Paramagnetic unlearning in neural network models

Kazuo Nokura*
Shonan Institute of Technology, Fujisawa 251, Japan
(Received 22 April 1996)

We study unlearning in the paramagnetic phase of neural network models. After many unlearning steps at temperature T, changes of synaptic interactions are expressed by the paramagnetic correlation function. Taking the second order terms of $\beta=T^{-1}$, we derive the evolution equation and find that the Hopfield model evolves into the pseudo-inverse model in some parameter region. As a second initial condition, we study the Hopfield model, which has varying pattern weights. In this case, unlearning works on the large weight patterns and removes the monopoly of them. [S1063-651X(96)13110-X]

PACS number(s): 87.10.+e, 05.90.+m, 75.10.Nr

I. INTRODUCTION

Recently, neural networks have been extensively studied by using the analogy to spin models of statistical physics. According to these studies, an associative memory can be achieved by spin dynamics when interactions among them are made according to the Hebb rule. For each spin interaction J_{ij} which connects site i and site j, the Hebb rule tells us the prescription $J_{ij}\rightarrow J_{ij}+\xi_i\xi_j/N$ to learn a pattern ξ_i, where N is the system size. With this replacement, the fixed point of spin dynamics is created in the configuration space and the system remembers this pattern from an imperfect pattern.

The Hebb rule is local in the sense that the change of J_{ij} is completely determined by the data on site i and site j of a newly arising pattern. After learning P patterns $\xi_i^\mu, \mu=1\cdots P$, $i=1\cdots N$, we obtain the spin model, that is, the Hopfield model [1], which is defined by

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

where

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

and S_i are Ising spin variables. We assume $\xi_i^\mu=\pm1$ with probability $\frac{1}{2}$. According to the study by the replica method [2], this model really has a retrieval phase for small $\alpha\equiv P/N$, but it also has a spin glass phase with a rather high phase transition point. This may imply that spin glass states dominate the configuration space. In the context of neural networks, spin glass states and other states which are different from learned patterns are called spurious states, which means that they are different from learned patterns.

The Hopfield model can be viewed as a point in the interaction space and it possibly has better models nearby. In fact, several years ago, some biologists have suggested a very interesting local algorithm called *unlearning* which improves the Hopfield model gradually [3–5]. The basic idea is to remove spurious states from the spin configuration space by unlearning them. They have also suggested that this procedure corresponds to REM sleep, which is widely observed among mammals. However, we have no idea about the resulting model, probably because it is quite difficult to study the ensemble of spurious states.

In this paper, we introduce and study unlearning of paramagnetic configurations, in which the system unlearns the spin configurations generated by paramagnetic dynamics of neural networks [6]. In Sec. II, we present the basic idea of unlearning of paramagnetic configurations. In Sec. III, the evolution equation is studied by using the high-temperature expansion. In Sec. IV, we apply our formulation to the generalized Hopfield models. The problems that remain to be studied are discussed in Sec. V.

II. UNLEARNING IN THE PARAMAGNETIC PHASE

Let us imagine S_i^d are generated by paramagnetic Monte Carlo dynamics with interactions J_{ij}^d and temperature T. After a whole updating of S_i^d, the next interactions are defined by the iterative equations given by

$$J_{ij}^{d+1}=(1+\overline{\mu})J_{ij}^d-\overline{\epsilon}S_i^dS_j^d. \tag{3}$$

The initial interactions J_{ij}^0 are assumed to be the Hopfield interactions. $\overline{\mu}$ and $\overline{\epsilon}$ are positive constants, which will be specified later.

According to statistical mechanics, S_i^d obeys the Maxwell-Boltzmann distribution defined by the energy function of the dth model. If spurious states have lower energies than embedded patterns, they are expected to appear very frequently in this dynamics. It is known that the spin glass states become global minimum for $\alpha>0.05$. Thus we expect the similar unlearning effect in this scheme. Actually, we have found some affirmative results of numerical simulations in the previous paper.

Using Eq. (3), the interactions after d_0 unlearning are formally given by

$$J_{ij}^{d+d_0}=(1+\overline{\mu})^{d_0}J_{ij}^d-\overline{\epsilon}\sum_{d'=d}^{d+d_0-1}(1+\overline{\mu})^{d+d_0-1-d'}S_i^{d'}S_j^{d'}. \tag{4}$$

*Electronic address: nokura@cosmos.la.shonan-it.ac.jp

In the small $\overline{\mu,\epsilon}$ limit and if $d_0\to\infty$ with small fixed $\mu\equiv\overline{\mu d}_0$ and $\epsilon\equiv\overline{\epsilon d}_0$, Eq. (4) reduces to

$$J_{ij}^{d+d_0}=(1+\mu)J_{ij}^{d}-\epsilon\langle S_iS_j\rangle_{J^d} \tag{5}$$

to the first order of ϵ and μ, where $\langle S_iS_j\rangle_{J^d}$ is a paramagnetic correlation function of the dth model, which is defined by

$$\langle S_iS_j\rangle_{J^d}=\sum_{\{S\}} S_iS_j\exp(-\beta H_d)/Z, \tag{6}$$

where $Z=\Sigma_{\{S\}}\exp(-\beta H_d)$ and $\beta=T^{-1}$. The summation $\Sigma_{\{S\}}$ is over spin configurations. H_d is the energy function of the dth model. The upper bound of $\overline{\epsilon}$ is given by the following argument. Roughly speaking, each term in the sum in Eq. (4) is nearly random ±1. Thus the average is of order $1/\sqrt{d_0}$. But actually, it should tend to the correlation function for large d_0. Therefore $1/d_0\ll(\langle S_iS_j\rangle_{J^d})^2$ should be satisfied. This gives the inequality $\overline{\epsilon}\ll(\langle S_iS_j\rangle_{J^d})^2$ if we take $\epsilon=1$. This implies $\overline{\epsilon}\ll\beta^2J_0^2/N$ for small β, where J_0 is a constant of order 1.

III. EVOLUTION IN HIGH TEMPERATURE

Let us discuss the solution of Eq. (5) to the second order of β, which gives the first nontrivial effects. In the following argument, it is convenient to use $t\equiv\overline{\epsilon d}$ as a time variable. Then the above equation becomes

$$\begin{aligned}J_{ij}(t+\Delta t)&=(1+\theta\Delta t)J_{ij}(t)-\Delta t\langle S_iS_j\rangle_{J(t)}\\&=(1+\theta\Delta t)J_{ij}(t)-\Delta t\Big[\beta J_{ij}(t)\\&\quad+\beta^2{\sum}'J_{ik}(t)J_{kj}(t)\Big],\end{aligned} \tag{7}$$

to the second order of β, where $\theta\equiv\overline{\mu/\epsilon}$ and $\Delta t\equiv\overline{\epsilon d}_0$. In the site sum Σ' in Eq. (7), the terms with $k=i$ or j are excluded. The important observation on this equation is that, when the Hopfield interactions are assumed for $J_{ij}(t)$ including fictitious diagonal interactions $J_{ii}=\Sigma_\mu\xi_i^\mu\xi_i^\mu/N=\alpha$, the pattern correlation matrix $C^{\mu\nu}\equiv\Sigma_i\xi_i^\mu\xi_i^\nu/N$ appears if the site sum is extended to all sites. In the following studies, it is convenient to introduce the interaction matrix with fictitious diagonal elements $J_{ii}(t)$, which will be defined as a natural extention of $J_{ij}(t)$. These diagonal elements are independent of i for the Hopfield model. We assume this for general t by setting $J_{ii}(t)\sim J_d(t)$, where $J_d(t)$ is the ξ average of $J_{ii}(t)$. Then we reach the evolution equation given by

$$J_{ij}(t+\Delta t)=(1+p\Delta t)J_{ij}(t)-q\Delta t{\sum}J_{ik}(t)J_{kj}(t), \tag{8}$$

where $p\equiv\beta\delta+2\beta^2J_d(t)$, $q\equiv\beta^2$, and $\beta\delta=\theta-\beta$. Equation (8) can be regarded as off-diagonal parts of the evolution equation for matrix $J(t)$. Thus it is natural to extend Eq. (8) to $i=j$, which define diagonal parts $J_{ii}(t)$. In this way, we get the evolution equation for $J(t)$. The evolution equation with quadratic terms of interactions was suggested in different algorithms [7–9], yet the relation to paramagnetic spin configurations has not been discussed clearly.

If p is a constant, we can easily obtain the solution of Eq. (8), which has the form given by

$$\begin{aligned}J_{ij}(t)&=s{\sum}\,J_{ik}\left(\frac{1}{1+rJ}\right)_{kj},\\&=\frac{s}{N}\sum_{\mu\nu}\xi_i^\mu\left(\frac{1}{1+rC}\right)^{\mu\nu}\xi_j^\nu,\end{aligned} \tag{9}$$

where J is an initial interaction matrix, $s=\exp(pt)$, and $r=(q/p)[\exp(pt)-1]$. Although t dependence of p is not so strong, we can take it into account in the following way. Using Eqs. (8) and (9), we obtain a set of differential equations for s and r given by

$$\begin{aligned}\frac{ds}{dt}&=ps,\\\frac{dr}{dt}&=qs.\end{aligned} \tag{10}$$

The initial conditions for s and r are 1 and 0, respectively, which represent the Hopfield model. When t is small or large, we can use the limiting behavior of J_d, that is, $J_d\to s\alpha$ for $r\to0$ and $J_d\to\alpha s/r$ for $r\to\infty$. Then, to the first order of t, we obtain

$$\begin{aligned}s&=1+bt+\cdots,\\r&=\beta^2t+\cdots,\end{aligned} \tag{11}$$

where $b=\beta\delta+2\alpha\beta^2$. This solution implies that the amplitude of beginning interactions are mainly controlled by δ. When t is large, we assume s and r become large for large enough δ. Using $J_d\sim\alpha s/r$ and assuming s/r tends to a positive constant a, we find the solution given by

$$\begin{aligned}s&=c\,\exp(\beta^2at),\\r&=c\,\frac{1}{a}\exp(\beta^2at),\end{aligned} \tag{12}$$

where $a=\delta/[\beta(1-2\alpha)]$ and c is some positive constant. This solution is valid when $a>0$, which implies $\delta>0$ for $\alpha<0.5$. The singularity at $\alpha=0.5$ is the artifact of the second order approximation. Solution (12) implies that the running model tends to the pseudoinverse model [10] for $t\gg(a\beta^2)^{-1}$. It was shown by the replica method that the pseudoinverse model has higher memory capacity and a lower spin glass transition point than the Hopfield model [11]. Model (9) was also studied by the replica method, showing larger capacity than the Hopfield model [7]. Thus, even for $\delta<0$, paramagnetic unlearning really improves the Hopfield model as long as s does not become too small.

Unlearning of spurious states has been studied for $\overline{\mu}=0$, which corresponds to $\delta=-1$. In this situation, there is a numerical study about the optimal number of unlearning $D_{\rm opt}$ which achieves the best model [5,8]. In our formulation, the amplitude of $J_{ij}(t)$ becomes very small for negative δ and the condition $\overline{\epsilon}\ll\langle S_iS_j\rangle_{J^d}{}^2$ will be violated eventually. Thus we define $D_{\rm opt}$ as the number of Monte Carlo steps when $J_{ij}(t)$ becomes very close to zero. According to Eq. (10), s becomes very close to zero when $t\sim|b|^{-1}$ for negative b, which leads to $D_{\rm opt}\sim1/(|b|\overline{\epsilon})$ for paramagnetic unlearning.

IV. APPLICATIONS TO THE GENERALIZED HOPFIELD MODELS

For the Hopfield model, there are various versions, each of which reflects the properties of patterns and the situation of learning. How unlearning works for such generalized Hopfield models is an interesting problem. In this section, we first discuss the model with varying pattern weights, and then give some comments on the model with correlated patterns.

For the Hopfield model with varying pattern weights, the initial interactions are given by

$$J_{ij}=\frac{1}{N}\sum_{\mu} a_{\mu}\xi_i^{\mu}\xi_j^{\mu}, \tag{13}$$

where a_{μ} depends on μ, for which we take $\exp(-g\mu/N)$ for convenience. This model was studied with regard to a working memory [12]. Although there is no limit on the sum over patterns, the number of patterns the model is expected to remember is $P_e \equiv N/g$. In the following, we imagine that $P_e/N=1/g$ is small enough and the system remembers about P_e patterns. In this situation, unlearning is expected to work not on spin glass states but on P_e strongly memorized states since they are located deeply in energy valleys. Note $J_{ij} \sim 1/\sqrt{2gN}$ for $i \neq j$ and $J_{ii}=1/g$.

To the second order of β of the evolution equation, the very same argument as for the usual Hopfield model formally gives the solution

$$J_{ij}(t)=\frac{1}{N}\sum_{\mu\nu}\xi_i^{\mu}\left(A\,\frac{s}{1+rCA}\right)^{\mu\nu}\xi_j^{\nu}, \tag{14}$$

where $A^{\mu\nu}=\delta^{\mu\nu}a_{\mu}$. The functions s and r obey Eqs. (10) with J_d derived from Eq. (14). For small t, the solution is given by Eq. (11) with α replaced by g^{-1}. For large t, r is expected to become large for large enough δ. Does this imply that Eq. (14) tends to the pseudoinverse model? We should be careful about this point when a_{μ} varies. That is, for some diagonal elements of $(1+rCA)^{\mu\nu}$, the matrix 1 cannot be neglected because ra_{μ} with large μ can be still very small.

Let us assume that r becomes much larger than 1 after many unlearning steps. This will be true for large enough δ. To see what Eq. (14) implies in such a situation, it is convenient to divide a group of patterns into two groups by introducing μ_0 defined by $ra_{\mu_0}=1$ and assume $ra_{\mu>\mu_0} \ll 1$. Then every matrix is decomposed into four submatrices depending on whether pattern indices are larger or smaller than μ_0. $A^{\mu\nu}$ are decomposed into A_0+A_1, where A_0 is a submatrix made of $a_{\mu\leq\mu_0}$ and A_1 made of $a_{\mu>\mu_0}$. Similarly, $C^{\mu\nu}$ are decomposed into four parts: C_0 for $\mu,\nu\leq\mu_0$, C_1 for $\mu>\mu_0$ and $\nu\leq\mu_0$, C_1^T for $\mu\leq\mu_0$ and $\nu>\mu_0$, and C_2 for $\mu,\nu>\mu_0$. For $\mu,\nu\leq\mu_0$, the matrix 1 can be neglected in $1+rCA$. In this way, we assume

$$(1+rCA)\sim\begin{pmatrix} P & e \\ Q & 1+f \end{pmatrix}, \tag{15}$$

where $P=rC_0A_0$, $Q=rC_1A_0$, $e=rC_1^TA_1$, and $f=rC_2A_1$. The terms with e or f are proportional to rA_1. To the first order of rA_1, we obtain

$$\begin{aligned} A(1+rCA)^{-1} &\sim \begin{pmatrix} A_0P^{-1}(1+eQ) & -A_0P^{-1}e \\ -A_1QP^{-1} & A_1 \end{pmatrix}, \\ &= \begin{pmatrix} r^{-1}C_0^{-1}(1+eQ) & -C_0^{-1}C_1^TA_1 \\ -A_1C_1C_0^{-1} & A_1 \end{pmatrix}, \end{aligned} \tag{16}$$

where we have used the relation $A_0P^{-1}=r^{-1}C_0^{-1}$. Most elements with A_1 are very small because of the assumption $ra_{\mu>\mu_0} \ll 1$. On the other hand, the part with $\mu,\nu\leq\mu_0$ is independent of μ since $a_{\mu<\mu_0}$ disappears. This part grows as t increases and dominates the interaction matrix eventually. Thus, for large t, we obtain the approximated form given by

$$J_{ij}(t)\sim\frac{1}{N}\frac{s}{r}\sum_{\mu,\nu\leq\mu_0}\xi_i^{\mu}C_0^{-1\mu\nu}\xi_j^{\nu}, \tag{17}$$

which is the pseudoinverse model made of the first μ_0 patterns. Using Eq. (17), we can determine the t dependence of μ_0 for large t. Let $s/r \equiv a(t)$ and $\mu_0/N \equiv \alpha(t)$, then the approximated solutions for s and r are given by

$$\begin{aligned} s &= c\,\exp\int p(t)dt, \\ r &\sim c\,\frac{\beta^2}{p(t)}\exp\int p(t)dt, \end{aligned} \tag{18}$$

where

$$p(t)\equiv\beta\delta+2\beta^2a(t)\alpha(t) \tag{19}$$

is assumed to depend on t weakly. Equation (18) yields $s/r=p(t)/\beta^2$, which gives the relation

$$a(t)=\frac{\delta}{\beta[1-2\alpha(t)]}. \tag{20}$$

On the other hand, by definition of $\alpha(t)$, we obtain

$$\begin{aligned} \alpha(t) &\equiv \frac{1}{g}\ln[r(t)], \\ &\sim \frac{1}{g}\int p(t)dt. \end{aligned} \tag{21}$$

Integration after differentiating both sides gives

$$\alpha^2(t)-\alpha(t)+\frac{\beta\delta}{g}(t-t_0)=0, \tag{22}$$

where t_0 is the time when the right hand side of Eq. (17) starts to dominate $J_{ij}(t)$. For small $\alpha(t)$, we obtain $\alpha(t)\sim\beta\delta(t-t_0)/g$. In the original model, the number of effectively memorized patterns is about $1/g$ for large g, while it can be of order 1 for Eq. (17) since $\alpha(t)$ can be of order of 1 for $t-t_0\sim g/(\beta\delta)$. However, there is an upper limit of $t-t_0$ since $\alpha(t)$ should not be too close to 0.5. Otherwise, $\beta\alpha(t)$ becomes too large to use the high-temperature expansion.

The disappearance of $\alpha_{\mu \leqslant \mu_0}$ in Eq. (17) is a very interesting phenomena and clearly reflects the unlearning of patterns which are located deeply in energy valleys. That is, with unlearning of paramagnetic configurations, the energies of large weight patterns become higher than small weight patterns do and these patterns make the pseudoinverse interactions successively. Biologically, the system removes a few patterns' monopoly by unlearning them and releases the potential capacity of other memorized patterns.

Another interesting generalization is the model with correlated patterns. It is well known that the original Hopfield model does not work well to memorize correlated patterns. In more general interactions, it is known that the solution to this problem is given by the pseudoinverse model, which is surely what paramagnetic unlearning brings into. Besides, unlearning of low-energy states seems quite natural since two correlated patterns are expected to create a deep energy valley. Let us discuss the simplest model made of two correlated patterns ξ_i^1 and ξ_i^2 with $\Sigma_i \xi_i^1 \xi_i^2/N = m > 0$. By the signal-noise analysis, we see that the Hopfield model does not have either ξ_i^1 or ξ_i^2 as a fixed point of spin dynamics, especially for m close to 1. This means that the mixed energy valley, which is different from embedded patterns, is created by the Hebb learning of two correlated patterns. This naturally leads to the idea of unlearning of this state. In our formulation, the study of unlearning in such a situation is reduced to the calculation of the inverse of $(1+rC)^{\mu\nu}$. A short calculation gives the interactions proportional to $(\xi_i^1 \xi_j^1 - m\xi_i^1 \xi_j^2 - m\xi_i^2 \xi_j^1 + \xi_i^2 \xi_j^2)/N$ for large r. This interaction surely has no noise for either pattern 1 or pattern 2. The inverse of $C^{\mu\nu}$ in the large-r interactions implies that the same thing is true for large r for many correlated patterns.

V. DISCUSSION

Let us give some concluding remarks. We have studied paramagnetic unlearning of neural networks by using the high-temperature expansion. Our study suggests that high-temperature configurations already contain nontrivial information about how to change the system to remove spurious states. In studying the evolution equation expressed by the correlation function, we have restricted ourselves to α which is smaller and not close to 0.5. To discuss $\alpha \sim 0.5$ or larger, the higher order terms of β should be studied. This will be an interesting problem especially for the generalized Hopfield model.

Having studied two generalized models, one may think of unlearning in the models which have learned many correlated and different-weight patterns. This will be interesting since, for a set of patterns with different weights and different correlations, one may ask which patterns the system will tend to memorize. We can also think of the model which evolves by alternate learning and unlearning. These situations will lead to quite complicated interactions, which will not arise only by learning.

The concept of unlearning is quite attractive in the sense that it describes the change of interactions using the information generated by the system itself. Another interesting suggestion of this nature is learning by selection using the spin glass model [13]. In this suggestion, initial synaptic interactions are assumed to be random and the system enforces the preexisting low-energy states selectively by learning. Although the system is influenced by input patterns in this case, it will be interesting to study how the paramagnetic phase works in this scheme.

The last subject we want to comment on is the application to optimization problems. It is known that some optimization problems can be formulated as a search for low-energy states of spin models [14,15]. If low-energy states appear more frequently than high-energy states in the paramagnetic configurations, searching for these states will become easier after learning paramagnetic configurations. The very same argument as was given in the text implies that the interactional changes by ''paramagnetic learning'' are also reduced to the correlation function in high temperature. Thus we can use the high-temperature expansion to find the new problem, which is expected to be easier than the original problem. This is an interesting possibility in the study of optimization problems.

After finishing this study, I strongly feel that the noises, which are generated either thermally or dynamically, can be quite essential in information processing in neural networks. Although, in the literature, temperature seems to be introduced rather formally in neural networks, the idea of thermal noise deserves to be studied in the very context of the functions of neural networks. I hope that our study adds another point of view to bridge statistical physics and neuroscience.

[1] J. J. Hopfield, Proc. Natl. Acad. Sci. USA, **79**, 2554 (1982).
[2] D. J. Amit, H. Gutreund, and H. Sompolinsky, Ann. Phys. **173**, 30 (1987).
[3] F. Crick and G. Mitchison, Nature **304**, 111 (1983).
[4] J. J. Hopfield, D. I. Feinstein, and R. G. Palmer, Nature **304**, 158 (1983).
[5] J. L. V. Hemmen, L. B. Ioffe, R. Kühn, and M. Vaas, Physica **163**, 386 (1990).
[6] K. Nokura, J. Phys. A **29**, 3871 (1996).
[7] V. Dotsenko, N. D. Yarunin, and E. A. Dorotheyev, J. Phys. A **24**, 2419 (1991).
[8] S. Wimbauer, N. Klemmer, and J. L. van Hemmen, Neural Networks **7**, 261 (1994).
[9] A. Y. Plakhov and S. A. Semenov, J. Phys. France I **4**, 253 (1994).
[10] L. Personnaz, I. Guyon, and G. Deyfus, J. Phys. (Paris) Lett. **46**, L-359 (1985).
[11] I. Kanter and H. Sompolinsky, Phys. Rev. A **35**, 380 (1987).
[12] M. Mézard, J. P. Nadal, and G. Toulouse, J. Phys. (France) **47**, 1457 (1986).
[13] G. Toulouse, S. Dehaene, and J. P. Changeux, Proc. Natl. Acad. Sci. USA, **83**, 1695 (1986).
[14] M. Mézard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[15] J. J. Hopfield and D. W. Tank, Biol. Cybern. **52**, 141 (1985).