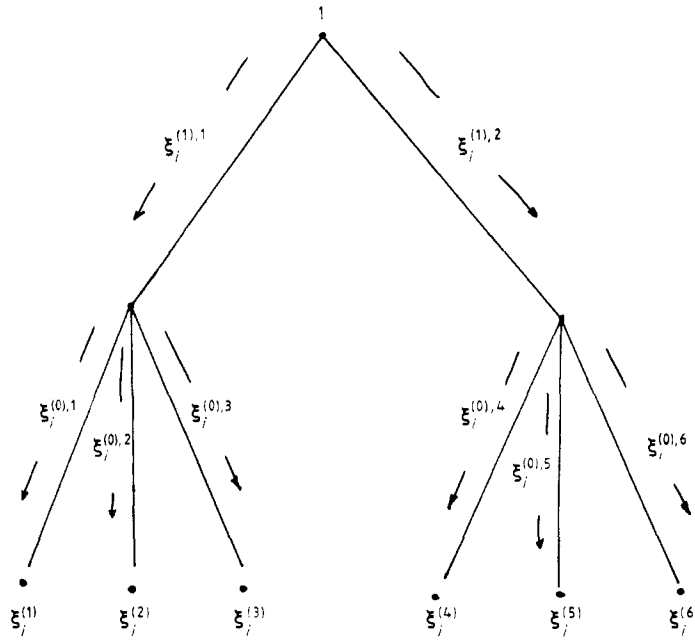

Figure 1. Two-level hierarchy with $z_1 = 2$ and $z_0 = 3$. The $\xi_i^{(k),\mu}$ are only auxiliary variables to define the correlations between the patterns $\{\xi_i^\mu\}$.

Note that only these patterns $\{\xi_i^\mu\}$ are to be stored, the $\{\xi_i^{(k),\mu}\}$ are just the auxiliary variables to fix the statistics of the $\{\xi_i^\mu\}$. The overlap matrix

$$C_{\mu\omega} = \frac{1}{N}\sum_i \xi_i^\mu \xi_i^\nu \tag{2.5}$$

is of Parisi type

$$C_{\mu\nu} = \begin{cases} 1 & \text{if } \mu = \nu \\ m_0^2 & \text{if } \mu \neq \nu,\ [\mu/z_0] = [\nu/z_0] \\ m_0^2 m_1^2 & \text{if } [\mu/z_0] \neq [\nu/z_0] \end{cases} \tag{2.6}$$

as follows from (2.3) and (2.4). The strongest correlations are those within a class on the lowest level and described by the parameter m_0. Similar definitions of pattern hierarchies have been used by other authors (Parga and Virasoro 1986, Rammal *et al* 1986, Feigelman and Ioffe 1987, Cortes *et al* 1987, Gutfreund 1988). For our purposes

it is important that $\xi_i^{(k),\mu} = \pm 1$ for all k. Finally we mention that patterns with low level of activity or 'magnetisation' (Amit *et al* 1987a, Gardner 1988) defined by

$$P(\xi_i^\mu) = \frac{1+m}{2}\delta(\xi_i^\mu - 1) + \frac{1-m}{2}\delta(\xi_i^\mu + 1) \tag{2.7}$$

can be interpreted as a 'one-level hierarchy' according to (2.3), (2.4).

3. Storage capacity for a two-level hierarchy

The storage capacity can be determined with the help of a projection operator $\chi(\{J_{ij}\})$ in the phase space of interactions J_{ij} of the form (Gardner 1988)

$$\chi(\{J_{ij}\}) = \prod_{\mu=1}^{p} \theta\left(\xi_i^\mu \frac{1}{N^{1/2}} \sum_j J_{ij}\xi_j^\mu - \kappa\right). \tag{3.1}$$

χ is equal to one for all points $\{J_{ij}\}$ which fulfil (2.1) for all patterns at one given neuron i and is zero otherwise. Since all the J_{ij} are independent it is sufficient to consider only one neuron.

The main idea of Gardner's approach consists in calculating the quantity

$$q = \left\langle\!\!\left\langle \frac{\int_s \prod_j \mathrm{d}J_{ij}^{(1)} \int_s \prod_j \mathrm{d}J_{ij}^{(2)} (1/N) \sum_j J_{ij}^{(1)} J_{ij}^{(2)} \chi(\{J_{ij}^{(1)}\})\chi(\{J_{ij}^{(2)}\})}{\int_s \prod_j \mathrm{d}J_{ij}^{(1)} \int_s \prod_j \mathrm{d}J_{ij}^{(2)} \chi(\{J_{ij}^{(1)}\})\chi(\{J_{ij}^{(2)}\})} \right\rangle\!\!\right\rangle \tag{3.2}$$

which gives the typical overlap between two different solutions $J_{ij}^{(1)}$ and $J_{ij}^{(2)}$ of the stability problem (2.1) and characterises the similarity between $J_{ij}^{(1)}$ and $J_{ij}^{(2)}$. For small values of α we will find $q \simeq 0$ since very different solutions of $\{J_{ij}\}$ are possible. With increasing α the solutions become more and more correlated and for $\alpha \to \alpha_c$ we find $q \to 1$ signalling the uniqueness of the solution of (2.1). The integrals in (3.2) are restricted to the sphere with radius $N^{1/2}$ in order to meet the constraint (2.2). The average over the patterns in (3.2) must not be done for the numerator and denominator separately and can be performed using the replica trick.

The quantity q arises in a natural way as a saddle-point variable in the calculation of the nth power of the partial volume V of points for which $\chi = 1$ averaged over the statistics of the patterns in the limits $N \to \infty$, $n \to 0$, which is therefore the central quantity of interest. Using appropriate integral representations of the θ-functions in (3.1) we have the starting expression as Gardner (1988):

$$\begin{aligned}\langle\!\langle V^n \rangle\!\rangle = \int \prod_\alpha \prod_j \mathrm{d}J_{ij}^\alpha \prod_\alpha \delta\left(\textstyle\sum_j (J_{ij}^\alpha) - N\right) \\ \times \left\langle\!\!\left\langle \int_\kappa^\infty \prod_{\mu,\alpha} \frac{\mathrm{d}\lambda_\mu^\alpha}{2\pi} \int \prod_{\mu,\alpha} \mathrm{d}x_\mu^\alpha \exp\left\{ \mathrm{i} \sum_{\mu,\alpha} x_\mu^\alpha \lambda_\mu^\alpha - \frac{\mathrm{i}}{N^{1/2}} \sum_{\mu,\alpha} x_\mu^\alpha \xi_i^\mu \sum_j J_{ij}^\alpha \xi_j^\mu \right\} \right\rangle\!\!\right\rangle \\ \times \left[\int \prod_\alpha \prod_j \mathrm{d}J_{ij}^\alpha \prod_\alpha \delta\left(\sum_j (J_{ij}^\alpha)^2 - N \right) \right]^{-1}. \end{aligned} \tag{3.3}$$

Here α is a replica index and runs from 1 to n and the δ-functions ensure the normalisation (2.2). The disorder-independent part in (3.3) can be handled exactly as in the case of uncorrelated patterns (Gardner 1988). In order to calculate the average

over the patterns we use (2.3) and first average over $\xi_j^{(0),\mu}$ shifting the average over $\xi_i^{(0),\mu}$ to the end of the calculation. To leading order in N we find using (2.4)

$$\left\langle\!\!\left\langle \exp\left\{-\frac{\mathrm{i}}{N^{1/2}}\sum_{\mu,\alpha} x_\mu^\alpha \xi_i^{(1),\mu'}\xi_i^{(0),\mu}\sum_j J_{ij}^\alpha \xi_j^{(1),\mu'}\xi_j^{(0),\mu}\right\}\right\rangle\!\!\right\rangle_{\xi_j^{(0)\mu}}$$

$$= \exp\left\{-\mathrm{i}m_0\sum_{\mu,\alpha}\xi_i^{(0),\mu}x_\mu^\alpha\frac{1}{N^{1/2}}\xi_i^{(1),\mu'}\sum_j J_{ij}^\alpha\xi_j^{(1),\mu'}\right.$$

$$\left.-\frac{1-m_0^2}{2}\sum_{\mu,\alpha,\beta}x_\mu^\alpha x_\mu^\beta\frac{1}{N}\sum_j J_{ij}^\alpha J_{ij}^\beta\right\}. \tag{3.4}$$

Introducing the abbreviations

$$\Lambda_{\mu'}^\alpha = \frac{1}{N^{1/2}}\xi_i^{(1),\mu'}\sum_j J_{ij}^\alpha\xi_j^{(1),\mu'} \tag{3.5}$$

and

$$q^{\alpha\beta} = \frac{1}{N}\sum_j J_{ij}^\alpha J_{ij}^\beta \qquad \alpha\neq\beta \tag{3.6}$$

we get for the dominant terms in the numerator of (3.3)

$$\int\prod_{\alpha,j}\mathrm{d}J_{ij}^\alpha\int\prod_{\mu',\alpha}\mathrm{d}\Lambda_{\mu'}^\alpha\int\prod_{\alpha<\beta}\mathrm{d}q^{\alpha\beta}\prod_\alpha\delta\left(\prod_j(J_{ij}^\alpha)^2-N\right)$$

$$\times\left\langle\!\!\left\langle\prod_{\mu',\alpha}\delta\left(\Lambda_{\mu'}^\alpha-\frac{1}{N^{1/2}}\xi_i^{(1),\mu'}\sum_j J_{ij}^\alpha\xi_j^{(1),\mu'}\right)\right\rangle\!\!\right\rangle_{\xi_i^{(1),\mu'},\xi_j^{(1),\mu'}}$$

$$\times\prod_{\alpha<\beta}\delta\left(q^{\alpha\beta}-\frac{1}{N}\sum_j J_{ij}^\alpha J_{ij}^\beta\right)$$

$$\times\left\langle\!\!\left\langle\int_\kappa^\infty\prod_{\mu,\alpha}\frac{\mathrm{d}\lambda_\mu^\alpha}{2\pi}\prod_{\mu,\alpha}\mathrm{d}x_\mu^\alpha\exp\left\{\mathrm{i}\sum_{\mu,\alpha}x_\mu^\alpha\lambda_\mu^\alpha-\mathrm{i}m_0\xi_i^{(0),\mu}\sum_{\mu,\alpha}x_\mu^\alpha\Lambda_{\mu'}^\alpha\right.\right.\right.$$

$$\left.\left.\left.-\frac{1-m_0^2}{2}\sum_{\mu,\alpha}(x_\mu^\alpha)^2-\frac{1-m_0^2}{2}\sum_\mu\sum_{(\alpha,\beta)}x_\mu^\alpha x_\mu^\beta q^{\alpha\beta}\right\}\right\rangle\!\!\right\rangle_{\xi_i^{(0),\mu}} \tag{3.7}$$

The λ_μ^α, x_μ^α integrals factorise in μ and introducing an integral representation of the second δ-function in (3.7) we get

$$\int\prod_{\alpha,j}\mathrm{d}J_{ij}^\alpha\int\prod_{\alpha<\beta}\mathrm{d}q^{\alpha\beta}\prod_\alpha\delta\left(\sum_j(J_{ij}^\alpha)^2-N\right)\prod_{\alpha<\beta}\delta\left(q^{\alpha\beta}-\frac{1}{N}\sum_j J_{ij}^\alpha J_{ij}^\beta\right)$$

$$\times\left\langle\!\!\left\langle\int_{-\infty}^\infty\prod_{\mu',\alpha}\frac{\mathrm{d}\Lambda_{\mu'}^\alpha}{2\pi}\int\prod_{\mu',\alpha}\mathrm{d}X_{\mu'}^\alpha\exp\left\{\mathrm{i}\sum_{\mu',\alpha}X_{\mu'}^\alpha\Lambda_{\mu'}^\alpha\right.\right.\right.$$

$$\left.\left.\left.-\frac{1}{N^{1/2}}\sum_{\mu',\alpha}\xi_i^{(1),\mu'}X_{\mu'}^\alpha\sum_j J_{ij}^\alpha\xi_j^{(1),\mu'}+z_0\sum_{\mu'}G^{(0)}(\bar\Lambda_{\mu'},\bar q)\right\}\right\rangle\!\!\right\rangle_{\xi_i^{(1),\mu'},\xi_j^{(1),\mu'}} \tag{3.8}$$

$$G^{(0)}(\bar\Lambda_\mu,\bar q) = \ln\left\langle\!\!\left\langle\int_\kappa^\infty\prod_\alpha\frac{\mathrm{d}\lambda^\alpha}{2\pi}\int\prod_\alpha\mathrm{d}x^\alpha\exp\left\{\mathrm{i}\sum_\alpha x^\alpha\lambda^\alpha-\mathrm{i}m_0\xi_i^{(0)}\sum_\alpha x^\alpha\Lambda_{\mu'}^\alpha\right.\right.\right.$$

$$\left.\left.\left.-\frac{1-m_0^2}{2}\sum_\alpha(x^\alpha)^2-\frac{1-m_0^2}{2}\sum_{(\alpha,\beta)}x^\alpha x^\beta q^{\alpha\beta}\right\}\right\rangle\!\!\right\rangle_{\xi_i^{(0)}}. \tag{3.9}$$

Comparing (3.8) with the numerator of (3.3) one realises that the expressions are similar in their mathematical structure. Averaging over the lowest-level auxiliary variables $\xi_j^{(0),\mu}$ leaves us with an expression for the partial volume $\langle\langle V'^n \rangle\rangle$ of a hierarchy with only one level and z_1 patterns $\{\xi_i^{(1),\mu'}\}$. Hence we expect that the Λ_μ^α dependence of the last term in (3.8) produces a renormalised lower cut-off κ' for the Λ_μ^α integral. These properties, which arise from the self-similar structure of the pattern hierarchy, will be useful in particular for studying infinite hierarchies. For the present case of a two-level hierarchy they just mean that the average over the $\xi_j^{(1),\mu'}$ is to be taken analogously to that over the $\xi_j^{(0),\mu}$.

In this way we get from (3.8):

$$\int \prod_{\alpha,j} \mathrm{d}J_{ij}^\alpha \int \prod_{\alpha<\beta} \mathrm{d}q^{\alpha\beta} \prod_\alpha \delta\Big(\sum_j (J_{ij}^\alpha)^2 - N\Big) \prod_{\alpha<\beta} \delta\Big(q^{\alpha\beta} - \frac{1}{N}\sum_j J_{ij}^\alpha J_{ij}^\beta\Big)$$
$$\times \Big\langle\!\Big\langle \int \prod_{\mu',\alpha} \frac{\mathrm{d}\Lambda_{\mu'}^\alpha}{2\pi} \prod_{\mu',\alpha} \mathrm{d}X_{\mu'}^\alpha \exp\Big\{ \mathrm{i} \sum_{\mu',\alpha} X_{\mu'}^\alpha \Lambda_{\mu'}^\alpha + z_0 G^{(0)}(\bar{\Lambda}_{\mu'}, \bar{\bar{q}})$$
$$- \mathrm{i}m_1 \sum_{\mu',\alpha} \xi_i^{(1),\mu'} X_{\mu'}^\alpha \frac{1}{N^{1/2}} \sum_j J_{ij}^\alpha - \frac{1-m_1^2}{2} \sum_{\mu',\alpha} (X_{\mu'}^\alpha)^2$$
$$- \frac{1-m_1^2}{2} \sum_{\mu',(\alpha,\beta)} X_{\mu'}^\alpha X_{\mu'}^\beta q^{\alpha\beta} \Big\} \Big\rangle\!\Big\rangle_{\xi_i^{(1),\mu}} \tag{3.10}$$

By analogy with (3.5) we introduce

$$M^\alpha = \frac{1}{N^{1/2}} \sum_j J_{ij}^\alpha \tag{3.11}$$

(since a two-level hierarchy is characterised by $\xi_i^{(2),\mu''} = 1$). Moreover, we define similarly to (3.9) a function

$$G^{(1)}(\bar{M}, \bar{\bar{q}}) = \ln \Big\langle\!\Big\langle \int \prod_\alpha \frac{\mathrm{d}\Lambda^\alpha}{2\pi} \int \prod_\alpha \mathrm{d}X^\alpha \exp\Big\{ \mathrm{i} \sum_\alpha X^\alpha \Lambda^\alpha - \mathrm{i}m_1 \xi_i^{(1)} \sum_\alpha X^\alpha M^\alpha$$
$$- \frac{1-m_1^2}{2} \sum_\alpha (X^\alpha)^2 - \frac{1-m_1^2}{2} \sum_{(\alpha,\beta)} X^\alpha X^\beta q^{\alpha\beta} + z_0 G^{(0)}(\bar{\Lambda}, \bar{\bar{q}}) \Big\} \Big\rangle\!\Big\rangle_{\xi_i^{(1)}}. \tag{3.12}$$

Including the disorder-independent terms in the same way as Gardner we get finally for the numerator of (3.3):

$$\int \prod_{\alpha<\beta} \mathrm{d}q^{\alpha\beta} \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \prod_{\alpha<\beta} \frac{\mathrm{d}F^{\alpha\beta}}{2\pi/N} \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \prod_\alpha \frac{\mathrm{d}E^\alpha}{4\pi} \int \prod_\alpha \mathrm{d}M^\alpha \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \prod_\alpha \frac{\mathrm{d}K^\alpha}{2\pi/N^{1/2}}$$
$$\times \exp\Big\{ N \Big[- \sum_{\alpha<\beta} q^{\alpha\beta} F^{\alpha\beta} + \tfrac{1}{2} \sum_\alpha E^\alpha + \frac{1}{N^{1/2}} \sum_\alpha M^\alpha K^\alpha$$
$$+ g(\bar{\bar{F}}, \bar{E}, \bar{K}) + \frac{z_1}{N} G^{(1)}(\bar{M}, \bar{\bar{q}}) \Big] \Big\} \tag{3.13}$$

where

$$g(\bar{\bar{F}}, \bar{E}, \bar{K}) = \ln \int \prod_\alpha \mathrm{d}J^\alpha \exp\Big\{ \sum_{\alpha<\beta} F^{\alpha\beta} J^\alpha J^\beta - \tfrac{1}{2} \sum_\alpha E^\alpha (J^\alpha)^2 - \sum_\alpha K^\alpha J^\alpha \Big\}. \tag{3.14}$$

The integrals in (3.13) are performed for $N \to \infty$ by the saddle-point method which requires that also the last term in the bracket is O(1). We investigate two different possibilities to ensure this. The first is defined by $z_0 = O(1)$ and $z_1 = O(N)$ and will be referred to as the *universalist* case, since the system stores some items in each of very many classes. The opposite case of a *specialist* is given by $z_0 = O(N)$ and $z_1 = O(1)$ where few classes containing very many patterns are stored. In this case we find from (3.12) $G^{(1)}(\bar{M}, \bar{q}) = O(N)$ and the last term in (3.13) is again O(1). Interestingly the maximal storage capacity is markedly different in these two cases.

The saddle point is assumed to be replica symmetric which simplifies the expressions significantly. After standard manipulations (Gardner 1988) we get from (3.13), (3.14), (3.12) and (3.9)

$$\langle\langle V^n \rangle\rangle = \exp\left\{ Nn \left[\operatorname*{extr}_{M,q} \left(\frac{z_1}{N} G^{(1)}(M, q) + \tfrac{1}{2} \ln(1-q) + \frac{1}{2} \frac{q}{1-q} \right) + O(n, N^{-1}) \right] \right\} \tag{3.15}$$

with

$$G^{(1)}(M, q) = \left\langle\!\!\left\langle \int \mathrm{D}t_1 \ln \int \mathrm{d}\Lambda [2\pi(1-m_1^2)(1-q)]^{-1/2} \times \exp\left\{ -\frac{(\Lambda - m_1 \xi_i^{(1)} M + (1-m_1^2)^{1/2} q^{1/2} t_1)^2}{2(1-m_1^2)(1-q)} + z_0 G^{(0)}(\Lambda, q) \right\} \right\rangle\!\!\right\rangle_{\xi_i^{(1)}} \tag{3.16}$$

and

$$G^{(0)}(\Lambda, q) = \left\langle\!\!\left\langle \int \mathrm{D}t_0 \ln H \left(\frac{\kappa - m_0 \xi_i^{(0)} \Lambda + (1-m_0^2)^{1/2} q^{1/2} t_0}{[(1-m_0^2)(1-q)]^{1/2}} \right) \right\rangle\!\!\right\rangle_{\xi_i^{(0)}}. \tag{3.17}$$

As usual we have used the notations

$$\int \mathrm{D}t \ldots = \int_{-\infty}^{\infty} \frac{\mathrm{d}t}{(2\pi)^{1/2}} \mathrm{e}^{-t^2/2} \ldots \qquad \text{and} \qquad H(x) = \int_x^{\infty} \mathrm{D}t.$$

As discussed at the beginning of this section the critical storage capacity $\alpha_c = (z_0 z_1)_c / N$ is given by the self-consistent equation for q corresponding to (3.15) in the limit $q \to 1$. Using the asymptotic expansion

$$H(x) \sim (2\pi)^{-1} x^{-1} \exp\left\{ -\frac{x^2}{2} \right\} \qquad x \to \infty$$

we get from (3.17)

$$G^{(0)}(\Lambda, q \to 1) \sim -\frac{1}{2(1-q)} \left\langle\!\!\left\langle \int_{-\kappa'}^{\infty} \mathrm{D}t_0 (\kappa' + t_0)^2 \right\rangle\!\!\right\rangle_{\xi_i^{(0)}} \tag{3.18}$$

where we have retained the most divergent term only and have introduced

$$\kappa' = \frac{\kappa - m_0 \xi_i^{(0)} \Lambda}{(1-m_0^2)^{1/2}}. \tag{3.19}$$

For the specialist case we find from (3.16)

$$G^{(1)}(M, q) = z_0 G^{(0)}(\Lambda^{(s)}, q) \tag{3.20}$$

and $\Lambda^{(s)}$ is given by

$$\left. \frac{\partial G^{(0)}}{\partial \Lambda} \right|_{\Lambda = \Lambda^{(s)}} = 0 \tag{3.21}$$

(the limit $N \to \infty$ is to be taken before $q \to 1$). Hence $G^{(1)}(M, q)$ does not depend on M and m_1 in this case.

For a hierarchy of universalist type we note that the first part of the Λ integral in (3.16) becomes a δ-function for $q \to 1$. Therefore we get

$$G^{(1)}(M, q \to 1) \sim -\frac{z_0}{2(1-q)} \left\langle\!\!\left\langle \int \mathrm{D}t_1 \int_{-\kappa''}^{\infty} \mathrm{D}t_0 (\kappa'' + t_0)^2 \right\rangle\!\!\right\rangle_{\xi_i^{(0)} \xi_i^{(1)}} \tag{3.22}$$

with

$$\kappa'' = (1 - m_0^2)^{-1/2} [\kappa - m_0 \xi_i^{(0)} m_1 \xi_i^{(1)} M + m_0 \xi_i^{(0)} (1 - m_1^2)^{1/2} t_1]. \tag{3.23}$$

Inserting (3.20), (3.18) and (3.22) into (3.15) we find the self-consistent equations for the order parameters in the limit $q \to 1$, which determine α_c. For the specialist case these are

$$\frac{1}{\alpha_c} = \left\langle\!\!\left\langle \int_{-\kappa'}^{\infty} \mathrm{D}t_0 (\kappa' + t_0)^2 \right\rangle\!\!\right\rangle_{\xi_i^{(0)}} \Bigg|_{\Lambda = \Lambda^{(s)}} \tag{3.24}$$

where κ' is given by (3.19) and $\Lambda^{(s)}$ is to be determined from

$$0 = \left\langle\!\!\left\langle \int_{-\kappa'}^{\infty} \mathrm{D}t_0 (\kappa' + t_0) \xi_i^{(0)} \right\rangle\!\!\right\rangle_{\xi_i^{(0)}} \Bigg|_{\Lambda = \Lambda^{(s)}}. \tag{3.25}$$

For the universalist we get

$$\frac{1}{\alpha_c} = \left\langle\!\!\left\langle \int \mathrm{D}t_1 \int_{-\kappa''}^{\infty} \mathrm{D}t_0 (\kappa'' + t_0)^2 \right\rangle\!\!\right\rangle_{\xi_i^{(0)} \xi_i^{(1)}} \Bigg|_{M = M^{(s)}} \tag{3.26}$$

where κ'' is given by (3.23) and $M^{(s)}$ follows from

$$0 = \left\langle\!\!\left\langle \int \mathrm{D}t_1 \int_{-\kappa''}^{\infty} \mathrm{D}t_0 (\kappa'' + t_0) \xi_i^{(1)} \xi_i^{(0)} \right\rangle\!\!\right\rangle_{\xi_i^{(1)} \xi_i^{(0)}} \Bigg|_{M = M^{(s)}}. \tag{3.27}$$

Equations (3.24) and (3.25) are exactly the same as obtained by Gardner for patterns with magnetisation m_0, therefore we find

$$\alpha_c^{\mathrm{spec}}(m_0, m_1, \kappa) = \alpha_c^{\mathrm{Ga}}(m_0, \kappa). \tag{3.28}$$

This is a remarkable result and means that if one can store p patterns with magnetisation m_0 one can also store $z_1 = \mathrm{O}(1)$ classes of p/z_1 patterns, where all patterns within one class have again mutual overlap m_0^2, *irrespective of the overlap between different classes.* In particular we find similar to Gardner for $m_0 \to 1$

$$\alpha_c^{\mathrm{spec}}(m_0, m_1, \kappa = 0) \sim \frac{1}{(1 - m_0)\,|\ln(1 - m_0)|} \to \infty \tag{3.29}$$

i.e. strong correlations within the classes facilitate the storage. Recently Ioffe *et al* introduced a symmetric learning rule which asymptotically realises half of the maximal value given by (3.29) and which is therefore probably the best-suited symmetric rule (Ioffe *et al* 1989). All the other learning rules which were proposed (Parga and Virasoro 1986, Feigelman and Ioffe 1987, Cortes *et al* 1987, Bös *et al* 1988, Gutfreund 1988) are, with $\alpha_c = 0.14, \ldots, 0.2$, far from optimal; in particular their values for α_c do not depend on m_0. The result (3.28) has also been reported by Virasoro for the special case $m_1 = 0$ and $\kappa = 0$ (Virasoro 1988). Here we see that the equivalence holds true for any value of κ, hence one should not expect enlarged basins of attractions due to the hierarchical organisation.

In order to determine α_c for the universalist case one has to perform the t_0 and t_1 integrations and the remaining averages in (3.26) and (3.27). This can be done analytically and comparing the results with the equations following from (3.24) and (3.25) one realises

$$\alpha_c^{\text{univ}}(m_0, m_1, k) = \frac{1-m_0^2}{1-m_0^2 m_1^2}\,\alpha_c^{\text{Ga}}(m_0 m_1, \kappa). \tag{3.30}$$

The prefactor in (3.30) characterises the correlations in the hierarchy since m_0^2 gives the overlap between patterns within one class and $m_0^2 m_1^2$ that of patterns of different classes. The storage capacity is very different from the specialist case; in particular we find for $m_0 \to 1$

$$\alpha_c^{\text{univ}}(m_0, m_1, \kappa) \to 0 \tag{3.31}$$

in marked contrast to (3.29). If one deals with infinitely many classes strong correlations within the classes make a perfect storage very difficult. It might be that by allowing for a small percentage of errors in the retrieval (Gardner and Derrida 1988) one could increase α_c significantly as in the Hopfield model (Amit *et al* 1987b); however, this seems not very likely. Note also that for $m_0 = 1$ one must have

$$\alpha_c^{\text{univ}}(1, m_1, \kappa) = z_0 \alpha_c^{\text{Ga}}(m_1, \kappa)$$

since one has to store the first level ancestors only.

Hence $\alpha_c^{\text{univ}}(m_0, m_1, \kappa)$ is discontinuous at $m_0 = 1$; it is the occurrence of *almost* identical patterns which make the storage so difficult. In this context it might be helpful to recall that it is impossible (without self-couplings) to store two patterns exactly which differ only by one bit. It is remarkable that all these complications do not arise in the specialist case. Figure 2 shows $\alpha_c^{\text{univ}}(m_0, m_1, 0)$ as a function of m_0 for different values of m_1; figure 3 illustrates the dependence on m_1 with m_0 as a parameter. For $m_1 \to 1$ we get from (3.30)

$$\alpha_c^{\text{univ}}(m_0, m_1, \kappa) \simeq \alpha_c^{\text{Ga}}(m_0, \kappa)$$

since we come back to the 'one-level hierarchy' of patterns with magnetisation m_0.

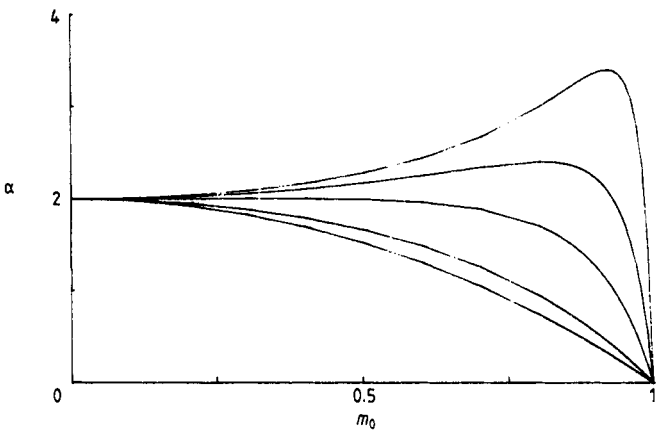

Figure 2. Storage capacity for a two-level hierarchy of universalist type for $\kappa = 0$ as a function of m_0 for $m_1 = 0.95, 0.9, 0.8, 0.5, 0.2$ (from top to bottom).

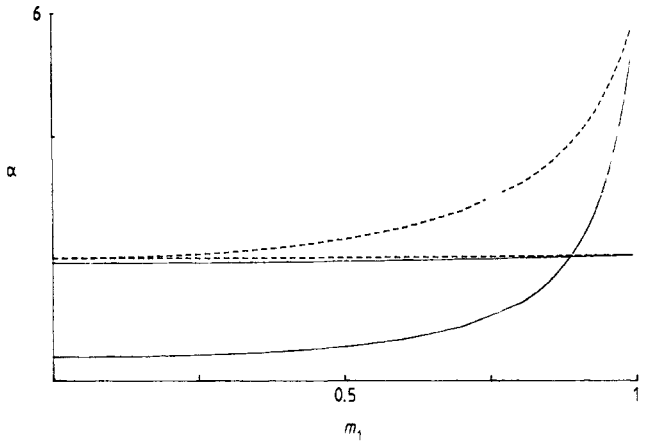

Figure 3. Storage capacity for a two-level hierarchy of universalist type for $\kappa = 0$ as a function of m_1 for $m_0 = 0.2$ (left top) and $m_0 = 0.9$. The dotted curve is $\alpha_c^{Ga}(m_0 m_1, \kappa = 0)$.

Using an explicit learning rule Gutfreund (1988) finds

$$\alpha_c \sim (1 - |m_0|)^2 \qquad m_0 \to 1 \tag{3.32}$$

which is clearly below the actual asymptotics given by (3.30). One should note, moreover, that with this learning rule one misses the difference between specialist and universalist and for the former (3.32) is very poor.

Finally we note that (3.30) is again valid for all values of κ so that there are again probably no improved attraction basins due to the hierarchical organisation of the patterns.

4. Hierarchies with infinitely many levels

The results for a two-level hierarchy obtained in the preceding section should be representative for hierarchies with more but finitely many levels. In particular the branching ratio z_k has to diverge for one or more k in order to ensure $p = \prod_{k=0}^{K-1} z_k = O(N)$. The statistics of the higher levels will then be again irrelevant for α_c as in the specialist case of a two-level hierarchy.

The situation is different if one allows for a divergence of the number of levels K with $N \to \infty$ keeping all branching ratios finite. The methods developed in the last section, in particular the renormalisation group procedure to determine $\langle\langle V^n \rangle\rangle$, can be used to estimate α_c also for this case. It will turn out that we find again $p = O(N)$ and if all branching ratios z_k are of the same order of magnitude this gives

$$K \sim \ln N \tag{4.1}$$

for the number of levels in the hierarchy.

The patterns are defined by

$$\xi_i^\mu = \prod_{k=0}^{K} \xi_i^{(k),\mu_k} \qquad \mu_k = [\mu_{k-1}/z_{k-1}] = 1, \ldots, z_k \tag{4.2}$$

with

$$P(\xi_i^{(k),\mu_k}) = \frac{1+m_k}{2}\,\delta(\xi_i^{(k),\mu_k}-1) + \frac{1-m_k}{2}\,\delta(\xi_i^{(k),\mu_k}+1) \tag{4.3}$$

by analogy with (2.3) and (2.4). The calculation of $\langle\!\langle V^n \rangle\!\rangle$ is a generalisation of the procedure used in section 3. The disorder-independent terms present no problems. The rest is first averaged over $\xi_j^{(0),\mu_0}$. Introducing the abbreviations

$$\lambda_{\mu_k}^{k,\alpha} = \frac{1}{N^{1/2}}\,{}^{(k)}\xi_i^{\mu_k} \sum_j J_{ij}^{\alpha}\,{}^{(k)}\xi_j^{\mu_k} \tag{4.4}$$

one finds for $k=1$ an expression similar to (3.8) with $\Lambda_\mu^\alpha \to \lambda_{\mu_1}^{1,\alpha}$ and $X_\mu^\alpha \to x_{\mu_1}^{1,\alpha}$. In (4.4) we have used the notation

$${}^{(k)}\xi_i^{\mu_k} = \prod_{l=k}^{K} \xi_i^{(l),\mu_l}$$

for the 'rest' of ξ_j^μ after averaging over the k lowest levels of the hierarchy. At this stage we find for the disorder-dependent part of the result (cf (3.8))

$$\left\langle\!\!\left\langle \int \prod_{\mu_k,\alpha} \frac{\mathrm{d}\lambda_{\mu_k}^{k,\alpha}}{2\pi} \int \prod_{\mu_k,\alpha} \mathrm{d}x_{\mu_k}^{k,\alpha} \exp\left\{ \mathrm{i} \sum_{\mu_k,\alpha} \lambda_{\mu_k}^{k,\alpha} x_{\mu_k}^{k,\alpha} - \frac{1}{N^{1/2}} \sum_{\mu_k,\alpha} {}^{(k)}\xi_i^{\mu_k} x_{\mu_k}^{k,\alpha} \sum_j J_{ij}^{\alpha}\,{}^{(k)}\xi_j^{\mu_k} \right.\right.\right.$$

$$\left.\left.\left. + z_{k-1} \sum_{\mu_k} G^{(k-1)}(\bar{\lambda}_{\mu_k}^{k}, \bar{\bar{q}}) \right\} \right\rangle\!\!\right\rangle_{{}^{(k)}\xi_i^{\mu_k},{}^{(k)}\xi_j^{\mu_k}} \tag{4.5}$$

with the recursion relation for the G functions:

$$G^{(k)}(\bar{\lambda}_{\mu_{k+1}}^{k+1}, \bar{\bar{q}}) = \ln\left\langle\!\!\left\langle \int \prod_\alpha \frac{\mathrm{d}\lambda^\alpha}{2\pi} \int \prod_\alpha \mathrm{d}x^\alpha \exp\left\{ \mathrm{i} \sum_\alpha x^\alpha \lambda^\alpha - \mathrm{i} m_k \xi_i^{(k)} \sum_\alpha x^\alpha \lambda_{\mu_{k+1}}^{k+1,\alpha} \right.\right.\right.$$

$$\left.\left.\left. - \frac{1-m_k^2}{2} \sum_\alpha (x^\alpha)^2 - \frac{1-m_k^2}{2} \sum_{(\alpha,\beta)} x^\alpha x^\beta q^{\alpha\beta} + z_{k-1} G^{(k-1)}(\bar{\lambda}, \bar{\bar{q}}) \right\} \right\rangle\!\!\right\rangle_{\xi_i^{(k)}} \tag{4.6}$$

similar to (3.12). The second term in the exponent of (4.6) describes the coupling of level k to level $(k+1)$, the last term to level $(k-1)$. If all averages are performed we introduce

$$M^\alpha = \lambda_1^{k,\alpha} = \frac{1}{N^{1/2}} \sum_j J_{ij}^\alpha$$

and find an expression for $\langle\!\langle V^n \rangle\!\rangle$ analogous with (3.13) where only the last term is to be replaced by $(z_{k-1}/N) G^{(K-1)}(\bar{M}, \bar{\bar{q}})$. As will become clear immediately we have $G^{(K-1)}(\bar{M}, \bar{\bar{q}}) = \mathrm{O}(\Pi_{k=1}^{K-1} z_k)$ and with (4.1) the remaining integrals can be calculated for $N \to \infty$ by the saddle-point method. Assuming replica symmetry similar simplifications as for the two-level hierarchy are possible and we get

$$\langle\!\langle V^n \rangle\!\rangle = \exp\left\{ Nn \left[\operatorname*{extr}_{M,q} \left(\frac{z_{k-1}}{N} G^{(K-1)}(M,q) + \tfrac{1}{2}\ln(1-q) + \frac{1}{2}\frac{q}{1-q} \right) + \mathrm{O}(n, N^{-1}) \right] \right\} \tag{4.7}$$

as a generalisation of (3.15). Moreover the recursion relation (4.6) simplifies to

$$G^{(k)}(\lambda^{k+1}_{\mu_{k+1}}, q) = \left\langle\!\!\left\langle \int \mathrm{D}t_k \ln\left[\int \mathrm{d}\lambda^k [2\pi(1-m_k^2)(1-q)]^{-1/2} \right.\right.\right.$$
$$\times \exp\left\{-\frac{(\lambda^k - m_k\xi_i^{(k)}\lambda^{k+1}_{\mu_{k+1}} + (1-m_k^2)^{1/2}q^{1/2}t_k)^2}{2(1-m_k^2)(1-q)}\right.$$
$$\left.\left.\left. + z_{k-1}G^{(k-1)}(\lambda^k, q)\right\}\right]\right\rangle\!\!\right\rangle_{\xi_i^{(k)}} \tag{4.8}$$

Finally we have to take the limit $q \to 1$ in order to obtain α_c. Since all z_k are of order one the λ^k integrals in (4.8) can be done by using that the first part of the integrand becomes a δ-function for $q \to 1$. As a result we find

$$G^{(k)}(\lambda^{k+1}, q \to 1) = z_{k-1}\left\langle\!\!\left\langle \int \mathrm{D}t_k G^{(k-1)}(m_k\xi_i^{(k)}\lambda^{k+1} - (1-m_k^2)^{1/2}t_k, q \to 1)\right\rangle\!\!\right\rangle_{\xi_i^{(k)}} \tag{4.9}$$

which after iteration justifies the use of the saddle-point method. With the help of (4.9) we can easily express $G^{(k-1)}(M, q \to 1)$ in terms of $G^{(0)}(\lambda^1, q \to 1)$ which after replacing Λ by λ^1 is given by (3.18). In this way we get from (4.7) the following equations for the determination of α_c:

$$\frac{1}{\alpha_c} = \left\langle\!\!\left\langle \int \mathrm{D}t_{K-1} \ldots \int \mathrm{D}t_1 \int_{-\kappa_K}^{\infty} \mathrm{D}t_0 (\kappa_K + t_0)^2 \right\rangle\!\!\right\rangle_{\xi_i^{(0)}, \ldots, \xi_i^{(K-1)}} \tag{4.10}$$

$$0 = \left\langle\!\!\left\langle \int Dt_{K-1} \ldots \int \mathrm{D}t_1 \int_{-\kappa_K}^{\infty} \mathrm{D}t_0 (\kappa_K + t_0)\xi_i^{(0)} \ldots \xi_i^{(K-1)} \right\rangle\!\!\right\rangle_{\xi_i^{(0)}, \ldots, \xi_i^{(K-1)}} \tag{4.11}$$

$$\kappa_K = (1-m_0^2)^{-1/2}(\kappa - m_0\xi_i^{(0)}\{m_1\xi_i^{(1)}[\ldots(m_{K-1}\xi_i^{(K-1)}M - (1-m_{K-1}^2)^{1/2}t_{K-1}) - \ldots]$$
$$-(1-m_1^2)^{1/2}t_1\}). \tag{4.12}$$

These equations generalise (3.26), (3.27), (3.23) of the previous section. The remaining averages over $\xi_i^{(k)}$ and t_k can again be performed analytically where one starts with $\xi_i^{(0)}$, t_0 and finds again a renormalisation-group-like behaviour due to the structure of κ_K. For simplicity we consider the case $\kappa = 0$ only. Having performed the averages from $\xi_i^{(0)}$, t_0 to $\xi_i^{(l)}$, t_l, $0 < l < K-1$, we find from (4.10)-(4.12):

$$\frac{1-m_0^2}{\alpha_c} = \left\langle\!\!\left\langle \int \mathrm{D}t_{K-1} \ldots \int \mathrm{D}t_{l+1}\left\{[(1-\bar{m}_l^2) + \bar{m}_l^2(1-m_{l+1}^2)(t_{l+1}+\kappa_{l+1})^2]\right.\right.\right.$$
$$\times\left[\frac{1+\bar{m}_l}{2} - \bar{m}_l H\left(\frac{\bar{m}_l(1-m_{l+1}^2)^{1/2}}{(1-\bar{m}_l^2)^{1/2}}(t_{l+1}+\kappa_{l+1})\right)\right]$$
$$+\bar{m}_l^2(1-\bar{m}_l^2)^{1/2}(1-m_{l+1}^2)^{1/2}(t_{l+1}+\kappa_{l+1})\frac{1}{(2\pi)^{1/2}}$$
$$\left.\left.\left.\times\exp\left\{-\frac{\bar{m}_l^2(1-m_{l+1}^2)(t_{l+1}+\kappa_{l+1})^2}{2(1-\bar{m}_l^2)}\right\}\right]\right\rangle\!\!\right\rangle_{\xi_l^{(l+1)}, \ldots, \xi_l^{(K-1)}} \tag{4.13}$$

$$0 = \left\langle\!\!\left\langle \int \mathrm{D}t_{K-1} \ldots \int \mathrm{D}t_{l+1}\left[\frac{(1-m_{l+1}^2)^{1/2}}{(1-\bar{m}_l^2)^{1/2}}\left[\frac{1+\bar{m}_l}{2} - \bar{m}_l H\left(\frac{\bar{m}_l(1-m_{l+1}^2)^{1/2}}{(1-\bar{m}_l^2)^{1/2}}(t_{l+1}+\kappa_{l+1})\right)\right]\right.\right.\right.$$
$$\left.\left.\left.+\frac{1}{(2\pi)^{1/2}}\exp\left\{-\frac{\bar{m}_l^2(1-m_{l+1}^2)(t_{l+1}+\kappa_{l+1})^2}{2(1-\bar{m}_l^2)}\right\}\right]\right\rangle\!\!\right\rangle_{\xi_l^{(l+1)}, \ldots, \xi_l^{(K-1)}} \tag{4.14}$$

$$\kappa_l = \frac{(-1)^{l+1} m_l \xi_i^{(l)}}{(1-m_l^2)^{1/2}} - m_{l+1}\xi_i^{(l+1)}[\ldots(m_{K-1}\xi_i^{(K-1)}M - (1-m_{K-1}^2)^{1/2}t_{K-1}) - \ldots]$$

$$-(1-m_{l+1}^2)^{1/2}t_{l+1} \tag{4.15}$$

where

$$\bar{m}_l = \prod_{k=0}^{l} m_k. \tag{4.16}$$

Hence the mathematical structure of these equations is the same for all l; only the parameters κ_l and $\bar{m}_l$ vary according to (4.15) and (4.16). With the additional definition $m_{-1}=1$, (4.10)-(4.12) also fit into this scheme.

Comparing the result for $l = K-1$ with the equation of Gardner (cf (3.24), (3.25)) we get the final result

$$\alpha_c(m_0, m_1, \ldots, m_{K-1}, \kappa = 0) = \frac{1-m_0^2}{1-m^2}\,\alpha_c^{\mathrm{Ga}}(m, \kappa = 0) \tag{4.17}$$

with

$$m = \bar{m}_{K-1} = \prod_{k=0}^{K-1} m_k. \tag{4.18}$$

Therefore a hierarchy with very many levels has storage properties similar to a two-level hierarchy of universalist type (cf (3.30)). Only the largest and the smallest correlations in the system, characterised by m_0 and m respectively, are essential for α_c. Since one assumes $m_k < 1$ for all k we have $m \to 0$ for $K = O(\ln N) \to \infty$ and find from (4.17) the remarkably simple result

$$\alpha_c(m_0, m_1, \ldots; \kappa = 0) = 2(1-m_0^2). \tag{4.19}$$

5. Information content of a pattern hierarchy

Already in the discussion of patterns with low level of activity or magnetisation it was noted that the storage capacity does not adequately characterise the efficiency of storage since the patterns contain less information if their correlation increases (Amit *et al* 1987a, Gardner 1988). In order to compare the storage of hierarchically correlated patterns with that of patterns with other statistical properties it is therefore necessary to calculate the information content of a pattern hierarchy. Since there are no correlations between different neurons it is sufficient to consider one neuron; the index i is therefore dropped. The information I of a sequence $\{\xi^\mu\}$, $\mu = 1, \ldots, p$, is given by

$$I_{\mathrm{seq}} = -\operatorname*{Tr}_{\{\xi^\mu\}} P(\{\xi^\mu\}) \ln P(\{\xi^\mu\}) \tag{5.1}$$

where $P(\{\xi^\mu\})$ denotes the probability of the sequence.

For simplicity we consider again a two-level hierarchy; the generalisation to hierarchies with more levels creates no problems. According to the definitions (2.3) and (2.4) the sequence consists of z_1 independent words, each containing z_0 letters which

are not independent of each other. All the words contain the same information, hence we have

$$I_{\text{seq}} = z_1 I_{\text{word}} = -z_1 \operatorname*{Tr}_{\{\xi^\mu\}} P(\{\xi^\mu\}) \ln P(\{\xi^\mu\}) \tag{5.2}$$

where $\{\xi^\mu\}$, $\mu = 1, \ldots, z_0$, now stands for one word.

From (2.3) we have

$$P(\{\xi^\mu\}) = \text{Prob}\{\xi^{(1)} = 1\} \prod_\mu \text{Prob}\{\xi^{(0),\mu} = \xi^\mu\} + \text{Prob}\{\xi^{(1)} = -1\} \prod_\mu \text{Prob}\{\xi^{(0),\mu} = -\xi^\mu\}. \tag{5.3}$$

Since all the $\xi^{(0),\mu}$ are independent, $P(\{\xi^\mu\})$ only depends on the number k of patterns with $\xi^\mu = +1$, i.e. we can write

$$\operatorname*{Tr}_{\{\xi^\mu\}} \ldots = \sum_{k=0}^{z_0} \binom{z_0}{k} \ldots$$

and from (5.2), (5.3) and (2.4) we get

$$I_{\text{word}} = -\sum_{k=0}^{z_0} \binom{z_0}{k} \left\{ \left[\frac{1+m_1}{2} \left(\frac{1+m_0}{2}\right)^k \left(\frac{1-m_0}{2}\right)^{z_0-k} + \frac{1-m_1}{2} \left(\frac{1+m_0}{2}\right)^{z_0-k} \left(\frac{1-m_0}{2}\right)^k \right] \times \ln \left[\frac{1+m_1}{2} \left(\frac{1+m_0}{2}\right)^k \left(\frac{1-m_0}{2}\right)^{z_0-k} + \frac{1-m_1}{2} \left(\frac{1+m_0}{2}\right)^{z_0-k} \left(\frac{1-m_0}{2}\right)^k \right] \right\}. \tag{5.4}$$

Using (5.4) we can calculate the whole information content per synapse of a two-level pattern hierarchy as:

$$I = \frac{N I_{\text{seq}}}{N^2} = \frac{z_1 I_{\text{word}}}{N} = \frac{\alpha_c}{z_0} I_{\text{word}}. \tag{5.5}$$

For $m_0 = 0$ we get $I_{\text{word}} = z_0 \ln 2$ independent of m_1, for $m_0 \to \pm 1$ only the $k = 0$ and $k = z_0$ terms in (5.4) survive yielding $I_{\text{word}} = \frac{1}{2}(1+m_1) \ln \frac{1}{2}(1+m_1) + \frac{1}{2}(1-m_1) \ln \frac{1}{2}(1-m_1)$. Moreover we find

$$I_{\text{word}} = -z_0 \left[\frac{1-m_0}{2} \ln \frac{1+m_0}{2} + \frac{1-m_0}{2} \ln \frac{1-m_0}{2} \right] \qquad \text{if } m_1 \to \pm 1 \tag{5.6}$$

which with (5.5) reproduces the result for patterns with magnetisation m_0 (Amit *et al* 1987a, Gardner 1988).

Let us now consider the differences of the information content of a two-level and a 'one-level' hierarchy. Because of (3.28) we compare a hierarchy of specialist type with patterns with magnetisation m_0 and infer from (5.6) for $m_1 \in (-1, 1)$

$$I_{\text{word}} \geqslant -z_0 \left[\frac{1+m_0}{2} \ln \frac{1+m_0}{2} + \frac{1-m_0}{2} \ln \frac{1-m_0}{2} \right] \tag{5.7}$$

and hence a two-level hierarchy contains indeed more information than one with one level only. However, taking into account (5.5) we see that because of $z_0 = O(N)$ this increase in information is only $O(1/N)$ and therefore negligible for $N \to \infty$. For a universalist we get from (5.4) and (5.5) that the information is maximal for $z_0 = 1$ which is just the case of patterns with magnetisation (m_0, m_1). This is demonstrated in figure 4 where I is plotted as a function of m_0 for two values of m_1 and $z_0 = 1, 2, 3$.

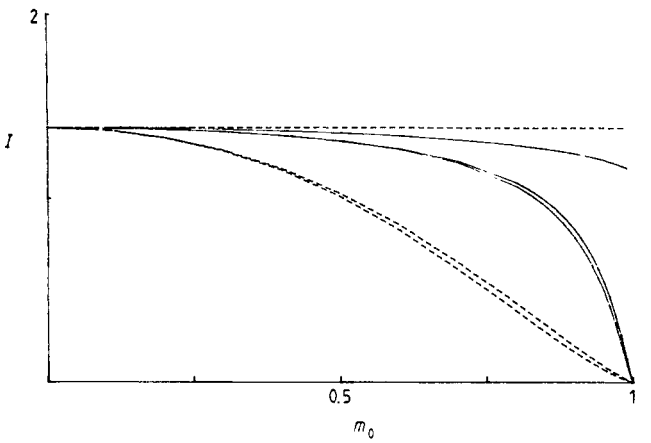

Figure 4. Information capacity of a network storing a two-level pattern hierarchy of universalist type as a function of m_0 for $m_1 = 0.95$ (full curve) and $m_1 = 0.2$ (broken curve). The branching ratio z_0 takes the values 1, 2, 3 (from top to bottom).

6. Summary

In this paper we have investigated the maximal storage capacity $\alpha_c = p_c N$ of an attractor neural network storing hierarchically correlated patterns. This has been done using the methods developed by Gardner for statistical mechanics in the phase space of synaptic interactions and therefore no reference to a special learning rule was necessary. Consequently the results for α_c depend indeed only on the statistical properties of the pattern hierarchies under consideration.

In the case of a two-level hierarchy a distinction according to the order of magnitude of the two branching ratios z_0 and z_1 was necessary. We have always $\alpha_c = O(1)$ but the actual value of α_c depends strongly on whether $z_0 = O(N)$ and $z_1 = O(1)$, which was called the *specialist* case, or $z_0 = O(1)$ and $z_1 = O(N)$ which was referred to as the *universalist* case. A specialist stores many items which belong to only a few classes; a universalist knows some items out of very many different classes. This distinction may easily be missed if one uses special learning rules and only requires $(z_0 z_1) = O(N)$ (Gutfreund 1988, Bös *et al* 1988).

The specialist case is characterised by the same storage capacity α_c as determined by Gardner for patterns with magnetisation (Gardner 1988) which can be interpreted as a 'one-level hierarchy'

$$\alpha_c^{\text{spec}}(m_0, m_1, \kappa) = \alpha_c^{\text{Ga}}(m_0, \kappa). \tag{6.1}$$

In particular one finds $\alpha_c^{\text{spec}}(m_0, m_1, \kappa) \to \infty$ if $m_0 \to 1$, i.e. strong correlations inside the classes facilitate the storage. This is very different from the universalist case, where one finds

$$\alpha_c^{\text{univ}}(m_0, m_1, \kappa) = \frac{1 - m_0^2}{1 - m_0^2 m_1^2} \alpha_c^{\text{Ga}}(m_0 m_1, \kappa) \tag{6.2}$$

and hence $\alpha_c(m_0, m_1, \kappa) \to 0$ if $m_0 \to 1$. Now strong correlations inside the classes requires storage of very many classes of almost identical patterns, which is very difficult and results in a strong decrease of α_c.

All values for α_c produced so far with special learning rules are, in part, considerably smaller than the results (6.1) and (6.2) (Parga and Virasoro 1986, Feigelman and Ioffe 1987, Cortes *et al* 1987, Bös *et al* 1988, Gutfreund 1988). The best result was reported recently by Ioffe *et al* which saturates with a symmetric learning rule asymptotically half of the optimal value (6.1) for a hierarchy of specialist type (Ioffe *et al* 1989). The methods developed in section 3 for the discussion of a two-level hierarchy allows for a generalisation to an arbitrary number of levels; in particular also the case where the number of levels diverge with the number N of neurons tending to infinity can be treated. The result for α_c is similar to the universalist case (cf (4.17)) and depends only on the largest and the smallest correlations in the system. Therefore the higher levels $k \geq 2$ of the hierarchy influence α_c only insofar as they produce the first level ancestors of the patterns. For infinitely many levels we get

$$\alpha_c = 2(1 - m_0^2) \tag{6.3}$$

as for a two-level hierarchy of universalist type with uncorrelated ancestors ($m_1 = 0$).

These results show that one cannot improve the storage capacity with the help of a regular hierarchical organisation of the patterns as defined in section 2 beyond the value for a 'one-level hierarchy' of patterns with magnetisation. Since (6.1) and (6.2) are valid for all values of κ there is probably also no enlargement of the typical basins of attraction of the patterns. Due to the correlations between patterns forming a hierarchy the storage capacity α_c gives only a rough estimate of the amount of information stored. Calculating the information content of a two-level hierarchy one finds that a hierarchy of specialist type contains more information than the corresponding set of patterns with magnetisation; however, the increase is only of order $1/N$. A universalist hierarchy contains less information than patterns with magnetisation $m_0 m_1$ because of the additional correlations inside the lowest classes. The same argument holds for hierarchies with infinitely many levels.

In conclusion, pattern hierarchies of the type mostly discussed so far do not fulfil the hopes which were linked with a hierarchical organisation of the information to be stored. Generalisations may include irregular hierarchies or those where all patterns of one class coincide for a fixed set of neurons, which is a stronger constraint than the requirement of equal mutual overlap used in our definition. Another possibility is opened by learning procedures in randomly prestructured networks which possess from the start exponentially many hierarchically organised attractors (Toulouse *et al* 1986).

Acknowledgments

This work was started during a stay at the Limburgs-Universitair Centrum in Diepenbeek, Belgium. I am indebted to Professors M Bouten, C van den Broeck and R Serneels for their kind hospitality and many interesting discussions. I would like to thank Dr H Herzel for clarifying discussions about the information of letters and words.

References

Amit D J, Gutfreund H and Sompolinsky H 1987a *Phys. Rev.* A **35** 2293
—— 1987b *Ann. Phys., NY* **173** 30

Bös S, Kühn R and van Hemmen J L 1988 *Z. Phys.* B **71** 261
Cortes C, Krogh A and Hertz J A 1987 *J. Phys. A: Math. Gen.* **20** 4449
Dotsenko V S 1985 *J. Phys. C: Solid State Phys.* **18** L1017
—— 1986 *Physica* **140A** 410
Feigelman M V and Ioffe L B 1987 *Int. J. Mod. Phys.* B **1** 51
Gardner E 1988 *J. Phys A: Math. Gen.* **21** 257
Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
Gutfreund H 1988 *Phys. Rev.* A **37** 570
Ioffe L B, Kühn R and van Hemmen J L 1989 *J. Phys. A: Math. Gen.* **22** L1037
Kanter I and Sompolinsky H 1987 *Phys. Rev.* A **35** 380
Mézard M, Parisi G, Sourlas N, Toulouse G and Virasoro M A 1984 *J. Physique* **45** 843
Mézard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)
Parga N and Virasoro M A 1986 *J. Physique* **47** 1857
Rammal R, Toulouse G and Virasoro M A 1986 *Rev. Mod. Phys.* **58** 765
Sourlas N 1988 *Europhys. Lett.* **7** 749
Sutton J P, Beis J S and Trainor L E H 1988 *J. Phys. A: Math. Gen.* **21** 4443
Toulouse G, Dehaene S and Changeux J P 1986 *Proc. Natl. Acad. Sci. USA* **83** 1695
Virasoro M A 1988 *Europhys. Lett.* **7** 293