Unlearning in the paramagnetic phase of neural network models

1996 J. Phys. A: Math. Gen. 29 3871

(http://iopscience.iop.org/0305-4470/29/14/013)

J. Phys. A: Math. Gen. **29** (1996) 3871–3891. Printed in the UK

Unlearning in the paramagnetic phase of neural network models

Kazuo Nokura†

Shonan Institute of Technology, Fujisawa 251, Japan

Received 30 May 1995, in final form 11 March 1996

Abstract. We propose and study unlearning in the paramagnetic phase of neural network models. More precisely, we introduce the evolution equation of synaptic interactions given by $J_{ij}^{d+1} = J_{ij}^{d} - \bar{\epsilon} S_i^d S_j^d$, where S_i^d, the dth dream, is generated by the paramagnetic dynamics of the J^d model. When $\bar{\epsilon}$ is much smaller than $\overline{\langle S_i S_j \rangle_{J^d}^2}$, we obtain the interactions $J_{ij}^{d+d_0} = J_{ij}^{d} - \epsilon \langle S_i S_j \rangle_{J^d}$ after d_0 dreams, where d_0 is a large integer, ϵ is $\bar{\epsilon} d_0$ which should be smaller than 1 and $\langle S_i S_j \rangle_{J^d}$ is the paramagnetic correlation function of the J^d model. The introduction of the paramagnetic correlation functions opens the possibilities for some analytic studies of unlearning. In this paper, we present two studies about the second equation by using the high-temperature expansion. In the first study, the J^d model is assumed to be the Hopfield model and the signal-to-noise ratio r for the J^{d+d_0} model, which is called the J' model, is studied. r is evaluated to the infinite order of $\beta = T^{-1}$ in the thermodynamic limit, giving affirmative results for paramagnetic unlearning. When the interactional changes are large, the J' model becomes a poor approximation for the resulting models. In the second study, the above equation is regarded as the iterative equation for every d_0 dreams. An expansion rate $\bar{\mu}$ for J_{ij}^d is introduced to control the amplitude of interactions for the large interactional changes. We find that, to the second order of β, the fixed-point interaction is given by the pseudo-inverse type for $\bar{\mu} > \beta\bar{\epsilon}$. For both studies, some results of numerical simulations are presented, which are consistent with the analytic results. Our analytic and numerical studies imply that pattern correlations hidden in the correlation function appear naturally in interactions through paramagnetic unlearning.

1. Introduction

During this decade, many interesting ideas about neural networks have been developed from the point of view of statistical physics. In particular, the studies of associative memory provide us with many interesting insights into the cooperative phenomena of neural networks [1, 2]. Among many neural network models, the Hopfield model is important, since it is simple enough to allow analytic studies and yet it has many interesting aspects as an associative memory.

The Hopfield model is an infinite-range spin model which has interactions prescribed by the Hebb rule. In the Hebb rule, learning is implemented by enforcing synaptic interactions between neurons. That is, when a pattern to be learnt has a datum ξ_i on neuron i and ξ_j on neuron j, the change of the interaction between neuron i and neuron j is assumed to be proportional to $\xi_i \xi_j$. We should note that this rule is local in the sense that an interactional change between two certain neurons is determined only by the temporary data on these two neurons.

† E-mail address: nokura@cosmos.la.shonan-it.ac.jp

0305-4470/96/143871+21$19.50

The ability of an associative memory is mainly characterized by the capacity, the number of patterns the model can memorize, and the quality of retrieval. In this respect, the Hopfield model has several aspects which remain to be improved, i.e. it has a small capacity and many spurious states. These two aspects may be closely connected since many spurious states occupy a large proportion of the configuration space.

Several authors have suggested methods to improve the capacity and quality of retrieval. Among them, the pseudo-inverse model shows perfect retrieval and a remarkable increase of the capacity [3, 4]. However, this model does not satisfy the locality of learning since interactions among neurons are characterized by the pattern correlation matrix. Thus the pseudo-inverse model has been studied mainly for technological interests. If we can find a local evolution rule which brings the Hopfield model into the pseudo-inverse model, it will also become relevant to biological studies of neural networks.

Several years ago, some biologists suggested a very interesting idea which improves the properties of a neural network, that is, unlearning of spurious states [5, 6]. The main idea of unlearning is to destabilize spurious states by the anti-Hebb rule. In this method, a spurious state is found by random shooting and zero-temperature spin dynamics. Some simulations shows that the improvement is really observed by iterations of unlearning [7, 8]. A biological assumption is that this procedure corresponds to rapid eye movement (REM) sleep found widely among mammals and spurious states being unlearned are dreams one sees during sleep. We can find many illuminating observations about REM sleep and neural networks in [5].

In this paper, we propose and study another version of unlearning, which seems more natural from the point of view of statistical physics. That is, we assume that dreams to be unlearned are spin configurations generated by the paramagnetic dynamics of the neural network model. This idea is inspired by learning with thermal noise suggested in [9]. Statistical mechanics tells us that the configurations generated by the paramagnetic dynamics obey the Maxwell–Boltzmann distribution. Thus spurious states or configurations close to them appear very frequently in such dynamics if they have lower energy than the embedded patterns. Therefore we expect that an effect similar to unlearning by random shooting also appears in our version. In addition, we can study the resulting models by using the standard methods of statistical mechanics.

In section 2, we introduce unlearning with a finite temperature and describe the formulation in terms of paramagnetic correlation functions. In section 3, the signal-to-noise analysis of the approximated model, which we call the J' mode, is presented. To do this study, we need some results about the high-temperature expansion of the Hopfield correlation function, which is discussed in appendix A. Some numerical results about the J' model are presented in section 4. In section 5, we discuss some generalizations of the iterative equation especially to treat large changes of interactions. Section 6 is devoted to some discussions.

2. Unlearning in the paramagnetic phase

In this section, we define unlearning at a finite temperature and describe the formulation in terms of the correlation function. Formally, unlearning at a finite temperature is achieved by replacing spurious states to be unlearned with configurations generated by the finite-temperature dynamics. We will show that, when the number of dreams is large enough, and interactional changes are small enough, total interactional changes can be described by the paramagnetic correlation function of the initial model. These conditions impose an upper bound on the interactional change of each unlearning step.

Let us first introduce some notations and describe the basic properties of the Hopfield model. P memorized patterns are given by random quenched variables $\xi_i^\mu = \pm 1$, where $\mu = 1, 2, \ldots, P$ is a pattern index and $i = 1, 2, \ldots, N$ is a site index. A site in this paper means a neuron. The Hopfield model is described by the Hamiltonian

$$H = -\tfrac{1}{2} \sum_{i \neq j} J_{ij} S_i S_j \tag{1}$$

where interactions J_{ij} are defined by

$$J_{ij} = \frac{1}{N} \sum_{\mu} \xi_i^\mu \xi_j^\mu . \tag{2}$$

S_i are Ising spin variables which take ± 1. In this paper, we concentrate on uncorrelated patterns, i.e. ξ_i^μ are quenched variables which take ± 1 with probability $\frac{1}{2}$.

The ratio $\alpha = P/N$ is an important parameter which measures how much the system is loaded with memories. For small enough α, the model works as an associative memory, that is, spin configurations close enough to a certain pattern evolve into that pattern under suitable spin dynamics. The thermodynamic study of the model gave us some ideas about states of the model [2]. It was shown that the phase diagram is rather complicated in the space of α and temperature T. For $\alpha < \alpha_{\mathrm{C}} \sim 0.14$, the model shows three phases: paramagnetic, spin-glass and retrieval phase as the temperature T decreases. The spin-glass transition takes place at $T = T_{\mathrm{g}} \equiv 1 + \sqrt{\alpha}$. The retrieval phase appears at much lower T. When T is equal or close to zero, the model has spin-glass states, mixed states and retrieval states. Mixed states are the mixture of embedded patterns. The spin-glass states have the lowest energy for $\alpha > 0.05$. For $\alpha > \alpha_{\mathrm{C}}$, there is no retrieval state even as a metastable state, yet spin-glass states remain. We call mixed states and spin-glass states together spurious states. The existence of too many spurious states is not desirable for an associative memory. In addition, rather high T_{g} implies that spin-glass states dominate the configuration space.

The main idea of unlearning is to destabilize these states by the following procedure. Let us take the system which is loaded with too many patterns to work as an associative memory. Imagine S_i are set to be random, that is, random shooting. With $T = 0$ spin dynamics, the random configuration evolves into a fixed point η_i which usually has little correlation with embedded patterns. This means that the dynamics found a spurious state. This state certainly corresponds to a dream if we assume that unlearning really happens during REM sleep [8]. Thus, by this analogy, the configurations to be unlearned are also called a dream. Then the system unlearns this dream by the replacement, $J'_{ij} = J_{ij} - \bar{\epsilon} \eta_i \eta_j$, where $\bar{\epsilon}$ is a small positive constant, which we call a unlearning parameter. Some simulations shows that the improvement of the capacity and retrieval quality are really observed by the iteration of this procedure if $\bar{\epsilon}$ is properly chosen. We should note here that random shooting corresponds to $T = \infty$ spin dynamics if we assume that spins are always driven by some finite-temperature dynamics. We call the scheme described here random shooting (RS) unlearning.

Our suggestion in this paper is that the dreams to be unlearned are simply generated by the paramagnetic dynamics of the neural network. This assumption means that η_i are replaced by spin configurations S_i generated by the high-temperature dynamics. Here we do not discuss a biological origin of the paramagnetic dynamics. Ordinary stochastic dynamics characterized by a temperature is sufficient for our argument. We call this scheme paramagnetic (PM) unlearning.

Let us assume that S_i are generated by Monte Carlo (MC) dynamics [10] with a temperature T higher than T_{g} and we do unlearning of S_i for every MC step. Then the dth

dream is given by

$$S_i^d = \begin{cases} S_i^{d-1} & \text{Prob } 1 - p(\beta \Delta H_d(i)) \\ -S_i^{d-1} & \text{Prob } p(\beta \Delta H_d(i)) \end{cases} \tag{3}$$

for all i asynchronously, and interactions are modified to

$$J_{ij}^{d+1} = J_{ij}^d - \bar{\epsilon} S_i^d S_j^d \tag{4}$$

after a whole sweep of the system, where $\beta = T^{-1}$ and $\Delta H_d(i)$ is the change of the dth model energy caused by the spin flip on site i, which is given by $2S_i^{d-1} \sum_{j \neq i} J_{ij}^d S_j^{d-1}$. $p(\beta \Delta H_d(i))$ is a probability of the flip of spin i of the paramagnetic dynamics. We assume that d starts from 1, S_i^0 is a random initial configuration and J_{ij}^1 is the Hopfield interaction. We take T which is not close to T_g. The explicit form of $p(x)$ depends on the type of MC simulation. $\bar{\epsilon}$ should be proportional to $1/N$. Below, we will give an argument about the suitable value of $\bar{\epsilon}$ for PM unlearning.

The evolution equation defines the sequence of models stochastically. The dth model is called the J^d model. To see what the resulting models will be, let us first study the relation between the J^d model and the J^{d+d_0} model. This relation is formally given by

$$J_{ij}^{d+d_0} = J_{ij}^d - \bar{\epsilon} \sum_{d'=d}^{d+d_0-1} S_i^{d'} S_j^{d'}. \tag{5}$$

Let us assume that the change of the interactions is small enough. This means that $\epsilon \equiv \bar{\epsilon} d_0$ is small enough. Then we can approximate the spin configurations in the sum to the ones generated by the J^d model. Further, in the $d_0 \to \infty$ limit, the time average reduces to the thermal average with the Maxwell–Boltzmann distribution of the J^d model. Therefore, to the first order of ϵ, the J^{d+d_0} model has interactions given by

$$J_{ij}^{d+d_0} = J_{ij}^d - \epsilon \langle S_i S_j \rangle_{J^d} \tag{6}$$

where $\langle S_i S_j \rangle_{J^d}$ is a paramagnetic correlation function of the J^d model, which is defined by

$$\langle S_i S_j \rangle_{J^d} = \sum_{\{S\}} S_i S_j \exp(-\beta H_d)/Z \tag{7}$$

where $Z = \sum_{\{S\}} \exp(-\beta H_d)$. The summation $\sum_{\{S\}}$ is over spin configurations. H_d is the energy function of the J^d model. Since J_{ij}^d and $\langle S_i S_j \rangle_{J^d}$ are the same order of magnitude in (6) (see below), ϵ should be some positive constant much smaller than 1.

In the above argument, $\bar{\epsilon}$ is assumed to be small enough. The value of suitable $\bar{\epsilon}$ is important especially in numerical simulations. Let us clarify the condition that $\bar{\epsilon}$ is small enough. The second term of (5) is a sum of ± 1 which are nearly random. In general, a random sequence of ± 1 of d_0 length has an average of order $1/\sqrt{d_0}$. This value should be much smaller than $\langle S_i S_j \rangle_{J^d}$ to have correlation effects in the sum. Then we get the condition $1/d_0 \ll \overline{(\langle S_i S_j \rangle_{J^d})^2}$, where $\overline{\cdots}$ is the ξ average of $\cdots$. Therefore we obtain

$$\bar{\epsilon} \ll \overline{(\langle S_i S_j \rangle_{J^d})^2} \tag{8}$$

where ϵ was set to 1 since it is irrelevant in this inequality. The right-hand side can be estimated by the high-temperature expansion. If we take the first-order term of β, $\overline{(\langle S_i S_j \rangle_{J^d})^2}$ is given by $\beta^2 \overline{(J_{ij}^d)^2} \sim \beta^2 J_0^2/N$, where J_0 is of order 1. It is interesting that the value suggested for RS unlearning [8] satisfies the condition with a moderate β, although their scheme is different from ours.

We study two versions of equation (6) in this paper. The first is that, when d is set to be 1, equation (6) is regarded as a definition of the model which appears after d_0 dreams

starting from the Hopfield model. This model is called the J' model. In this case, we can study the correlation function explicitly since it is defined by the Hopfield model. Sections 3 and 4 are devoted to the study of the J' model. When ϵ becomes large, the J' model will be a poor approximation of the resulting models. In the second study, equation (6) is regarded as an iterative equation for every d_0 dreams. In this method, the large change of interactions can be treated. However, when J_{ij}^d changes greatly, we should be careful about the magnitude of J_{ij}^d since our arguments are based upon the paramagnetic phase of neural networks. In addition to this, the inequality (8) can be violated when J_{ij}^d become too small. In section 5, to avoid this problem, we introduce the expansion rate for J_{ij}^d to control their amplitudes and study the generalized iterative equation which shows better performance.

Our discussion so far is rather formal. Everything is hidden in the paramagnetic correlation function. In the next section, we discuss the signal-to-noise analysis of the J' model by using the high-temperature expansion of $\langle S_i S_j \rangle_J$.

3. Signal-to-noise analysis of the J' model

In this section, we discuss the signal-to-noise analysis of the J' model defined by the interactions

$$J'_{ij} = J_{ij} - \epsilon \langle S_i S_j \rangle \tag{9}$$

where $\langle S_i S_j \rangle$ is the paramagnetic correlation function of the Hopfield model. This model is a special case of (6). The main concern in this section is whether the second term in (9) really improves the signal-to-noise ratio of the Hopfield model or not.

To begin with, we describe the high-temperature expansion of $\langle S_i S_j \rangle$. Following the common procedure, it can be expanded in terms of $\tanh \beta J_{ij} \sim \beta J_{ij}$, where the higher-order terms of βJ_{ij} are dropped since $J_{ij} \sim 1/\sqrt{N}$. The result is given formally by

$$\langle S_i S_j \rangle = \beta J_{ij} + \beta^2 \sum{}' J_{ik} J_{kj} + \beta^3 \sum{}' J_{ik} J_{kl} J_{lj} + \cdots . \tag{10}$$

Each term is represented by a diagram which has vertices for sites and edges for interactions. In the sum $\sum'$, no two indices are equal to each other since a loop of edges should be factorized to cancel the denominator Z. This point is important for the Hopfield model since the number of loops is relevant for ξ averages (see appendix A). The spin-glass transition temperature is given by the point at which $\overline{\langle S_i S_j \rangle^2}$ diverges. In appendix A, we describe the evaluation using diagrammatic representations. The result is

$$\overline{\langle S_i S_j \rangle^2} = \frac{1}{N} \frac{A}{1 - A} \tag{11}$$

where

$$A \equiv \frac{\alpha \beta^2}{(1 - \beta)^2} . \tag{12}$$

$\overline{\langle S_i S_j \rangle^2}$ diverges at $T = 1 \pm \sqrt{\alpha}$. The higher temperature should be adopted as the transition point. In this way, we obtain $T_g = 1 + \sqrt{\alpha}$, which is the same result as the replica method. At the end of this section, we obtain another derivation of (11) as a by-product of the signal-to-noise analysis.

Now we discuss the signal-to-noise analysis of the J' model. Let us study the stability of pattern 1. The local field on site i for this configuration is given by

$$\begin{aligned} h_i &= \sum_{j\neq i} J'_{ij}\xi_j^1 \\ &= \sum_{j\neq i} J_{ij}\xi_j^1 - \epsilon \sum_{j\neq i}\langle S_i S_j\rangle \xi_j^1. \end{aligned} \tag{13}$$

If $\xi_i^1 \times h_i$ is positive for all or almost all sites, pattern 1 is expected to be stable. To see this, h_i is decomposed into a signal part h_{s} which is proportional to ξ_i^1 and a noise part h_{n} which is not correlated with ξ_i^1. The first term in (13), i.e. the local field of the Hopfield model, is decomposed into

$$\sum_{j\neq i} J_{ij}\xi_j^1 = \xi_i^1 + \sum_{j\neq i,\mu\neq 1} j_{ij}^{\mu}\xi_j^1 \tag{14}$$

where $j_{ij}^{\mu} = \xi_i^{\mu}\xi_j^{\mu}/N$. The first term is a signal and the second term is a noise, which we call a Hopfield noise. The signal-to-noise ratio $|h_{\mathrm{n}}/h_{\mathrm{s}}|$ in this case becomes $\sqrt{\alpha}$ in the $N\to\infty$ limit, which is small for small α. Thus the local fields are parallel with ξ_i^1 for almost all sites for small enough α. For the Hopfield model, it was shown by the replica method that the upper limit of the signal-to-noise ratio which allows the retrieval phase is $\sqrt{\alpha_{\mathrm{C}}}\sim 0.37$. We take this value as a reference for the J' model for the possible retrieval phase.

Let us now study the second term of (13). According to the high-temperature expansion, the coefficient of β^n is given by

$$\sum_{j\neq i}{\sum_{}}' J_{ik}J_{kl}\cdots J_{lj}\xi_j^1 \tag{15}$$

where each term is a product of n Js. In this expression, there are signal terms, Hopfield noise terms and other kinds of noise terms which are absent in the Hopfield model. To be specific, let us concentrate on the second-order terms, which become

$$\sum_{j\neq i}\sum_{k\neq i,j} J_{ik}J_{kj}\xi_j^1 = \sum_{j\neq i}\sum_{k\neq i,j}\sum_{\mu}\sum_{\nu} j_{ik}^{\mu}j_{kj}^{\nu}\xi_j^1 \tag{16}$$

after putting $J_{kl} = \sum_{\mu} j_{kl}^{\mu}$. Note here that the site indices are all different, while the pattern indices are free from any restriction. If we neglect the restriction on the site indices, the correlation matrix $C^{\mu\nu} \equiv \sum_k \xi_k^{\mu}\xi_k^{\nu}/N$ appears after the k sum. This fact was found in [9] in a different context. Thus it is natural to expect that this term changes the signal-to-noise ratio of the Hopfield model.

To find the signal part and the noise part of (16), we should group the terms in (16) according to the correlation under an ξ-average. The number of different pattern and site indices is a good guide for this purpose. Let us concentrate on the pattern indices. The formulae $\sum_k j_{ik}^a j_{kl}^a = j_{il}^a$ and $\sum_k j_{ik}^a \xi_k^a = \xi_i^a$, which are valid for $N\to\infty$, are convenient in the following evaluation. The contribution to the signal part is given by the term $\mu=\nu=1$, giving ξ_i^1, whereas the noise part is decomposed into several uncorrelated elements. The Hopfield noise comes from $\mu=\nu\neq 1$ and $\mu\neq\nu=1$, which make $2\times\sum_{j\neq i,\mu\neq 1} j_{ij}^{\mu}\xi_j^1$. We should note that each contribution corresponds to a position at which pattern indices switch when we follow the expression (16) from one end to another. There are other types of noise, the term with $\mu\neq\nu\neq 1$, where μ can be 1. In general, we can also group the higher-order terms according to the number of positions the pattern indices switch.

These observations imply that after putting $J_{kl} = \sum_\mu j_{kl}^\mu$ in equation (15), terms are categorized according to the number of positions where the sequence of pattern indices changes. In this way we reach the expression for the nth order term

$$\sum_{j\neq i}{\sum}' J_{ik}J_{kl}\cdots J_{zj}\xi_j^1 = \xi_i^1 + \sum_{p=1}^{n}\binom{n}{p}\sum_{j\neq i}{\sum}'' j_{ij}^{(p)}\xi_j^1 \tag{17}$$

where we have introduced the abbreviation $j_{ij}^{(p)} = j_{ik}^\mu j_{kl}^\nu \cdots j_{zj}^\eta$. The indices other than i or j are dropped in $j_{ij}^{(p)}$ since they always appear in the sum $\sum''$. The sum $\sum''$ means that all site indices are different and two neighbouring pattern indices are also different. As we discuss in appendix B, $\sum'' j_{ij}^{(p)}\xi_j^1$ of different p are not correlated to each other in the $N\to\infty$ limit. In this sense, they work like a set of basis functions on ξ-space. Using this expression, we obtain

$$\begin{aligned}\sum_{j\neq i}\langle S_iS_j\rangle\xi_j &= \left(\sum_{n=1}^{\infty}\beta^n\right)\xi_i^1 + \sum_{p=1}^{\infty}\left(\sum_{n=p}^{\infty}\binom{n}{p}\beta^n\right)\sum_{j\neq i}{\sum}'' j_{ij}^{(p)}\xi_j^1 \\ &= \frac{\beta}{1-\beta}\xi_i^1 + \sum_{p=1}^{\infty}\frac{\beta^p}{(1-\beta)^{p+1}}\sum_{j\neq i}{\sum}'' j_{ij}^{(p)}\xi_j^1.\end{aligned} \tag{18}$$

In this expression, the signal term has a coefficient $\beta/(1-\beta)$, while the Hopfield noise term, the term with $p=1$, has a coefficient $\beta/(1-\beta)^2$. That is, their ratio $1/(1-\beta)$ is different from 1. This is the reason why the correlation function changes the signal-to-noise ratio of the Hopfield model. The amplitudes of other types of noise are evaluated by using the formula

$$\overline{\left(\sum_{j\neq i}{\sum}'' j_{ij}^{(p)}\xi_j^1\right)\left(\sum_{j'\neq i}{\sum}'' j_{ij'}^{(p')}\xi_{j'}^1\right)} = \delta_{pp'}\alpha^p \tag{19}$$

which is valid in the $N\to\infty$ limit, where $\delta_{pp'}$ is a Kronecker delta. See appendix B for a derivation. Using this formula, we finally obtain the expression

$$h_i = \sum_{j\neq i}J'_{ij}\xi_j^1 = h_{\rm s}\xi_i^1 + h_{\rm n} \tag{20}$$

where

$$h_{\rm s} = 1 - \frac{\epsilon\beta}{1-\beta} \tag{21}$$

$$|h_{\rm n}| \sim \sqrt{\left(1-\frac{\epsilon\beta}{(1-\beta)^2}\right)^2\alpha + \frac{\epsilon^2}{(1-\beta)^2}\frac{A^2}{1-A}}. \tag{22}$$

The ratio $r \equiv |h_{\rm n}/h_{\rm s}|$ has a minimum at some positive ϵ since $|h_{\rm n}|$ decreases more rapidly than $h_{\rm s}$ as ϵ increases from zero. The study of the minimum of r is straightforward. We discuss this point in the next section.

To conclude this section, we sketch another derivation of $\overline{\langle S_iS_j\rangle^2}$. By following the same procedure as above, $\langle S_iS_j\rangle$ is written in the form

$$\langle S_iS_j\rangle = \sum_{p=1}^{\infty}\frac{\beta^p}{(1-\beta)^p}{\sum}'' j_{ij}^{(p)}. \tag{23}$$

Using the relation $\overline{(\sum'' j_{ij}^{(p)})(\sum'' j_{ij}^{(p')})} = \delta_{pp'}\alpha^p/N$, which is derived in appendix B, we obtain the same result as in appendix A.

In the next section, we discuss the behaviour of r and compare it with the results of computer simulations.

4. Numerical study of the J' model

In this section, we study the behaviour of r and present some numerical results of simulations of the J' model. Let us first study the value of ϵ which minimizes r. By differentiating r^2 with respect to ϵ and solving the resulting equation, we find that r takes the minimum

$$r_m = \frac{A\sqrt{\alpha + (1-\beta)^2(1-A)}}{\beta^2 + A - \beta^2 A} \tag{24}$$

at

$$\epsilon_m = \frac{(1-\beta)^2(1-A)}{1-(1-\beta)(1-A)}. \tag{25}$$

r_m decreases as β decreases and tends to the smallest value $\alpha/\sqrt{1+\alpha}$ when $\beta \to 0$. In this limit, ϵ_m tends to $(1/\beta) - (2+\alpha)$ and the interactions of the J' model reduce to

$$\overline{J}_{ij} = \beta\left\{(2+\alpha)J_{ij} - \sum_{k\neq i,j} J_{ik}J_{kj}\right\}. \tag{26}$$

The reason that a small β gives a small r_m is that the factor $A^2/(1-A)$ of the second term in $|h_{\mathrm{n}}|^2$ becomes rather small for small β. If we adopt $\sqrt{\alpha_{\mathrm{C}}} \sim 0.37$ as a critical value of r, the best critical capacity α'_{C} of the J' model is determined by the equation $\alpha/\sqrt{1+\alpha} = \sqrt{\alpha_{\mathrm{C}}}$. This yields $\alpha'_{\mathrm{C}} \sim 0.42$, which is about three times the critical capacity of the Hopfield model. In the computer simulations, however, too small β is not desirable since $\langle S_i S_j\rangle$ becomes very small and more than $(\overline{\langle S_i S_j\rangle^2})^{-1} \sim N/A \sim N/(\alpha\beta^2)$ MC steps are required to have a correlation effect as discussed in section 2.

Two kinds of simulation are presented here. In the first case, the J' model is explicitly made by evaluating $\langle S_i S_j\rangle$ by MC simulations. In the second case, the iterative equation for J^d_{ij} suggested in section 2 is simulated directly. For spin dynamics, the Metropolis function, $p(x) = 1$ for $x \leqslant 0$ and $p(x) = \exp(-x)$ for $x > 0$, was used. In both cases, we want to know whether the embedded patterns are stable or not. This stability is a necessary condition for associative memory. Thus $m^\mu = \sum \xi^\mu_i \sigma^\mu_i / N$, where σ^μ_i is obtained by $T = 0$ spin dynamics starting from ξ^μ_i, is evaluated for every some MC steps. The average of m^μ over patterns and some samples is denoted by m.

Let us study the first case. To be specific, we mainly discuss the simulations for $(\alpha, \beta) = (0.2, 0.5)$. These values give $\epsilon_m = \frac{1}{3}$ and $r_m = \sqrt{0.1} \sim 0.32$, which is smaller than $\sqrt{\alpha_{\mathrm{C}}} \sim 0.37$. The small value of ϵ_m is desirable if we want the J' model to be the approximation of the original iterative equation. System size N is 200. For these parameters, N/A is of order 10^3. The MC steps d_0 should be much larger than this value. To make the J' model, $\langle S_i S_j\rangle$ were numerically obtained by the Monte Carlo simulation with mainly 10^5 MC steps. We found that, for these values of parameters, $\langle S_i S_j\rangle^2$ takes a value close to (11). After obtaining $\langle S_i S_j\rangle$, the interaction $J'_{ij} = J_{ij} - \epsilon\langle S_i S_j\rangle$ is assigned to the Hamiltonian. Figure 1 shows the ϵ dependence of m for $(\alpha, \beta) = (0.2, 0.5)$ with the signal-to-noise ratio r. Figure 2 shows the results for $(0.3, 0.3)$ with various MC steps. The maximum of m clearly tends to ϵ_m as the number of MC steps increases. Note that $N/A \sim 10^4$ for this (α, β), which is the lower bound of MC steps to have correlation effects in the sum (5). This explains the improvement from 10^4 to 10^5 MC steps. We suppose that some amount of random noise is induced in $\langle S_i S_j\rangle$ by the numerical evaluation with finite d_0. For example, we will get a maximum of m near $\epsilon \sim 0$ if the MC steps are too small to give the thermal average. This explains the reason why the maxima of m are always placed at ϵ smaller than ϵ_m and they shift to ϵ_m as the number of MC steps increases in

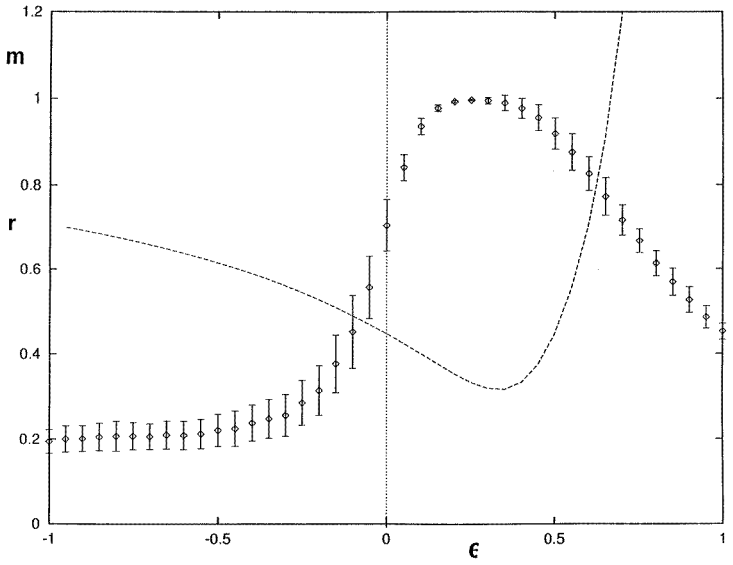

Figure 1. The overlap m evaluated for the J' model with $(\alpha, \beta) = (0.2, 0.5)$ and $N = 200$. The signal-to-noise ratio $r = |h_{\rm n}/h_{\rm s}|$ is also depicted by a broken curve. r takes the minimum $r_m = \sqrt{0.1} \sim 0.32$ at $\epsilon_m = \frac{1}{3}$ for these parameters. The averages m and their sample fluctuations, which are denoted by error bars, are evaluated for 10 samples.

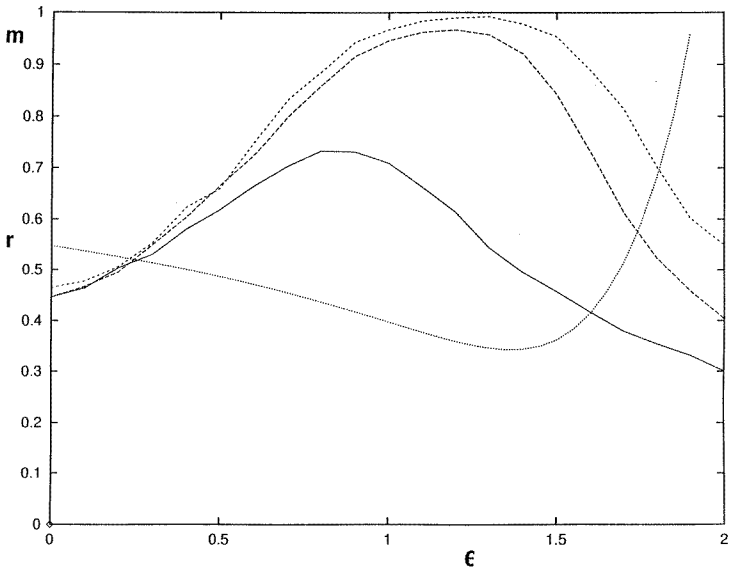

Figure 2. The overlap m evaluated for the J' model with $(\alpha, \beta) = (0.3, 0.3)$ and the signal-to-noise ratio $r = |h_{\rm n}/h_{\rm s}|$ which is depicted by a dotted curve. $N = 200$. Each curve of m corresponds to a different number of MC steps to evaluate $\langle S_i S_j \rangle$, which are 10^4, 10^5 (average of 10 samples), and 10^6 (one sample) MC steps from the bottom. r takes the minimum $r_m \sim 0.34$ at $\epsilon_m \sim 1.37$.

figure 2. Except for this aspect, the numerical results seem to be in good agreement with the signal-to-noise ratio r.

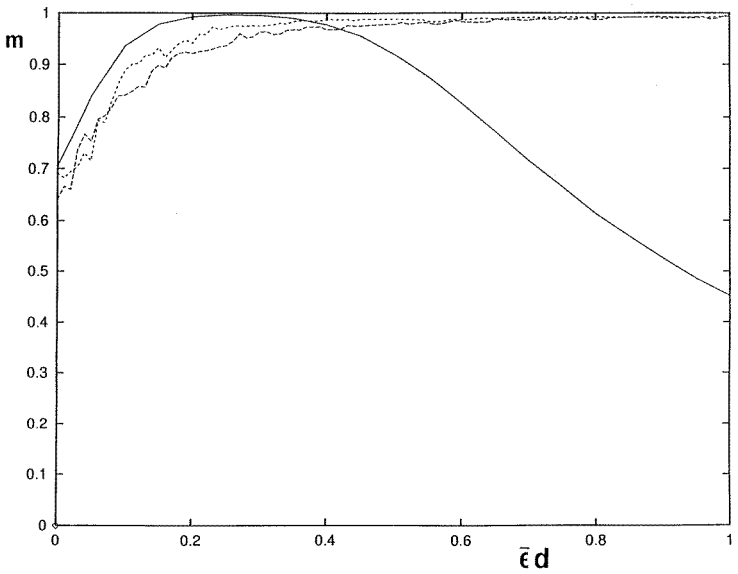

Figure 3. The evolution of m evaluated by the iterative equation of the J^d model for two samples. $N = 200$. The full curve shows the same m as in figure 1, which are connected by lines to guide the eye. In the case of the J^d model, the horizontal axis means $t \equiv \bar{\epsilon}d$, where $\bar{\epsilon} = 10^{-5}$.

In the second case, the original iterative equation (4) was simulated for $(\alpha, \beta) = (0.2, 0.5)$. We adopted $\bar{\epsilon} = 10^{-5}$, which is about the inverse of the MC steps of the first case. The numerical results for m are presented in figure 3 with the result of the first case. Note that the horizontal axis for the second case means $\epsilon \equiv \bar{\epsilon}d$. For ϵ smaller than 0.4, the two simulations give a similar behaviour of m as expected. The case of $(0.3, 0.3)$ was also studied in the same way. In this case, m also keeps increasing but shows poor results around the minimum of r since ϵ_m is rather large.

When α is not so small or when a better retrieval property is demanded, we should study a rather large change of interactions. This requires a lot more MC steps for the iterative equation. In the next section, we will discuss this situation.

5. Study for large interactional changes

This section is devoted to the study of large interactional changes caused by PM unlearning. When we have to deal with large interactional changes, the amplitude of interactions becomes an important problem. That is, if it gets smaller and smaller as unlearning proceeds, the upper bound on $\bar{\epsilon}$ in section 2 will be violated since $\overline{(\langle S_i S_j \rangle_{J^d})^2} \sim \beta^2 \overline{(J_{ij}^d)^2}$ becomes very small. On the other hand, when interactions get larger and larger, our scheme of unlearning breaks down since the spin configuration will be trapped in a certain valley of the energy function. Here we suggest introducing the expansion rate $\bar{\mu}$ for interactions to control their amplitude. Then the iterative equation which is studied in this section is given by

$$J_{ij}^{d+1} = (1 + \bar{\mu}) J_{ij}^d - \bar{\epsilon} S_i^d S_j^d . \tag{27}$$

S_i^d obey the same dynamics as in section 2. We expect that the expansion rate $\bar{\mu}$ will balance the contribution from the second terms.

5.1. Study of the iterative equation

Let us discuss the J^{d+d_0} model defined by (27) using the formulation introduced in section 2. The situation here is a bit different since $\bar{\mu}$ induces terms like $(1+\bar{\mu})^{d'-d} S_i^{d'} S_j^{d'}$. However, the factors $(1+\bar{\mu})^{d'-d} - 1 \sim \bar{\mu}(d'-d)$ only give next-order corrections to the correlation function. Thus we can set $\bar{\mu} = 0$ in the sum over paramagnetic configurations. Therefore in the large-d_0 limit with a small fixed $\Delta t \equiv \bar{\epsilon} d_0$, we obtain

$$J_{ij}(t+\Delta t) = (1+\theta \Delta t) J_{ij}(t) - \Delta t \langle S_i S_j \rangle_{J(t)} \tag{28}$$

to the first order of Δt, where we have introduced the time variable $t \equiv \bar{\epsilon} d$ and the ratio $\theta \equiv \overline{\mu}/\bar{\epsilon}$.

Let us study this equation by high-temperature expansion. In this paper, to make the arguments as simple as possible, we restrict ourselves to the second order of β, which gives the first non-trivial effect. The studies including higher-order terms will be done in a similar way. To the second order of β, we obtain

$$J_{ij}(t+\Delta t) = (1+\beta\delta\Delta t) J_{ij}(t) - \Delta t \beta^2 \sum_{k \neq i,j} J_{ik}(t) J_{kj}(t) \tag{29}$$

where $\delta \equiv (\theta - \beta)/\beta$. The fixed point is given by setting $J_{ij}(t+\Delta t) = J_{ij}(t) \equiv J_{ij}^{\mathrm{F}}$. If the two terms with $k = i, j$ are added and subtracted in the sum and assuming $J_{ii}^{\mathrm{F}} = J_{jj}^{\mathrm{F}}$, we find that one solution for the fixed point is the pseudo-inverse-type interaction $J_{ij}^{\mathrm{F}} = a T_{ij}$, where

$$T_{ij} = \frac{1}{N} \sum_{\mu\nu} \xi_i^\mu C^{-1\mu\nu} \xi_j^\nu \tag{30}$$

and $C^{\mu\nu}$ is a pattern correlation matrix. The amplitude a is determined by (29), which becomes

$$\delta = (1-2\alpha) a\beta. \tag{31}$$

This yields

$$a\beta = \frac{\delta}{1-2\alpha} \tag{32}$$

where we have used $J_{ii}^{\mathrm{F}} \sim a\alpha$. This solution is valid when $a\beta$ is positive and small enough. The singularity at $\alpha = 0.5$ is an artefact of the second-order approximation. Higher-order terms of β should be taken into account when α is around 0.5. Even with higher-order terms, we suppose that the solution of the fixed-point equation will also be given by (30), since, in every order of the expansion in terms of $\beta J_{ij}^{\mathrm{F}}$, each C^{-1} is always associated with the matrix C. The questions are how the restrictions on site indices affect this simple structure and whether a can be positive or not. In this paper, we restrict ourselves to $\alpha < 0.5$ and not close to 0.5.

Let us discuss the solution of (29). As was done for the fixed point equation, it is convenient to introduce diagonal interactions $J_{ii}(t)$ in the site sum. The t dependence of $J_{ii}(t)$ will be specified later. Thus $\beta^2 \Delta t (J_{ii}(t) + J_{jj}(t)) J_{ij}(t)$ is added in the second term and it is subtracted from the first term. We further assume $J_{ii}(t)$ to be J_d for all i, which is the site and ξ average of $J_{ii}(t)$. Then we obtain

$$J_{ij}(t+\Delta t) = (1+p\Delta t) J_{ij}(t) - q\Delta t \sum_k J_{ik}(t) J_{kj}(t) \tag{33}$$

where $p = \beta\delta + 2\beta^2 J_d$ and $q = \beta^2$. Now it is natural to define $J_{ii}(t)$ by the elements with $i = j$ in (33). Then equation (33) can be regarded as a matrix differential equation for the interaction matrix $J(t)$. Note that $J_{ii}(0)$ should be $\sum_\mu \xi_i^\mu \xi_i^\mu / N = \alpha$, the fictitious diagonal elements of the Hopfield interactions.

If we neglect p, the solution of (33) is given by the model discussed in [9]. Although the t dependence of p will not be so strong, we can take it into account by assuming the same form as in [9] with t-dependent parameters. Thus we assume

$$\begin{aligned} J_{ij}(t) &= s \sum_k J_{ik} \left(\frac{1}{1 + rJ} \right)_{kj} \\ &= \frac{s}{N} \sum_{\mu\nu} \xi_i^\mu \left(\frac{1}{1 + rC} \right)^{\mu\nu} \xi_j^\nu \end{aligned} \tag{34}$$

where J is an initial interaction matrix, whose elements are the Hopfield interactions. s and r are functions of t. From the first to second line, we have used the relation $\sum_k J_{ik} J_{kj} = \xi_i^\mu C^{\mu\nu} \xi_j^\nu / N$ etc. Although J_d defined by (34) is a complicated function of s and r, we can use the approximated forms in the limiting situations. That is, $J_d \to s\alpha$ for $r \to 0$ and $J_d \to \alpha s / r$ for $r \to \infty$. The differential equations for s and r are obtained by replacing $s \to s + \Delta s$ and $r \to r + \Delta r$ in (34) and comparing it with (33). In this way, we obtain

$$\frac{\mathrm{d}s}{\mathrm{d}t} = p(s, r)s \qquad \frac{\mathrm{d}r}{\mathrm{d}t} = qs \tag{35}$$

where $p(s, r) \equiv p$. The initial conditions are $s = 1$ and $r = 0$, which give the Hopfield model. If $p(s, r)$ does not depend on t, the solutions of (35) have a simple exponential form. The corresponding solution (34) can be obtained from (33) directly.

The solution of (35) is explicitly obtained for small t or large t. For small t, we obtain $s = 1 + bt + \cdots$ and $r = \beta^2 t + \cdots$, where $b = \beta\delta + 2\alpha\beta^2$. For large t, we expect that $r \to \infty$ and that s/r becomes some constant for positive δ. Actually, using $J_d \to \alpha s / r$ and this assumption, we find

$$s = c \exp(\beta^2 a t) \qquad r = c \frac{1}{a} \exp(\beta^2 a t) \tag{36}$$

where a is defined by (32) and c is a positive constant. This solution is valid when $a > 0$, which imposes the condition $\delta > 0$. These results imply that $J_{ij}(t)$ tend to the pseudo-inverse interactions for $t \gg t_0 \equiv (\beta^2 a)^{-1}$ for $\delta > 0$. Note the factor $(\beta\delta)^{-1}$ in t_0, which controls the MC steps when the crossover takes place. This aspect is relevant in numerical simulations even when β is not so small. For $\delta < 0$, the limiting forms (36) are unphysical. To discuss the solution for intermediate t, we need to know J_d for moderate r. As we will see, the numerical simulations for $\delta < 0$ imply that the amplitude of interactions becomes so small that the condition on $\bar{\epsilon}$ will be violated eventually. Our arguments are no longer vaild in such a situation.

Let us give some comments about the results. First, actually, $J_{ii}(t)$ depend on i for finite systems, especially for small α. This site dependence can create some noise in the iterative equation. Second, the arguments so far are based on the limit $\bar{\epsilon} \to 0$. However, $\bar{\epsilon}$ and MC step d are finite in computer simulations. As discussed in section 4, unlearning terms will induce some random noise in simulations with finite $\bar{\epsilon}$. These two kinds of noise can cause a large deviation from the above results after many iterations. We also note that, when the amplitude of the fixed-point model is too small, the condition $\bar{\epsilon} \ll \overline{(\langle S_i S_j \rangle_{J(t)})^2}$ can be violated before reaching the fixed point. Even with these points, we think that the studies in this section are a good guide to understanding the results of numerical simulations.

5.2. Numerical study of the iterative equation

Now let us turn to the results of computer simulations for the evolution equation (27), where spin variables are driven by paramagnetic MC dynamics. To see what properties the running models have, we have studied the overlap m, the amplitude of interactions $\sqrt{N}|J(t)|$, signal-to-noise ratio r_h, and Q_J which is the overlap between $J_{ij}(t)$ and T_{ij}. The latter two quantities are defined by

$$r_h = \sqrt{(\Delta h^2)_a}/h_a$$
$$Q_J = J(t) \bullet T/|J(t)||T|$$

where $X \bullet Y \equiv \sum_{i\neq j} X_{ij}Y_{ij}/N(N-1)$ and $|X| \equiv \sqrt{X \bullet X}$. h_a is the average of absolute values of the local fields, which is defined by (13), and $\sqrt{(\Delta h^2)_a}$ is its variance. Both averages are evaluated over all sites and embedded patterns. r_h is zero for the pseudo-inverse model and $\sqrt{\alpha}$ for the Hopfield model with small α. By using $|T| = \sqrt{\alpha-\alpha^2}/\sqrt{N}$, $|J| = \sqrt{\alpha}/\sqrt{N}$ and $T \bullet J = (\alpha-\alpha^2)/N$, we see that Q_J is $\sqrt{1-\alpha}$ when the $J(t)$ model is the Hopfield model. At the fixed point, $\sqrt{N}|J(t)|$ is given by

$$\sqrt{N}|J^{\mathrm{F}}| = \frac{\sqrt{\alpha-\alpha^2}}{(1-2\alpha)\beta}\delta \tag{37}$$

to the second order of β, while it is $\sqrt{\alpha}$ for the Hopfield model.

In figure 4(a)–(d), the time developments of these quantities are presented for $(\alpha, \beta) = (0.2, 0.5)$ for every 1000 MC steps. N and $\bar{\epsilon}$ are 100 and 10^{-5}, respectively. δ are -0.2, 0.2 and 0.4. When $(\alpha, \beta) = (0.2, 0.5)$, we have $t_0 = 6/(5\delta)$, which becomes 3×10^5 MC steps for $\bar{\epsilon} = 10^{-5}$ and $\delta = 0.4$. In all the simulations, we have studied up to 5×10^5 MC steps. The results of the simulation for $(\alpha, \beta) = (0.6, 0.5)$ with $\delta = 0.4$ are also depicted for comparison. We have studied several samples in the same manner and found that the sample fluctuations are not so large except for the details of m, which are shown in figure 3. Let us first look through the results of $\alpha = 0.2$.

In figure 4(a) and (b), the time developments of m and r_h are depicted. The behaviour of m can be understood in terms of that of r_h. In general, as r_h decreases, m increases and becomes close to 1 when $r_h \sim \sqrt{\alpha_{\mathrm{C}}}$. For $\alpha = 0.2$, m becomes 1 after about 10^5 MC steps and stay there except for the case of $\delta = -0.2$, for which m starts to decrease at about 4×10^5 MC steps. This decrease is so small that m barely gets separated from the $m = 1$ line in figure 4(a). On the other hand, in figure 4(b), the difference between various δ is rather clear. r_h for $\delta = -0.2$ starts to increase at 2.5×10^5 MC steps, while there are no such increases of r_h for positive δ. At a given MC step, r_h is smaller for larger δ. This implies that the improvement of the models is quicker for larger δ as discussed in section 5.1, yet they do not reduce to zero in the studied MC steps. The increase of r_h for negative δ means that unlearning deteriorates the model as an associative memory. For $\bar{\mu} = 0$, which corresponds to $\delta = -1.0$, we found that this happens in earlier MC steps, although the model is improved in the beginning. As we have discussed in section 5.1, the evolution with negative δ does not seem to have a fixed-point model at least in our framework.

Figure 4(c) shows the evolution of $\sqrt{N}|J(t)|$ for the same parameters. For $\delta = -0.2$, $\sqrt{N}|J(t)|$ keeps decreasing, while it tends to some constant values for $\delta = 0.2$ and 0.4. To the second order of β, the theoretical values for the fixed-point model are given by (37), which reduces to $\frac{4}{3}\delta$ in this case. They are marked by $\triangle$ for each positive δ on the right vertical axis. Note $\sqrt{N}|J(t=5)|$ for $\delta = -0.2$ is about half of $\sqrt{N}|J(t=0)|$. This means that $\overline{(\langle S_i S_j\rangle_{J(t)})^2}$ becomes about one-fourth of the original value. Thus the condition (8) tends to be violated.

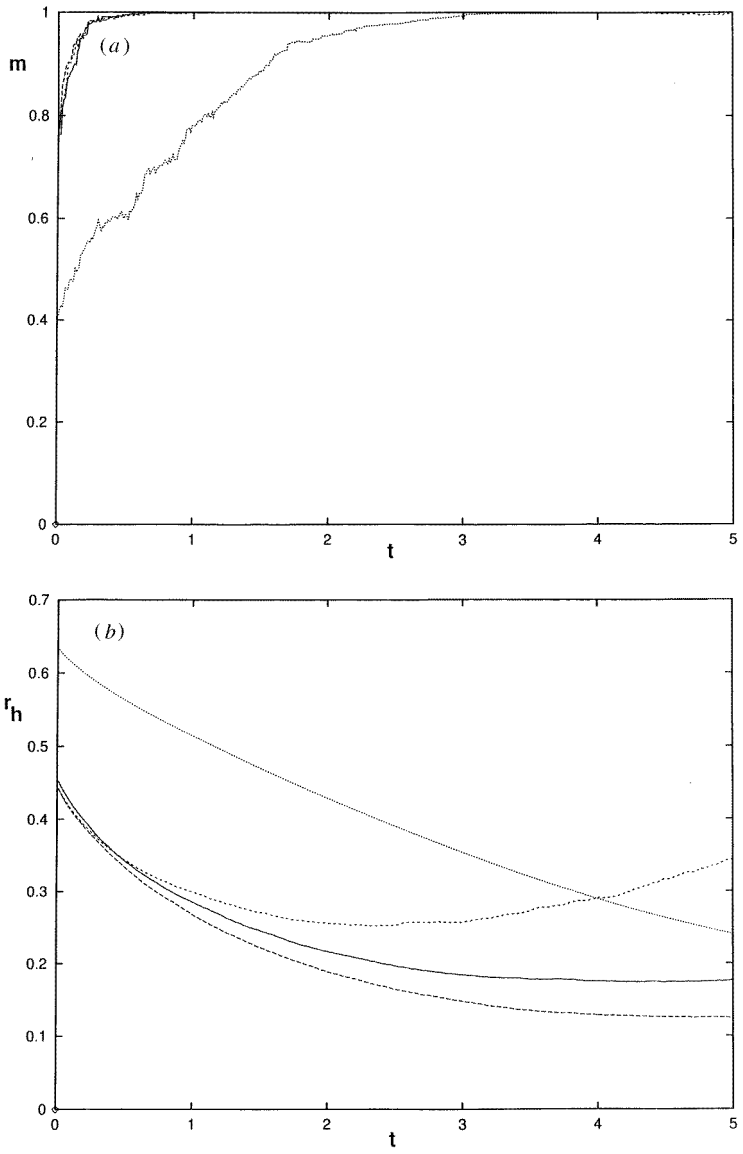

Figure 4. The evolutions of (*a*) m, (*b*) r_h, (*c*) $\sqrt{N}|J(t)|$ and (*d*) Q_J obtained by the iterative equation (27). The parameters are $(\alpha, \beta) = (0.2, 0.5)$ with $\delta = -0.2$ (short-broken curves), 0.2 (full curves) and 0.4 (long-broken curves). We took the same sample for $\delta = -0.2$ and 0.4. In each graph, the evolutions for $(\alpha, \beta) = (0.6, 0.5)$ with $\delta = 0.4$ are also depicted by dotted curves for comparison. All simulations were done with $\bar{\epsilon} = 10^{-5}$ and $N = 100$. On the horizontal axis $t \equiv \bar{\epsilon}d$. (*a*) The evolution of m. For $\alpha = 0.2$, it becomes very close to 1 in less than 1×10^5 MC steps, while it takes about 4×10^5 MC steps to become close to 1 for $\alpha = 0.6$. (*b*) The evolution of r_h. It is clearly shown how the models are improved. Note the increase of r_h for $\delta = -0.2$. (*c*) The evolution of $\sqrt{N}|J(t)|$. For $\alpha = 0.2$ and positive δ, we have the approximated values of the fixed-point model, $\frac{4}{3}\delta$, given by (37). They are marked by $\triangle$ for each δ on the right vertical axis. (*d*) The evolution of Q_J. For $(\alpha, \beta) = (0.2, 0.5)$, it first increases then starts to decrease in the middle of the simulation. For $(\alpha, \beta) = (0.6, 0.5)$ with $\delta = 0.4$, it keeps increasing during the MC steps studied.

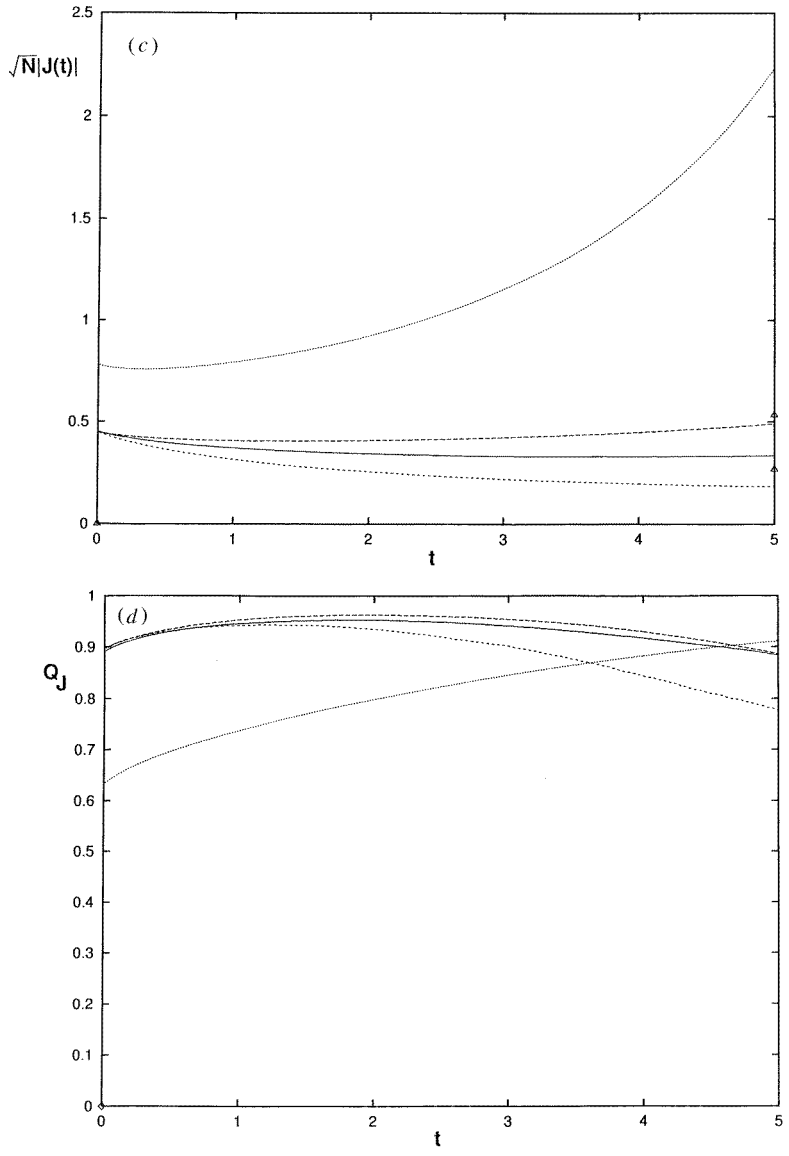

Figure 4. (Continued)

Figure 4(a)–(c) imply that the critical point of the evolution should be between $\delta = -0.2$ and 0.2. Although the observed transition is not so sharp, this supports that $\bar{\epsilon}\beta$ is a critical value of $\bar{\mu}$. Note that β in this expression appears as a result of the high-temperature expansion of the correlation function. To affirm this β dependence, we have studied the time developments for $\beta = 0.3$ with various δ and obtained similar δ dependence of the evolutions.

Figure 4(d) shows the behaviour of Q_J. It directly measures how similar the evolving models are to the pseudo-inverse model. For $\alpha = 0.2$, Q_J start at $\sqrt{0.8} \sim 0.89$ and increases to about 0.96, but then it keeps decreasing in the last half of the MC steps. Among the

studied δ, $\delta = 0.4$ achieves the largest value. Although the decrease of Q_J seems different from what we have studied in section 5.1, we can think of several reasons for this. In section 5.1, we have discussed two sources which create noise in the evolution equation. These noises can cause large effects near the fixed point where the drive by the evolution equation becomes very small. In addition, $\delta = 0.2$ seems rather close to the critical point of the evolution. This can be another reason for noise. Actually, maximal Q_J becomes closer to 1 as δ increases in figure 4(d).

Although we do not have any analytic results for large α, the numerical simulations can be done for arbitrary α. In figure 4(a)–(d), the results for $(\alpha, \beta) = (0.6, 0.5)$ are also depicted by dotted curves with the same N and $\bar{\epsilon}$. δ was assumed to be 0.4, which shows a faster improvement than smaller δ. From figure 4(a) and (b), we see that r_h keep decreasing and m becomes 1 around 4×10^5 MC steps. In figure 4(c), $\sqrt{N}|J(t)|$ increases more rapidly than the case of $\alpha = 0.2$. $\sqrt{N}|J(t=0)|$ is $\sqrt{0.6} \sim 0.775$ and increases to 2.2 at the last MC step. There, the acceptance rate of Monte Carlo spin flip, which is usually more than 50% of N, becomes about 20% of N. In figure 4(d), Q_J starts at $\sqrt{0.4} \sim 0.632$ and keeps increasing during the studied MC steps. Other samples with $\alpha = 0.6$ showed similar behaviour for these quantities.

To find the upper bound of α which allows $m = 1$, we did the simulations with $\alpha = 0.8$. In this case, however, m becomes only about 0.9 and we did not find any sets of parameters which achieve $m = 1$. Thus the critical capacity of PM unlearning is expected to be larger than 0.6 but smaller than 0.8, yet this may change if we take different $\bar{\epsilon}$.

6. Discussions

Unlearning of spurious states is a very interesting subject in the study of neural networks. The algorithm is local and the improvement is rather impressive. Biologically it is related to an interesting hypothesis on 'the function of dream sleep' as discussed in [5].

In this paper, we studied unlearning in the paramagnetic phase of neural network models. After the unlearning of many dreams, the changes of interactions are expressed by the paramagnetic correlation function $\langle S_i S_j \rangle$ of the initial model. The condition for this is that the unlearning parameter $\bar{\epsilon}$ is much smaller than $\overline{\langle S_i S_j \rangle^2}$. Using the high-temperature expansion to study $\langle S_i S_j \rangle$, we found that this condition is consistent with that suggested for RS unlearning. We defined the J' model by taking the Hopfield model as an initial model. The signal-to-noise analysis of the J' model was performed to the infinite order of β. The result supports the idea that our algorithm actually improves the Hopfield model. Briefly, $\langle S_i S_j \rangle$ of the Hopfield model contains the correlation matrix among embedded patterns, which changes the signal-to-noise ratio of the Hopfield model.

When interactional changes are large, the expansion rate $\bar{\mu}$ was introduced to control the amplitude of interactions. When $\bar{\mu} = 0$, interactions become relatively small after much unlearning. Then the condition on $\bar{\epsilon}$ will be violated eventually and unlearning terms begin to work as nothing more than random noise. We suppose that this may also happen in the simulations reported in some papers. In the $\bar{\mu}, \bar{\epsilon} \to 0$ limit with a suitable $\bar{\mu}/\bar{\epsilon}$, the iterative equation to the second order of β has the pseudo-inverse model as a fixed point, at least for small α. The appearance of the pseudo-inverse model is important since it can memorize a set of strongly correlated patterns, for which the Hopfield model does not work well. The simulations showed that the overlap Q_J between the evolving model and the pseudo-inverse model actually increases close to 1, but does not reach 1. We suppose that $\bar{\epsilon}$ should be much smaller than our value to get a better agreement with the theory.

The studies in this paper were mainly addressed to small α and the stability of embedded patterns. This is because, first of all, we wanted to see if our analysis using correlation functions agrees with the simulations or not. The retrieval properties and large-α fixed-point models are the next subjects of study. To discuss the case where $\alpha \sim 0.5$ or larger, we have to take into account higher-order terms in the expansion of $\langle S_i S_j \rangle_{J(t)}$. The study of transient models will become important if there is no physical fixed point.

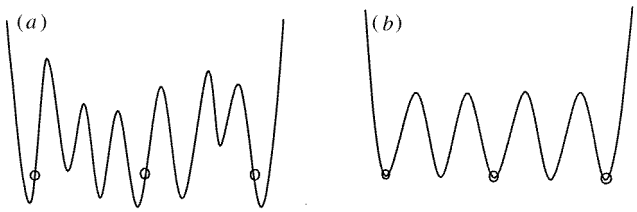

Figure 5. These diagrams illustrate the change of energy landscape under PM unlearning. ○ denote the embedded patterns. The energies of the low-energy states are expected to become equal after unlearning since lower-energy states are unlearned more according to the Maxwell–Boltzmann weights.

To summarize the study, we present intuitive pictures in figure 5, which illustrates how PM unlearning works. In figure 5(a), the energy landscape of the overloaded Hopfield model is depicted. No embedded patterns are at the bottom of valleys. Instead, there are many spurious states at the bottom of valleys having lower energy than the embedded patterns. In this situation, spurious states are expected to appear rather frequently in the paramagnetic dynamics since the probability of appearance obeys the Maxwell–Boltzmann distribution. However, unlearning of them make their energy higher. As unlearning proceeds, their energy will become equal to that of the embedded patterns as shown in figure 5(b). This situation is very similar to the results of replica studies for the pseudo-inverse model [4]. Under this situation, spurious states and embedded patterns are expected to appear with an equal probability in the dynamics. If unlearning goes further, there will be three possibilities: (i) $|J(t)|$ decreases, (ii) $|J(t)|$ is rather stable or (iii) $|J(t)|$ increases. Unlearning of paramagnetic configurations tends to make the energy landscape flat. This corresponds to case (i). The expansion of interactions compensates for this effect. Thus case (ii) happens when the expansion rate $\bar{\mu}$ is chosen properly. When $\bar{\mu}$ is larger, case (iii) happens and the spin configuration will be trapped in a certain energy valley. The problem here is that, as seen in (32), the range of $\bar{\mu}$ where (ii) is realized seems to become narrower or disappear as α increases. It may be possible to think of some dynamics of β which keeps $\beta\sqrt{N}|J(t)|$ of order 1. However, it may spoil the locality of unlearning. Probably the simplest way to avoid (i) and (iii) is just to stop unlearning.

Our formulation using correlation functions is quite general. Here we point out two possible extensions of our study. Firstly, it is possible to apply the idea to various versions of the Hopfield model which learn either memories with different weights or memories with strong correlations. These kinds of patterns seem more natural since the environment around us seems to give a great number of correlated patterns rather than a limited number of uncorrelated patterns. Besides, their weights, the frequencies of learning, depend on patterns. It will be quite interesting to study how unlearning works in such difficult situations. Secondly, by using correlation functions, the effect of unlearning can be formulated for the systems which have no energy function. If we have a suitable Hebb rule for the system, interactional changes by unlearning will be represented by a suitable correlation function in

the same way as in section 2. The expansion of the correlation function in terms of some sort of temperature will also be useful to study the effect of unlearning for such systems.

In our study, the concept of temperature plays a very important role. For ordinary spin models, the system is assumed to be in contact with a heat reservoir and is naturally described by a temperature. However, there is no such reservoir for neural networks. Then we may ask what the temperature in our case means. In this respect, the relation between RS unlearning and PM unlearning is an interesting subject. Note that RS unlearning can also be formulated in terms of a correlation function made of spurious states if interactional changes are very small, yet we have no method to estimate the correlation function except a numerical one. Here we point out another point of view. That is, instead of addressing the relation, we should rather think of some intrinsic mechanism which causes random dynamics and generates dreams. Then the problem is to study to what extent this dynamics is simulated by the usual thermodynamics or random shooting with relaxation. As far as unlearning is concerned, details of the dynamics will not matter and any dynamics will work well if undesirable states appear more frequently than the embedded patterns. In any case, the problem is to evaluate the correlation function in terms of interactions and to see how it affects the original interactions.

Finally, let us comment on the general aspect of our study. That is, we can regard our study as a special case of paramagnetic evolution of complex systems. If paramagnetic configurations reflect a low-energy energy landscape, we may naturally ask what happens to the systems which have complex energy landscapes after being modified by paramagnetic learning or unlearning. This question may sound rather academic, however, I think it deserves to be studied. For one thing, this idea can be helpful in studying optimization problems since learning about low energy states may make the search for them easier. Also, if the initial models are random models, it will be possible to introduce models which have correlated irregular interactions rather than uncorrelated random interactions. Thus it will be quite interesting and meaningful to study random spin models using our formulation.

Acknowledgments

The author is grateful to Dr Y Kabashima and Dr T Uezu for valuable discussions.

Appendix A

In this appendix, we describe the high-temperature expansion of $\overline{\langle S_i S_j\rangle^2}$ for the Hopfield model. A related study is found in [11]. In the large-α limit, the results should reduce to those of the infinite range spin-glass model, which has a simple high-temperature expansion as was discussed in [12]. In our case, the diagrams which contribute after the ξ-average look like a cross between the ferromagnetic model and the spin-glass model.

The correlation function $\langle S_i S_j\rangle$ is defined by

$$\langle S_i S_j\rangle = \sum_{\{S\}} S_i S_j \exp(-\beta H)/Z \tag{A1}$$

where $Z = \sum_{\{S\}} \exp(-\beta H)$. The summation $\sum_{\{S\}}$ is over spin configurations. $\langle S_i S_j\rangle$ is formally expanded in terms of $\tanh \beta J_{ij} \sim \beta J_{ij}$, giving

$$\langle S_i S_j\rangle = \beta J_{ij} + \beta^2 {\sum}' J_{ik} J_{kj} + \beta^3 {\sum}' J_{ik} J_{kl} J_{lj} + \cdots. \tag{A2}$$

In the sum $\sum'$, no two site indices are equal to each other, since a loop of edges should be factorized to cancel the denominator Z. This point is important in the Hopfield model as

we will see in the following and appendix B. In general, a loop gives a factor P after the ξ-average. More precisely, when all site indices are different, a loop of J_{ij}'s gives

$$\overline{J_{ij}J_{jk}\cdots J_{zi}} = PN^{-L} \tag{A3}$$

where $\overline{\cdots}$ is the ξ-average and L is the number of J_{ij}'s.

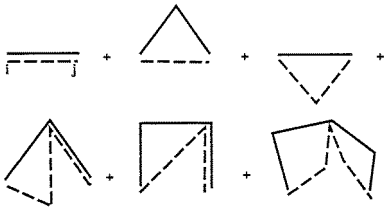

Figure A1. Some diagrams which appear in $\overline{\langle S_i S_j\rangle^2}$. The full lines represent terms which come from one $\langle S_i S_j\rangle$ and the broken lines from other $\langle S_i S_j\rangle$. The first row shows the diagrams with one loop. The second row shows those with two loops.

Now let us discuss $\overline{\langle S_i S_j\rangle^2}$. Each term in $\langle S_i S_j\rangle$ can be expressed by a zig-zag line which starts at site i, visits some different sites and ends at site j. No site is visited twice or more in $\langle S_i S_j\rangle$, while in $\langle S_i S_j\rangle^2$, two terms coming from different $\langle S_i S_j\rangle$ can share some sites other than i or j (see figure A1). Let that number be denoted by n_s. Although these sites impose restrictions over the site sum, they create loops, each of which gives a factor $P = \alpha N$. Among diagrams with fixed n_s, the diagrams which have the largest number of loops are the ladder type, which have $n_s + 1$ loops. We can see this as follows. For n_s shared sites, each zig-zag line is divided into $n_s + 1$ fragments. When the order of shared sites is different between two zig-zag lines, there is at least one loop which costs more than two fragments. This gives a number of loops smaller than $n_s + 1$. When the site order is the same, two fragments coming from different $\langle S_i S_j\rangle$ can make a loop, giving $n_s + 1$ loops for this diagram. Therefore the ladder diagrams give the leading contributions to $\overline{\langle S_i S_j\rangle^2}$. Let us first calculate the contribution of a loop. Imagine a certain loop of length $l = l_1 + l_2$, l_1 from one $\langle S_i S_j\rangle$ and l_2 from another $\langle S_i S_j\rangle$. Using equation (A3), we see that this loop yields $\beta^l PN^{-l}$ after the ξ-average. The summation over internal sites, the sites which are not shared by two $\langle S_i S_j\rangle$, yields a factor N^{l-2}. This cancels the factor N^{-l} above and l remains only as a power of β. Thus the sum over l_1 and l_2 gives $\beta/(1-\beta) \times \beta/(1-\beta)$. Putting these together, we obtain

$$\frac{1}{N}A \equiv \frac{1}{N}\frac{\alpha\beta^2}{(1-\beta)^2} \tag{A4}$$

as a contribution from the sum over diagrams of a single loop. Finally the summation over the number of loops and shared sites gives

$$\overline{\langle S_i S_j\rangle^2} = \frac{1}{N}\frac{A}{1-A}. \tag{A5}$$

This result is also obtained as a by-product of the signal-to-noise analysis of the J' model, which is discussed in section 4. Note that, when $\alpha \to \infty$ with fixed $J_0^2 \equiv \alpha\beta^2$, this expression reduces to the paramagnetic correlation function of the infinite-range spin-glass model with interactional variance $J_0/\sqrt{N}$. On the other hand, when $\alpha \to 0$, equation (A5) reduces to A/N, which is just the square of the ferromagnetic correlation function multiplied by the number of patterns P.

Appendix B

In this appendix, we derive the relation

$$\overline{\left(\sum{}'' j_{ij}^{(p)}\right)\left(\sum{}'' j_{ij}^{(p')}\right)} = \delta_{pp'}\alpha^p/N \tag{B1}$$

and

$$\overline{\left(\sum_{j\neq i}\sum{}'' j_{ij}^{(p)}\xi_j^1\right)\left(\sum_{j'\neq i}\sum{}'' j_{ij'}^{(p')}\xi_{j'}^1\right)} = \delta_{pp'}\alpha^p \tag{B2}$$

where

$$j_{ij}^{(p)} = j_{ik}^{\mu} j_{kl}^{\nu} \cdots j_{zj}^{\eta} \tag{B3}$$

and $j_{ik}^{\mu} = \xi_i^{\mu}\xi_k^{\mu}/N$. These formulae are valid in the $N \to \infty$ limit.

Let us first concentrate on (B1). In the sum $\sum''$, site indices are all different and neighbouring pattern indices are not equal to each other. The restriction on site indices implies that a certain site appears only once in $j_{ij}^{(p)}$, if it does at all. Further, on such a site, there are two ξ which have different pattern indices. Therefore two $j_{ij}^{(p)}$ in (B1) should have the same site indices to give a non-zero contribution except that the order of them can be different. For this reason, the case $p \neq p'$ gives zero on the right-hand side. In the $P, N \to \infty$ limit, the pairing of two terms of the same site order gives the leading term $N^{p-1}P^p \times (1/N)^{2p}$ after the site and pattern sum, where we used $(j_{kl}^{\mu})^2 = 1/N^2$. This is the right-hand side of (B1). Thus we should show that the products of two terms with different site order do not give leading contributions.

In the product of different site order terms, the number of free site indices and the number of j_{ij}^{μ}, which give factors N and $1/N$, respectively, are the same as above. However, the number of free pattern indices becomes smaller. Let us take a closer look at this point. To evaluate the number of free pattern indices, we should count the number of loops made of j_{ij}^{μ}. Note that $(j_{kl}^{\mu})^2$ is the simplest loop of length two, which gives a factor P after the μ-sum. In the product of different site orders, there are loops longer than two. These loops should have lines coming from different $j_{ij}^{(p)}$ alternately since neighbouring j_{ij}^{μ} in $j_{ij}^{(p)}$ do not have the same pattern index. A loop of this type costs four or more than four j_{ij}^{μ}. Therefore it gives non-leading terms in (B1). In other words, there appear additional restrictions on the pattern indices in this case. We should note that the situation is essentially parallel to the evaluation of $\overline{\langle S_i S_j\rangle^2}$.

Let us turn to (B2). When $j = j'$, we can follow the argument presented for (B1) and obtain the leading contributions, which make the right-hand side of (B2) after the j sum. When $j \neq j'$, a factor ξ_j^1 restricts the sum over pattern indices in $\sum'' j_{ij'}^{(p)}$ since there should be ξ_j^1 somewhere in $j_{ij'}^{(p)}$. The same is true for $\xi_{j'}^1$. The rest of the pattern indices should be paired between $j_{ij}^{(p)}$ and $j_{ij'}^{(p)}$ to give a non-zero contribution. Thus the number of free pattern indices is reduced by at least one. Therefore the case $j \neq j'$ does not give leading contributions to the right-hand side.

References

[1] Hopfield J J 1982 *Proc. Nat. Acad. Sci., USA* **79** 2554
[2] Amit D J, Gutreund H and Sompolinsky H 1987 *Ann. Phys.* **173** 30
[3] Personnaz L, Guyon I and Deyfus G 1985 *J. Physique Lett.* **46** L359
[4] Kanter I and Sompolinsky H 1987 *Phys. Rev.* A **35** 380

[5] Crick F and Mitchison G 1983 *Nature* **304** 111
[6] Hopfield J J, Feinstein D I and Palmer R G 1983 *Nature* **304** 158
[7] Kleinfeld D and Pendergraft D B 1987 *Biophys. J.* **51** 47
[8] Hemmen J L V, Ioffe L B, Kühn R and Vaas M 1990 *Physica* **163** 386
[9] Dotsenko V S, Yarunin N D and Dorotheyev E A 1991 *J. Phys. A: Math. Gen.* **24** 2419
[10] For example, Binder K and Heermann D W 1988 *Monte Carlo Simulation in Statistical Physics* (Berlin: Springer)
[11] Dotsenko V S, Feigel'mann M V and Ioffe L B 1990 *Spin Glasses and Related Problems* (New York: Harwood Academic) p 209
[12] Thouless D J, Anderson P W and Palmer R G 1977 *Phil. Mag.* **35** 593