J. Phys. A: Math. Gen. 27 (1994) 2351–2365. Printed in the UK

Partial annealing and overfrustration in disordered systems

Vik Dotsenko†, S Franz and M Mézard

Laboratoire de Physique Théorique de l'Ecole Normale Supérieure, Unité propre du CNRS, associée à l'Ecole Normale Supérieure et à l'Université de Paris Sud, 24 rue Lhomond, 75231 Paris Cedex 05, France

Received 16 November 1993

Abstract. We study disordered systems with the replica method keeping the number of replicas finite and negative. This is shown to bias the distribution of samples towards overfrustrated ones. General results on the thermodynamics of such a system is presented. The physical situation described by this finite-n approach is one where the usually quenched variables evolve on long timescales, their evolution being driven by the quasi-equilibrium correlations of the thermalized variables. In the case of neural networks this amounts to a coupled dynamics of neurons (on fast timescales) and synapses (on longer timescales). The storage capacity of the Hopfield model is shown to be substantially increased by these coupled dynamics.

1. Introduction

An essential feature of the physics of disordered systems is the existence of a wide separation of timescales between 'annealed variables' which evolve and eventually reach equilibrium on experimental timescales, and 'quenched varables' that can be considered frozen and highly out of equilibrium in experiments. For example, in metallic spin glasses one studies the evolution of the magnetic moment of impurities which have random position in a non-magnetic substrate. This induces a probability distribution on the values of the interactions, which can be identified as 'quenched variables'. The specific character of the interactions, which take both ferromagnetic and antiferromagnetic values leads to *frustration*. Frustation and quenched disorder are commonly thought of as the necessary ingredients in order to have complex spin-glass-like phenomena, such as ergodicity breaking and ageing. It is clear that on hypothetical timescales such that slow and fast variables equilibrate, frustation and all complex phenomena would disappear. This situation, in contrast to the quenched case is often referred to as 'annealed'.

In this paper we consider, for the specific case of SK spin glasses [1] and Hopfield neural networks [2], a situation somewhat intermediate between the completely quenched and the completely annealed cases. The 'slow' degrees of freedom—the interaction between spins—will be allowed to vary, evolving towards a partial equilibrium with fast degrees of freedom. It is quite natural, given the wide separation of timescales, to think of the dynamics of the interactions as a heat bath process driven by the free energy of the spin system. We call partial equilibrium a situation in which the slow as well as the fast variables are at thermal equilibrium, but have different temperatures. We concentrate here on the case in which the temperature of the slow variables is negative. The positive temperature case has been studied recently in papers by Penney *et al* [3]. For positive temperatures, the dynamics

† On leave from Landau Institute for Theoretical Physics, Russian Academy of Sciences, Moscow.

0305-4470/94/072351+15$19.50

of the interactions is such as to progressively reduce the frustration of the system. On the contrary, if the temperature is negative the system evolves towards configurations of higher and higher frustration. The problem can be analysed with the replica method, where the 'number of replicas' n, which goes to zero in the usual quenched case, has in this context the interpretation of the relative temperature between spins and couplings, $n = T_{\text{spins}}/T_{\text{couplings}}$. Early work on the replica method with non-zero n can be found in [4, 5]

In the case of SK spin glasses we find that for zero magnetic field the overfrustation does not macroscopically change the free energy of the system. In the Hopfield model, in which a coupling dynamics with negative temperature is reminiscent of the 'unlearning algorithm', we find a dramatic effect for the retrieval phase. The limit of capacity is increased from the AGS value $\alpha_c = 0.145$ [6] to $\alpha_c = 1$.

The general formalism is discussed in section 2. In section 3 we concentrate on the SK spin glass, both for zero and non-zero magnetic field. The Hopfield model, its retrieval and spin-glass phases, are discussed in section 4. Finally we draw brief conclusions.

2. Partial annealing and replicas

Let us consider a general spin system described by some Hamiltonian $H[J;\sigma] = -\sum_{i<j} J_{ij}\sigma_i\sigma_j - h\sum_i \sigma_i$, which depends on the spin variables $\{\sigma_i \; i = 1,\ldots,N\}$ and the spin–spin interactions J_{ij}. In the usual spin-glass problem, the interactions J_{ij} are quenched. The free energy for a given realization of the J_{ij}'s,

$$F[J] = -\frac{1}{\beta}\ln Z[J] \tag{1}$$

where

$$Z[J] = \sum_{\sigma} \exp(-\beta H[J;\sigma]) \tag{2}$$

is often known to be self-averaging. This means that $F[J]/N$ has a limit for large N for almost all realizations, but of course different realizations J_{ij} have non-extensive differences in their free-energies.

Now, let us assume that the spin–spin interactions are not perfectly quenched, so that they can also change their values, but the characteristic timescale of their changes is much larger than the timescale at which the spin degrees of freedom reach thermal equilibrium. We shall consider here the case in which the free energy (1) still makes sense, and becomes the energy function (the Hamiltonian) for the J_{ij}'s degrees of freedom. As we shall discuss later, this situation corresponds to the existence of a long-time dynamics of the couplings, depending on the correlation functions of the underlying spin system.

Besides this, the interactions J_{ij} could be of different kind (e.g. real, binary, etc). In the quenched case the nature of the J_{ij}'s is defined by some statistical distribution function $P[J]$. In the case of the partial annealing this function $P[J]$ can be interpreted as an internal potential for the J_{ij}'s.

Let us now assume that the spin and the interaction degrees of freedom *are not mutually equilibrated*, so that the interaction degrees of freedom have their own temperature T', which is different from that of the spin degrees of freedom T. In this case the total partition function of the system is

$$\mathcal{Z} = \int DJ\,P[J]\exp(-\beta' F[J]) = \int DJ\,P[J]\exp\left(\frac{\beta'}{\beta}\ln Z[J]\right) = \int DJ\,P[J](Z[J])^n \tag{3}$$

with the definition $n = T/T'$. Correspondingly, the total free energy of the system is

$$\mathcal{F} = -T' \ln\{\langle\langle (Z[J])^n \rangle\rangle\} \tag{4}$$

where

$$\langle\langle (Z[J])^n \rangle\rangle \equiv \int DJ \; P[J] \; (Z[J])^n \,. \tag{5}$$

We can evaluate $\mathcal{F}$ in (4) by means of the well known replica formalism, in which the 'number of replicas' $n = T/T'$, initially integer valued, has to be continued to finite (arbitrary real) values. A similar approach has recently been developed by Penney *et al* [3], who have also noticed this interpretation of the free energy for finite n. Our approach follows a different route and in particular we shall concentrate mainly on the case where n is negative.

To obtain the physical (self-averaging) free energy in the replica approach in the case of the quenched random J_{ij}'s one takes the limit $n \to 0$. From the point of view of the partial annealing considered here, this situation corresponds to the limit of infinite temperature T' in the system of J_{ij}'s. This is natural in the sense that in this case the thermodynamics of the spin degrees of freedom produces no effect on the distribution of the spin–spin interactions.

In the case where the spin and the interaction degrees of freedom are thermally equilibrated $T' = T$ $(n = 1)$, we get the case of purely annealed disorder, whatever the difference of the characteristic timescales of the J_{ij}'s and the spins. This is also natural because the thermodynamic description formally corresponds to the infinite times, and the characteristic timescales of the dynamics of the internal degrees of freedom become of no importance.

If $n \neq 0$ and $n \neq 1$, we are in the situation which we call partial annealing. This situation may not be as unusual as it looks at first sight. It describes the case of a stochastic dynamics of the couplings which could be, for instance, (in the case of continuous couplings) a Langevin dynamics:

$$\frac{\mathrm{d}J_{ij}}{\mathrm{d}t} = -\frac{\mathrm{d}}{\mathrm{d}J_{ij}} \left(\frac{-\ln P[J]}{\beta'} + F[J] \right) + \eta_{ij}(t) \,. \tag{6}$$

In the following we shall keep to the usual terminology and call a given realization of the interactions J_{ij} a sample. Of course one should not be misled by this terminology which is more adapted to the case of quenched systems: the long-time dynamics of the interactions now corresponds to a change of samples. We shall call 'global configuration' a given set of interactions and spins. $F[J]$ defined in (1) will be called the free energy of the sample $[J]$, and $\mathcal{F}$ will be called the total free energy.

In what follows some concrete systems will be considered. In particular, we are going to study the spin glasses and the neural networks in which the parameter n is *negative*. This just corresponds to the situation where the dynamics in the system of interactions drives the system towards some samples of *high* free energy. In the language of spin glasses it means that the J_{ij}'s are evolving in a direction such that the degree of frustration in the system is increasing (unlike the annealed disorder which is just trying to remove frustrations).

In the Hopfield model of auto-associative memory [2], introducing a partial annealing means that the stored patterns become (slow) dynamical variables. In conventional models of auto-associative memory the patterns, to be associated to the memory states, are quenched. In the present model the 'patterns', to be thought as the low-free-energy states, evolve with time, and as we will see in section 4, they eventually undergo diffusion in a certain space. At first sight this may look astonishing. Nevertheless, we believe that it does make sense, and in particular if the temperature in the system of the patterns is taken to be negative, one

finds that the 'patterns' move to become as orthogonal as possible. The 'patterns' can be interpreted as an internal representation of some information, which adapts itself towards internal representations which have as few correlations as possible. This will be shown to produce a substantial increase of the storage capacity up to $\alpha_c = 1$. The case of negative n presents some similarities with the unlearning algorithm [7], which is known to increase the storage capacity due to the reduction of the noisy interference effects among the patterns.

3. Spin glasses

Consider the Sherrington and Kirkpatrick (SK) model of spin glasses with long-range interactions [1]:

$$H = -\tfrac{1}{2}\sum_{i\neq j}^{N} J_{ij}\sigma_i\sigma_j - h\sum_{i}^{N}\sigma_i . \tag{7}$$

This system consists of N Ising spins $\{\sigma_i\}$ $(i = 1, 2, \ldots, N)$ taking values ± 1 which are placed in the vertices of some lattice, labelled by the index i. The spin–spin interactions J_{ij} in this system are random variables which are independent for each pair of sites (i, j), and their *a priori* distribution is Gaussian:

$$P[J_{ij}] = \prod_{i<j}\left[\sqrt{\frac{N}{2\pi}}\exp(-\tfrac{1}{2}J_{ij}^2 N)\right]. \tag{8}$$

For the case of quenched J_{ij}'s this model has been studied in detail (see e.g. [10]). In an attempt to get a better understanding of the analytic continuation to $n \to 0$, it has also been studied for small positive values of the replica parameter n [5]. Here we are going to follow the same traditional replica approach but keeping the replica parameter n finite and negative. As discussed in the introduction this amounts to the hypothesis that the slow dynamics of the couplings biases the distribution of samples towards overfrustrated ones.

The replica partition function of (5) is:

$$\langle\langle (Z[J])^n \rangle\rangle = \sum_{\sigma_i^a}\int DJ_{ij}\exp\left\{\beta\sum_{a=1}^{n}\sum_{i<j}^{N} J_{ij}\sigma_i^a\sigma_j^a + \beta h\sum_{a=1}^{n}\sum_{i}^{N}\sigma_i^a - \frac{1}{2}\sum_{i<j}J_{ij}^2 N\right\} \tag{9}$$

(here and everywhere in what follows all kinds of pre-exponential factors are omitted).

Let us note that the scaling of the J_{ij} in the *a priori* distribution (8): $J_{ij} \sim 1/\sqrt{N}$ is the usual one needed in the spin system to have energy and entropy of the same order $O(N)$ when $N \to \infty$. The distribution of interactions (8) contains an overall contribution to the global entropy

$$S_0 = \frac{N(N-1)}{2}\left[\frac{1}{2}\ln\sqrt{\frac{N}{2\pi}}\right] \tag{10}$$

which always gives the dominant contribution to the global free energy. This means that the corrections to the Gaussian distribution will be very small in all situations. Of course these corrections can have a dramatic effect on the spin system. For example, in the presence of a non-zero magnetic field h in the spin system, the J_{ij} acquire a non-zero mean value of order $1/N$, very small with respect to the dominant contribution which is of order $1/\sqrt{N}$, but which can cause non-zero magnetization. In what follows we will be concerned with the terms of order N in the free energy, and with the small correction to the J_{ij} statistics. Therefore we choose a normalization in (9) such that S_0 is subtracted from the free energy.

Standard calculations (see e.g. [10]) lead to the following form of the partition function:

$$\langle\langle (Z[J])^n \rangle\rangle = \int D\hat{Q} \exp(-\beta n N f[\hat{Q}]) \tag{11}$$

where

$$f[\hat{Q}] = -\tfrac{1}{4}\beta + \frac{\beta}{2n}\sum_{a<b}^{n} Q_{ab}^2 - \frac{1}{\beta n}\ln\left[\sum_{\sigma_a} \exp\left(\beta^2 \sum_{a<b}^{n} Q_{ab}\sigma_a\sigma_b + \beta h \sum_{a=1}^{n}\sigma_a\right)\right] \tag{12}$$

is the replica free energy and Q_{ab} is the matrix

$$Q_{ab} = \frac{1}{N}\sum_{i}^{N}\langle \sigma_i^a \sigma_i^b \rangle \qquad a < b\,. \tag{13}$$

The matrix Q_{ab} can be interpreted in the usual way [11], and used to reconstruct, for example, the probability distribution of the overlap between two real replicas of the spin system with the same J_{ij}'s, $P(q)$. Of course $P(q)$ becomes the distribution of overlaps of samples which are chosen with the probability distribution $P[J]Z[J]^n$, which means that these samples may be quite different from the ones considered in the quenched case where the distribution is $P[J]$. Moreover, Q_{ab} also admits a natural interpretation in terms of the statistics of the J_{ij}'s. Let us consider the sum of the values of the frustration on all the loops of order k ($k > 2$):

$$\left\langle \sum_{\substack{i_1,\ldots,i_k \\ \text{all different}}} J_{i_1 i_2} J_{i_2 i_3} \cdots J_{i_k i_1} \right\rangle \equiv \langle \mathrm{Tr}' J^k \rangle\,. \tag{14}$$

Observing that this is

$$\sum_{\substack{i_1,\ldots,i_k \\ \text{all different}}} \int \prod_{i<j} \mathrm{d}J_{ij} \frac{(-1)^k}{N^k} \left(\frac{\partial}{\partial J_{i_1 i_2}} \cdots \frac{\partial}{\partial J_{i_k i_1}} \exp(-\sum_{i<j} N J_{ij}^2/2) \right) Z^n(J) \tag{15}$$

and integrating by parts, we obtain

$$\langle \mathrm{Tr}' J^k \rangle = \beta^k \sum_{a_1,\cdots,a_k} \frac{1}{N^k} \sum_{\substack{i_1,\ldots,i_k \\ \text{all different}}} \langle \sigma_{i_1}^{a_1} \sigma_{i_2}^{a_1} \sigma_{i_2}^{a_2} \sigma_{i_3}^{a_2} \cdots \sigma_{i_k}^{a_k} \sigma_{i_1}^{a_k} \rangle \tag{16}$$

which is, to leading order in N,

$$\langle \mathrm{Tr}' J^k \rangle = \beta^k \mathrm{Tr}(I + Q)^k\,. \tag{17}$$

One should not be confused by the notations in this formula: the Tr′ is a sum over all distinct lattice sites defined in (14), while the Tr in the right-hand side is a sum over the n replica indices. The presence of the identity matrix results from the convention $Q_{aa} = 0$. If one includes in the sum (14) the terms with equal indices i, other contributions may be present, e.g. one sees that

$$\left\langle \sum_{i,j} J_{ij}^2 \right\rangle = \langle \mathrm{Tr} J^2 \rangle = N + \beta^2 \mathrm{Tr}(I + Q)^2\,. \tag{18}$$

The first term in the RHS of (18) is the usual main contribution coming from the $P[J]$, the second is the (very small) correction due to the coupling with the spin system. In the quenched case the second term is zero because it involves a sum over n ($\to 0$) replica indices.

In order to perform the analytic continuation to non-integer n, we make Parisi's ansatz for the matrix Q_{ab} [8, 10]. For generic n the matrix is parametrized in terms of a function

$q(x)$ in the interval $x \in [n, 1]$ ($x \in [1, n]$ if $n > 1$). Standard calculations [11] show that $P(q) = (1/(1-n))\mathrm{d}x/\mathrm{d}q$ for generic n.

In order to compute the global free energy we need to generalize the usual algebra of matrices invented by Parisi to the case where n is not zero. Let us remember that the linear space of Parisi matrices, when completed with the identity $I_{ab} = \delta_{ab}$, is closed with respect to the matrix product $(QP)_{ab} = \sum_c Q_{ac}P_{cb}$ and the Hadamard product $(Q \cdot P)_{ab} = Q_{ab}P_{ab}$, operation by means of which it is possible to build many polynomials which are invariant by permutations of replica indices†.

A generic matrix A in this space is parametrized by a diagonal element $\tilde{a}$ and a function $a(x)$. The linear invariants are $\mathrm{Tr}A = n\tilde{a}$ and $\sum_{ab} A_{ab} = -n\int_n^1 \mathrm{d}x\, q(x)$. Let A and B be two Parisi matrices parametrized respectively by $(\tilde{a}, a(x))$ and $(\tilde{b}, b(x))$. For finite n the two products take the following forms:

$$A \cdot B \to (\tilde{a}\tilde{b}, a(x)b(x)) \tag{19}$$

and $AB \to (\tilde{c}, c(x))$, with

$$\begin{aligned} &\tilde{c} = \tilde{a}\tilde{b} - \langle ab\rangle \\ &c(x) = -na(x)b(x) + (\tilde{a} - \langle a\rangle)b(x) + (\tilde{b} - \langle b\rangle)a(x) \\ &\qquad - \int_n^x \mathrm{d}y(a(x) - a(y))(b(x) - b(y)) \end{aligned} \tag{20}$$

where

$$\langle a\rangle = \int_n^1 \mathrm{d}x\, a(x)\,. \tag{21}$$

For the eigenvalues of a Parisi matrix A and their multiplicities one finds

$$\lambda_0 = \tilde{a} - \langle a\rangle \quad \text{with multiplicity 1} \tag{22}$$

$$\lambda(x) = \tilde{a} - xa(x) - \int_x^1 \mathrm{d}y\, q(y) \quad \text{with multiplicity } -\frac{n\,\mathrm{d}x}{x^2} \qquad x \in [n, 1]\,. \tag{23}$$

Therefore the frustration loops (14), (17) take the form

$$\beta^k\left((1 - \langle q\rangle)^k - n\int_n^1 \frac{\mathrm{d}x}{x^2}\left[1 - xq(x) - \int_x^1 \mathrm{d}y\, q(y)\right]^k\right). \tag{24}$$

Before giving a general statement about the behaviour of the replica symmetry breaking (RSB) solution for Q_{ab} at arbitrary values of the temperature and (negative) n, consider first, just for illustration, the situation near the critical temperature $T_c = 1$ for small values of n and external magnetic field h. Expanding the free energy (12) in powers of Q_{ab} one gets

$$f[Q] = -\tau\frac{1}{2n}\mathrm{Tr}(Q)^2 - \frac{1}{6n}\mathrm{Tr}(Q)^3 - \frac{1}{12n}\sum_{a\neq b} Q_{ab}^4 - \frac{1}{2n}h^2\sum_{a,b} Q_{ab} \tag{25}$$

where $\tau = (1 - T) \ll 1$.

Inserting the parametrization $Q \to (0, q(x))$ and using the rules (20) one easily gets

$$f[q(x)] = \tfrac{1}{2}\int_n^1 \mathrm{d}x\left[\tau q^2(x) - \tfrac{1}{3}xq^3(x) - q(x)\int_n^x \mathrm{d}y\, q^2(y) + \tfrac{1}{6}q^4(x) + h^2 q(x)\right]. \tag{26}$$

† An example of an invariant which is not in this class is $\sum_{a_1,a_2,a_3,a_4} Q_{a_1,a_2}Q_{a_1,a_3}Q_{a_1,a_4}Q_{a_2,a_3}Q_{a_2,a_4}Q_{a_3,a_4}$. Such invariants can also be computed at finite n, but they do not derive from the rules (20). Hereafter we shall mostly keep to the situations where such invariants do not appear, unless otherwise stated.

Variation of this expression with respect to the function $q(x)$ gives the following saddle-point equation:

$$2\tau q(x) - xq^2(x) - 2q(x)\int_x^1 dy\, q(y) - \int_n^x dy\, q^2(y) + \tfrac{2}{3}q^3(x) + h^2 = 0. \tag{27}$$

Before solving (27) let us note that to order zero in τ the values of the 'frustration loops' (24) are just given by $\mathrm{Tr}I = n$ for all $k << 1/\sqrt{\tau}$.

The solution of (27) is similar to that of the case $n = 0$. By differentiating (27) with respect to x, one finds that the only continuous solution is

$$q(x) = \begin{cases} q_0 & n \leqslant x \leqslant x_0 \\ \frac{1}{2}x & x_0 \leqslant x \leqslant x_1 \\ q_1 & x_1 \leqslant x \leqslant 1 \end{cases} \tag{28}$$

where

$$x_1 = 2q_1 \qquad x_0 = 2q_0 \tag{29}$$

and the values of q_0 and q_1 are defined by the equations

$$\tau - q_1 + q_1^2 = 0 \qquad \tfrac{4}{3}q_0^3 - nq_0^2 - h^2 = 0. \tag{30}$$

Let us consider separately the two cases $h = 0$ and $h \neq 0$.

(i) $h = 0$.

If n is positive, the solution of (30) (to leading order in τ and n) is

$$q_1 \simeq \tau \qquad q_0 = \tfrac{3}{4}n \tag{31}$$

and, correspondingly, $x_1 = 2\tau$ and $x_0 = \frac{3}{2}n$. The solution for $q(x)$ becomes replica symmetric if $q_1 = q_0$. This gives the critical temperature: $\tau(n) = \frac{3}{4}n$, as derived previously by Kondor [5].

If n is negative, the solution is:

$$q_1 \simeq \tau \qquad q_0 = 0 \tag{32}$$

and the critical temperature is always $\tau(n) = 0$. The structure of these solutions is shown in figures 1(*a*) and (*b*). The free energy F is independent of n for negative n and takes the same value as for $n = 0$. We shall see hereafter that this is a general situation.

(ii) $h \neq 0$.

In this case one still gets $q_1 \simeq \tau$. For $h \neq 0$ the equation for q_0 (equation (30)) always has a positive non-zero solution. In particular, if n is negative: $q_0 \simeq h/\sqrt{|n|}$, if $h \ll (|n|)^{3/2}$; and $q_0 \simeq h^{2/3}$, if $h \gg (|n|)^{3/2}$.

In the space (τ, h, n) we find RSB below the surface defined by the equation ($q_1 = q_0$)

$$\tfrac{4}{3}\tau^3 - n\tau^2 = h^2. \tag{33}$$

In general, for arbitrary values of T, n and h, the equation for the transition surface can be derived easily

$$T^2 = \frac{\langle\langle(\cosh\beta(\sqrt{q}z + h))^{n-4}\rangle\rangle}{\langle\langle(\cosh\beta(\sqrt{q}z + h))^{n}\rangle\rangle} \tag{34}$$

(here $\langle\langle\cdots\rangle\rangle$ means Gaussian averaging over the variable z with zero mean and unit variance). For $h = 0$ this equation coincides with that obtained in [5]. At $T \to 0$ and $h = 0$ one gets: $n(T) \simeq T\sqrt{2\ln(1/T)}$. Note that when replica symmetry holds, (24) gives

$$\mathrm{Tr}^l J^k = (1 - q)^k n. \tag{35}$$

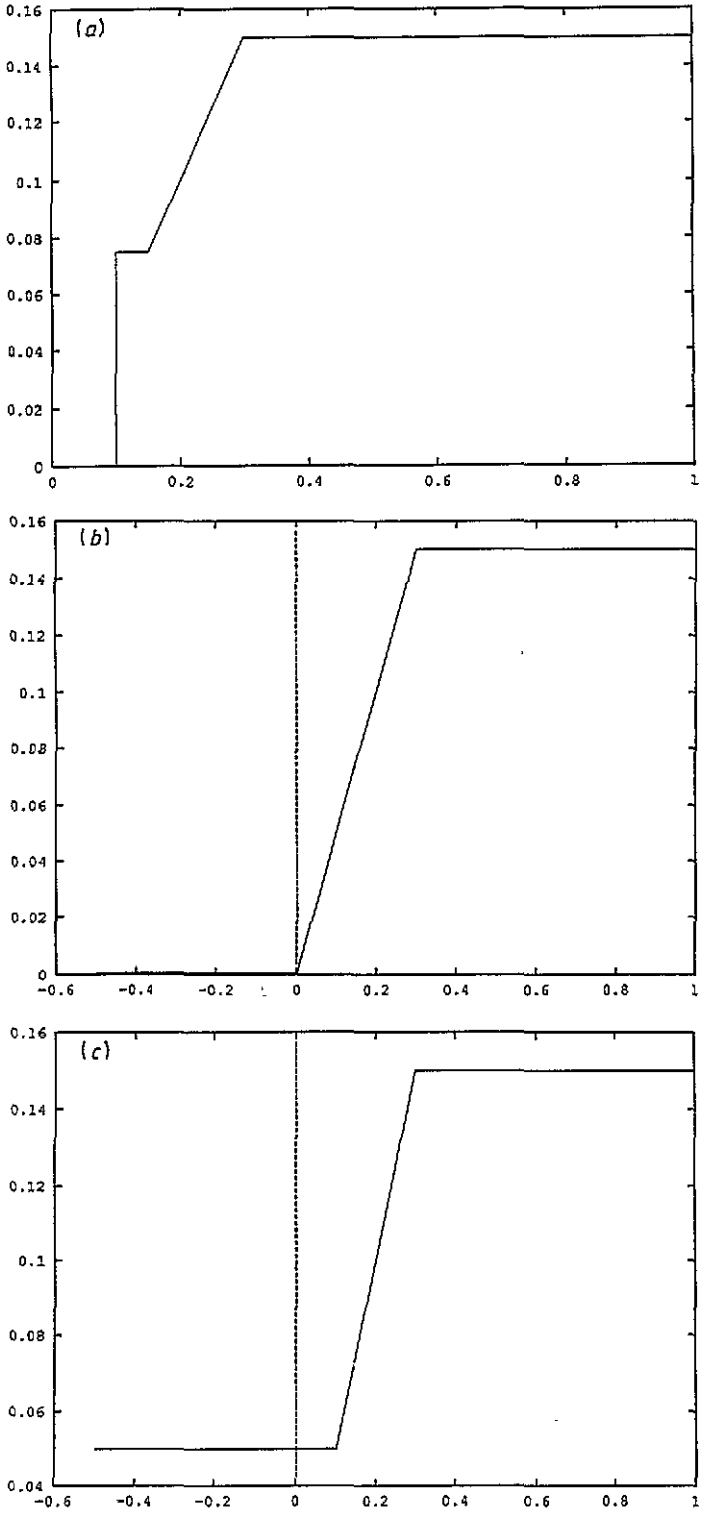

Figure 1. The behaviour of the function $q(x)$: (*a*) $h = 0$ and $n > 0$, (*b*) $h = 0$ and $n < 0$, (*c*) $h \neq 0$ and $n < 0$.

The sign of the 'frustration loops' is equal, as expected, to the sign of n.

We can now make a more general statement about the RSB solution at negative n. Let us consider any spin-glass system, for instance the SK model, characterized at zero n by an order parameter function $q_0(x)$. We shall assume that the saddle-point equations are polynomial in Q of arbitrary degree, and vanish at $Q = 0$, as in the SK model at zero field. Simple algebraic arguments can then be used to show that the function

$$q_n(x) = \begin{cases} 0 & n \leqslant x < 0 \\ q_0(x) & 0 \leqslant x \leqslant 1 \end{cases} \tag{36}$$

is a solution of the saddle-point equations with respect to Q for negative n and has the same total free energy as in the $n = 0$ case. This is easily obtained observing by a direct computation that the linear space of the matrices of the form (36) form a closed algebra with respect to the products mentioned above. Consequently, the saddle-point equations generalizing (27) take the form $g_n(x) = 0$ with

$$g_n(x) = \begin{cases} 0 & n \leqslant x < 0 \\ g_0(x) \equiv 0 & 0 \leqslant x \leqslant 1\,. \end{cases} \tag{37}$$

Moreover, again by direct inspection, it is easy to relate all the permutation invariants and in particular the free energy, to their values in the $n = 0$ case†.

In non-zero magnetic field, the above argument does not apply. In fact an inspection of the case $T \simeq T_c$ shows that the right solution $q_n(x)$ has a plateau that extends between n and x_0.

Let us finally comment about the physical interpretation of the solution. In zero field, assuming that the solution (36) is the correct one, one finds two striking results.

(i) The total free energy $\mathcal{F}$ is just equal to n times the free energy F_0 of the $n = 0$ system. This may seem a bit strange at first sight. As we argued in the introduction, taking a finite negative n introduces a bias in the sample distribution which favours overfrustrated samples which should have a large free energy. Yet we find that the typical free energy density of a generic sample is unchanged with respect to the usual $n = 0$ case. This is also totally different from what happens when one turns n to positive values, in which case the free energy is lowered. The reason for this phenomenon is that it is very difficult to find samples which have a free energy density larger than F_0. The case of the random energy model (REM) [9] is instructive in this respect. In the REM there are 2^N energy levels E_i which are independent random variables picked up at random from the distribution $P(E) = c^t \exp(-E^2/N)$. In the quenched case the thermodynamics below the critical temperature is dominated by the lowest energy levels which have a free energy $E_i = -N\sqrt{\ln(2)} + \epsilon_i$, where ϵ_i are small non-extensive fluctuations. If n is positive, the total partition functions is dominated by samples in which at least one level has a free energy extensively lower than $-N\sqrt{\ln(2)}$, say $E_1 = -N(\sqrt{\ln(2)} + \delta)$. The probability of such a sample in the original measure is exponentially small in N, but this is compensated by the gain in total free energy obtained because of the positive n in (3). Turning now to negative n, the situation is very different. In order to increase the total free energy density, we seek samples such that *all* the energy levels verify $E_i > -N(\sqrt{\ln(2)}) + N\delta)$, with $\delta > 0$. But the probability of such a sample is much smaller than exponential in N, as can easily be seen, and therefore this extremely small probability cannot be compensated by the gain

† Strictly speaking our proof holds only for those systems in which the free energy as a function of Q is expressed by invariant combinations of the two products mentioned in the text. Neverthless we think that the property should hold in general.

of order $\exp(-nN\delta)$ in the measure (3). A negative n does bias the sample distribution towards overfrustrated ones, as can be seen from the frustration loops, but it cannot change the free energy density.

(ii) The $P_n(q)$ takes the form

$$P_n(q) = \frac{1}{1-n}P_0(q) - \frac{n}{1-n}\delta(q). \tag{38}$$

A finite probability at the minimal value of q, namely $q = 0$, has appeared. Again this phenomenon can be understood by arguments similar to the ones above. While not changing the free energy density, a negative n does shift the free energy towards higher values. As a consequence the free energy of the states become less scattered and the probability of finding two low-lying, but different, pure states is increased.

4. Neural networks

In this section we study the Hopfield model of neural networks at negative n, focusing on the zero-temperature limit. Consider the usual Hopfield model [2], described by a system of Ising spins with the Hamiltonian

$$H = -\tfrac{1}{2}\sum_{j\neq i}^{N} J_{ij}\sigma_i\sigma_j \tag{39}$$

where

$$J_{ij} = \frac{1}{N}\sum_{\mu}^{P}\xi_i^\mu\xi_j^\mu \tag{40}$$

and $\{\xi_i^\mu\} = \pm 1$ are the stored patterns. We consider the case where the number of stored patterns P is proportional to N in the thermodynamic limit $N \to \infty$, so that the parameter $\alpha = P/N$ remains finite.

In terms of the standard replica formalism for the replica partition function

$$\langle\langle Z^n\rangle\rangle = \sum_{\xi=\pm 1}\sum_{\sigma=\pm 1}\exp\left\{\tfrac{1}{2}\beta N\sum_{a}^{n}\sum_{\mu}^{P}\left(\frac{1}{N}\sum_{i}^{N}\sigma_i^a\xi_i^\mu\right)^2\right\} \tag{41}$$

one gets (see, e.g., [6]):

$$\langle\langle Z^n\rangle\rangle = \int Dm_a\int D\hat{Q}\int D\hat{r}\exp\{-\beta nNF[m_a,\hat{Q},\hat{r}]\}. \tag{42}$$

In the 'condensed ansatz', in which only the overlap with one pattern is macroscopically different from zero, the replica free energy $F[m_a,\hat{Q},\hat{r}]$ is

$$F[m_a,\hat{Q},\hat{r}] = \frac{1}{2n}\sum_{a}^{n}(m_a)^2 + \frac{1}{2n}\alpha\beta\sum_{a\neq b}r_{ab}Q_{ab} + \frac{\alpha}{2\beta n}\mathrm{Tr}\ln(\hat{1}-\beta\hat{Q})$$
$$-\frac{1}{\beta n}\ln\left[\sum_{\sigma}\exp\left(\beta\sum_{a}^{n}m_a\sigma^a + \tfrac{1}{2}\alpha\beta^2\sum_{a\neq b}r_{ab}\sigma^a\sigma^b\right)\right]. \tag{43}$$

Here m_a is the overlap with the condensed pattern

$$m_a = \frac{1}{N}\sum_{i}^{N}\langle\sigma_i^a\rangle\xi_i^{\mu=1} \tag{44}$$

and Q_{ab} is the spin-glass order parameter:

$$Q_{ab} = \frac{1}{N}\sum_{i}^{N}\langle\sigma_i^a\sigma_i^b\rangle \tag{45}$$

($Q_{aa} \equiv 1$), r_{ab} gives the average value of the noisy overlaps with non-condensed patterns:

$$r_{ab} = \frac{1}{\alpha}\sum_{\mu=2}^{P} m_\mu^a m_\mu^b . \tag{46}$$

4.1. Replica symmetric solution

In the replica symmetric ansatz one takes

$$\begin{aligned} Q_{ab} &= q \qquad \text{for all } a \neq b \\ r_{ab} &= r \qquad \text{for all } a \neq b \\ m_a &= m \qquad \text{for all } a \end{aligned} \tag{47}$$

(the diagonal elements $Q_{aa} \equiv 1$). The standard calculations [6] result in the following expression for the free energy:

$$\begin{aligned} F[m,q,r] = {} & \frac{1}{2}m^2 + \frac{1}{2}\alpha\beta r(1-q) + \frac{n}{2}\alpha\beta r q \\ & + \frac{\alpha}{2\beta}\left[\ln(1-\beta+\beta q) + \frac{1}{n}\ln\left(1 - \frac{n\beta q}{1-\beta+\beta q}\right)\right] \\ & - \frac{1}{n\beta}\langle\langle\ln[2\cosh(\beta(m+\sqrt{\alpha r}z))]\rangle\rangle \end{aligned} \tag{48}$$

where $\langle\langle\cdots\rangle\rangle$ means Gaussian averaging over z:

$$\langle\langle(\cdots)\rangle\rangle = \int_{-\infty}^{+\infty}\frac{\mathrm{d}z}{\sqrt{2\pi}}(\cdots)\exp\left(-\frac{z^2}{2}\right). \tag{49}$$

The corresponding saddle-point equations for the parameters m, q, and r are

$$m = \frac{\langle\langle(\cosh\beta(m+\sqrt{\alpha r}z))^n\tanh[\beta(m+\sqrt{\alpha r}z)]\rangle\rangle}{\langle\langle(\cosh\beta(m+\sqrt{\alpha r}z))^n\rangle\rangle} \tag{50}$$

$$\beta(1-q) \equiv C = \beta\frac{\langle\langle(\cosh\beta(m+\sqrt{\alpha r}z))^{n-2}\rangle\rangle}{\langle\langle(\cosh\beta(m+\sqrt{\alpha r}z))^n\rangle\rangle} \tag{51}$$

$$r = \frac{q}{(1-C)(1-C-\beta n q)} . \tag{52}$$

In what follows we consider only the case of negative n in the limit of zero temperature. It is clear from (52) that if the parameter C remains finite (which will be shown to be the case), the parameter r must scale with the temperature as β^{-1}. Let us redefine: $r = r'/\beta$.

In the limit $\beta \to \infty$ one gets

$$\langle\langle(\cosh\beta(m+\sqrt{\alpha r}z))^{-|n|}\rangle\rangle \to \int_{-\infty}^{+\infty}\frac{\mathrm{d}z}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2} - |n|\,|\beta m + \sqrt{\alpha\beta r'}z|\right). \tag{53}$$

The main contribution to the above integral comes from the saddle point which is defined by the equation

$$z^* = -|n|\sqrt{\alpha\beta r'}\,\mathrm{sign}(\beta m + \sqrt{\alpha\beta r'}z^*) . \tag{54}$$

The solution of this equation is

$$z^* = \begin{cases} -|n|\sqrt{\alpha\beta r'} & \text{if } |n|\alpha r' < m \\ -m\sqrt{\beta/\alpha r'} \equiv z_0 & \text{if } |n|\alpha r' > m\,. \end{cases} \tag{55}$$

In the second case ($|n|\alpha r' > m$) the point z_0 is actually not the saddle point: this is the situation when the main contribution to the integral in (53) comes from the vicinity of the point at which $|\beta m + \sqrt{\alpha\beta r'}z| = 0$, such that the Gaussian part $-z^2/2$ is becoming irrelevant.

The result of the integration is

$$\langle\langle(\cosh\beta(m+\sqrt{\alpha r}z))^{-|n|}\rangle\rangle \simeq \begin{cases} \exp\{-\beta|n|(m-\frac{1}{2}|n|\alpha r')\} & \text{if } |n|\alpha r' < m \\ \exp\{-\frac{1}{2}\beta m^2/\alpha r'\} & \text{if } |n|\alpha r' > m\,. \end{cases} \tag{56}$$

Let us consider the two cases separately.

(i) $|n|\alpha r' < m$. In this case from (51) one gets:

(a) if $(|n|+2)\alpha r' > m$:

$$C \simeq \beta\exp\left\{-\frac{\beta}{2\alpha r'}(m-|n|\alpha r')^2\right\} \to 0 \tag{57}$$

(b) if $(|n|+2)\alpha r' < m$:

$$C \simeq \beta\exp\{-2\beta(m-(|n|+1)\alpha r')\} \to 0\,. \tag{58}$$

Therefore, from (52) one obtains

$$r' = \frac{1}{|n|}\,. \tag{59}$$

One can easily see that, in this case, ($|n|\alpha r' < m$) $|z_0| \ll |z^*|$ (equation (55)). Therefore, from (50) one gets

$$m = \frac{\langle\langle(\cosh(\beta m+\sqrt{\alpha\beta r'}z))^{-|n|}\mathrm{sign}(z+z_0)\rangle\rangle}{\langle\langle(\cosh(\beta m+\sqrt{\alpha\beta r'}z))^{-|n|}\rangle\rangle} \simeq \frac{\langle\langle(\cosh(\beta m+\sqrt{\alpha\beta r'}z))^{-|n|}\rangle\rangle}{\langle\langle(\cosh(\beta m+\sqrt{\alpha\beta r'}z))^{-|n|}\rangle\rangle} = 1\,. \tag{60}$$

According to the condition $|n|\alpha r' < m$, the obtained retrieval solutions $r' = 1/|n|$ (or $r = 1/\beta|n|$) and $m = 1$ exist in the domain $\alpha < \alpha_c = 1$.

Note that in the case $|n| \ll 1$, the 'perfect retrieval' state that we have found exists only at temperatures such that $\beta|n| \gg 1$. Otherwise, if $\beta|n| \ll 1$, the equations are becoming equivalent to those of the usual Hopfield model with quenched patterns ($n = 0$). Therefore, in the system under consideration the limits $T \to 0$ and $n \to 0$ do not commute.

Note also that the finite-temperature corrections to the obtained values of m, r and $q = 1$ are exponentially small: $\sim \exp(-\mathrm{const}\,\beta)$.

(ii) $|n|\alpha r' > m$. In this case the main contribution to the Gaussian integration over z comes from the vicinity of the point $z = z_0$, and one immediately sees from (60) that $m = 0$. Therefore, in this case the system is in the spin-glass state. However, it can easily be shown that the symmetric ansatz gives a pathological solution for the spin-glass state. Indeed, from the result (56) for the parameter C (equation (51)) one gets:

$$C = \mathrm{const}\;\beta \to \infty\,. \tag{61}$$

In view of what we have seen before for the spin-glass solutions in the SK model with negative n, it is actually quite natural that the considered RS ansatz can also not be applied for the spin-glass state in the Hopfield model.

4.2. Replica symmetry breaking

For the SK model with negative n and zero field we have learned that the RSB solution for the function $q(x)$ coincides with that of the model with $n = 0$ on the interval $0 \leqslant x \leqslant 1$, and $q(x) \equiv 0$ on the interval $-|n| \leqslant x \leqslant 0$. The same general arguments can be used for the Hopfield model. Therefore, in the limit $T \to 0$ where the functions $q(x)$ and $r(x)$ are getting almost 'flat' (replica symmetric) on the interval $0 \leqslant x \leqslant 1$, for the model with negative n we shall consider the following simple ansatz: $q(x) = q$ and $r(x) = r$ in the interval $0 \leqslant x \leqslant 1$, and $q(x) = r(x) = 0$ in the interval $-|n| \leqslant x \leqslant 0$.

Using the general expression for the free energy (43) one gets (for $m = 0$):

$$F[q,r] = \frac{1}{2}\alpha\beta r(1-q) + \frac{\alpha}{2\beta}\left[\ln(1-\beta+\beta q) - \frac{\beta q}{1-\beta+\beta q}\right] - \frac{1}{\beta}\langle\langle \ln[2\cosh(\beta\sqrt{\alpha r}z)]\rangle\rangle . \tag{62}$$

This free energy coincides with the replica-symmetric one of the usual Hopfield model (with $n = 0$). Therefore the parameters q and r of this solution coincide with the ones of the RS spin-glass solution of the usual Hopfield model with $n = 0$, which are (see, e.g., [6])

$$r = \left(1 + \sqrt{2/\pi\alpha}\right)^2 \qquad C = \frac{\sqrt{2/\pi\alpha}}{1+\sqrt{2/\pi\alpha}} . \tag{63}$$

To conclude this technical analysis, the peculiar point of the Hopfield model with negative n is that at zero temperature its retrieval state is given be the replica-symmetric solutions of the mean-field equations, and this retrieval solution exists up to $\alpha_c = 1$. In the whole interval $0 \leqslant \alpha < \alpha_c$ we find perfect retrieval, $m = 1$. On the other hand, the spin-glass state is described by the RSB solution. In the limit of zero temperature we have found one such solution which becomes nearly a one-step breaking. Although we have not proved that this is the only solution, it seems to be a reasonable one in view of the discussion of the previous section on spin glasses.

The physical interpretation of this model must be understood along the same lines as explained in the introduction. It describes a coupled dynamics of neurons and synapses, taking place on two very different timescales. But now the synapses are constrained to be of the Hebb type (40), so their dynamics is constrained to a certain subspace, and it can be understood as a slow dynamics of the patterns. In the retrieval phase, starting from an initial configuration of the neurons which is close to one of the memorized patterns, one will first see a fast dynamics of the neurons towards the pattern, and superimposed on it the patterns, which should rather be called here the internal representations of the original patterns, will drift slowly. This drift will tend to overfrustrate the system. In this context it is reasonable to believe that it actually corresponds to some small changes of the internal representations tending to orthogonalize them. While we have not really proven that this interpretation is the correct one, it is in agreement with the above computations. The orthogonalization of the patterns is consistent with the fact that the parameter r goes to zero at low temperatures. It also agrees with the new value of the storage capacity $\alpha_c = 1$, which is the maximal number of patterns that can be orthogonalized exactly.

The situation that we have studied here is a very special one. However it is interesting to see that the coupled dynamics of neurons and synapses, taking place on two very different timescales, can be amenable to an analytic treatment with the replica method at negative n. Such dynamics have received much attention in recent years [3, 12]. In our case the synapses dynamics was constrained to its Hebbian subspace. It would be interesting to

generalize this approach, firstly by constraining the internal representations to stay close to the original stored patterns, secondly by allowing the synapses to take values outside the Hebbian subspace.

5. Conclusions

We have considered spin systems in which the interactions between spins, as well as the spins themselves are dynamical variables. Spins and interactions characteristic scales are widely separated. We have assumed that the spins completely equilibrate before the interactions change by a finite amount; conversely, the interactions evolve in a kind of 'consistent field' created by the spins. The dynamics is such that spins and interactions do not tend to mutual equilibrium at a temperature T. Each kind of variables thermalizes at different temperatures, respectively T and T'. For negative T' the spin system tends to induce overfrustration. The analysis of the frustration loops confirms this picture. We have shown that in the case of the SK model in zero field, overfrustration has a very weak effect for zero magnetic field. Due to the constraints imposed by the *a priori* distribution, the interactions can not differ too much from a typical quenched sample. As a result the free energy of the spin system does not change extensively compared with the quenched case. Neverthless overfrustation has a consequence on the organization of pure states: the $P(q)$ developes a delta-function peak for $q = 0$ and equilibrium states are more likely to be far apart than in the quenched case.

In the Hopfield model, the results for the spin-glass phase are similar to those for the SK. More dramatic effects are observed on the retrieval phase. Overfrustration, which in the context of neural networks is reminescent of the unlearning algorithm, pushes the patterns towards mutual orthogonalization. This leads to a net increase of the capacity from the value of 0.145 to 1. This last value is typical of the 'pseudo-inverse learning rule' [13] where the patterns are orthogonalized by hand. A criticism that can be applied to the use of this approach as a learning algorithm is that the patterns, once they have reached thermal equilibrium, are still free to diffuse. It is not clear what is the correlation between the initial patterns one wanted to store in the system and those found in it for long times. Another interesting open question concerns the basins of attraction of the 'patterns'.

Acknowledgements

We thank D O'Kane for making us aware of [3] prior of its publication. SF acknowledges the hospitality of The Department of Theoretical Physics of Oxford, where this work was taken to accomplishment.

References

[1] Sherrington D and Kirkpatrick S 1975 *Phys. Rev. Lett.* **35** 1972
[2] Hopfield J J 1982 *PNAS USA* **79** 2554
[3] Penney R W, Coolen T and Sherrington D 1993 *J. Phys. A: Math. Gen.* **26** 3681
Coolen T, Penney R W and Sherrington D 1994 *Phys. Rev.* B to be published
Sherrington D, Penney R W and Coolen T 1993 Complexity in the coupled dynamics of fast neurons and slow synapses *Proc. 1993 Blois rencontres on Chaos and Complexity* to appear
[4] Sherrington D 1980 *J. Phys. A: Math. Gen.* **13** 637
[5] Kondor I 1983 *J. Phys. A: Math. Gen.* **16** L127
[6] Amit D, Sompolinsky H and Gutfreund H 1987 *Ann. Phys.* **173** 30

[7] Kleinfeld D and Pendergraft D B 1987 *Biophys. J.* **51** 47
van Hemmen J L, Ioffe L B, Kuhn R and Vaas M 1989 *Physica* **163A** 386
[8] Parisi G 1979 *Phys. Lett.* **73A** 203; 1980 *J. Phys. A: Math. Gen.* **13** L115, 1101
[9] Derrida B 1981 *Phys. Rev.* B **24** 2613
[10] Mézard M, Parisi G and Virasoro M A 1987 *Spin-glass Theory and Beyond* (Singapore: World Scientific)
[11] Parisi G 1983 *Phys. Rev. Lett.* **50** 1946
[12] Jonker H J J and Coolen A C C 1991 *J. Phys. A: Math. Gen.* **24** 4219
[13] Kantor I and Sompolinsky H 1986 *Phys. Rev.* A **35** 380