Thermodynamics of the training process

D. É. Fel'dman

L.D. Landau Institute of Theoretical Physics, Russian Academy of Sciences, 142432 Chernogolovka, Moscow Region, Russia

(Submitted 1 September 1994)
Zh. Éksp. Teor. Fiz. **107**, 880–893 (March 1995)

The training rule for Hopfield neural networks has been investigated with the help of a simple thermodynamic method which allows one to describe both the training stage and the stage of information reconstruction. Both the spins and the synaptic weights are considered to be dynamic variables in contact with heat reservoirs, where the temperatures of the system of spins and the system of weights are assumed to be different. A phase diagram of the system is constructed and the maximum number of patterns is determined which the system can remember for different values of the parameters. It is shown that the storage capacity grows as a result of training. © *1995 American Institute of Physics.*

1. INTRODUCTION

In recent years there has been an upsurge of interest in the Hopfield model of associative memory.[1] From the viewpoint of statistical mechanics this model is described by an Ising Hamiltonian

$$H=-\frac{1}{2}\sum_{i,j}^{N} J_{ij}\sigma_i\sigma_j, \tag{1}$$

in which N spins $\sigma_i=\pm 1$ play the role of the neurons, and the coupling constants J_{ij} play the role of the synaptic weights. In the original model, the synaptic weights were determined by Hebb's training rule,[2] which allows P patterns to be remembered after P step-presentations. Here the maximum storage capacity (i.e., the maximum number of retrievable patterns) is achieved at zero temperature, i.e., in the absence of noise, and the largest number of N-bit words that the system can remember, P_c, turns out to be equal to $\sim 0.14N$ (N is the total number of spins).[3] A number of other training rules were developed later (see, e.g., Refs. 4–9), which made it possible to increase the number of retrievable patterns and reduce the number of errors in their reconstruction.

Hopfield and coworkers proposed an untraining algorithm[4] consisting of successive modifications of the synaptic weights, starting with their Hebbian values

$$J_{ij}(t=0)=J_{ij}^{(H)}\equiv\frac{1}{N}\sum_{\mu}^{P}\xi_i^{\mu}\xi_j^{\mu}, \tag{2}$$

where $\{\xi_i^{\mu}=\pm 1\}$, $\mu=1,\ldots,P\equiv\alpha N$ are the retrievable patterns. At each step the weights change according to the rule

$$J_{ij}(t+1)=J_{ij}(t)+\epsilon\sigma_i^*\sigma_j^*, \tag{3}$$

where the spin values σ_i^* are chosen from one of the minima of the Hamiltonian (1), and ϵ is some small *negative* parameter. Such dynamics leads to the destruction of the potential minima and an increase in the energy of the system. However, it is primarily the shallower spin-glass minima that are annihilated and, as numerical modeling has shown,[8,9] for a moderate number of steps of the algorithm the storage capacity grows.

In this paper we consider an algorithm related to unlearning, which may be called underlearning.[10] At zero temperature, the dynamics of underlearning differs from that of unlearning (3) only in the sign of the parameter ϵ. Such a direction of evolution of the synaptic weights leads to a deepening of the energy valleys. If a spin state is located near one of the patterns $\{\xi_i^{\mu}\}$, then the energy minimum associated with that pattern deepens. As a result, the storage capacity should grow and, as we will see, the retrieval region in the phase diagram increases in comparison with the Hebb case.[3]

We stress that in spite of the similarity of the underlearning and unlearning algorithms, their mechanisms are substantially different. In the former, the effect is attained thanks to a deepening of the "useful" energy minima, whereas in the latter it is attained as a result of destruction of "spurious" energy valleys.

In a system with noise, the underlearning rule can be formulated[10] in the form of the Langevin equation, assuming the dynamics of the weights to be slow in comparison with the dynamics of the spins:

$$\frac{dJ_{ij}(t)}{dt}=\langle\sigma_i\sigma_j\rangle_{J(t),T}+\eta_{ij}(t) \tag{4}$$

or

$$\frac{dJ_{ij}(t)}{dt}=-\frac{\delta}{\delta J_{ij}}F[J(t),T]+\eta_{ij}(t). \tag{5}$$

Here T is the temperature of the spin system, $\langle\ldots\rangle_{J(t),T}$ denotes the thermal average for given values of the synaptic weights J_{ij}, F is the free energy of the spin system

$$F=-T\ln\sum_{\sigma=\pm 1}\exp(-\beta H[J(t),\sigma]), \tag{6}$$

 1063-7761/95/030500-08$10.00

and η_{ij} is thermal noise corresponding to the temperature of the weights $T' \neq T$. The parameter ϵ (3) is absent from Eqs. (4) and (5) thanks to the appropriate choice of units for measuring time.

It is traditional in the construction and analysis of training rules to devote principal attention to the ideal noise-free case, and nonzero temperatures are often not considered at all,[8] for, as it turns out, there exists a maximum temperature above which retrieval is impossible.[3,6] For Hebb's rule the critical temperature $T_c = 1$ (Ref. 3). We will see below that in our algorithm,[10] training is possible, in principle, at any temperature T, and for the appropriate choice of parameters the efficacy of training does not depend on the temperature.

The approach which we shall use for studying systems in which both the spin degrees of freedom and the couplings are in contact with a heat reservoir was proposed in Refs. 11–13. The idea is based on the traditional replica method,[11,14] but the number of replicas n does not tend to zero; rather it is taken to be equal to T/T'. The replica method is used in this approach to calculate the free energy of the system of synaptic weights. This free energy describes the neural network in a state of thermodynamic equilibrium, i.e., at infinite times. Hence it is possible to extract information about the limiting result of carrying out a large number of steps of the algorithm. Hebb's rule for variable patterns $\{\xi_i^\mu\}$ was examined from this point of view in Ref. 15.

In order to examine the result of carrying out a finite number of steps when using the approach proposed in Refs. 11–13, one needs to restrict the possible range of the coupling constants J_{ij}. Toward this end, we add a term to the potential $F[J(t),T]$ in the equation of motion (5) that depends on the distance between the current and initial values of the couplings. It is convenient to choose it in the form

$$W[J] = \frac{NT'}{2J_0^2} \sum_{i<j} (J_{ij} - J_{ij}^{(H)})^2, \tag{7}$$

where N is the total number of spins, T' is the temperature of the system of synaptic weights, and J_0 controls the range of values of the coupling constants J_{ij}. Now the effective Hamiltonian controlling the dynamics of the weights has the form

$$H_J = F[J,T] + W[J]. \tag{8}$$

References 13 and 15 considered an unlearning algorithm[4,8,9] on the basis of the method of Refs. 11–13. A significant increase in the storage capacity was found in Refs. 8 and 9 on the basis of computer modeling. However, in Refs. 13 and 15 the patterns $\{\xi_i^\mu\}$ were assumed to be dynamic variables, whereas in reality the patterns should be fixed and only the synaptic weights should move around. These studies therefore represent only a rough attempt at an analytic explanation of the properties of unlearning. In our treatment the patterns $\{\xi_i^\mu\}$ are fixed, as they should be. This leads to the appearance in the formalism of two replica indices instead of one as in Refs. 11–13, and 15. This more realistic approach allows us to study the underlearning process, and opens up a new approach to the construction of a thermodynamic theory of unlearning.

In the present study we investigate the dependence of the storage capacity on the temperatures T and T', and the parameter J_0 corresponding to the number of steps of the algorithm. In the second section we derive equations for the order parameters of a system with Hamiltonian (8). In the third section we discuss limiting cases. The fourth section is devoted to a construction of the phase diagram, and the fifth section contains the conditions for replica symmetry breaking. The Conclusion gives a qualitative discussion of the results and possibilities for further study.

2. ORDER PARAMETERS

For a fixed realization of disorder $\{\xi\}$, the partition function of the system of weights has the form

$$Z[\xi] = \int DJ_{i<j} \exp\left\{ -\frac{N}{2J_0^2} \sum_{i<j} (J_{ij} - J_{ij}^{(H)})^2 - \frac{F[J]}{T'} \right\}, \tag{9}$$

where F is the free energy of the spins (6). Introducing the notation $n = T/T'$ and introducing n replicas of the spin system, we represent the partition function in the form

$$Z[\xi] = \int DJ_{i<j} \sum_{\sigma^b} \exp\left\{ -\frac{N}{2J_0^2} \sum_{i<j} (J_{ij} - J_{ij}^{(H)})^2 + \beta \sum_{i<j} \sum_{b}^{n} J_{ij} \sigma_i^b \sigma_j^b \right\}, \tag{10}$$

where the index b labels the replicas. In order to find the free energy of the system of weights averaged over the disorder $\{\xi\}$, we again make use of the replica method. But this time we introduce replicas for the synaptic weights and allow their number k to approach zero, as is customary.[11,14] The free energy of the weights is given by

$$f = \lim_{k \to 0} \frac{\langle\langle Z^k \rangle\rangle - 1}{k}. \tag{11}$$

Here $\langle\langle ... \rangle\rangle$ denotes averaging over realizations of the patterns $\{\xi\}$. The spin variables now carry two replica indices, since the index b distinguishes only copies of the system with the same J_{ij}. For the value of the partition function averaged over the disorder $\langle\langle Z^k \rangle\rangle$ we find

$$\langle\langle Z^k \rangle\rangle = \sum_{\xi} \int DJ_{i<j}^a \sum_{\sigma^{ab}} \exp\left\{ -\frac{N}{2J_0^2} \sum_{i<j,a} (J_{ij}^a - J_{ij}^{(H)})^2 + \beta \sum_{i<j} \sum_{a,b} J_{ij}^a \sigma_i^{ab} \sigma_j^{ab} \right\}. \tag{12}$$

In this formula the replica indices run over the values $a = 1,\ldots,k$ $(k \to 0)$ and $b = 1,\ldots,n$ $(n = T/T')$. Standard calculations (see, e.g., Ref. 3) lead to the result

$$\langle\langle Z^k \rangle\rangle = \int Dm_{ab} \int Dq_{ab}^{a'b'} \int Dr_{ab}^{a'b'} \times \exp\{ -\beta N k n f[m_{ab}, q_{ab}^{a'b'}, r_{ab}^{a'b'}] \}, \tag{13}$$

where

$$f[m_{ab},q_{ab}^{a'b'},r_{ab}^{a'b'}]=\frac{1}{2nk}\sum_{a=1}^{k}\sum_{b=1}^{n}m_{ab}^2+\frac{\alpha}{2\beta kn}\mathrm{Tr}\ln(1-\beta\hat{q})$$
$$+\frac{\alpha\beta}{2kn}\sum_{aa'}^{k}\sum_{bb'}^{n}q_{ab}^{a'b'}r_{ab}^{a'b'}$$
$$-\frac{\beta J_0^2}{4kn}\sum_{a=1}^{k}\sum_{b\neq b'}^{n}(q_{ab}^{a'b'})^2$$
$$-\frac{1}{\beta kn}\ln\left[\sum_{\sigma^{ab}}\exp\left\{\beta\sum_{ab}m_{ab}\sigma^{ab}\right.\right.$$
$$\left.\left.+\frac{\alpha\beta^2}{2}\sum_{aa',bb'}r_{ab}^{a'b'}\sigma^{ab}\sigma^{a'b'}\right\}\right] \quad (14)$$

is the replica free energy, the variable $\alpha=P/N$ is the ratio of the number of patterns ξ^μ to the total number of spins. Expression (14) contains three order parameters. The first of these is the overlap with the given pattern ξ^1, which we want to reconstruct:

$$m^{ab}=\frac{1}{N}\sum_{i}^{N}\langle\langle[\langle\sigma_i^{ab}\rangle]\xi_i^{\mu=1}\rangle\rangle. \quad (15)$$

The second is the spin-glass overlap matrix of the various thermodynamic states:

$$q_{ab}^{a'b'}=\frac{1}{N}\left\langle\left\langle\sum_{i}^{N}[\langle\sigma_i^{ab}\sigma_i^{a'b'}\rangle]\right\rangle\right\rangle. \quad (16)$$

The third is defined by the small (in the thermodynamic limit) overlaps with the remaining patterns:

$$r_{ab}^{a'b'}=\frac{1}{\alpha}\sum_{\mu>1}^{P}\left\langle\left\langle\left[\left(\frac{1}{N}\sum_{i}^{N}\langle\sigma_i^{ab}\xi_i^\mu\rangle\right)\right.\right.\right.$$
$$\left.\left.\left.\times\left(\frac{1}{N}\sum_{j}^{N}\langle\sigma_j^{a'b'}\xi_j^\mu\rangle\right)\right]\right\rangle\right\rangle+\delta_{aa'}\frac{J_0^2}{\alpha}q_{ab}^{ab'}. \quad (17)$$

Here $\langle\ldots\rangle$ denotes the thermal average for the case of fixed weights, $[\ldots]$ denotes the average over realizations of the weights for fixed patterns, and $\langle\langle\ldots\rangle\rangle$, as was already noted, is the average over the realizations of the patterns.

In the replica-symmetric approximation

$$q_{ab}^{a'b'}=q(a\neq a'),\quad q_{ab}^{ab'}=Q(b\neq b'),$$
$$r_{ab}^{a'b'}=r(a\neq a'),\quad r_{ab}^{ab'}=R(b\neq b'),\quad m_{ab}=m. \quad (18)$$

The stability of the solution in the form (18) will be investigated in Sec. 5. After substituting this ansatz into the expression for the free energy (14), the latter takes the form

$$f=\frac{1}{2}m^2+\frac{\alpha}{2\beta}\left\{\ln[1-\beta(1-Q)]\right.$$
$$\left.-\frac{\beta q}{1-\beta(1-Q)-\beta'(Q-q)}+\frac{1}{n}\ln\left[1-\frac{\beta'(Q-q)}{1-\beta(1-Q)}\right]\right\}$$
$$-\frac{\beta J_0^2(n-1)}{4}Q^2+\frac{\alpha\beta R}{2}(1-Q)+\frac{\alpha\beta n}{2}(QR-qr)$$
$$-\frac{1}{\beta n}\int\frac{dz}{\sqrt{2\pi}}e^{-z^2/2}$$
$$\times\ln\left[\int dx e^{-x^2/2}(\cosh\beta(m+\sqrt{\alpha r}z+\sqrt{\alpha(R-r)}x))^n\right]. \quad (19)$$

Differentiating the free energy, we arrive at equations for the order parameters. They simplify after making the replacement $R-r\to R$ and have the form

$$r=\frac{q}{(1-C-\beta'(Q-q))^2}, \quad (20)$$

$$R=\frac{Q}{\alpha}J_0^2+\frac{Q-q}{(1-C)(1-C-\beta'(Q-q))}, \quad (21)$$

$$C\equiv\beta(1-Q)=\beta\int Dz\frac{\int Dx\cosh^{n-2}(M)}{\int Dx\cosh^n(M)}, \quad (22)$$

$$q=\int Dz\left(\frac{\int Dx\cosh^n(M)\tanh(M)}{\int Dx\cosh^n(M)}\right)^2, \quad (23)$$

$$m=\int Dz\frac{\int Dx\cosh^n(M)\tanh(M)}{\int Dx\cosh^n(M)}, \quad (24)$$

where we have introduced the notation

$$M=\beta(m+\sqrt{\alpha r}z+\sqrt{\alpha R}x),$$
$$Dz=\frac{e^{-z^2/2}dz}{\sqrt{2\pi}}.$$

Recall that the parameter n is equal to the ratio of the temperatures β'/β.

In terms of the spin variables σ and patterns ξ, the order parameters can be written as

$$Q=\frac{1}{N}\sum_{i}\langle\langle[\langle\sigma_i\rangle^2]\rangle\rangle,\quad q=\frac{1}{N}\sum_{i}\langle\langle[\langle\sigma_i\rangle]^2\rangle\rangle,$$
$$m=\frac{1}{N}\sum_{i}\langle\langle[\langle\sigma_i\rangle\xi_i^1]\rangle\rangle,\quad r=\frac{1}{\alpha}\sum_{\mu>1}\langle\langle[m_\mu]^2\rangle\rangle, \quad (25)$$
$$R-\frac{1}{\alpha}J_0^2Q+r=\frac{1}{\alpha}\sum_{\mu>1}\langle\langle[m_\mu^2]\rangle\rangle,$$

where

$$m_\mu=\frac{1}{N}\sum_{i}^{N}\langle\sigma_i\rangle\xi_i^\mu.$$

Recall once again that we are using three types of averaging: $\langle\ldots\rangle$ is the thermal average over the "fast" spin variables at

temperature T, [...] is the thermal average over the "slow" coupling constants at temperature T', and $\langle\langle ... \rangle\rangle$ is the average over the disorder $\{\xi\}$.

In what follows we will be interested in the case $T'<T$, i.e., $n>1$. In this case the tendency to reach the energy minimum dominates the effect of thermal noise, and we will see from the analysis of Eqs. (20)–(24) that the storage capacity grows.

Below we examine the phase diagram of the system. There arise three types of phases: a paramagnetic phase (PM) with $Q=q=R=r=m=0$, a spin-glass phase (SG) with $Q\neq 0$, and a phase with memory (FM), in which $m\neq 0$. In addition, we will study the question of replica symmetry breaking (RSB).

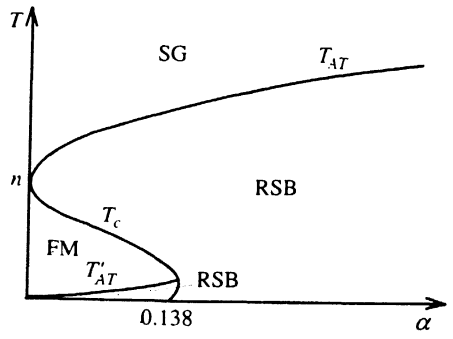

FIG. 1. Phase diagram for $J_0\to\infty$. T_c is the temperature of the transition to the phase with memory. T_{AT} and T'_{AT} are the curves of replica symmetry breaking in the spin-glass phase and in the phase with memory.

3. LIMITING CASES

There is a simply way to construct the phase diagram of the system in the limiting cases $J_0\to\infty$ or $T\to 0$. The first of these corresponds to an infinite number of underlearning steps. Integrating over J_{ij}^{α} in Eq. (12), we find

$$\langle\langle Z^k\rangle\rangle = \text{const} \sum_{\xi} \sum_{\sigma^{ab}} \exp \times \left\{ \frac{\beta^2 J_0^2 N}{2} \sum_{b<b'} \sum_{a} \left(\frac{1}{N} \sum_{i} \sigma_i^{ab} \sigma_i^{ab'} \right)^2 + \beta \sum_{i<j} \sum_{ab} J_{ij}^{(H)} \sigma_i^{ab} \sigma_j^{ab} \right\}. \quad (26)$$

If $\beta^2 J_0^2 \gg \beta$, then the main contribution to the sum over σ in Eq. (26) comes from terms with

$$\left(\frac{1}{N} \sum_{i} \sigma_i^{ab} \sigma_i^{ab'} \right)^2 = 1, \quad (27)$$

i.e., with maximum possible overlap between the replicas (ab) and (ab'). Taking only these terms into account, we obtain

$$\langle\langle Z^k\rangle\rangle = \text{const} \sum_{\xi} \sum_{\sigma^a} \exp\left\{ \beta n \sum_{i<j} \sum_{a} J_{ij}^{(H)} \sigma_i^a \sigma_j^a \right\}. \quad (28)$$

This expression differs from the case of the Hopfield model with Hebbian coupling[3] only in the replacement of β by $n\beta$.

Now, in the limit $J_0\to\infty$, the phase diagram (Fig. 1) is obtained from the phase diagram of the Hopfield model[3] by stretching it by a factor of n along the temperature axis. In particular, the boundary $T_c(\alpha)$ of the retrieval region $(m\neq 0)$ is given by

$$T_c(\alpha) = nT_c^{(H)}(\alpha), \quad (29)$$

where $T_c^{(H)}(\alpha)$ determines the boundary of the retrieval region in the Hopfield model.[3] We see that the dimensions of the retrieval region increase at all temperatures except $T=0$, where the storage capacity does not change. In the limiting case $n\to\infty$, $\alpha_c(T)$ coincides at all temperatures with the critical value in the Hopfield model,[3] $\alpha_c^{(H)}(T)\approx 0.145$. The phase transition curves in the Hopfield model between the paramagnetic and spin-glass states for us correspond to the second-order transition curve between the two spin-glass phases: with $Q=1$ and $q=0$ and with $Q=1$ and $q\neq 0$. [Recall that $Q\equiv 1$ by virtue of Eq. (27).] Its equation has the form

$$T_q^{\alpha} = n(1+\sqrt{\alpha}). \quad (30)$$

Below this curve replica symmetry is broken, i.e., the matrix elements $q_{ab}^{a'b''}$ are different for different values of the indices $a\neq a'$. The condition of replica symmetry breaking in the phase with memory also follows easily by comparing with the Hopfield model.

If we limit ourselves to the region in which replica symmetry is not broken and we apply ansatz (18), then formulas (29) and (30) can be derived on the basis of Eqs. (20)–(24) \for the order parameters. To this end, it is necessary to bear in mind that as $J_0\to\infty$, the order parameter $R\to\infty$ and the parameter $Q\to 1$. Next we must integrate over x in expressions (23) and (24). Then the equations for the order parameters take the same form as in the usual Hopfield model,[3] but with n times lower temperature. We emphasize, however, that the approach that we have used is independent of the degree of replica symmetry breaking.

All of the above considerations are also valid, except for the limit $J_0\to\infty$, at low temperatures, where we again arrive at formula (29).

The second simple limiting case corresponds to $J_0\to 0$. Here we return to the Hopfield model with Hebbian weights, consistent with the sense of the parameter J_0.

4. PHASE DIAGRAM

In this section we investigate the phase diagram in the replica symmetry approximation (18). The typical form of the phase diagram for $n<2$ and $n>2$ is shown qualitatively in Figs. 2(a) and 2(b).

4.1. Spin-glass region

1) $1<n<2$. In order to find the curve of the second-order transition from the paramagnetic phase with $q, Q, r, R, m=0$ to the spin-glass phase with $Q, R\neq 0$, we expand ex-

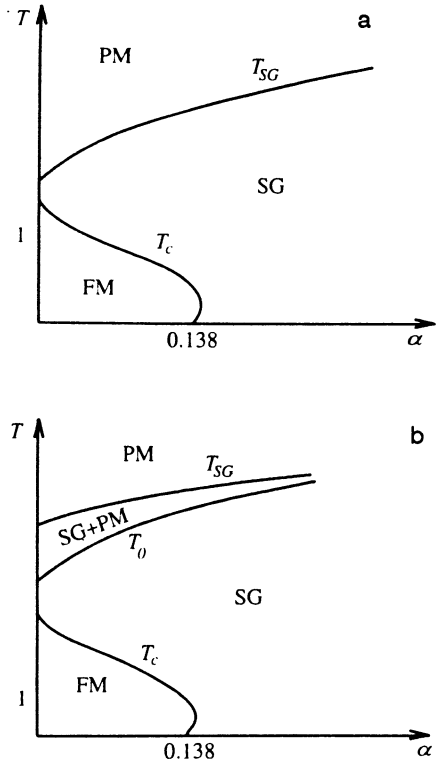

FIG. 2. Typical form of the phase diagram: a) for $n<2$, b) for $n>2$ (for large enough J_0). T_{SG} is the temperature of the transition to the spin-glass state, T_c is the temperature of the transition to the phase with memory. T_0 is the curve at which the paramagnetic state disappears for $n>2$.

pressions (20) and (21) near the transition curve in the small parameters Q and R. As a result, we find for the transition temperature T_{SG}

$$\frac{\alpha}{(T_{SG}-1)^2}+\frac{J_0^2}{T_{SG}^2}=1. \tag{31}$$

For large α the temperature T_{SG} behaves like $\sqrt{\alpha}$. For small α

$$T_{SG}(\alpha\to 0)=1, \quad J_0<1, \tag{32}$$

$$T_{SG}(\alpha\to 0)=J_0, \quad J_0>1. \tag{33}$$

2) $n>2$. Analysis of the stability of the paramagnetic solution with respect to perturbations with $m=0$ (the calculations are analogous to those in Ref. 3) shows that the paramagnetic phase is unstable below the transition curve (31). Here the phase transition to the spin-glass state is first-order and takes place at a higher temperature than T_{SG} from Eq. (31). We will determine this temperature in the limit $\alpha\to 0$. In this limit $\alpha r\to 0$ and $\alpha R\to J_0^2 Q$, and Eqs. (22)–(24) for the order parameters take the form

$$Q(\alpha=0)=1-\frac{\int Dx\,\cosh^{n-2}(\beta m+\beta J_0\sqrt{Q}x)}{\int Dx\,\cosh^{n}(\beta m+J_0\sqrt{Q}x)}, \tag{34}$$

$$m(\alpha=0)=\frac{\int Dx\,\cosh^{n}(\beta m+\beta J_0\sqrt{Q}x)\tanh(\beta m+\beta J_0\sqrt{Q}x)}{\int Dx\,\cosh^{n}(\beta m+\beta J_0\sqrt{Q}x)}, \tag{35}$$

$$q=m^2. \tag{36}$$

In these formulas we have introduced the notation

$$Dx=\frac{e^{-x^2/2}dx}{\sqrt{2\pi}}.$$

In the spin-glass region $m=0$. Therefore Eq. (34) transforms into the equation for the order parameter in the Sherrington–Kirkpatrick model with $n\neq 0$ (Refs. 12, 13, and 16) (with β replaced by βJ_0). Sherrington shows[16] in this model that there is a second-order transition for $n>2$. Numerical results for the temperature $T_{SG}^*(n)$ of this transition and the corresponding values $Q^*(n)$ of the order parameter at the transition point are presented in Ref. 12. For us, the transition temperature $T_{SG}=J_0T_{SG}^*(n)$ corresponds to this temperature. However, our arguments do not apply for all J_0. Indeed, below the transition point, the order parameter is

$$q=m^2=0. \tag{37}$$

At the same time, the expression for the free energy (19) makes sense only for

$$q>\frac{1-T+Q(n-1)}{n}. \tag{38}$$

Conditions (37) and (38) can be fulfilled simultaneously for $T=T_{SG}$ only for

$$J_0>J_0^*\equiv\frac{1+(n-1)Q^*}{T_{SG}^*}. \tag{39}$$

Consequently, the transition from the paramagnetic state to the spin-glass state takes place at the temperature

$$T_{SG}(\alpha\to 0)=J_0T_{SG}^*(n) \tag{40}$$

only for $J_0>J_0^*(n)$. For smaller J_0, the equations (34)–(36) no longer hold, since αr is nonzero. In this case it is possible to assert that

$$T_{SG}(\alpha\to 0)\geq\max(1,J_0). \tag{41}$$

In the limit $\alpha\to\infty$, it is possible to neglect all terms in the expression for the free energy (14) that do not contain α. As a result, we arrive at the free energy of the Hopfield model with Hebbian weights.[3] This means that the transition from the paramagnetic phase to the spin-glass phase takes place at the temperature

$$T_{SG}\approx\sqrt{\alpha}. \tag{42}$$

Note that for $n=1$, the number of components of the order parameters $R_a^{bb'}$ and $Q_a^{bb'}$ vanishes for each a. Therefore these parameters drop out of the expression for the free energy, and we return to the Hopfield model.[1,3]

4.2. Phase with memory

Let us consider two cases: $T\to 0$ and $\alpha\to 0$. The first is studied in a way analogous to the case $J_0\to\infty$. For the transition from the spin-glass phase to the phase with memory, we obtain

$$\alpha_c(T)=\alpha_c(0)+\frac{C_0\alpha_c(0)}{n}T, \tag{43}$$

where $\alpha_c(0)\approx 0.138$ and $C_0\approx 0.18$ are the same constants as in the model with Hebbian weights.[3] We see that for $T=0$, the storage capacity does not change in comparison with the result obtained in Ref. 3.

In the case $\alpha=0$, the order parameters are determined by Eqs. (34) and (35). The very same equations describe the Sherrington–Kirkpatrick model with smeared interactions.[12] It was shown in Refs. 12 and 16 that for $J_0^2<1/(3n-2)$, a second-order transition takes place in this model from the paramagnetic state to the ferromagnetic ($m\neq 0$) state at a temperature

$$T_c=1. \tag{44}$$

For $J_0^2=1/(3n-2)$ the order of the transition changes. At large J_0, the ferromagnetic state appears as a result of a second-order transition from the spin-glass phase. The transition point can be found by expanding expressions (34) and (35) in x. As a result, we find that for $n>2$,

$$T_c=n-4(n-1)\exp\left(-2\frac{n-1}{n^2}J_0^2\right), \tag{45}$$

for $n=2$,

$$T_c=2-2\exp\left(-\frac{J_0^2}{2}\right), \tag{46}$$

and for $1<n<2$,

$$T_c=n-\frac{n(n-1)B\left(1-\frac{n}{2},1-\frac{n}{2}\right)}{\sqrt{2\pi J_0^2}}\exp\left(-\frac{J_0^2}{2}\right), \tag{47}$$

where

$$B(x,y)=\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

is the beta function.

Note that for $J_0^2<1/(3n-2)$, the equation of the transition to the state with memory, at small α, has the same asymptotic form that it has in the Hopfield model:[3]

$$T_c(\alpha)=1-\tau_c(J_0)\sqrt{\alpha}. \tag{48}$$

The expression for $\tau_c J_0$ has the form

$$\tau_c\approx 1.95-5.3(n-1)J_0^2, \quad J_0\to 0, \tag{49}$$

$$\tau_c\approx\frac{2}{3}\frac{3n-2}{\sqrt{n-1}}\sqrt{\frac{1}{3n-2}-J_0^2}, \quad J_0^2\to\frac{1}{3n-2}. \tag{50}$$

The calculation of $\tau_c J_0$ is given in the Appendix.

5. REPLICA SYMMETRY BREAKING

The ansatz (18) corresponds to Parisi structure in the matrices $q_{ab}^{a'b'}$ and $r_{ab}^{a'b'}$, which itself corresponds to replica symmetry breaking in one step. The stability criterion for states with one-step broken replica symmetry was obtained in Ref. 17. It can be written in the form

$$1-\frac{\lambda_i^q\lambda_i^r}{\alpha^2\beta^2}>0, \tag{51}$$

where the eigenvalues λ_i^q and λ_i^r correspond to the four replica modes obtained in Ref. 17. We display these eigenvalues for our model:

$$\lambda_1^q=-\frac{\alpha\beta}{[1-\beta(1-Q)]^2}, \tag{52}$$

$$\lambda_2^q=\lambda_1^q-\frac{\alpha\beta n}{1-\beta(1-Q)}K, \tag{53}$$

$$\lambda_3^q=2\lambda_2^q-\lambda_1^q-n^2\alpha\beta K^2, \tag{54}$$

$$\lambda_4^q=\lambda_1^q-\beta J_0^2, \tag{55}$$

$$\lambda_1^r=-\alpha^2\beta^3\left[1-2Q+\int Dz(\langle\langle\tanh^2(M)\rangle\rangle)^2\right], \tag{56}$$

$$\lambda_2^r=\lambda_1^r-n\alpha^2\beta^3\left[Q-\int Dz(\langle\langle\tanh^2(M)\rangle\rangle)^2-q+\int Dz(\langle\langle\tanh^2(M)\rangle\rangle)(\langle\langle\tanh(M)\rangle\rangle)^2\right], \tag{57}$$

$$\lambda_3^r=2\lambda_2^r-\lambda_1^r-n^2\alpha^2\beta^3\left[\int Dz(\langle\langle\tanh^2(M)\rangle\rangle)^2-2\int Dz(\langle\langle\tanh^2(M)\rangle\rangle)\right](\langle\langle\tanh(M)\rangle\rangle)^2+\int Dz(\langle\langle\tanh(M)\rangle\rangle)^4], \tag{58}$$

$$\lambda_4^r=-\alpha^2\beta^3\int Dz\langle\langle\cosh^{-4}(M)\rangle\rangle, \tag{59}$$

where we have used the notation

$$\langle\langle f(M)\rangle\rangle=\frac{\int Dx\,\cosh^n(M)f(M)}{\int Dx\,\cosh^n(M)},$$

$$M=\beta(m+\sqrt{\alpha r}z+\sqrt{\alpha R}x),$$

$$Dz=\frac{dz}{\sqrt{2\pi}}e^{-z^2/2}$$

and

$$K=\frac{\beta(Q-q)}{[1-\beta(1-Q)][1-\beta(1-Q)-n\beta(Q-q)]}. \tag{60}$$

The values of the order parameters Q, q, R, r, and m in formulas (52)–(60) should be taken from the solutions of the "replica-symmetric" equations (20)–(24).

Analysis of Eqs. (51)–(60) shows that for $\alpha=0$, the replica-symmetric solution for the order parameters is stable

at all temperatures in the phase with memory. Note that the conditions of replica symmetry breaking obtained in Sec. 3 for $J_0^2 \to \infty$ can be reproduced with the help of the general criterion (51)–(60).

6. CONCLUSION

We have examined the phase diagram in the replica-symmetric approximation (18) and derived the conditions of replica symmetry breaking. We gave special consideration to the case $J_0^2 \to \infty$, which corresponds to an infinite number of underlearning steps. In this limit, the phase diagram is obtained from the phase diagram of the Hopfield model by a simple rescaling of the temperature axis. We showed that underlearning leads to an increase in the storage capacity, as one might expect. At the same time, for $T=0$ the storage capacity does not change. This can be understood on the basis of the following arguments. At $T=0$, the spin system finds itself in one of the energy minima of the Hamiltonian (1), and thermal fluctuations are absent. Modification of the interactions according to rule (4) means adding a correction term to the Hamiltonian, whose minimum is reached in the same state in which the system is already found. As a result, the spin system will remain in its previous state at each step of the algorithm. Therefore nonzero overlap of the spin state with any of the patterns $\{\xi^\mu\}$ cannot arise if it was not there before the beginning of underlearning.

It would be interesting to apply the thermodynamic method used here to a study of the unlearning algorithm.[4,8,9] However, in this problem the method should be modified. The point here is that in contrast to underlearning, in unlearning an increase in the storage capacity is not related to a deepening of "ferromagnetic" valleys, but to a destruction of the spin-glass states. Therefore in the corrected approach, the system should find itself in the spin-glass region during training, and in the retrieval region only during retrieval of information.

The author is grateful to V. S. Dotsenko for formulation of the problem and valuable advice, and to E. A. Dorofeev for useful discussions. This work was carried out with the support of the International Scientific Foundation, Grant No. M5R000, and with the support of INTAS, Grant No. 1010-CT93-0027.

APPENDIX

Here we calculate the temperature for the transition to the phase with memory for small α in the region $J_0^2 < 1/(3n-2)$. We expand the free energy (19) out to the second-order term in the small parameter $\sqrt{\alpha}$. As a result we obtain

$$f=\frac{\alpha}{2}\left\{\left(1-\frac{1}{n}\right)\ln[Q^*-\tau]+\frac{1}{n}\ln[Q^*(1-n)+nq^*-\tau]-\frac{q^*}{Q^*(1-n)+nq^*-\tau}\right\}-\frac{3\alpha}{4}\tau y-\frac{\alpha n}{2}\left[\frac{3}{2}yR^*+r^*R^*+\frac{R^{*2}}{2}\right]+\frac{\alpha}{4}\left[\frac{3y^2}{4}+3yr^*+r^{*2}+3yR^*+2r^*R^*+R^{*2}\right]-\frac{nJ_0^2\alpha}{2}Q^*+\frac{\alpha J_0^2}{4}Q^{*2}-\frac{\alpha}{2}Q^*(R^*+r^*)+\frac{\alpha n}{2}(Q^*(R^*+r^*)-q^*r^*). \tag{61}$$

In this formula, in place of the order parameters (18) we have introduced the variables

$$y=\frac{2m^2}{3\sqrt{\alpha}},\quad Q^*=\frac{Q}{\sqrt{\alpha}},\quad q^*=\frac{q}{\sqrt{\alpha}},$$

$$R^*=R\sqrt{\alpha},\quad r^*=r\sqrt{\alpha},\quad \tau=\frac{1-T}{\sqrt{\alpha}}, \tag{62}$$

which are independent of α in the limit $\alpha \to 0$, and the definition of the parameter R is the same as in formulas (20)–(24). The equations for the order parameters (62) take the form

$$r^*=\frac{q^*}{[Q^*(1-n)+nq^*-\tau]^2}, \tag{63}$$

$$R^*=J_0^2Q^*+\frac{Q^*-q^*}{[Q^*-\tau][Q^*(1-n)+nq^*-\tau]}, \tag{64}$$

$$\frac{y}{2}=\tau+R^*(n-1)-r^*, \tag{65}$$

$$q^*=\frac{2y}{3}+r^*, \tag{66}$$

$$Q^*=q^*+R^*. \tag{67}$$

The variables q^*, Q^*, and r^* can be eliminated from these equations. As a result, we obtain

$$\frac{y^3}{2}-\tau^*y^2+y+\tau^*=0, \tag{68}$$

$$R^*\left[1-J_0^2-\frac{1}{y(y+nR^*)}\right]=(\tau^*+y)J_0^2, \tag{69}$$

where

$$\tau^*=\tau+R^*(n-1). \tag{70}$$

In the retrieval region the order parameter $y \neq 0$. Our problem reduces to finding the maximum temperature at which this inequality can still be satisfied. This temperature is given by Eq. (48) with τ_c replaced by the minimum possible value of τ for which $y \neq 0$. In order to find this extreme value of τ, we augment system (68)–(70) by the equation

$$\frac{d\tau}{d\tau^*}=0, \tag{71}$$

or equivalently,

$$\frac{dR^*}{d\tau^*}=\frac{1}{n-1}. \tag{72}$$

The regions $J_0^2 \to 0$ and $J_0^2 \to 1/(3n-2)$ correspond respectively to $\tau^* \to 1.95$ (Ref. 3) and $\tau^* \to \infty$. In these limits it

is possible to obtain approximate expressions for y from Eq. (68). After substituting them in Eq. (69), the combined solution of Eqs. (69) and (72) leads to formulas (49)–(50).

[1] J. J. Hopfield, Proc. Nat'l. Acad. Sci. **79**, 2554 (1982).
[2] D. O. Hebb, *The Organization of Behavior*, Wiley, New York (1949).
[3] D. Amit, H. Gutfreund, and H. Sompolinsky, Ann. Phys. **173**, 30 (1987).
[4] J. J. Hopfield, D. J. Feinstein, and R. G. Palmer, Nature **304**, 158 (1983).
[5] T. O. Kohonen, *Self-Organization and Associative Memory*, Springer, Berlin (1984).
[6] V. S. Dotsenko, N. D. Yarunin, and E. A. Dorofeev, J. Phys. A **24**, 2419 (1991).
[7] S. Diederich and M. Opper, Phys. Rev. Lett. **58**, 949 (1987).
[8] D. Kleinfeld and D. B. Pendergraft, Biophys. J. **51**, 47 (1987).
[9] J. L. van Hemmen, L. B. Ioffe, R. Kühn, and M. Vaas, Physica **163A**, 386 (1989).
[10] V. S. Dotsenko and D. E. Feldman, J. Phys. A **27**, L821 (1994).
[11] V. S. Dotsenko, Usp. Fiz. Nauk **163**, No. 6, 1 (1993) [Phys.-Usp. **36**, 455 (1993)].
[12] R. W. Penney, A. C. C. Coolen, and D. Sherrington, J. Phys. A **26**, 3681 (1993).
[13] V. S. Dotsenko, S. Franz, and M. Mézard, J. Phys. A **27**, 2351 (1994).
[14] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond*, World Scientific, Singapore (1987).
[15] D. E. Feldman and V. S. Dotsenko, J. Phys. A **27**, 4401 (1994).
[16] D. Sherrington, J. Phys. A **13**, 637 (1980).
[17] E. A. Dorofeev, J. Phys. A **25**, 5527 (1992).

Translated by Paul F. Schippnick