Partially annealed neural networks

1994 J. Phys. A: Math. Gen. 27 4401

(http://iopscience.iop.org/0305-4470/27/13/015)

J. Phys. A: Math. Gen. 27 (1994) 4401–4411. Printed in the UK

Partially annealed neural networks

D E Feldman and V S Dotsenko

Landau Institute for Theoretical Physics, Russian Academy of Sciences, Kosygina 2, Moscow 117940, Russia

Received 20 December 1993

Abstract. We consider the Hopfield model of neural networks in which the patterns, as well as the spins, are dynamical variables. The characteristic time scales of the dynamics of the spins and the patterns are assumed to be widely separated such that the spins completely equilibrate at the time scale at which the elementary changes in the patterns take place. We study the situation in which each type of variable thermalizes at different temperatures, respectively, T and T'. In this case, such a system is described in terms of the traditional replica formalism in which the number of replicas $n = T/T'$ is still the finite parameter.

The complete phase diagram of the model in the space of the parameters T, α and n is obtained. If the parameter n is negative, the model is argued to present some similarities with the unlearning training algorithm. In this case a substantial increase in size of the retrieval phase in the plane (T, α) is found.

1. Introduction

In the physics of disordered materials, the degrees of freedom describing an actual system are usually well separated into two essentially different types: the 'annealed' or 'dynamical' variables in terms of which the statistical mechanics is calculated; and the 'quenched' variables which enter the statistical mechanics as the fixed parameters. A traditional problem is then to calculate the self-averaging thermodynamic quantities, like the free energy, which require averaging over the statistical distribution of the quenched variables. At this second step, the quenched degrees of freedom are effectively 'annealed' (at infinite time scale), provided that their statistics is not affected by the annealed degrees of freedom. This is the typical situation for spin-glasses and statistical models of neural networks where, for the actual calculations, the so-called replica formalism is used (see e.g. [1]).

Here, we consider the specific situation in which the originally quenched variables are taken to be somewhat intermediate between the 'quenched' and 'annealed' cases. They are considered to be 'slow' dynamical variables evolving at a time scale which is much larger than the thermal-equilibration time of the annealed degrees of freedom. In this case, it would be natural to expect the dynamics of the slow variables to be described by the 'heat bath' random process in which the role of the effective potential is played by the free energy of the thermally equilibrated fast variables.

In what follows, we study the situation in which both the fast as well as the slow variables are at thermal equilibrium but at different temperatures. In this case, the two types of degrees of freedom are not mutually equilibrated (they relate to two different thermal baths).

Let us consider a system described by some Hamiltonian $H[\xi; \sigma]$ which depends on the fast variables $\{\sigma_i\}$ and the slow variables ξ. For the free energy of such a system with

0305-4470/94/134401+11$19.50

a given fixed realization of the ξs, one gets

$$F[\xi] = -\frac{1}{\beta}\log Z[\xi] \tag{1}$$

where

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

is the partition function.

If the ξ's change their values on a time scale which is much larger than the equilibration time of the σ's, the statistics of these fast variables is described by the free energy (1) where the ξ's remain effectively quenched. Then, this free energy becomes the energy function (the Hamiltonian) for the ξ's degrees of freedom. The space in which the variables ξ exist should be specified separately. In the quenched case, this space is defined by some statistical distribution function $P[\xi]$. In the partially annealed case, this function $P[\xi]$ has the meaning of an internal potential for the ξ's, which restricts the space of their values.

If the fast and slow degrees of freedom are not thermally equilibrated, such that the slow variables have temperature T' which is different from the temperature T of the fast variables, then for the total partition function of the system, one gets

$$\mathcal{Z} = \int D\xi\ P[\xi]\exp(-\beta' F[\xi]) = \int D\xi\ P[\xi]\exp\left(\frac{\beta'}{\beta}\log Z[\xi]\right) = \int D\xi\ P[\xi](Z[\xi])^n \tag{3}$$

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

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

where

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

In this way, we recover the well known replica formalism in which the 'number of replicas' $n = T/T'$ remains a *finite* parameter.

From the point of view of the partial annealing considered here, the quenched case corresponds to the limit of the infinite temperature T' of the slow variables. In this case, the thermodynamics of the fast degrees of freedom produces no effect on the distribution of the slow ones. If $T' = T$ ($n = 1$), we get the trivial case of purely annealed disorder whatever the difference in the characteristic time scales of the ξs and σs is.

In the case of partial equilibrium $n \neq 0$ and $n \neq 1$, the evolution of the slow variables could be described by a Langevin type stochastic dynamics

$$\tau_\xi \frac{d\xi}{dt} = -\frac{d}{d\xi}\left(F[\xi] - \frac{\log P}{\beta'}\right) + \eta(t) \tag{6}$$

where $\langle\langle \eta(t)\eta(t')\rangle\rangle = 2T'\delta(t-t')$ and the 'microscopic' time τ_ξ of the ξ's dynamics is assumed to be much larger than the thermal-equilibration time of the fast variables σ.

The above physical interpretation of replicas has been proposed by Penney *et al* [2], who studied the case of positive n in the SK model of spin-glasses [3], and by Dotsenko *et al* [4] who studied replica-symmetry breaking in the Sherrington–Kirkpatrick (SK) model and retrieval properties in neural networks for negative and positive n.

In this paper, we are going to study the Hopfield neural networks [5] for positive and negative values of the parameter n. Traditionally, the quenched variables in neural networks are the stored Ising 'patterns'. Partial annealing here means that the stored patterns become (slow) dynamical variables. Although neural networks with moving patterns look a bit tricky at first sight, we believe that it does make sense to remember the various 'training' procedures (see e.g. [6]) for modifying the synaptic interactions.

In the case of a negative value for the temperature T', the situation, to a certain extent, is reminiscent of the unlearning algorithm [7]. In the original formulation, this algorithm defines the discrete-time evolution of the spin–spin couplings J_{ij} in the form

$$J_{ij}(t+1) = J_{ij}(t) - \epsilon\sigma_i^*\sigma_j^* \tag{7}$$

where ϵ is some (numerically) small positive parameter and $\{\sigma^*\}$ is taken at a random spin configuration corresponding to one of the energy minima at given values of the couplings $J_{ij}(t)$. The point is that the above modification of the couplings (with the chosen sign of ϵ) makes the corresponding energy minima higher and, in general, the couplings evolve towards *maximum* energy. Taking the Hebb learning rule (11) for the initial couplings at $t = 0$, it was demonstrated that (presumably) due to the reduction in the noisy interference effects among the patterns, the above training procedure provides substantial increase in the storage capacity α_c.

On the other hand, considering the unlearning dynamics in a generalized form, namely, introducing finite temperature in the spin system and also finite thermal noise for the modification of the J_{ij}'s at each iteration step, one obtains the following (discrete-time) dynamics

$$\frac{\delta J_{ij}(t)}{\delta t} = -\langle\sigma_i\sigma_j\rangle_{J(t),T} + \eta_{ij}(t) \tag{8}$$

or

$$\frac{\delta J_{ij}(t)}{\delta t} = +\frac{\partial}{\partial J_{ij}} F[J(t), T] + \eta_{ij}(t). \tag{9}$$

Here, the thermal average $\langle\ldots\rangle_{J(t),T}$ and the free energy $F[J(t), T]$ are obtained for given values of the couplings $J_{ij}(t)$ and spin temperature T, and $\eta_{ij}(t)$ is the thermal white noise: $\langle\eta_{ij}(t)\eta_{kl}(t')\rangle = 2T'\delta_{(ij),(kl)}\delta(t-t')$. Equations (9) define the Langevin dynamics in the space of the spin couplings with the driving potential being the free energy $F[J(t), T]$ created by the thermally equilibrated spin system. The crucial point here is that according to equations (9), in thermal equilibrium, the system of couplings must be described by the corresponding Gibbs distribution with *negative* temperature.

The problem, however, is that, in Hopfield neural networks with finite (negative) replica parameter $n = T/T'$, the slow dynamical variables are the 'patterns' and not the synaptic couplings themselves (which are constrained to keep the Hebb structure in terms of the moving patterns). In this sense, the system considered here is not quite adequate for the unlearning procedure, but is only its distant analogy. Nevertheless, it exhibits interesting properties (section 3) and we believe that it might be valuable in its own right.

In particular, one finds that in the case of negative n, the patterns effectively move to become as orthogonal as possible. The 'patterns' here can be interpreted as internal representations of information which adapt themselves towards internal representations which have as little correlation as possible. At the zero temperature, this has been shown [4] to produce a substantial increase in the storage capacity, up to $\alpha_c = 1$ instead of $\alpha_c = 0.14$, in the usual Hopfield model [8]. At finite temperatures, we obtain a substantial increase in the size of the retrieval phase in the plane (T, α).

In the opposite case ($n > 0$ section 4), the 'patterns' move to become as parallel as possible. In this situation, the interference among the patterns increases, and the storage capacity decreases. In particular, for $n > 2/3$, the retrieval phase will be shown to disappear completely. Besides, at low enough temperatures ($T < n$), the system breaks down into an unusual 'superferromagnetic' phase in which the overlaps of the thermodynamic state with *all* the patterns become finite (the free energy in this phase becomes proportional to N^2, unlike the usual situation in which the free energy is of the order of N).

The full phase diagram of the model in the space of the parameters T and the reduced number of the stored patterns $\alpha = P/N$ for different values of the parameter n will be obtained. In particular, we also study the transitions into the spin–glass state which in the regions $1 < n < 2$ and $n > 2$ are shown to become quite peculiar. The stability of the obtained replica-symmetric (retrieval and spin–glass) states with respect to the replica-symmetry breaking is also studied and the corresponding Almeida–Thoulness (AT) lines both at $n < 0$ and $n > 0$ are calculated.

2. The model

Consider the usual Hopfield model [5], described by a system of Ising spins with the Hamiltonian

$$H = -\tfrac{1}{2}\sum_{j\neq i}^{N} J_{ij}\sigma_i\sigma_j \tag{10}$$

where

$$J_{ij} = \frac{1}{N}\sum_{\mu}^{P}\xi_i^\mu\xi_j^\mu \tag{11}$$

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

$$\langle\langle Z^n\rangle\rangle = \sum_{\xi=\pm1}\sum_{\sigma=\pm1}\exp\left\{\frac{1}{2}\beta N\sum_{a}^{n}\sum_{\mu}^{P}\left(\frac{1}{N}\sum_{i}^{N}\sigma_i^a\xi_i^\mu\right)^2\right\} \tag{12}$$

one gets (see, e.g. [8])

$$\langle\langle Z^n\rangle\rangle = \int Dm_a\int D\hat{Q}\int D\hat{r}\,\exp\{-\beta nNF[m_a,\hat{Q},\hat{r}]\}. \tag{13}$$

In the 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

$$F[m_a, \hat{Q}, \hat{r}] = \frac{1}{2n}\sum_a^n (m_a)^2 + \frac{1}{2n}\alpha\beta \sum_{a\neq b} r_{ab}Q_{ab} + \frac{\alpha}{2\beta n}\operatorname{Tr}\log(\hat{1} - \beta\hat{Q})$$
$$- \frac{1}{\beta n}\log\left[\sum_\sigma \exp\left(\beta\sum_a^n m_a\sigma^a + \frac{1}{2}\alpha\beta^2\sum_{a\neq b} r_{ab}\sigma^a\sigma^b\right)\right]. \tag{14}$$

Here m_a is the overlap with the condensed pattern

$$m_a = \frac{1}{N}\sum_i^N \langle\sigma_i^a\rangle\xi_i^{\mu=1} \tag{15}$$

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

$$Q_{ab} = \frac{1}{N}\sum_i^N \langle\sigma_i^a\sigma_i^b\rangle. \tag{16}$$

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

$$r_{ab} = \frac{1}{\alpha}\sum_{\mu=2}^P m_\mu^a m_\mu^b. \tag{17}$$

2.1. Replica-symmetric solution

In the replica-symmetric ansatz, one takes

$$\begin{aligned} Q_{ab} &= q \qquad \text{for all } a \neq b \\ r_{ab} &= r \qquad \text{for all } a \neq b \\ m_a &= m \qquad \text{for all } a \end{aligned} \tag{18}$$

(the diagonal elements $Q_{aa} \equiv 1$). The standard calculations [8] result in the following expression for the free energy:

$$F[m, q, r] = \frac{1}{2}m^2 + \frac{1}{2}\alpha\beta r(1-q) + \frac{n}{2}\alpha\beta r q$$
$$+ \frac{\alpha}{2\beta}\left[\log(1-\beta+\beta q) + \frac{1}{n}\log\left(1 - \frac{n\beta q}{1-\beta+\beta q}\right)\right]$$
$$- \frac{1}{n\beta}\log[\langle\langle(2\cosh(\beta(m+\sqrt{\alpha r}z)))^n\rangle\rangle] \tag{19}$$

where $\langle\langle\ldots\rangle\rangle$ means the Gaussian averaging over z

$$\langle\langle(\ldots)\rangle\rangle = \int_{-\infty}^{+\infty}\frac{\mathrm{d}z}{\sqrt{2\pi}}(\ldots)\exp(-\tfrac{1}{2}z^2). \tag{20}$$

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

$$m = \frac{\langle\langle(\cosh\beta(m+\sqrt{\alpha r}z))^n \tanh[\beta(m+\sqrt{\alpha r}z)]\rangle\rangle}{\langle\langle(\cosh\beta(m+\sqrt{\alpha r}z))^n\rangle\rangle} \tag{21}$$

$$\beta(1-q) \equiv C = \beta\frac{\langle\langle(\cosh\beta(m+\sqrt{\alpha r}z))^{n-2}\rangle\rangle}{\langle\langle(\cosh\beta(m+\sqrt{\alpha r}z))^n\rangle\rangle} \tag{22}$$

$$r = \frac{q}{(1-C)(1-C-\beta n q)}. \tag{23}$$

2.2. *Replica-symmetry breaking*

The region in which the replica-symmetric states (both the retrieval and spin–glass), defined by equations (21)–(23), become unstable with respect to the replica-symmetry breaking can be calculated in a standard way (see e.g. [1, 8]). In our case, when the replica parameter n is kept finite, this region could be easily shown to be defined by the condition

$$(T-1+q)^2 < \alpha \frac{\langle\langle(\cosh\beta(m+\sqrt{\alpha r}z))^{n-4}\rangle\rangle}{\langle\langle(\cosh\beta(m+\sqrt{\alpha r}z))^{n}\rangle\rangle}. \tag{24}$$

Correspondingly, the border of this region (determined by the above equation, where the $<$ sign is changed for $=$) defines the AT line, $T_{AT}(\alpha, n)$.

3. Negative n

For a given value of the parameter $n < 0$, the standard calculations of the solutions of the saddle-point equations (21)–(23) give the phase diagram in the space of the parameters T and α, qualitatively shown in figure 1.

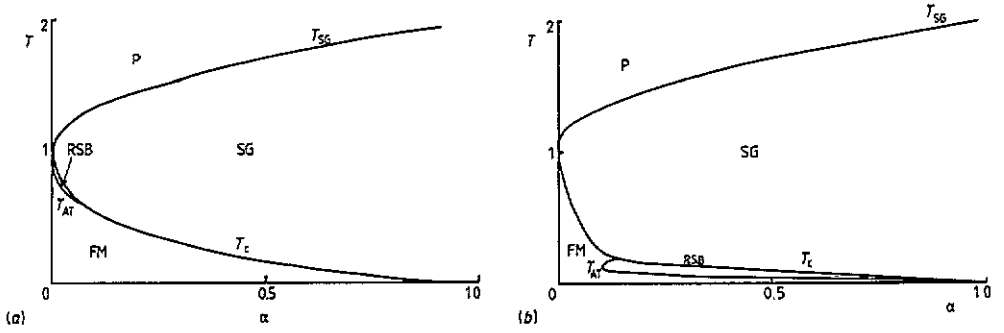

Figure 1. Phase diagram of the model with negative value of the replica parameter n: (*a*) $|n| \gg 1$; (*b*) $|n| \ll 1$.

3.1. *Retrieval state*

This phase diagram looks similar to that of the usual Hopfield model with quenched patterns [8]. The essential difference, however, is that the curve of the (first-order) phase transition into the retrieval state (with $m \neq 0$) $T_c(\alpha)$ starts from the point $\alpha_c = 1$ at $T = 0$ and not from $\alpha = 0.138$ as in the case of quenched patterns. This remains to be the universal point for any non-zero value of negative n.

In the limit $|n| \gg 1$ (which corresponds to the zero-temperature limit in the subsystem of patterns), the saddle-point equations (21)–(23) could be essentially simplified. From equation (23) for r one gets $r \simeq 1/(1-C)\beta|n|$. Then, as a result of a simple analysis of the Gaussian integration in equations (21) and (22), one could easily reduce the three equations (21)–(23) into one equation for the parameter m

$$m = \tanh\left[\frac{m}{T} - \frac{\alpha m}{T-1+m^2}\right] \tag{25}$$

which no longer depends on the replica parameter n. For the spin–glass order parameter one has $q = m^2$. The solution of equation (25) gives the phase diagram shown in figure 1(*a*).

In the opposite limit $|n| \ll 1$ (figure 1(*b*)), in the region of the temperatures $T \gg |n|$, the phase diagram coincides with that of the usual Hopfield model where the line $T_c(\alpha)$ moves towards the point $\alpha = 0.138$ at low temperatures. However, in the narrow region $T \ll |n|$, this line turns quickly to the point $\alpha_c = 1$ anyway. Therefore, the limits $T \to 0$ and $|n| \to 0$ do not commute in the considered system.

The behaviour of the line $T_c(\alpha)$ near the point $(T = 1, \alpha = 0)$ is also similar to that of the usual Hopfield model

$$T_c(\alpha) \simeq 1 - \tau(n)\sqrt{\alpha} \tag{26}$$

where the coefficient τ is some function of n. It could be easily proved that $\tau(|n| \to 0) \simeq 1.95$ and $\tau(|n| \to \infty) = \sqrt{3}$.

3.2. Replica-symmetry breaking

The solution of equation (24) shows that in the region restricted by the lines $T_{AT}(\alpha)$ and $T_c(\alpha)$ (figure 1) the obtained replica-symmetric solution for the retrieval state becomes unstable and the correct solution must be calculated in terms of the Parisi replica symmetry-breaking scheme [1] which we do not consider here.

It is interesting to note that in the limit $|n| \to \infty$, the region of the replica-symmetry breaking shrinks to zero near the point $(T = 1, \alpha = 0)$ (figure 1(*a*)).

In the opposite limit, $|n| \ll 1$ (figure 1(*b*)), this region moves to the right and below the retrieval region. Here the upper branch of the AT line (at which $T \gg |n|$) coincides with the corresponding AT line of the usual Hopfield model.

What is essential, however, is that in the low-temperature part of the retrieval phase (including the zero-temperature interval $0 \leqslant \alpha < 1$), the replica-symmetric solution is stable for both cases $|n| \gg 1$ and $|n| \ll 1$, unlike the situation in the usual Hopfield model and unlike the spin–glass solution in the considered model.

3.3. Spin–glass state

As in the usual Hopfield model, the spin–glass state $(m = 0, q, r \neq 0)$ is stable everywhere below the second-order phase-transition line $T_{SG}(\alpha) = 1 + \sqrt{\alpha}$. In this region, according to equation (24), the replica symmetry appears to be broken. Therefore, the SG state should be described in terms of the Parisi functions $q(x)$ and $r(x)$ such that $q(x)$ and $r(x)$ equal zero in the interval $-|n| \leqslant x \leqslant 0$, while in the interval $0 \leqslant x \leqslant 1$ they coincide with those of the spin–glass solution of the usual Hopfield model (see [4]).

4. Positive n

If the parameter n is positive then the phase diagram of the system becomes much more sophisticated. First of all, there exist several intervals for the values of n in which the phase diagrams are essentially different.

The other important point is that, in the low-temperature region at $T < n$, the system breaks down into a new 'superferromagnetic' (SF) phase in which all the overlaps m_μ of the thermodynamic state with the stored patterns ξ^μ become finite. This can be easily seen from equation (23) for r in which it is obvious that, because of the factor $-\beta n q$ in the

denominator at low enough temperatures the order parameter r will become divergent. This means, in turn, that the 'non-condensing' overlaps which compose r (equation (17)) are no longer of order $1/\sqrt{N}$ and the other order parameters in the calculation of the free energy should be used.

This can be done quite easily assuming that all the overlaps m_μ are finite and equal. For the partition function (12), in a standard way one gets

$$\langle\langle Z^n\rangle\rangle = \int \mathrm{d}m_a^\mu \exp\left\{-\frac{1}{2}\beta N \sum_{a=1}^{n}\sum_{\mu=1}^{P}(m_a^\mu)^2 + \log\left[\sum_{\xi}\sum_{\sigma}\exp\left(\beta\sum_{i,a,\mu} m_a^\mu \sigma_i^a \xi_i^\mu\right)\right]\right\}. \quad (27)$$

Assuming replica symmetry and substituting $m_a^\mu = m$, after summation over the ξs and σs for the free energy, one obtains

$$F(m) \equiv -\frac{1}{\beta n}\langle\langle Z^n\rangle\rangle = \frac{1}{2}\alpha N^2 m^2 - \frac{\alpha N^2}{\beta n}\log\cosh(\beta n m). \quad (28)$$

In the result for the order parameter m, one finds the usual mean-field equation

$$m = \tanh(\beta n m) \quad (29)$$

which gives the transition temperature $T_f = n$. Therefore at $T < T_f$ (whatever the (non-zero) value of α is), the system appears to be in the state described by non-zero-order parameter m, which is the value of the overlap of the thermodynamic state with all the patterns ξ^μ. The point is that, at low temperatures, the patterns, being free to move and tending to become as parallel as possible, condense into the state in which all of them have finite components parallel to the spin state of the system. Since the number of the patterns is macroscopic (αN), the total free energy of the system in this state becomes proportional to N^2 instead of N. This state could be conditionally called SF.

4.1. $0 < n < 2/3$

In this case, the qualitative phase diagram of the system is shown in figure 2. The line $T_c(\alpha)$ bounds the region where the retrieval state is stable. As usual, this is the line of the first-order phase transition.

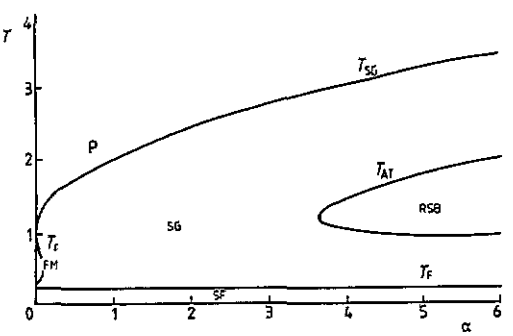

Figure 2. Phase diagram of the model with $0 < n < 2/3$.

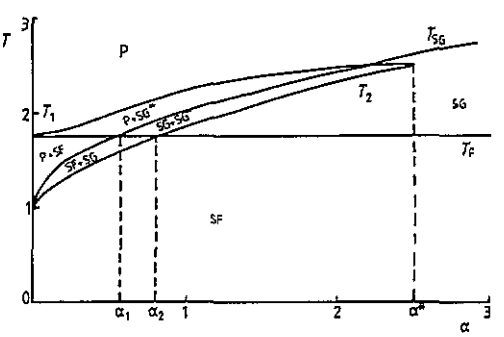

Figure 3. Phase diagram of the model with $1 < n < 2$.

If n is small, the point $T_0(n)$ at which the line $T_c(\alpha)$ starts at $\alpha \ll 1$, behaves as $T_0(n) \simeq n + 4\exp(-2/n)$. As $n \to 2/3$, $T_0 \to 1$.

Near the point $T = 1, \alpha = 0$ for the line $T_c(\alpha)$, one finds the usual behaviour: $T_c(\alpha) \simeq 1 - \tau(n)\sqrt{\alpha}$. However, the coefficient $\tau(n)$ diverges as $n \to 2/3$: $\tau(n) \simeq (1 - 3n/2)^{-1/2}$.

Therefore, the retrieval region shrinks to the point $T = 1, \alpha = 0$ as $n \to 2/3$.

Besides, everywhere below the second-order transition line $T_{SG}(\alpha) = 1 + \sqrt{\alpha}$ and above T_f there exists the usual spin–glass state. Solving equation (24), one finds that everywhere beyond the region restricted by the line $T_{AT}(\alpha)$, the replica-symmetric spin–glass solution is stable.

At $\alpha \gg 1$ the asymptotics of both the 'up' and 'down' branches of the AT line are proportional to $\sqrt{\alpha}$. At $n \ll 1$ these asymptotics are: $T_{AT}^{(up)} \simeq (1 - 3n/4)\sqrt{\alpha}$; $T_{AT}^{(down)} \simeq [n/\sqrt{2\log(1/n)}]\sqrt{\alpha}$. As $n \to n^* \simeq 0.3$, the replica-symmetry breaking region moves to infinity and disappears.

Below the line T_f, the system is in the SF phase described above.

4.2. 2/3 < n < 1

In this interval of n, the phase diagram is similar to that for $0 < n < 2/3$ (figure 2) with the only difference being that the retrieval phase is absent here.

4.3. 1 < n < 2

The phase diagram is qualitatively shown in figure 3. Unlike the previous case, two more transition lines are present here: $T_1(\alpha)$ and $T_2(\alpha)$, which intersect at $\alpha^*(n)$. This makes the phase structure of the system rather sophisticated.

In the region marked by P the only stable state is paramagnetic.

In the region marked by P + SG* below $T_1(\alpha)$, *in addition* to the paramagnetic state, the other stable state with *finite* value of $q = q^*(T, \alpha)$ appears.

As the temperature decreases, at $T = T_f$, the value of q^* becomes equal to one and this SG* state turns into the SF one below T_f. Here, in the region P + SF, the paramagnetic phase coexists with the SF state.

Below the line $T_{SG}(\alpha)$ in the region SG + SF the paramagnetic state becomes unstable turning into the usual SG state (via second-order phase transition at T_{SG}). In this region the spin–glass state coexists with the SF state.

At the line $T_2(\alpha)$ (in the interval $0 < \alpha < \alpha_2(n)$), the spin–glass state becomes unstable and first-order phase transition into the SF state takes place.

In the region SG + SG* (in the interval $\alpha_1(n) < \alpha < \alpha_*(n)$), the two spin–glass states coexist: the usual SG state which appears via the second-order phase transition at T_{SG} and the 'special' SG* state which appears with finite value $q = q^*(T, \alpha)$ below the line T_1. In the interval $\alpha_1(n) < \alpha < \alpha_2(n)$ at $T = T_f$, the value of q^* becomes equal to one and the SG* state turns into the SF one. In the interval $\alpha_2(n) < \alpha < \alpha_*(n)$, at the line $T_2(\alpha)$, the 'usual' SG state becomes unstable and down to T_f there is only one stable SG state. At $T = T_f$ the value of q is becoming one and the second-order phase transition into the SF state takes place.

4.4. n = 2

In this marginal case, the saddle-point equations (22) and (23) could be easily reduced to one very simple algebraic equation for q

$$q = \tanh\left(\frac{\alpha q}{(T-1)^2 - q^2}\right). \tag{30}$$

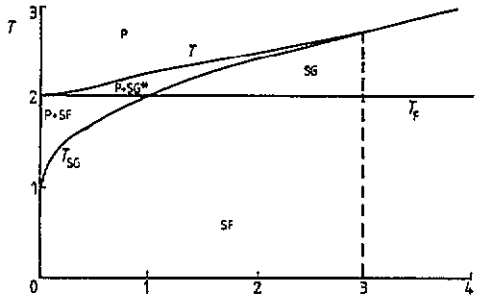

Figure 4. Phase diagram of the model with $n = 2$.

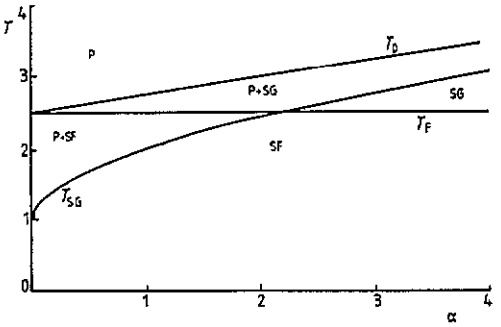

Figure 5. Phase diagram of the model with $n > 2$.

The solution of this equation gives the phase diagram shown in figure 4.

Similar to the $1 < n < 2$ case, in the region marked by P + SG* below the line $T_0(\alpha)$ in addition to the paramagnetic state, the other stable SG state with *finite* value of $q = q^*(T, \alpha)$ appears.

In the interval $0 < \alpha < 1$ at temperature $T_f = 2$, the value of q^* becomes equal to one and the SG* state turns (via a second-order phase transition) into the SF state. In the region P + SF, this SF state co-exists with the paramagnetic one. Below the line $T_{SG} = 1 + \sqrt{\alpha}$, the paramagnetic state becomes unstable and the only stable state is the SF one.

In the interval $1 < \alpha < 3$ below the line T_{SG}, the paramagnetic state becomes unstable and in the temperature region $T_f < T < T_{SG}$, the only stable state is the spin–glass one (which has appeared below $T_0(\alpha)$) with finite value of q. At T_f, again, the value of q becomes equal to one and below T_f the system appears to be in the SF state.

The two lines $T_0(\alpha)$ and $T_{SG}(\alpha)$ meet at $\alpha = 3$. At $\alpha > 3$ there is the usual second-order phase transition from paramagnetic to SG state at $T_{SG} = 1 + \sqrt{\alpha}$.

4.5. $n > 2$

The qualitative phase diagram in this case is shown in figure 5. The phase structure here is similar to the $n = 2$ case discussed above and the only difference is that the lines $T_0(\alpha)$ and $T_{SG}(\alpha)$ never now meet. For large α the two lines becomes asymptotically parallel. A special point here is that, in the limit $n \to 2 + 0$, the distance between the lines $T_0(\alpha)$ and $T_{SG}(\alpha)$ turns to zero at all $\alpha > 3$.

5. Conclusions

We have considered the Hopfield model of neural networks in which the patterns, as well as the spins, are dynamical variables. The characteristic time scales of the dynamics of the spins and the patterns are widely separated. It is assumed that the spins completely equilibrate at the time scale at which the elementary changes in the patterns take place. It is also assumed that the patterns evolve in a sort of self-consistent field created by the spins. We have studied the situation when each kind of the variables thermalizes at different temperatures T and T', respectively.

In the case of a negative value of the temperature T', the model presents some similarities with the unlearning training algorithm [7] which is known to increase the storage capacity due to a reduction in noisy interference effects among the patterns. We have demonstrated

a substantial increase in the size of the retrieval phase in the plane (T, α). In particular, at the zero temperature with n finite and negative, the dynamics of the patterns somehow pushes them towards mutual orthogonalization and this leads to an increase in capacity from the value of 0.14 to 1 [4]. This last value is typical of the 'pseudoinverse learning rule' [9] where the patterns are orthogonalized by hand.

We have also considered the opposite case $n > 0$ when the patterns move to become as parallel as possible. Growing interference among them produces a reduction in the storage capacity and for $n > 2/3$, the retrieval phase was shown to disappear completely. Besides, at low enough temperatures ($T < n$), the system appears to be in the 'superferromagnetic' phase in which the overlaps of the thermodynamic state with *all* the patterns become finite. The complete phase diagram of the model in the space of the parameters T, α and n was obtained. The stability of the obtained replica-symmetric solutions with respect to the replica-symmetry breaking was also studied and the corresponding AT lines both at $n < 0$ and $n > 0$ were calculated.

The principal problem of the present approach is that in neural networks with finite replica parameter $n = T/T'$, the slow dynamical variables are the 'patterns' and not the synaptic couplings themselves (which are constrained to keep the Hebb structure in terms of the moving patterns). In this sense, the system considered here (with n negative) is not quite adequate for the unlearning procedure. Moreover, the situation here is such that the patterns, once they have reached thermal equilibrium, are still free to diffuse and it is not clear what the correlation between the initial patterns one wants to store in the system and those found in it for long times is. Therefore, it would be very interesting to formulate such a system in which the slow dynamical variables would be the *synaptic couplings* (and not the patterns) which are somehow confined in a subspace keeping memory about truly quenched stored patterns.

Acknowledgments

The research described in this publication was supported in part by grant N MSR000 from the International Science Foundation, and by the INTAS grant N 1010-CT93-0027.

References

[1] Mézard M, Parisi G and Virasoro M A 1987 *Spin–Glass Theory and Beyond* (Singapore: World Scientific)
[2] Penney R W, Coolen T, Sherrington D 1993 *J. Phys. A: Math. Gen.* **26** 3681
[3] Sherrington D and Kirkpatrick S 1975 *Phys. Rev. Lett.* **35** 1972
[4] Dotsenko V, Franz S and Mezard M 1994 Partial annealing and overfrustration in the disordered systems *J. Phys. A: Math. Gen.* **27** 2351
[5] Hopfield J J 1882 *Proc. Natl. Acad. Sci. USA* **79** 2554
[6] Wong K Y M and Sherrington D 1993 Neural networks optimally trained with noisy data *Preprint* University of Oxford OUTP-93-16S
[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] Amit D, Sompolinsky H and Gutfreund H 1987 *Ann. Phys.* **173** 30
[9] Personaz L, Guyon I and Dreyfus G 1985 *J. Physique Lett.* **46** L359
Kanter I and Sompolinsky H 1987 *Phys. Rev.* A **35** 380