Dynamics and statistical mechanics of the Hopfield model

1987 J. Phys. A: Math. Gen. 20 2909

(http://iopscience.iop.org/0305-4470/20/10/035)

J. Phys. A: Math. Gen. **20** (1987) 2909-2934. Printed in the UK

Dynamics and statistical mechanics of the Hopfield model

A D Bruce, E J Gardner and D J Wallace

Department of Physics, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JZ, UK

Received 17 September 1986

Abstract. We present a study of the Hopfield model of the memory characteristics of a network of interconnected two-state neuron variables. The fraction of nominated configurations which the model stores without error is calculated analytically as a function of the number, N, of neurons and the number, n, of the nominated configurations. The calculation is tested by computer simulation. The noise-free (zero-temperature) phase diagram of the model is determined within a replica-symmetric solution of the mean-field equations. The model exhibits a phase transition at $\alpha(\equiv n/N)=\alpha_c\approx 0.069$; at this point the thermodynamic states having macroscopic overlap with the nominated configurations disappear, implying a discontinuous change in the fraction of bits (of any nominated configuration) recalled correctly. Large scale Monte Carlo simulations using a distributed array processor provide some support for the existence of a phase transition close to the predicted value.

1. Introduction

This paper describes a study of a model assembly composed of a large number of interacting two-state variables. The physical motivation for the model lies in neurobiology: with appropriate interpretations the model may be viewed as displaying properties ('memory', 'learning') characteristic of neural networks (Little 1974, Hopfield 1982). The model† has, however, attracted considerable attention in the physics community (Amit *et al* 1985a, b, 1986, Kinzel 1985, Dotsenko 1985, Parisi 1986), spurred by the recognition that the problems it poses are similar to those raised by the Sherrington-Kirkpatrick (1975) model of a spin glass, which it generalises.

The model is composed of a set of N variables V_i $(i=1,\ldots,N)$ each of which may adopt either of two values 1 or 0. In the neuro-physiological interpretation each variable characterises the state ('firing' or 'not firing') of a particular neuron. The collective behaviour of the model has two distinct controlling ingredients.

Firstly, we define a *configurational energy* which serves to prescribe the effective environment of each variable. This energy function can be written in the form

$$E(\{V\})=-\tfrac{1}{2}\sum_{i,j}T_{ij}V_iV_j+\sum_i U_iV_i. \tag{1.1a}$$

The lower energy state of any particular variable V_i (given a particular configuration of all the other variables) is then $V_i=1$ (or $V_i=0$) according as the effective local field

$$H_i=\sum_j T_{ij}V_j-U_i \tag{1.1b}$$

† We are aware that this model has its origins in much earlier work and can be regarded as a limiting case of other models. However we believe that it is appropriate to associate it with the name of Hopfield who appreciated its prototypical simplicity and the important role of the energy function (1.1*a*).

0305-4470/87/102909+26$02.50

is positive (or negative). In the neural context the interaction coefficients T_{ij} represent the excitatory ($T_{ij}>0$) or inhibitory ($T_{ij}<0$) effects of the activity of neuron j upon the activity of neuron i, mediated by their synaptic connection. The local field U_i represents the effect of a threshold barrier against the excitation of neuron i.

Secondly we define a *configurational dynamics* which prescribes the manner in which the local variables change in response to their environment. In keeping with the significance accorded to the model parameters the dynamics is chosen so that, with time, the assembly evolves to a configuration in which each local variable occupies the state favoured by the effective field (1.1*b*) it then experiences. These stationary states constitute local minima of the function $E(\{V\})$ with respect to the switching of any (single) variable. They may be thought of as representing patterns of 'neural' activity in which a neuron continues to fire (remains inactive) according to whether or not the net synaptic potential set up by the activity of other neurons (the first term on the right-hand side of (1.1*b*)) exceeds the local threshold potential.

There remains the crucial question of the choice of interaction strengths and threshold potentials. These parameters together serve to define the energy surface, E, in the space $\{V\}$ and thus, in particular, the set of local minima in that space. The essential (neural) aim of the model is realised by tuning these parameters to bring these minimum energy configurations (or at least a subset of them) into coincidence with (or as close as possible to) a set of n nominated configurations $V^{(r)}$ ($r=1,\ldots,n$). Within the neural context the nominated configurations are supposed to represent particular 'images', each expressed as an N-bit word. The strategy of encoding these images (via the model parameters) in the energy surface realises a content-addressable form of storage ('memory'): a particular image may be identified ('recalled') from one of its fragments to the extent that the fragment configuration lies within the basin of attraction associated with the nominated image, and to the extent that the minimum of that basin does indeed accurately represent that image.

We will not review here the arguments which suggest that this model of neuron function, though clearly oversimplified in detail, is plausible in structure. The interested reader is referred to Hopfield (1982) and Hinton and Anderson (1981). We proceed rather to address the more detailed issues necessary to define the concerns of the present study, and to set it in the context of recent work.

Let us consider then, in more specific terms, the assignment of the model parameters (the 'storage prescription'). The simplest prescription realising (in some measure) the aims set out above is defined by (Hopfield 1982)

$$T_{ij}=\frac{1}{N}\sum_r (2V_i^{(r)}-1)(2V_j^{(r)}-1) \qquad i\neq j$$

$$T_{ii}=0 \tag{1.2a}$$

$$U_i=0. \tag{1.2b}$$

The rationale of this choice is revealed by substitution into (1.1*b*). For a particular configuration $V^{(r)}$ one finds that the effective local field can be written as the sum of two terms

$$H_i^{(r)}=(2V_i^{(r)}-1)\frac{1}{N}\sum_{j\neq i} V_j^{(r)}+\frac{1}{N}\sum_{j\neq i}\sum_{s\neq r} V_j^{(r)}(2V_i^{(s)}-1)(2V_j^{(s)}-1). \tag{1.3}$$

If, as we shall suppose throughout this paper, the nominated images are described by random N-bit words, the first term in (1.3) has (average) magnitude $\frac{1}{2}$ and a sign which

is such as to *stabilise* (the ith component of) the nominated image, $V^{(r)}$. The second term, on the other hand, is independent of the value of $V_i^{(r)}$, may be stabilising or destabilising, and has mean zero and standard deviation $N^{-1}[(N-1)(n-1)/2]^{1/2}$. To the extent that N is large compared to n, the first ('signal') term may thus be expected to dominate the second ('interference') term. To the extent that one can, in addition, neglect correlations amongst the elements of the local effective field one may then also expect that the set of vectors $V^{(r)}$ will display the desired stability.

There is a second closely related storage prescription which makes the spin glass analogy more explicit. In this prescription the configurations are most naturally represented by a set of variables $\{S\}$ where $S_i \equiv 2V_i - 1$ takes on values ± 1. The model parameters are chosen to be (Amit *et al* 1985a, b, 1986)

$$T_{ij} = \frac{1}{N} \sum_r S_i^{(r)} S_j^{(r)} \qquad i \neq j \tag{1.4a}$$

$$T_{ii} = 0$$

$$U_i = \tfrac{1}{2} \sum_j T_{ij}. \tag{1.4b}$$

Equation (1.4a) merely recasts (1.2a): the nominated configurations $S^{(r)}$ have elements taken to be ± 1 at random. The choice for the threshold potentials (1.4b) allows the total configurational energy (1.1a) to be written in the form (to within a configuration-independent constant)

$$E(\{S\}) = -\tfrac{1}{2} K \sum_{ij} T_{ij} S_i S_j \tag{1.5}$$

where K ($=\frac{1}{4}$) is a constant whose value is irrelevant to the ('zero-temperature') behaviour studied here. Equation (1.5) has the form of the configurational energy for a model of an Ising spin glass with long-range interactions coupling all the 'spin' coordinates. However, in contrast to the spin glass model of Sherrington and Kirkpatrick (1975, 1978) the 'exchange constants' (the elements of the T_{ij} matrix) are not drawn randomly from some distribution but are correlated in a fashion dictated by the storage prescription (1.4a).

The 'signal plus interference' picture outlined above suggests that this second storage prescription (the 'S model') functions in a very similar way to the first prescription (the 'V model'). The two models are not, however, isomorphic (or, at least, not obviously so). The V model was studied in the seminal paper by Hopfield (1982) with the aid of computer simulations of networks with $N = 30$ and $N = 100$ nodes. More recently, the S model has been studied in considerable detail by Amit *et al* (1985a, b, 1986) using both analytic and computer simulation techniques. In this paper (preliminary versions of which have been reported by Wallace (1985, 1986)), we extend the study of both models. Using the ICL Distributed Array Processor (DAP) we have performed simulations, of the V model, using the much larger networks (up to 4096 nodes) which, we have found, are essential if one is to cope with the substantial finite size effects apparent in the model's rich cooperative behaviour. In tandem with this numerical work we have performed analytic calculations, extending the analysis of the S model by Amit *et al* (1985a, b, 1986) and generalising it to deal also with the V model.

Our specific concerns here are restricted to the effectiveness of the storage prescription (1.2a, b) and, in particular, its dependence upon the storage ratio $\alpha \equiv n/N$. There are at least two rather different criteria for an effective storage prescription. Firstly it

is desirable that the typical minimum energy state $\tilde{V}^{(r)}$, singled out by the configurational flow emanating from a nominated state $V^{(r)}$, should lie as close as possible to that nominated state: recall is then 'accurate'. Secondly it is desirable that the basin of attraction of each such minimum, $V^{(r)}$, should be as large as possible: recall is then indeed by address of content. In this paper we shall consider only the former criterion. The effectiveness (in this sense) of the storage prescription may then be measured by the function $p(D)$ prescribing the distribution p of the fractional Hamming distance D defined by

$$D = \frac{1}{N} \sum_i |\tilde{V}_i^{(r)} - V_i^{(r)}| \tag{1.6}$$

and characterising the difference between nominated states $V^{(r)}$ and the minimum energy states $\tilde{V}^{(r)}$ to which they are linked by the configurational flow. Clearly the storage prescription operates accurately to the extent that the weight of this function is concentrated near $D = 0$. One may characterise the degree to which this aim is realised through either of two parameters.

Firstly we may define

$$\bar{F}_I \equiv p(D = 0) \tag{1.7a}$$

giving the mean fraction of nominated configurations which are stable against the configurational dynamics (the fraction of images which are stored without error). In § 2 we describe an analytic calculation of $\bar{F}_I$, as a function of N and n, which refines the 'signal plus interference' picture by incorporating the correlations which that picture neglects. Complementary Monte Carlo calculations are reported in § 4 and agree well with the analytic theory.

The collective behaviour of the model is, however, more explicitly displayed in the parameter

$$\bar{F}_B \equiv 1 - \sum Dp(D) \equiv 1 - \bar{D} \tag{1.7b}$$

giving the mean fraction of the local variables of a nominated configuration $V_i^{(r)}$ coinciding with those of the associated local minimum, $\tilde{V}_i^{(r)}$, i.e. the fraction of bits which are recalled without error. The storage prescription is then effective to the extent that $\bar{F}_B$ exceeds 0.5, its value in the limit in which there is no coherent overlap between the configurations $V^{(r)}$ and $\tilde{V}^{(r)}$. We have studied this quantity both by direct numerical simulation and by somewhat less direct analytic calculation.

The analytic studies, following those of Amit *et al* (1986) for the S model, and reported in § 3, use the replica techniques of spin glass theory to identify the stable states of the model. The analysis reveals the existence of a critical value, α_c, of the storage ratio. Above α_c (within this framework) the model possesses no stable states having macroscopic overlaps with the nominated configurations; below α_c there do exist such states, the associated overlap remaining finite (indeed very large) as $\alpha \to \alpha_c^-$. The value α_c at which this phase transition occurs is located within the approximation that replica symmetry breaking is ignored.

It is tempting but (for reasons touched on below and elaborated in §§ 4 and 5) not obviously correct to infer that this phase transition will be signalled by a jump discontinuity in the fraction $\bar{F}_B$. In § 4 we present the results of numerical simulations designed to check this inference. There is, indeed, evidence for a phase transition occurring at a value of α close to that predicted by the analytic theory.

The link between the analytic and the numerical studies is, however, a subtle one. The analytic calculations effectively identify the states $\tilde{V}^{(r)}$ (1.6) with minima of the equilibrium free energy of the model, with configurational energy (1.1*a*), in the zero temperature limit. The calculation makes no reference to the configurational dynamics: the results are uniquely prescribed by the energy function. On the other hand, the numerical calculations determine the states $\tilde{V}^{(r)}$ (in direct accordance with the way we have defined them) by following the configurational flow emanating from the state $V^{(r)}$. In this case the results clearly *can* reflect the details of the dynamics. It is not clear what dynamic prescription (if any) will effectively fulfil the same averaging processes as are implicit in the 'equilibrium' theory. We have studied two prescriptions. In the first, the updating of the local variables is *completely asynchronous* and the energy a strictly decreasing function of time. In the second the updating is *partially concurrent* and the energy is only (much) more likely to decrease than increase. The results, it transpires, are perhaps surprisingly insensitive to these differences.

2. The stability of nominated configurations: analytic studies

2.1. Preliminaries

In this section we calculate the dependence of the mean error-free-image fraction $\bar{F}_1$ (1.7*a*) upon the number, N, of nodes in the network and the number n ($\equiv \alpha N$) of nominated images. The calculation is carried out for a model which is somewhat more general than the S and V models introduced in the preceding section and which includes these models as special cases. Specifically we suppose that the local effective field at site i has the form

$$H_i = \sum_j T_{ij} \tfrac{1}{2}[(1-\lambda)+(1+\lambda)(2V_j-1)] \tag{2.1}$$

keeping the convention that the stable state of a neuron at site i has $V_i = 1$ ($V_i = 0$) for $H_i > 0$ ($H_i \leqslant 0$). Setting the control parameter λ to have value $\lambda = 0$, or $\lambda = 1$, one recovers (1.1*b*) with the V model (1.2) or the S model (1.4). The requirement that a particular nominated image, $V^{(r)}$, is stable against the configuration flow is equivalent to the condition that the quantity

$$R_i^{(r)} \equiv (2V_i^{(r)}-1)H_i^{(r)} \tag{2.2}$$

is positive for each site i, where $H_i^{(r)}$ signifies the effective field (2.1) for the particular configuration $V^{(r)}$. Explicitly, the nominated image is stable (unstable) according as to whether

$$P_r \equiv \prod_{i=1}^{N} \theta(R_i^{(r)}) \tag{2.3}$$

is 1 (or 0). The fraction of a particular set of n nominated images which will then be stored without error is

$$F_1 = \frac{1}{n}\sum_{r=1}^{n} P_r \tag{2.4}$$

and the mean error-free-image fraction is

$$\bar{F}_1(N, \alpha) \equiv \langle F_1 \rangle \tag{2.5}$$

where the average is taken over an ensemble of nominated configurations.

Throughout this paper we shall be concerned with the limit in which N and n are large with $\alpha = n/N$ finite. It turns out that, in this limit, the correlations between terms appearing in the sum (2.4) can be neglected so that (to within fluctuation corrections $O(1/\sqrt{n})$) the fraction F_1 for a particular nominal set (2.4) coincides with the ensemble average (2.5). However, correlations between the terms in the product (2.3) *are* important: the effective fields at different sites are correlated by virtue of the correlations amongst the elements of the T_{ij} matrix introduced by the storage prescription, most obviously through the symmetry which it imposes upon this matrix.

2.2. Approximate treatment

We first discuss the approximation in which the correlations identified above are neglected. The 'signal plus interference' picture outlined in the preceding section may then be developed in a straightforward way (Hopfield 1982, Weisbuch and Fogelman-Soulie 1985). Specifically, making the decomposition (1.3) (for the generalised model considered here) in (2.2) we find

$$R_i^{(r)} = \frac{1}{2N}\sum_{j\neq i}[(1+\lambda)+(1-\lambda)(2V_j^{(r)}-1)]$$
$$+\frac{1}{2N}(2V_i^{(r)}-1)\sum_{s\neq r}(2V_i^{(s)}-1)\sum_{j\neq i}(2V_j^{(s)}-1)$$
$$\times[(2V_j^{(r)}-1)(1+\lambda)+(1-\lambda)]. \tag{2.6}$$

The distribution of the first (signal) term (over an ensemble of nominated states) has mean $(N-1)(1+\lambda)/2N$ and standard deviation of order $N^{-1/2}$. The distribution of the second (interference) term is Gaussian with mean zero and standard deviation $[2(N-1)(n-1)(1+\lambda^2)]^{1/2}/2N$. The probability that a particular element $V_i^{(r)}$ is recalled correctly is then (in the prescribed limit)

$$\langle\theta(R_i^{(r)})\rangle = \int_{-[(1+\lambda)^2/2\alpha(1+\lambda^2)]^{1/2}}^{\infty} d\mu \frac{\exp(-\mu^2/2)}{(2\pi)^{1/2}}. \tag{2.7}$$

The neglect of the intersite correlations in the product (2.3) then allows one to write

$$\bar{F}_1(N,\alpha) = \langle\theta(R_i^{(r)})\rangle^N$$
$$\equiv \exp(-Nf_0(\lambda,\alpha)) \tag{2.8a}$$

with

$$f_0(\lambda,\alpha) = -\ln\left\{\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{(1+\lambda)^2}{4\alpha(1+\lambda^2)}\right)^{1/2}\right]\right\} \tag{2.8b}$$

where

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z \exp(-y^2)\,dy. \tag{2.8c}$$

We shall see that the functional form (2.8*a*) is preserved in the exact calculation; the specific form of the function $f_0(\lambda,\alpha)$, given in (2.8*b*), is not.

A measure of the storage capacity can be obtained by expanding f_0 for small α, to give

$$f_0(\lambda, \alpha) \approx \left(\frac{\alpha}{\pi}\frac{(1+\lambda^2)}{(1+\lambda)^2}\right)^{1/2} \exp\left(-\frac{(1+\lambda)^2}{4(1+\lambda^2)\alpha}\right). \tag{2.9}$$

Since $\bar{F}_1$ remains finite provided $\alpha < [(1+\lambda)^2/4(1+\lambda^2)](1/\ln N)$ the maximum number of nominal vectors for which there is a finite probability of perfect storage is

$$n_{\max} \sim \frac{N}{\ln N}\frac{(1+\lambda)^2}{4(1+\lambda^2)}. \tag{2.10}$$

Expression (2.9) will turn out to be correct in the exact calculation, so that (2.10) is indeed a true measure of the storage capacity. The probability of perfect recall (of a given image) is therefore zero for any finite value of α and the optimal storage capacity is obtained for the S model ($\lambda = 1$).

Finally, we note that, according to this approximate calculation, the dependence of f_0 upon λ and α has the simple scaling form

$$f_0(\lambda, \alpha) = f_0\left(0, \frac{1+\lambda^2}{(1+\lambda)^2}\alpha\right). \tag{2.11}$$

We shall see that, when correlations are included, this scaling behaviour holds only in the $\alpha \to 0$ limit.

2.3. Exact calculation

In order to calculate $\bar{F}_1(N, \alpha)$ exactly, the following integral representation for the θ functions will be used:

$$\theta\left(\frac{\lambda+1}{2}+x\right) = \int_{-(\lambda+1)/2}^{\infty}\frac{\mathrm{d}\mu}{2\pi}\int_{-\infty}^{\infty}\mathrm{d}\tau\,\exp[\mathrm{i}\tau(\mu - x)]. \tag{2.12}$$

Then

$$\bar{F}_1(N, \alpha) = \left\langle \prod_{i=1}^{N}\int_{-(\lambda+1)/2}^{\infty}\frac{\mathrm{d}\mu_i}{2\pi}\int_{-\infty}^{\infty}\mathrm{d}\tau_i\,\exp\left\{\mathrm{i}\sum_i \tau_i\left[\mu_i - \left(R_i^r - \frac{(1+\lambda)}{2}\right)\right]\right\}\right\rangle. \tag{2.13}$$

The sums over i and j in the 'interference' term in (2.6) can be decoupled by adding and subtracting the missing $i=j$ term and using the integral representation

$$\exp(-\mathrm{i}A_sB_s/N) = \int_{-\infty}^{\infty}\frac{\mathrm{d}a_s}{(2\pi/N)^{1/2}}\int_{-\infty}^{\infty}\frac{\mathrm{d}b_s}{(2\pi/N)^{1/2}}$$
$$\times \exp\left(\mathrm{i}\frac{N}{2}(a_s^2 - b_s^2) - \frac{\mathrm{i}}{\sqrt{2}}A_s(a_s + b_s) - \frac{\mathrm{i}}{\sqrt{2}}B_s(a_s - b_s)\right) \tag{2.14}$$

with

$$A_s = \sum_i \tau_i(2V_i^{(s)} - 1)(2V_i^{(r)} - 1) \tag{2.15}$$

and

$$B_s = \sum_j (2V_j^{(s)} - 1)\tfrac{1}{2}[(2V_j^{(r)} - 1)(1+\lambda) + (1-\lambda)]. \tag{2.16}$$

Then, after doing the trace over the random variables $V_i^{(r)}$ and $V_i^{(s)}$,

$$\bar{F}_1(N,\alpha)=\prod_{s\neq r}\int\frac{\mathrm{d}a_s}{(2\pi/N)^{1/2}}\int\frac{\mathrm{d}b_s}{(2\pi/N)^{1/2}}\exp\left(\mathrm{i}\frac{N}{2}\sum_{s\neq r}(a_s^2-b_s^2)\right)$$
$$\times\prod_i\left\{\int_{-(\lambda+1)/2}^{\infty}\frac{\mathrm{d}\mu_i}{2\pi}\int_{-\infty}^{\infty}\mathrm{d}\tau_i\exp(\mathrm{i}\tau_i\mu_i)\right.$$
$$\times\frac{1}{2}\left[\prod_{s\neq r}\cos\left(\tau_i\frac{(a_s+b_s)}{\sqrt{2}}+\frac{(a_s-b_s)}{\sqrt{2}}\right)\right.$$
$$\times\exp\left(-\frac{\mathrm{i}}{N}\frac{(1-\lambda)}{2}\sum_{j\neq i}\tau_j+\mathrm{i}\frac{(n-1)}{N}\tau_i\right)$$
$$+\prod_{s\neq r}\cos\left(\tau_i\frac{(a_s+b_s)}{\sqrt{2}}+\lambda\frac{(a_s-b_s)}{\sqrt{2}}\right)$$
$$\left.\left.\times\exp\left(\frac{\mathrm{i}}{N}\frac{(1-\lambda)}{2}\sum_{j\neq i}\tau_j+\mathrm{i}\frac{(n-1)}{N}\lambda\tau_i\right)\right]\right\}. \tag{2.17}$$

Since from equation (2.14) a_s and b_s are of the order of the overlap between pairs of input vectors, which is $\sim 1/\sqrt{N}$ for α finite, only terms in expression (2.17) up to second order in a_s and b_s need be kept (for large N).

Letting

$$a=\frac{1}{\alpha}\sum_{s\neq r}\frac{(a_s+b_s)^2}{2}$$
$$b=\frac{\mathrm{i}}{\alpha}\sum_{s\neq r}\frac{(a_s+b_s)(a_s-b_s)}{2}+1$$
$$c=\frac{1}{\alpha}\sum_{s\neq r}\frac{(a_s-b_s)^2}{4}$$
$$d=\frac{1-\lambda}{2}\left(\sum_i\tau_i\right)N^{-1} \tag{2.18}$$

and introducing Lagrange multipliers A, B, C, D, respectively, conjugate to the constraints in (2.18) so that, for example,

$$1=\int\frac{\mathrm{d}a\,\mathrm{d}A}{2\pi/N\alpha}\exp\left[\mathrm{i}NA\left(\alpha a-\sum_s\frac{(a_s+b_s)^2}{2}\right)\right] \tag{2.19}$$

the integrals over τ_i and over a_s and b_s are decoupled and so can be done easily. $\bar{F}_1(N,\alpha)$ is then given for large N by the saddle point of an integral over the variables a, b, c, d, A, B, C and D. The saddle-point equations for A, B, C, c, d are algebraic and so the integrals over these variables can be done explicitly. One then finds that, as $N\to\infty$, $\bar{F}_1(N,\alpha)$ can be written in the form

$$\bar{F}_1(N,\alpha)=\int\frac{\mathrm{d}D\,\mathrm{d}a\,\mathrm{d}b}{(2\pi/N\alpha)(2\pi/N)^{1/2}}\exp[NG(a,b,D)] \tag{2.20}$$

where

$$G(a,b,D)=\alpha\Bigg[b+\tfrac{1}{2}\ln\left(\frac{a}{\frac{1}{2}(1+D)+\frac{1}{2}\lambda^2(1-D)}\right)$$
$$-\frac{1}{2}+\frac{(1-b)^2}{2a}[\tfrac{1}{2}(1+D)+\tfrac{1}{2}(1-D)\lambda^2]+\tfrac{1}{2}(1+D)$$
$$\times\left[\left[\ln\frac{1}{1+D}\int_{[-\frac{1}{2}(1+\lambda)-\frac{1}{2}D(1-\lambda)]/(a\alpha)^{1/2}}^{\infty}\frac{d\mu}{(2\pi)^{1/2}}\exp\left\{-\tfrac{1}{2}\left[\mu+b\left(\frac{\alpha}{a}\right)^{1/2}\right]^2\right\}\right]\right]$$
$$+\tfrac{1}{2}(1-D)\ln\left[\left[\frac{1}{1-D}\int_{[-\frac{1}{2}(1+\lambda)-\frac{1}{2}D(1-\lambda)]/(a\alpha)^{1/2}}^{\infty}\frac{d\mu}{(2\pi)^{1/2}}\right.\right.$$
$$\left.\left.\times\exp\left\{-\tfrac{1}{2}\left[\mu+b\lambda\left(\frac{\alpha}{a}\right)^{1/2}\right]^2\right\}\right]\right] \tag{2.21}$$

and so, as $N\to\infty$,

$$\bar{F}_1(N,\alpha)=\exp\{-N[f(\lambda,\alpha)+O(1/N)]\} \tag{2.22}$$

where $f(\lambda,\alpha)$ is the extremum over a, b and D of $G(a,b,D)$. The leading finite size correction to (2.22) is effectively a constant (N-independent) prefactor which originates in the determinants associated with the integrations around the saddle point. The function $f(\lambda,\alpha)$ can be calculated analytically as $\alpha\to 0$ by solving the saddle-point equations for a, b and D, with the result

$$f(\lambda,\alpha)=\left(\frac{\alpha}{\pi}\frac{(1+\lambda^2)}{(1+\lambda)^2}\right)^{1/2}\exp\left(-\frac{(1+\lambda)^2}{4(1+\lambda^2)\alpha}\right)\left(1-\frac{2(1+\lambda^2)\alpha}{(1+\lambda)^2}\right)$$
$$\times\left[1-\frac{1}{16}\frac{(1+\lambda)^3}{(1+\lambda^2)^{3/2}}\exp\left(-\frac{(1+\lambda)^2}{4(1+\lambda^2)\alpha}\right)\pi^{-1/2}\alpha^{-5/2}\right] \tag{2.23}$$

as $\alpha\to 0$.

The leading term in this expression agrees with the approximate calculation (2.9) and satisfies the scaling relation (2.11). However, the correction terms break this scaling form, showing that (except in the $\alpha\to 0$ limit) the function $\bar{F}_1(N,\alpha)$ does not have a universal (λ-independent) form.

It is also possible to solve for f as $\alpha\to\infty$. In this limit the signal term vanishes and so $\bar{F}_1(N,\alpha)$ is simply equal to the total number of metastable states divided by 2^N. For $\lambda=1$, we recover the result for the Sherrington-Kirkpatrick model (Bray and Moore 1980, De Dominicis *et al* 1981)

$$f(1,\infty)=0.4939 \tag{2.24}$$

and for $\lambda=0$ we have

$$f(0,\infty)=0.5886. \tag{2.25}$$

For general values of α, f can be determined by solving the saddle-point equations numerically. In figure 1 we show the results for the S model ($\lambda=1$) and the V model ($\lambda=0$); the results are plotted as functions of the scaled storage ratio $\tilde{\alpha}\equiv\alpha(1+\lambda^2)/(1+\lambda)^2$ to reveal the extent of the departure from the scaling form suggested by the approximate analysis (§ 2.2), the explicit result of which is also shown. In § 4.2 we will compare these results (for the V model) with those yielded by computer simulation for various N values.

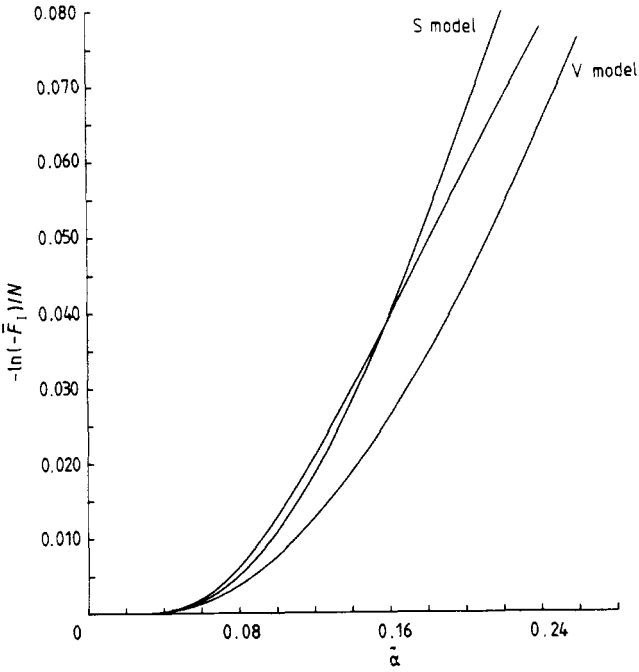

Figure 1. The logarithm of the error-free-image fraction $\bar{F}_1(N, \alpha)$ for the S model, the V model and as given by the approximate calculation discussed in § 2.2. The abscissa is the scaled storage ratio $\tilde{\alpha} \equiv \alpha(1+\lambda^2)/(1+\lambda)^2$.

In carrying out the above calculations we have effectively supposed that the correlations between different P_r in (2.4) are negligible. This assumption can be justified by repeating the calculation using replicas; corrections come from overlaps between different nominated configurations and are down by factors of order $1/\sqrt{\pi}$ on the quoted result.

One can also consider other ways of changing details of the storage prescription. For example, a diagonal term $T_{ii} = \alpha x$ can be added to T_{ij}. For $\lambda = 1$, the leading term as $\alpha \to 0$ is given by the approximate model

$$f(1, \alpha, x) = f_0\left(1, \frac{\alpha}{(1+\alpha x)^2}, 0\right) \tag{2.26}$$

and again non-leading terms do not satisfy this relation.

In general, although increase in the diagonal term or tuning of λ can lead to an increase in storage capacity, this is compensated by an increase in the number of spurious metastable states (Gardner 1986). This means that associative memory should decrease: it is necessary to start nearer the nominated configuration in order to iterate to the associated minimum.

3. Stable configurations: a mean field theory

In this section we will describe the thermodynamics of the model defined by the configurational energy of (1.1*a*). At zero temperature this will allow us to calculate

the Hamming distance between energy minima and nominated configurations and thence (with certain caveats) the error-free-bit fraction defined in § 1. The stability of these energy minima with respect to noise can then be studied by examining the thermodynamics at finite temperature.

The free energy of the model is given by its quenched average over all possible sets of nominated configurations $\{V_i^{(r)}\}$ and can be calculated using the replica method:

$$-\beta\mathscr{F}=\lim_{l\to 0}\frac{\langle Z^l\rangle-1}{l} \tag{3.1}$$

where the partition function Z is given by

$$Z=\operatorname*{Tr}_{\{V_i\}}\exp[-\beta E(\{V_i\})] \tag{3.2}$$

where β is the inverse temperature and where $E(\{V_i\})$ is the configurational energy defined in (1.1a). The method requires the analytic continuation from positive integer to zero number of replicas l. For positive integer values of l

$$\langle Z^l\rangle=\left\langle \operatorname*{Tr}_{\{V_i^\gamma\}}\exp\left(\frac{\beta}{2}\sum_{\gamma=1}^{l}\sum_{i\neq j}T_{ij}V_i^\gamma V_j^\gamma\right)\right\rangle. \tag{3.3}$$

The trace is over all possible configurations of the V_i^γ for each replica γ, and T_{ij} is given by (1.2).

The sites can be decoupled by introducing variables m_r^γ for each nominated configuration r and each replica γ,

$$\langle Z^l\rangle=\int_{-\infty}^{\infty}\prod_{\gamma,r}\frac{\mathrm{d}m_r^\gamma}{(2\pi/\beta N)^{1/2}}\exp\left(-\frac{\beta N}{2}\sum_{r,\gamma}(m_r^\gamma)^2\right)$$
$$\times\left\langle\left[\operatorname*{Tr}_{\{V_i^\gamma\}}\exp\left(-\frac{\beta\alpha}{2}\sum_i\sum_\gamma V_i^\gamma+\beta\sum_{\gamma,r}m_r^\gamma\sum_i(2V_i^{(r)}-1)V_i^\gamma\right)\right]\right\rangle. \tag{3.4}$$

We define the overlap of the thermal expectation, $\bar{V}_i^\gamma$, of the V_i^γ for the replica γ with a nominated configuration r by

$$k_r^\gamma=\frac{1}{N}\sum_i(2V_i^{(r)}-1)\bar{V}_i^\gamma. \tag{3.5}$$

This overlap can be macroscopic (of order 1) for only a finite number (in comparison with N) of values of r. Otherwise it is microscopic (of order $1/\sqrt{N}$).

The sum over r inside the trace in (3.4) can be separated into two parts—a finite subset Γ of values of r which have a macroscopic overlap with at least one of the replicas γ and a subset with a macroscopic number (of order n) of elements which have microscopic overlaps with each replica γ. In order to identify the extensive part of the free energy it is necessary to expand the exponential inside the trace in (3.4) up to second order in $m_r^\gamma(2V_i^{(r)}-1)V_i^\gamma$ for each $r\notin\Gamma$ and at each site i. Then, after averaging over the nominated configurations $\{V_i^{(r)}\}$ for $r\notin\Gamma$, (3.4) becomes

$$\langle Z^l\rangle=\int_{-\infty}^{\infty}\prod_{\gamma,r}\frac{\mathrm{d}m_r^\gamma}{(2\pi/\beta N)^{1/2}}\exp\left(-\frac{\beta N}{2}\sum_{r,\gamma}(m_r^\gamma)^2\right)$$
$$\times\left\langle\operatorname*{Tr}_{\{V^\gamma\}}\exp\left(-\frac{\beta\alpha}{2}\sum_\gamma V^\gamma+\beta\sum_\gamma\sum_{r\in\Gamma_\gamma}m_r^\gamma(2V^{(r)}-1)V^\gamma\right.\right.$$
$$\left.\left.+\frac{\beta^2}{2}\sum_{r\notin\Gamma}\sum_{\gamma,\delta}m_r^\gamma m_r^\delta V^\gamma V^\delta\right)\right\rangle^N \tag{3.6}$$

where Γ_γ denotes the subset of values of r which have a macroscopic overlap with the replica γ and the average $\langle\ \rangle$ in (3.6) now represents an average over nominated configurations $\{V_i^{(r)}\}$ with $r \in \Gamma$. The contribution of terms proportional to m_r^γ where $r \in \Gamma$ but $r \notin \Gamma_\gamma$ can be neglected since they are of order $1/\sqrt{N}$ and since the number of these terms is finite.

The integrals over the m_r^γ for $r \in \Gamma$ can be done by introducing variables x^γ, y^γ for each γ and $q^{\gamma\delta}$, $t^{\gamma\delta}$ for each pair γ and δ with $\gamma < \delta$. Then

$$\langle Z^l \rangle = \prod_{\gamma<\delta} \int_{-i\infty}^{i\infty} \left(\frac{N\beta^2\alpha}{2\pi}\right) dq^{\gamma\delta} \int_{-\infty}^{\infty} dt^{\gamma\delta} \prod_\gamma \int_{-i\infty}^{i\infty} \left(\frac{N\beta^2\alpha}{4\pi}\right) dx^\gamma$$
$$\times \int_{-\infty}^{\infty} dy^\gamma \prod_{\gamma, r\in\Gamma_\gamma} \int \frac{dm_r^\gamma}{(2\pi/N\beta)^{1/2}} \exp[NG(\{m_r^\gamma, x^\gamma, y^\gamma, q^{\gamma\delta}, t^{\gamma\delta}\})] \tag{3.7}$$

where

$$G(\{m_r^\gamma, x^\gamma, y^\gamma, q^{\gamma\delta}, t^{\gamma\delta}\})$$
$$= -\frac{\beta}{2}\sum_\gamma \sum_{r\in\Gamma_\gamma} (m_r^\gamma)^2 - \beta^2\alpha \sum_{\gamma<\delta} q^{\gamma\delta}t^{\gamma\delta} - \frac{\beta^2\alpha}{2}\sum_\gamma y^\gamma x^\gamma - \tfrac{1}{2}\alpha\ \mathrm{Tr}_\gamma \ln(1-\beta X)$$
$$+ \ln\left\langle \mathop{\mathrm{Tr}}_{\{V^\gamma\}} \exp\left(-\frac{\beta\alpha}{2}\sum_\gamma V^\gamma + \beta \sum_\gamma \sum_{r\in\Gamma_\gamma} m_r^\gamma (2V^{(r)}-1)V^\gamma + \frac{\beta^2\alpha}{2}\sum_\gamma y^\gamma V^\gamma \right)\right.$$
$$\left. \times \exp\left(\beta^2\alpha \sum_{\gamma<\delta} t^{\gamma\delta} V^\gamma V^\delta\right)\right\rangle \tag{3.8a}$$

where X is an $l \times l$ matrix with

$$X^{\gamma\gamma} = x^\gamma$$
$$X^{\gamma\delta} = X^{\delta\gamma} = q^{\gamma\delta} \qquad \gamma < \delta. \tag{3.8b}$$

The first trace in (3.8a) is over the replica indices γ and comes from the determinant in the Gaussian integrals over m_r^γ for $r \notin \Gamma$.

In the limit $N \to \infty$, mean field theory should become exact and the free energy is determined by the saddle point of (3.8a) with respect to the variables $q^{\gamma\delta}$, m_r^γ for $r \in \Gamma_\gamma$, $t^{\gamma\delta}$, x^γ and y^γ, in the limit that the number of replicas $l \to 0$. The saddle-point values of these quantities constitute order parameters, whose identifications (within the context of the replicated system) are as follows:

$$m_r^\gamma = \langle (2V_i^{(r)} - 1)\overline{V_i^\gamma} \rangle \qquad r \in \Gamma_\gamma \tag{3.9}$$

$$q^{\gamma\delta} = \langle \overline{V_i^\gamma}\ \overline{V_i^\delta} \rangle \qquad \gamma < \delta \tag{3.10a}$$

$$t^{\gamma\delta} = \frac{1}{\alpha}\left\langle \sum_{r\notin\Gamma} \overline{m_r^\gamma}\ \overline{m_r^\delta} \right\rangle \qquad \gamma < \delta \tag{3.10b}$$

$$x^\gamma = \frac{1}{N}\left\langle \sum_i \overline{V_i^\gamma} \right\rangle \tag{3.11a}$$

$$y^\gamma = \frac{1}{\alpha}\left\langle \sum_{r\notin\Gamma} (\overline{m_r^\gamma})^2 \right\rangle. \tag{3.11b}$$

We shall seek for a saddle point of (3.8) only within the replica-symmetric space for which

$$\begin{aligned} m_r^\gamma &= m_r \\ q^{\gamma\delta} &= q \qquad \gamma < \delta \\ t^{\gamma\delta} &= t \qquad \gamma < \delta \\ x^\gamma &= x \\ y^\gamma &= y. \end{aligned} \tag{3.12}$$

Physically the replica-symmetric ansatz expresses the assumption that there is only one thermodynamically relevant free energy valley associated with each nominated configuration. The validity of this assumption will be discussed later in this section.

Within the replica-symmetric framework one can now proceed to identify the physical significance of the order parameters, through their $l \to 0$ limits. Firstly we see that

$$m_r = \langle (2V_i^{(r)} - 1)\bar{V}_i \rangle \tag{3.13}$$

measures the overlap of the nominated configuration r with the thermodynamic state. Secondly the order parameters

$$q = \frac{1}{N}\sum_i \langle (\bar{V}_i)^2 \rangle \tag{3.14a}$$

and

$$x = \frac{1}{N}\sum_i \langle \overline{V_i} \rangle \tag{3.14b}$$

together characterise the mean and thermal fluctuations of the 'spin glass' order of the thermodynamic state. Finally

$$t = \frac{1}{\alpha}\sum_{r \notin \Gamma_r} \left(\frac{1}{N}\sum_i (2V_i^{(r)} - 1)\bar{V}_i \right)^2 \tag{3.15a}$$

and

$$y = \frac{1}{\alpha}\sum_{r \notin \Gamma_r} \overline{\left(\frac{1}{N}\sum_i (2V_i^{(r)} - 1)V_i \right)^2} \tag{3.15b}$$

together characterise the mean and thermal fluctuations of the overlap between the thermodynamic state and those nominated configurations with which its overlap is microscopic. The order parameters m_r, q and t are similar to those appearing in the S model (Amit *et al* 1985a). The order parameters x and y are peculiar to the V model, reflecting the lack of symmetry between the two neuron states in this system.

In formulating the saddle-point equations we shall impose two further conditions. Firstly we will assume that all replicas have a macroscopic overlap only with the nominated configuration $r = 1$ (so that only m_1 is non-zero); higher energy mixture states can be obtained by assuming macroscopic overlaps with more than one of the nominated configurations. Secondly we will restrict our explicit analysis to the zero-temperature ($\beta \to \infty$) limit most immediately relevant to the noise-free Hopfield model; the behaviour at finite temperature will be discussed in qualitative terms.

In the zero-temperature limit, then we find the following saddle-point equations for the five order parameters m_1, q, t, x and y:

$$t = \frac{q}{[1-\beta(x-q)]^2} \tag{3.16}$$

$$\beta(y-t) = \frac{1}{1-\beta(x-q)} \tag{3.17}$$

$$x = \tfrac{1}{2}(1+\mathrm{erf}(z_+)-\mathrm{erf}(z_-)) \tag{3.18}$$

$$m_1 = \tfrac{1}{2}(\mathrm{erf}(z_+)+\mathrm{erf}(z_-)) \tag{3.19}$$

and

$$\beta(x-q) = \frac{1}{2(2\pi\alpha t)^{1/2}}[\exp(-z_+^2)+\exp(-z_-^2)] \tag{3.20}$$

where

$$z_\pm = (m_1 \pm \tfrac{1}{2}(\beta(y-t)-1)\alpha)(2\alpha t)^{-1/2} \tag{3.21}$$

and $\mathrm{erf}(z)$ is as in (2.8*c*).

The energy (per site) is given by

$$E = -\tfrac{1}{2}m_1^2 - \tfrac{1}{2}\alpha(t-x). \tag{3.22}$$

The saddle-point equations (3.16)-(3.22) can be solved numerically. In the $\beta \to \infty$ limit the thermal fluctuations measured by the differences $x-q$ and $y-t$ vanish; the products $\beta(x-q)$ and $\beta(y-t)$ remain finite. The parameters with the most immediate physical significance are m_1, q and t and it is on these that we shall focus.

There are two kinds of solutions. For all values of α there exists a solution with $m_1 = 0$ but with values of q and t which differ from their high-temperature fully disordered limits ($q = t = \frac{1}{4}$). This is a spin glass solution with no macroscopic overlap with any nominated configuration. For $\alpha < \alpha_c \simeq 0.069$ (cf figure 2) there exists a further solution, characterised by a non-zero value of m_1. For this 'ferromagnetic' solution the thermodynamic state has a macroscopic overlap with the pattern $r = 1$. For $\alpha_c > \alpha > \alpha_1 = 0.025$ the ferromagnetic solution has higher energy than the spin glass solution; for $\alpha < \alpha_1$ the ferromagnetic solution represents the state of minimum energy.

The value of the order parameter m_1 provides the most immediate measure of the efficiency with which the model acts as an associative memory. Specifically, to the extent that one may regard the thermodynamic state as representative of the typical dynamically stable state singled out by the configurational flow from a nominated image (cf remarks at the end of § 1 and in § 4 below) one may identify the error-free-bit fraction (1.7*b*) as

$$\bar{F}_B = \tfrac{1}{2} + m_1. \tag{3.23}$$

The mean field results (figure 2) show that $\bar{F}_B$ remains close to unity throughout the region of metastability of the ferromagnetic phase, right up to the critical storage ratio for which $\bar{F}_B(\alpha_c) \simeq 0.984$. Accordingly, for $\alpha < \alpha_c$, one may expect that the V model does indeed provide an effective form of associative memory.

At zero temperature the order parameter $q(=x)$ measures the mean fraction of bits with value 1. Throughout the ferromagnetic (accurate retrieval) phase this parameter remains close to 0.5 (figure 2); for $\alpha \to \alpha_c^-$ one finds $q \to q(\alpha_c^-) \simeq 0.501$, implying only a small asymmetry between the two neuron states.

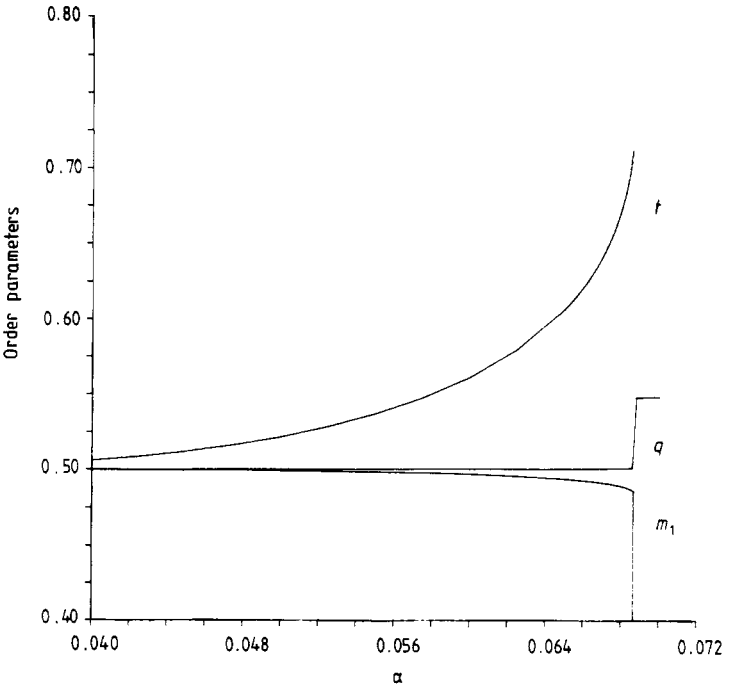

Figure 2. The order parameters t (3.15a), q (3.14a) and m_1 (3.13) as functions of the storage ratio α, according to the replica-symmetric mean field theory. For $\alpha > \alpha_c$, $m_1(\alpha) \equiv 0$; $t(\alpha_c^-) \simeq 0.7$.

The most obvious pretransitional behaviour as $\alpha \to \alpha_c$ is displayed by the order parameter $t = y$ which measures the mean square overlap between the thermodynamic state and nominated configurations other than $r = 1$. For small α this parameter approaches the limit ($t = \frac{1}{2}$) appropriate if the overlaps involved are purely random; as α increases, t grows, signalling a growing correlation between the thermodynamic state and the other ($r \neq 1$) configurations; for $\alpha \to \alpha_c$ one finds $t \to t(\alpha_c^-) \simeq 0.7$; as α_c is approached from above, through the spin glass phase, $t \to t(\alpha_c^+) \simeq 5.1$.

Equations (3.16)-(3.20) are similar to those obtained by Amit *et al* (1985b) for the S model. At zero temperature there is again an approximate scaling relation between the two models which becomes exact only as α tends to zero. Specifically, if one writes $q = x = \frac{1}{2}$, $\beta(y-t)-1=0$, then (3.16), (3.18) and (3.19) are the same as those of the S model provided one replaces t by $t/2$, α by $\alpha/2$ and m_1 by $m_1/2$. Since, for small α, q and x are indeed close to $\frac{1}{2}$ while $[\beta(y-t)-1]\alpha$ is small for the ferromagnetic solution, the two models are indeed nearly equivalent in this regime (given the replacement $\alpha \to \alpha/2$). For the spin glass solution, q and x are again close to $\frac{1}{2}$ and $[\beta(y-t)-1]\alpha$ is again small and so the models are again virtually equivalent.

The mean field equations at finite temperature can also be derived from (3.8). The phase diagram of the V model is similar to that of the S model. At high temperature the system is paramagnetic (with $q = \frac{1}{4}$, $x = \frac{1}{2}$, $t = 1/(4-\beta)$, $y = (4+\beta)/\beta(4-\beta)$). At a temperature $T_g(\alpha)$ the system freezes into a spin glass phase. For all values of α in this phase, there is a (replica-symmetric) spin glass solution ($q \neq \frac{1}{4}$, $x \neq \frac{1}{2}$ and $m_1 = 0$) while for sufficiently small values of $\alpha (< \alpha_c(T))$ there is also a ferromagnetic solution having a finite correlation with the nominated configuration ($m_1 \neq 0$). For $\alpha < \alpha_1(T)$

this solution has lower free energy than the spin glass solution; for $\alpha_1(T) < \alpha < \alpha_c(T)$ the ferromagnetic solution is metastable.

All these results have been obtained in the context of the replica symmetric ansatz (3.12). However for low temperatures (Amit *et al* 1985a) this solution becomes unstable and so replica symmetry must be broken for both the ferromagnetic and the spin glass solutions. Since, for the ferromagnetic solution, the instability line $T = T_R(\alpha)$ is very close to the zero temperature axis for $\alpha < \alpha_c(0)$ ($T_R(\alpha) \sim \sqrt{\alpha} \exp(-\frac{1}{4}\alpha)$ as $\alpha \to 0$) it is reasonable to expect that the effect on α_c is small.

In the following section we present numerical studies of the V model, conducted in the light of the predictions of the mean field theory, assembled in this section, notably the existence of a critical storage ratio α_c. As we have already remarked, the connection between the $T = 0$ thermodynamics and the observed dynamic behaviour of the model is a subtle one; further discussion is deferred until after the presentation of the numerical results.

4. Numerical studies

4.1. Preliminaries

In tandem with the analytic work described in the preceding sections we have performed a Monte Carlo study of the V model. The simulations have been carried out using the ICL Distributed Array Processor (DAP) which is particularly well suited to the study of the configurational dynamics of single-bit (logical) variables of interest here (see, e.g., Bowler and Pawley 1984). We have addressed two issues. Firstly, to complement the calculations reported in § 2, we have studied the error-free-image fraction $\bar{F}_I$. Secondly we have studied the error-free-bit fraction $\bar{F}_B$ (and related quantities), with particular emphasis on the 'phase transition' which (according to § 3) it may be expected to exhibit as the storage ratio is increased.

These two quantities have a somewhat different status. The former is clearly independent of the specific way in which the configurational dynamics is defined: a nominated configuration is stable if and only if $(2V_i^{(r)} - 1) = \operatorname{sgn} H_i^{(r)}$. In contrast, the second quantity does, in principle, reflect the configurational dynamics: the fraction $\bar{F}_B$ characterises the typical configuration $\tilde{V}^{(r)}$ singled out by the configurational flow from $V^{(r)}$. We have investigated two configuration updating schemes. In the first scheme (as adopted by Hopfield (1982)) the switching of the local variables to the states favoured by the local effective field (1*b*) takes place randomly and asynchronously: one local variable is selected at a time and switched (if necessary) to the state then favoured. In this scheme the assembly energy (1*a*) is a strictly decreasing function of time. In the second scheme the state of each local variable is checked and such switching as is necessary is simultaneously implemented at each site with probability one-half (so that, within any one updating operation, on average one-half of those variables in unfavoured states will be switched). In this scheme the energy is not guaranteed to decrease at each step but will do so with a probability which differs from unity only through a small finite size effect.

4.2. The stability of nominated configurations

We have calculated the error-free image fraction $\bar{F}_I$ (1.7*a*) for assemblies with N ranging from 54 to 512, and for a range of values of the storage ratio $\alpha = n/N$. The

implied stability check is efficiently implemented in the parallel architecture of the DAP: for $N = 512$ and $n = 24$ the stability check of one complete set of nominated configurations required 1.1 s.

The results are shown in figure 3. They are consistent with those (for $N = 30$ and $N = 100$) reported by Hopfield (1982) but are considerably more precise. (The statistical uncertainties are smaller than the symbol sizes.)

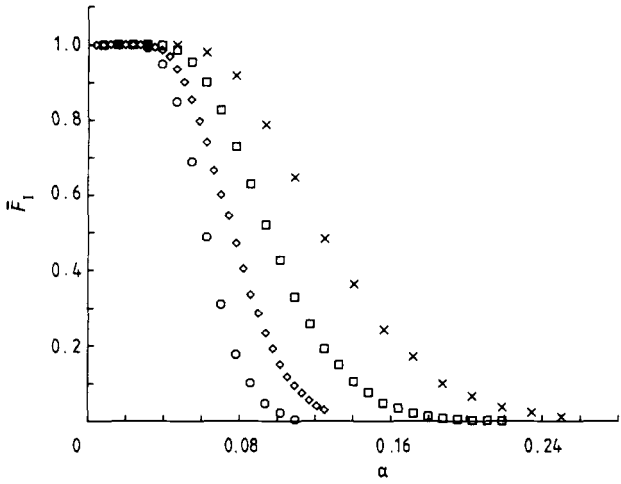

Figure 3. The error-free-image fraction $\bar{F}_1(N, \alpha)$ for the V model, according to the numerical simulations for systems of various sizes (×, $N = 64$; □, $N = 128$; ◇, $N = 256$; ◯, $N = 512$). The statistical uncertainties are smaller than the symbol sizes.

Comparison with the analytic theory developed in § 2.3 is effected in figure 4 which shows a plot of $-N^{-1} \ln \bar{F}_1$ against α. For the larger assemblies the data are consistent with the form emerging from the asymptotically exact calculation (§ 2.3). The full curve represents the function $f(\lambda = 0, \alpha)$ defined in (2.22).

4.3. Stable configurations

We now turn to a study of the stable configurations of the V model singled out by the configurational flow from nominated configurations. Our studies have been conducted on assemblies with N ranging from 128 to 4096, with particular emphasis on values of n yielding storage ratios in the vicinity of the critical value, α_c, identified in the replica-symmetric mean field theory. For each pair of values, N and n, we generated a number N_s of sets each consisting of n random N-bit configurations. The existence of large set-to-set fluctuations in the observables of interest (particularly in the vicinity of α_c) necessitated compensatingly large N_s values; for example, for $N = 1024$ and $\alpha = 0.066$ we studied $N_s = 50$ different sets. Each image in each set was used as the starting configuration for one or other of the forms of updating discussed in § 4.1: in the data presented below the random serial updating was used for all N except for $N = 1024$ for which we used the random parallel updating scheme. In each case the configurational flow was followed to completion. The flow times were found to depend sensitively upon the storage ratio (as well as the system size and the updating scheme): thus, for example, with $N = 1024$ the typical CPU time per vector (to iterate to completion) was 0.42 s for $\alpha = 0.047$ and 3.7 s for $\alpha = 0.07$. The Hamming distance D (1.6)

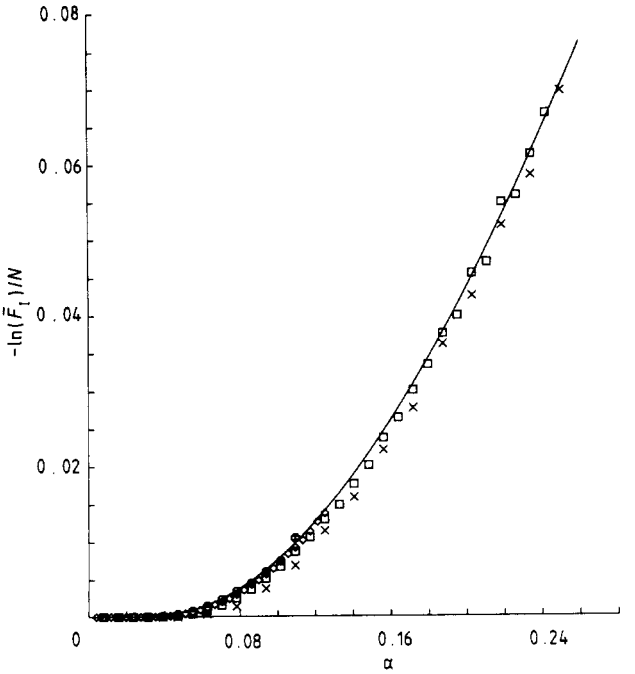

Figure 4. Comparison of the numerical simulations and the analytic theory for the error-free-image fraction. (——, analytical theory; ◯, $N = 512$; ◇, $N = 256$; □, $N = 128$; ×, $N = 64$.)

between each initial configuration $V^{(r)}$ and the corresponding stable image $\tilde{V}^{(r)}$ was determined and the $N_s n$ observations for each pair N, n were used to determine the Hamming distance distribution $p(D; \alpha, N)$.

The Hamming distance distribution typically consists of two components. One component lies close to $D = 0$ and is thus associated with stable configurations $\tilde{V}^{(r)}$ correspondingly close to their progenitors $V^{(r)}$. The other component lies close to $D = 0.5$ and is thus associated with configurations $\tilde{V}^{(r)}$ whose overlap with the starting configurations $V^{(r)}$ is correspondingly close to random. Both these statements require refinement, which we defer until the end of this section.

As noted by Amit *et al* (1985a) in the context of the S model, the evolution of $p(D; \alpha, N)$ with N shows qualitative differences according to the size of the storage ratio α. For small enough α (figure 5) the weight in the high D peak is transferred to the low D peak with increasing N; for large enough α (figure 6), in contrast, the weight shifts from the low D peak with increasing N.

The phenomenon is revealed more quantitatively and systematically in figure 7 where we show the weight W in the low D peak, as a function of $1/N$ for various values of α. (In practice we took W to measure the fraction of the distribution associated with fractional Hamming distances $D < \frac{1}{4}$; since the two peaks in the distribution are well separated the value assigned to W is rather insensitive to the details of this prescription.) The data are consistent with a bifurcation at a value $\alpha = \alpha_0$ in the vicinity of 0.068, with W assuming limiting values 0 or 1 according to the sign of $\alpha - \alpha_0$. The close correspondence between this suggested bifurcation point and the critical storage ratio α_c identified in the mean field theory is noteworthy. However, even after some 300 h of DAP time, we cannot exclude the possibility that

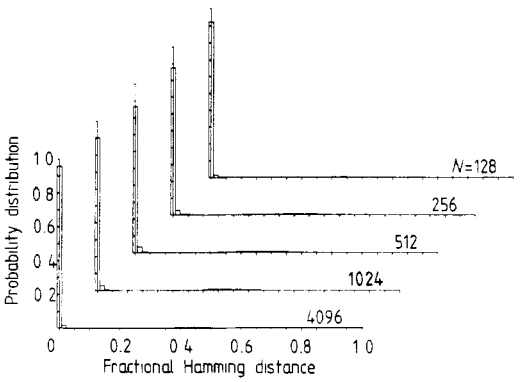

Figure 5. Histogram representing the Hamming distance distribution $p(D)$ for $\alpha = 0.0625$.

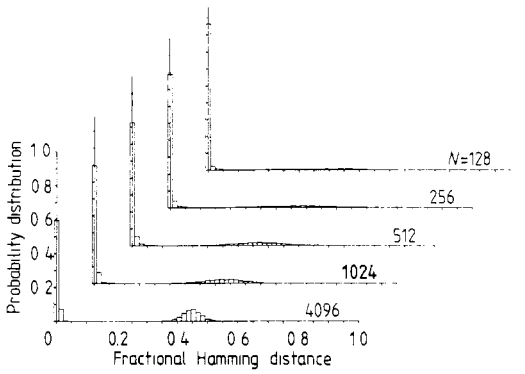

Figure 6. Histogram representing the Hamming distance distribution $p(D)$ for $\alpha = 0.0703$.

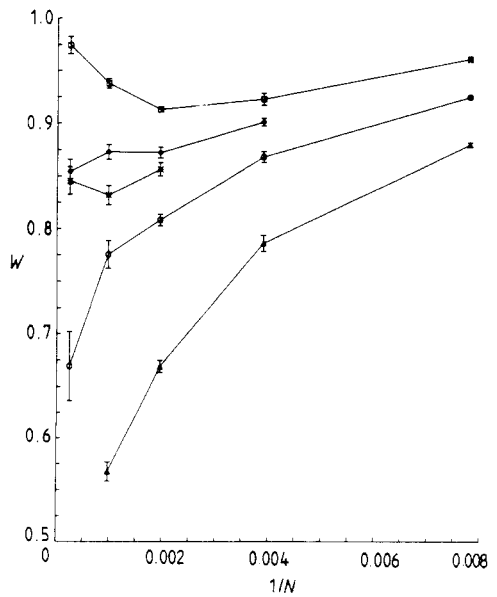

Figure 7. The weight W in the low Hamming distance peak of the distribution $p(D)$ for various α as a function of $1/N$ (□, 0.0625; ◇, 0.0664; *, 0.0683; ◯, 0.0703; ◇, 0.0781).

W remains a continuous function of α as N increases; the computational difficulty can be gauged from the large sample to sample fluctuations revealed by the histogram of peak weights of 46 individual samples for $\alpha = 0.0664$ (270 nominated configurations with 4096 nodes) shown in figure 8. In the absence of a deeper understanding of the relevant finite size effects we thus prefer not to make a more definitive claim regarding the existence and location of α_0. The assignment $\alpha_0 = 0.0731 \pm 0.005$ made in a preliminary presentation of this work (Wallace 1986) was based on an extrapolation procedure which relied much more heavily on theoretical prejudices. This value is, we now believe, too high: moreover, the quoted statistical error does not reflect the systematic uncertainties associated with the prescription for the extrapolation from the finite systems.

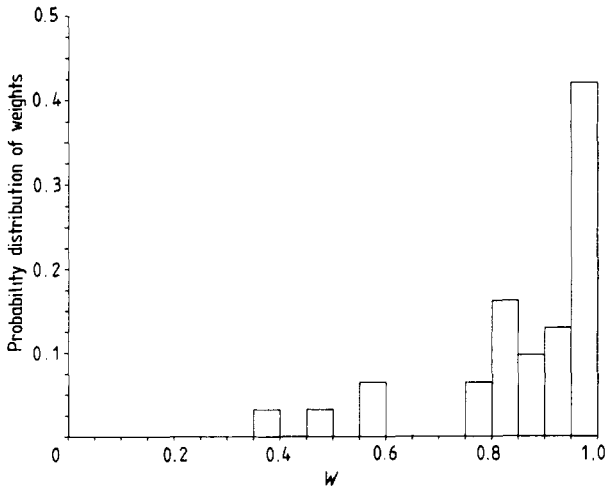

Figure 8. Distribution of peak weights, W, for individual samples with $n = 270$, $N = 4096$.

In this context we note also that, in their analysis of the S model, Amit *et al* (1985a, 1986) found that their peak weight data could be well represented by the form $W = A \exp[BN(\alpha_0 - \alpha)]$ (for $\alpha > \alpha_0$) and made this the basis of their assignment of α_0 for that model. However, as is evident from the strong curvature in the logarithmic plots shown in figure 9, our data cannot be satisfactorily represented in this way.

A further comparison between the thermodynamic and dynamic analyses is provided in figure 10. We show the error-free-bit fraction $\overline{F_B}$ determined from the simulation through (1.7*b*) as a function of α, for various N, together with the result of the mean field calculation based on (3.23). Again there is substantial support for the existence of singular behaviour at a value of α close to the predictions of the mean field theory. Quantitatively, however, it is clear that in the regime of good recall ($\alpha < \alpha_0$) the simulations exhibit a substantially greater fraction of errors (i.e. a fraction $\overline{F_B}$ differing from unity by substantially more) than is suggested by the mean field theory. In a large measure this is due to those finite size effects which are expressed in the existence of the peak in the distribution at large Hamming distance, the weight of which, we have suggested, vanishes in the thermodynamic limit (for $\alpha < \alpha_0$). Indeed, if one calculates the mean fractional Hamming distance $\bar{D}$, and thence the fraction $\overline{F_B}$, utilising *only* those configurations associated with the *low* Hamming distance peak

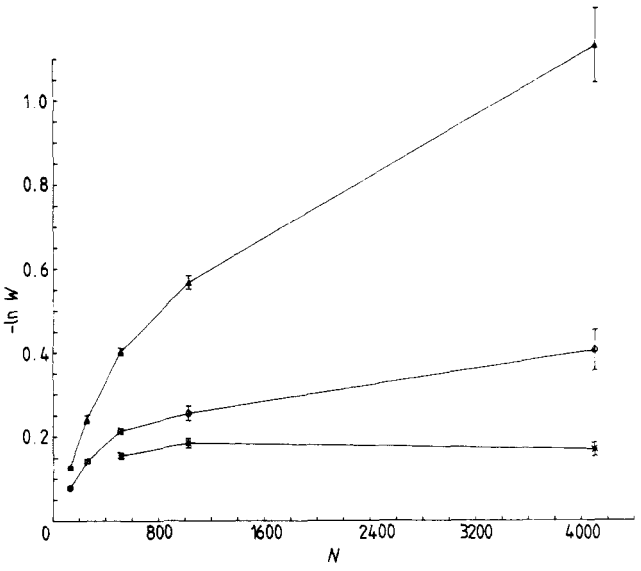

Figure 9. The logarithm of the low Hamming distance peak weight, for various α, as a function of N (*, 0.0683; ◯, 0.0703; △, 0.0781).

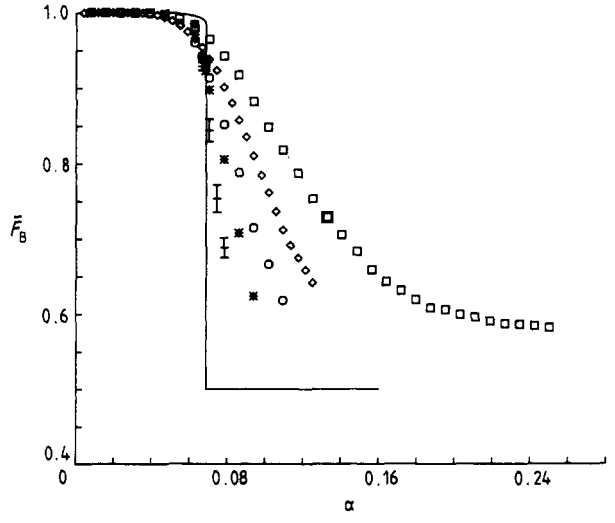

Figure 10. The error-free-bit fraction $\bar{F}_B(N, \alpha)$ determined from the mean field theory (through (3.23)) and as given by the numerical simulations (through (1.7*b*)) for various N (□, 128; ◇, 256; ◯, 512; *, 1024; +, 4096, —— mean field theory), as a function of α.

(and thus eliminating this gross finite size effect) one finds substantially better accord with the mean field prediction (cf figure 11).

Two issues merit further elaboration. Firstly we return to the question of the dependence of the results upon the details of the updating scheme. As stated, all but the $N = 1024$ data presented here were accumulated on the basis of the asynchronous Hopfield algorithm. The partially parallel algorithm is, of course, particularly well suited to the DAP and we have used it in the case of the $N = 1024$ network. The data thus generated match smoothly onto the data yielded by the asynchronous study of other-sized networks. Moreover, separate studies at selected N and n values using

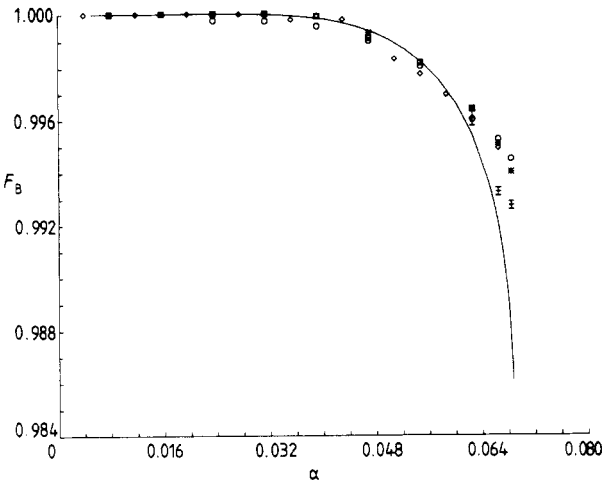

Figure 11. The error-free-bit fraction $\bar{F}_B(N, \alpha)$ obtained from the numerical simulations including *only* the states associated with the low Hamming distance peak in $p(D)$, compared with the mean field prediction (3.23) (□, $N = 128$; ◇, $N = 256$; ◯, $N = 512$; *, $N = 1024$; +, $N = 4096$, ——, mean field theory).

both schemes yield results which are quantitatively consistent with one another. We conclude that, at the level of statistical accuracy realised in the present study, any difference between the two schemes is not apparent.

Finally we return to the structure of the Hamming distance distribution. First we remark that, when viewed on a finer scale than that used in figures 5 and 6, one finds (cf figure 12) that the low Hamming distance peak is actually centred on a non-zero value of D. Secondly (cf figure 13) we note that the high Hamming distance peak is centred not on $D = \frac{1}{2}$ but on a value of D (weakly dependent upon α) somewhat lower than this. We shall take up these issues in the concluding section to which we now turn.

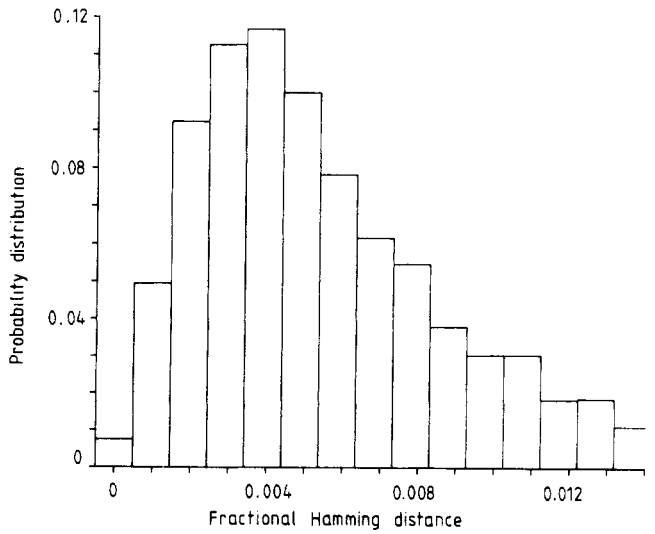

Figure 12. The low D structure of the Hamming distance distribution for $N = 4096$, $\alpha = 0.0683$.

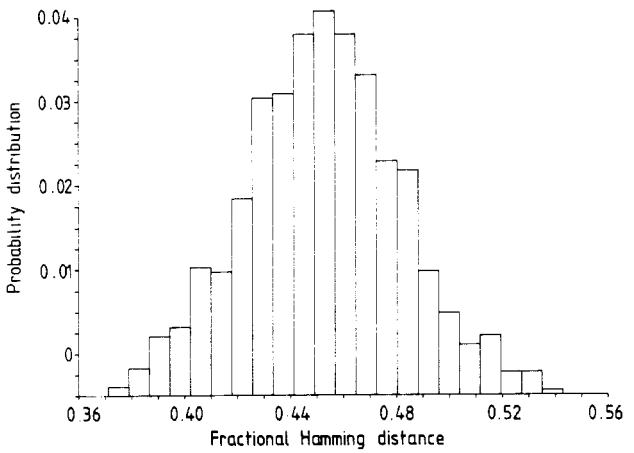

Figure 13. The high D structure of the Hamming distance distribution for $N = 4096$, $\alpha = 0.0703$.

5. Discussion and conclusions

We must now address the problems inherent in the comparison of the numerical work described in the previous section with the analytic studies of § 3. The numerical work faithfully realises the model of interest. The analytic work does so only to the extent that we may trust three key assumptions which we now identify.

Firstly, and most obviously, the steepest descent arguments effectively assume the thermodynamic limit ($N, n \to \infty$ with $\alpha = n/N$ fixed). Secondly, the replica-symmetric ansatz (3.12) assumes that there is only one equilibrium state with a macroscopic overlap with a given nominated configuration. Thirdly, in interpreting the results of the mean field theory we effectively assume that the flow from the nominated configuration will always terminate in that thermodynamic state with which it has macroscopic overlap, as long as it exists (i.e. for $\alpha < \alpha_c$) and will otherwise terminate in the thermodynamic (spin glass) state which has no memory of the initial configuration.

Within these approximations the Hamming distance distribution studies in the preceding section may be expected to consist of a δ function centred on the mean Hamming distance, whose location undergoes a discontinuous change at the value $\alpha_c \simeq 0.069$ determined by the replica-symmetric theory.

Of the three approximations it is the first (the presumed thermodynamic limit) which, we believe, accounts for the bulk of the mismatch between the analytic and the numerical work. Although we have not attempted to refine the arguments of § 3 to incorporate finite size effects in a formal way, intuitively their consequences seem clear: for finite n and N the fluctuations in the Hamming distance will no longer be small on the scale of its mean and the sharp (δ-function) structure of the distribution will be replaced by a structure whose width, one may anticipate, will be of order $1/\sqrt{N}$. Moreover, one would expect the sharp transition at α_c to be replaced by a regime of phase coexistence whose α width vanishes in the thermodynamic limit. These expectations are in qualitative accord with the observations recorded in the preceding section (cf especially figures 5, 6 and 7).

We now consider the second approximation. As noted in § 3 the replica-symmetric solution to the (zero-temperature) mean field equations is actually unstable. It seems

unlikely that the quantities studied in § 4 will be particularly sensitive to this instability. Nevertheless the effect should, in principle, manifest itself in an upward shift of α_c with respect to the replica-symmetric prediction. Indeed, in a recent paper Crisanti *et al* (1986) have estimated the effect for the S model and find that the implied shift (from $\alpha_c \simeq 0.138$ to $\alpha_c \simeq 0.144$) leads to better agreement with the transition point suggested by their numerical studies (although the shift is actually smaller than the uncertainty they associate with the numerical prediction). They also find that the replica symmetry breaking leads to a decrease in the limiting ($\alpha \to \alpha_c^-$) value of the mean Hamming distance (an increase in the error-free-bit fraction). In our case it is clear that the uncertainties in locating the dynamic instability point of the model, α_0, are likely to be large compared with the likely shift in α_c associated with replica symmetry breaking. There is some suggestion (figure 11) that, close to the instability, the error-free-bit fraction is indeed higher than the replica-symmetric prediction, but residual finite size effects make it difficult to interpret this difference.

Finally we turn to the third key assumption. The central issue here is the extent to which the consequences of an essentially non-ergodic dynamics can be captured by equilibrium statistical mechanics. In fact it is clear that the thermodynamic (minimum energy) states captured by the mean-field calculation do not exhaust the spectrum of dynamic-equilibrium states which can represent the endpoints of configuration flow. Analysis shows (Gardner 1987a) that, clustered around each of the two thermodynamic states envisaged in the mean-field calculations of § 3 (the 'correlated' state with mean Hamming distance close to zero and the 'uncorrelated', spin glass, state with mean Hamming distance 0.5) there exists a band of other states which, though higher in energy, are stable against one-spin-flip dynamics. As in the case of the spin glass (Bray and Moore 1980) the number of states in each of these bands is exponentially large in N and so it is likely that it is amongst these states (rather than the thermodynamic states) that the configurational flow will terminate. The quantitative implications of this picture remain to be determined. However, two qualitative consequences may be usefully identified.

Firstly it is possible that the transition point α_0 (the bifurcation point in the configurational flow) may not coincide with the transition point α_c (the limit of stability of the correlated equilibrium state or its replica symmetry broken counterpart). In particular α_0 might exceed α_c if the non-equilibrium states making up the correlated band were to persist above α_c where the thermodynamic correlation state no longer exists. (It is also possible that $\alpha_0 < \alpha_c$ since the existence of the correlated band does not guarantee that it will capture the configurational flow.) As with the effects of replica symmetry breaking the data presented in the preceding section only allow us to draw a tentative conclusion: any difference between α_0 and α_c seems likely to be small.

Secondly it is clear that the Hamming distance distribution must differ from that implied by the thermodynamic argument. Specifically, although we anticipate that the distribution will remain sharp (in the thermodynamic limit) its peaks (associated with the two bands of states) will not in general coincide with the thermodynamic equilibrium states. While our data provide no unambiguous evidence for a shift in the peak associated with the correlated states, there is clear evidence (figure 13) that the peak associated with the band of nominally uncorrelated states is shifted below the thermodynamic prediction ($D = \frac{1}{2}$) implying a remanent overlap between initial and final states (Kinzel 1986) even above α_0.

It is clear that there remain many questions meriting further study. First, detailed study of the size of the basins of attraction associated with the nominated configurations,

and of the time taken to iterate to them, would give a more explicit indication of how well associative memory works in the Hopfield model. (Some aspects of the associative functioning of the S model have been explored by Kinzel (1985a).) Second, there are interesting questions to be resolved regarding the possibly hierarchical ('ultrametric') structure of the equilibrium states of the model (Mézard *et al* 1984). Third, it would be interesting to study quantities more sensitive to the existence of replica symmetry breaking.

There are also many proposed generalisations of the model (some of which actually predate it: see Cohen and Grossberg (1983)). It is possible to increase the storage capacity by utilising multiple-neuron interactions (Gardner 1987, Psaltis and Park 1986, Chen *et al* 1986, Maxwell *et al* 1986). It may also be possible to produce a dramatic improvement in the storage capacity by imposing a hierarchical organisation upon the configurations to be stored (Dotsenko 1985, Parga and Virasoro 1987). The role of time-dependent neuron interactions in the learning process has also attracted much recent attention (Toulouse *et al* 1986, Personnaz *et al* 1986, Mézard *et al* 1986, Parisi 1986). Finally, there are extensions of the perceptron learning algorithm (Minsky and Papert 1969), which are based on iterative application of the storage prescription, and important generalisations with hidden neurons which should capture more complicated correlations between patterns (Ackley *et al* 1985, Rumelhart *et al* 1985).

These and related issues are the subject of continuing study.

Acknowledgments

We thank David Bounds and David Willshaw for stimulating discussions. This work is supported in part by SERC grants NG14840, NG15908 and GRC80431.

References

Ackley D H, Hinton G E and Sejnowski T J 1985 *Cognitive Sci.* **9** 147
Amit D J, Gutfreund H and Sompolinsky H 1985a *Phys. Rev. Lett.* **55** 1530
—— 1985b *Phys. Rev.* A **32** 1007
—— 1986 *Europhys. Lett.* **2** 337
Bowler K C and Pawley G S 1984 *Proc. IEEE* **72** 42
Bray A J and Moore M A 1980 *J. Phys. C: Solid State Phys.* **13** L469
Chen H H, Lee Y C, Sun G Z, Lee H Y, Maxwell T and Giles C L 1986 *Proc. Conf. on Neural Networks For Computing, Snowbird, UT* ed J S Denker (New York: AIP) p 86
Cohen M A and Grossberg S 1983 *IEEE Trans. Systems, Man and Cybernetics* **SMC-13** 815
Crisanti A, Amit D J and Gutfreund H 1986 *Preprint* Racah Institute of Physics
De Dominicis C, Gabay M, Garel T and Orland H 1981 *J. Physique* **41** 923
Dotsenko V 1985 *J. Phys. C: Solid State Phys.* **18** L1017
Gardner E 1986 *J. Phys. A: Math. Gen* **19** L1047
—— 1987 *J. Phys. A: Math. Gen.* **20** 3453
Hebb D O 1949 *The Organisation of Behaviour* (New York: Wiley)
Hinton G E and Anderson J A 1981 *Parallel Models of Associate Memory* (Hillsdale, NJ: Lawrence Erlbaum)
Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554
Kinzel W 1985 *Z. Phys.* B **60** 205
—— 1986 *Phys. Rev.* B **33** 5086
Little W A 1974 *Math. Biosci.* **19** 101
Maxwell T, Giles C L, Lee Y C and Chen H H 1986 *Proc. Conf. on Neural Networks For Computing, Snowbird, UT* ed J S Denker (New York: AIP) p 299

McCulloch W S and Pitts W A 1943 *Bull. Math. Biophys.* **5** 115
Mézard M, Parisi G, Sourlas N, Toulouse G and Virasoro M 1984a *Phys. Rev. Lett.* **52** 1146
—— 1984b *J. Physique* **45** 843
Mézard M, Nadal J P and Toulouse G 1986 *J. Physique* **47** 1457
Minsky M L and Papert S 1969 *Perceptrons* (Cambridge, MA: MIT)
Parga N and Virasoro M 1987 *J. Physique* to be published
Parisi G 1986 *J. Phys. A: Math. Gen.* **19** L617
Personnaz L, Guyon I, Dreyfus G and Toulouse G 1986 *J. Stat. Phys.* **43** 411
Psaltis D and Park C H 1986 *Proc. Conf. on Neural Networks For Computing, Snowbird, UT* ed J S Denker (New York: AIP) p 370
Rumelhart D E, Hinton G E and Williams R J 1985 *Parallel Distributed Processing Explorations in the Microstructure of Cognition* vol 1 *Foundations* (Cambridge, MA: Bradford Books/MIT) to be published
Sherrington D and Kirkpatrick S 1975 *Phys. Rev. Lett.* **35** 1792
—— 1978 *Phys. Rev.* B **17** 4384
Toulouse G, Dehane S and Changeux J P 1986 *Proc. Natl Acad. Sci. USA* **83** 1695
Wallace D J 1985 *Advances in Lattice Gauge Theory* ed D W Duke and J F Owens (Philadelphia: World Scientific) p 326
——1986 *Lattice Gauge Theory—A Challenge in Large Scale Computing* ed B Bünk and K H Mutter (New York: Plenum) p 313
Weisbuch G and Fogelman-Soulie F 1985 *J. Physique Lett.* **46** L623