Unlearning of Mixed States in the Hopfield Model – Extensive Loading Case –

Kao Hayashi[1], Chinami Hashimoto[2], Tomoyuki Kimoto[3], and Tatsuya Uezu[1] *

[1]*Graduate School of Humanities and Sciences, Nara Women's University, Nara 630, Japan,*
[2]*Faculty of Science, Nara Women's University, Nara 630, Japan,*
[3]*National Institute of Technology, Oita College, Oita 870-0152, Japan*

We study unlearning of mixed states in the Hopfield model for the extensive loading case. Firstly, we focus on the case I that several embedded patterns are correlated with each other, whereas the rest are uncorrelated. Secondly, we study the case II that patterns are divided into clusters in such a way that patterns in any cluster are correlated but those in two different clusters are not correlated. By using the replica method, we derive the saddle point equations for order parameters under the ansatz of the replica symmetry. The same equations are derived by the self-consistent signal-to-noise analysis as well in case I. In both cases I and II, we find that when the correlation between patterns is large, the network loses its ability to retrieve the embedded patterns, and depending on parameters, a confused memory which is a mixed state and/or the spin glass state emerges. By unlearning of the mixed state, the network gets the ability to retrieve the embedded patterns again in some parameter regions. We find that to delete the mixed state and to retrieve the embedded pattern, the coefficient of unlearning should be chosen appropriately. We perform the Markov chain Monte Carlo simulations and find that the simulation results and theoretical ones agree quite well, except for the SG solution in some parameter region due to the replica symmetry breaking. Furthermore, we find that the existence of many correlated clusters reduces the stabilities of both embedded patterns and mixed states.

1. Introduction

Since an associative memory model, the Hopfield model,[1)] has been proposed, many studies on attractor neural network models have been done.[2–4)] Among others, There is a paper on dreaming in which the authors consider that one role of dreaming is to regulate memories,

*uezu@cc.nara-wu.ac.jp

that is, unnecessary memories are considered to be deleted by dreaming.[5] In this context, studies on unlearning have been done.[6–9, 11–13]

In dreaming, many memories appear, and each memory is usually not strange itself, but dreams themselves are sometimes unrealistic and strange. In the Hopfield model, undesirable memories are the mixed states and spin glass states. A spin glass state does not have any correlation with any embedded patterns, and it seems that it does not appear in dreaming. On the other hand, a mixed state is a mixture of several embedded patterns, and is considered to appear in dreaming.

Let us consider methods to regulate memories in the Hopfield model. One method is to enhance the stability of the embedded patterns, and usually the Hebbian rule is used. However, for this method, the stability of the mixed states composed of strengthened patterns are also enhanced, and thus this method is not appropriate to regulate memories. Another method is to reduce the stability of the mixed states. Unlearning is considered to reduce or lose a memory, and reducing the stability of the mixed states is regarded as unlearning of mixed states. The anti-Hebbian rule, which is the Hebbian learning multiplied by negative coefficient, has been used as the method of unlearning. For example, unlearning at high temperature[12, 13] which is the anti-Hebbian learning of spin glass states and unlearning of the product of local fields by the anti-Hebbian rule[10] have been studied. In the present paper, we also adopt the anti-Hebbian rule for unlearning of mixed states.

Previously we studied unlearning of mixed states in the Hopfield model for the case that the number of patterns p is much smaller than that of neurons, N.[14] By unlearning of mixed states, we showed that there is an unlearning region of parameters where all patterns are retained and all mixed states are deleted, although tuning the parameters in this region is more difficult as p increases since the region shrinks.

It is interesting to study the extensive loading case and also the effect of correlations between embedded patterns. It is known that, when there is a correlation between patterns, mixed states also appear for a rather wide range of parameters, and there are parameter regions where mixed states exist but the memory states do not. Thus, in this paper, we study unlearning of the mixed state in the Hopfield model for the extensive loading case with and without correlations between patterns. In particular, we study two cases, case I that several patterns are correlated and others are uncorrelated, and case II that patterns in the same cluster are correlated and patterns in two different clusters are uncorrelated. We derive the saddle point equations (SPEs) by the replica method for both cases and by self-consistent signal-to-noise analysis (SCSNA) for case I. We show that by unlearning of a mixed state we can

eliminate the mixed state and retrieve a memory pattern in some parameter regions.

This paper is organized as follows. In sect. 2, we study case I. We formulate the problem by the replica method and describe the saddle point equations (SPEs) of overlaps. We compare the theoretical results and results by the Markov chain Monte Carlo simulations (MCMCs) for the case that temperature T is positive and the number of correlated patterns is 3. We also study the de Almeida-Thouless (AT) stabilities,[15] derive the 1-step replica symmetry breaking (1RSB) solution.[16] However, we could not find solutions of the SPE for 1RSB solution numerically. Furthermore, we treat unlearning by the SCSNA, and study the case that the number of correlated patterns is 9 at $T = 0$, and compare the theoretical and numerical results. In sect. 3, we study case II and make a similar analysis to case I. Section 4 contains a summary and discussion of the results. In Appendix A, we derive the free energy and the SPEs for case I. The SPEs of the 1RSB solution is described in Appendix B for case I. In Appendices C and D, we derive the free energy and the SPEs, and the formula for the AT stability for case II, respectively.

2. Case I. Several correlated patterns and others uncorrelated

2.1 Formulation

The Hopfield model is a recurrent network of N neurons and all neurons interact with each other. The state of the ith neuron is represented by s_i. $s_i = 1$ or $s_i = -1$ corresponds to a firing or rest state, respectively. Let $\boldsymbol{\xi}^\mu = (\xi_1^\mu, \xi_2^\mu, ..., \xi_N^\mu)$ be μth pattern, where $\mu = 1, 2, \cdots, p$. We assume that ξ_j^μ takes values of ± 1. Furthermore, we assume that the three patterns $\xi_j^1, \xi_j^2, \xi_j^3$ are correlated with each other, and other patterns are not correlated with each other and with these three patterns. We generate three patterns as follows. Let ${\xi_j}^{(\mathrm{m})}$ be a mother pattern for a fixed j. This takes ± 1 with the probability 1/2. Then, we generate a pattern ξ_j^μ, which takes the value ${\xi_j}^{(\mathrm{m})}$ with the probability $P = \frac{1+\sqrt{a}}{2}$ and $-{\xi_j}^{(\mathrm{m})}$ with the probability $1 - P$. We denote the average of ξ_j^μ with the mother pattern ${\xi_j}^{(\mathrm{m})}$ fixed by $\langle\rangle$ and we have

$$\langle \xi_j^\mu \rangle = {\xi_j}^{(\mathrm{m})} \sqrt{a}, \tag{1}$$

$$\langle \xi_j^\mu \xi_k^\nu \rangle = a\, {\xi_j}^{(\mathrm{m})}\, {\xi_k}^{(\mathrm{m})}\ , \ (\mu \neq \nu). \tag{2}$$

Furthermore, if we take the average over the mother pattern, we have

$$[\xi_j^\mu] = 0, \ \ [\xi_j^\mu \xi_k^\nu] = a\, \delta_{j,k}, (\mu \neq \nu). \tag{3}$$

Here, we denote the average over $\{\xi_j^\mu\}$ and $\{\xi_j^{(m)}\}$ by $[\cdot]$. Now, let us consider unlearning of the mixed state $\xi_j^{\mathrm{mix}} = \mathrm{sgn}(\xi_j^1 + \xi_j^2 + \xi_j^3)$. Denoting the interaction of the Hopfield model by

$J_{jk}^{(\mathrm{H})} = \frac{1}{N}\sum_{\mu=1}^{p}\xi_j^\mu\xi_k^\mu$, we adopt the following interaction,

$$J_{jk} = J_{jk}^{(\mathrm{H})} - \frac{h}{N}\xi_j^{\mathrm{mix}}\xi_k^{\mathrm{mix}} = \frac{1}{N}\sum_{\mu=1}^{p}\xi_j^\mu\xi_k^\mu - \frac{h}{N}\xi_j^{\mathrm{mix}}\xi_k^{\mathrm{mix}}. \tag{4}$$

The Hamiltonian is given as

$$\begin{aligned} H &= -\sum_{j<k}J_{jk}s_js_k = -\frac{N}{2}\sum_{\mu=1}^{p}(\frac{1}{N}\sum_{j=1}^{N}\xi_j^\mu s_j)^2 \\ &\quad +\frac{hN}{2}(\frac{1}{N}\sum_{j=1}^{N}\xi_j^{\mathrm{mix}}s_j)^2 + \frac{p-h}{2}. \end{aligned} \tag{5}$$

Now, we study the system by the replica method. We consider the situation that some of correlated patterns and mixed states composed of these patterns are retrieved. We obtain the same results by the SCSNA. The partition function is $Z = \mathrm{Tr}_{\{s_j\}}\exp\{-\beta H\}$, where T is temperature and $\beta = 1/T$. Regarding patterns as quenched variables, we calculate the free energy $F = -T\ln Z$ averaged over patterns, $[F]$. In order to take the average, we use the replica method and introduce the replica index ρ, $(s_1^\rho, s_2^\rho, \cdots, s_N^\rho)$, $\rho = 1, \cdots, n$. By using the standard recipe, we obtain the free energy per neuron $f = [F]/N = -T[\ln Z]/N$ as

$$\begin{aligned} f &= \frac{1}{n}\sum_{\tau\le 3,\rho}\frac{(m_\rho^\tau)^2}{2} - h\frac{1}{n}\sum_{\rho}\frac{(m_\rho^{\mathrm{mix}})^2}{2} + \frac{\alpha}{2n\beta}\mathrm{Tr}\ln\Big((1-\beta)E - \beta Q\Big) + \frac{\alpha\beta}{2n}\sum_{\rho\neq\sigma}r_{\rho\sigma}q_{\rho\sigma} \\ &\quad -\frac{1}{n\beta}\langle\ln\mathrm{Tr}_{\{s^\rho\}}\exp(\beta H_\eta)\rangle_{\{\eta^1,\eta^2,\eta^3\}}, \end{aligned} \tag{6}$$

$$\beta H_\eta = \sum_{\rho\neq\sigma}\frac{\alpha\beta^2}{2}r_{\rho\sigma}s^\rho s^\sigma + \beta\sum_{\tau\le 3,\rho}\eta^\tau s^\rho m_\rho^\tau - \beta h\sum_\rho \eta^{\mathrm{mix}}s^\rho m_\rho^{\mathrm{mix}}. \tag{7}$$

Here, $\langle\cdot\rangle_{\{\eta^1,\eta^2,\eta^3\}}$ means the average over η^1, η^2, η^3 where η^μ takes ± 1 with the probability $\frac{1\pm\sqrt{a}}{2}$. See Appendix A for the derivation. Here, $\eta^{\mathrm{mix}} = \mathrm{sgn}(\eta^1+\eta^2+\eta^3)$ and we define the so-called spin glass order parameter $q_{\rho\sigma}(\rho\neq\sigma)$ and $n\times n$ matrices, E, K and Q as

$$q_{\rho\sigma} \equiv \frac{1}{N}\sum_j s_j^\rho s_j^\sigma, \tag{8}$$

$$E_{\rho\sigma} \equiv \delta_{\rho\sigma}, \tag{9}$$

$$K_{\rho\sigma} \equiv \delta_{\rho\sigma} - \beta q_{\rho\sigma}, \tag{10}$$

$$Q_{\rho\sigma} \equiv \begin{cases} q_{\rho\sigma}, & \text{for } \rho\neq\sigma \\ 0, & \text{for } \rho = \sigma. \end{cases} \tag{11}$$

Note that $q_{\rho\rho} = 1$ in the above expression. The following relations hold.

$$m_\rho^\mu = \frac{1}{N}\sum_j \xi_j^\mu s_j^\rho, \tag{12}$$

$$r_{\rho\sigma} = \frac{1}{\alpha}\sum_{\mu\geq 4} m_\rho^\mu m_\sigma^\mu. \tag{13}$$

2.2 Replica symmetric solution

We assume the replica symmetry.

$$m_\rho^\tau = m^\tau,\ q_{\rho\sigma} = q\ (\rho\neq\sigma),\ r_{\rho\sigma} = r\ (\rho\neq\sigma). \tag{14}$$

The replica symmetric free energy $f_{\rm RS}$ is obtained as

$$\begin{aligned} f_{\rm RS} &= \sum_{\tau\leq 3}\frac{(m^\tau)^2}{2} - h\frac{(m^{\rm mix})^2}{2} + \frac{\alpha}{2\beta}\Big(\ln(1-\beta+\beta q) - \frac{\beta q}{1-\beta+\beta q}\Big) + \frac{\alpha\beta}{2}r(1-q) \\ &\quad -\frac{1}{\beta}\langle\int Dz \ln\Big(2\cosh\Big\{\beta(\sqrt{\alpha r}z + \sum_{\tau\leq 3}\eta^\tau m^\tau - h\eta^{\rm mix}m^{\rm mix})\Big\}\Big)\rangle_{\{\eta^1,\eta^2,\eta^3\}}. \end{aligned} \tag{15}$$

From this, the SPEs are obtained as

$$m^\tau = \int Dz\langle\eta^\tau\tanh\Big\{\beta(\sqrt{\alpha r}z + \sum_{\nu\leq 3}\eta^\nu m^\nu - h\eta^{\rm mix}m^{\rm mix})\Big\}\rangle_{\{\eta^1,\eta^2,\eta^3\}},\ \tau = 1,2,3. \tag{16}$$

$$m^{\rm mix} = \int Dz\langle\eta^{\rm mix}\tanh\Big\{\beta(\sqrt{\alpha r}z + \sum_{\nu\leq 3}\eta^\nu m^\nu - h\eta^{\rm mix}m^{\rm mix})\Big\}\rangle_{\{\eta^1,\eta^2,\eta^3\}}, \tag{17}$$

$$q = \int Dz\langle\tanh^2\Big\{\beta(\sqrt{\alpha r}z + \sum_{\nu\leq 3}\eta^\nu m^\nu - h\eta^{\rm mix}m^{\rm mix})\Big\}\rangle_{\{\eta^1,\eta^2,\eta^3\}}, \tag{18}$$

$$r = \frac{q}{(1-\beta+\beta q)^2}. \tag{19}$$

There are four kinds of solutions.

Retrieval (R) state $m^1 > 0, m^2 = m^3 < m^1, q > 0$.

Mixed (M) state $m^1 = m^2 = m^3, q > 0$.

Spin glass (SG) state $m^\tau = 0, q > 0,\ (\tau = 1,2,3)$.

Para (P) state $m^\tau = 0, q = 0,\ (\tau = 1,2,3)$.

The transition temperature $T_{\rm SG}$ from the P to SG states is the same as in the Hopfield model and is given by

$$T_{\rm SG} = 1 + \sqrt{\alpha}. \tag{20}$$

2.3 AT stability

Here, we study AT stability.[15] We investigated the replicon mode, and derived a similar formula to that in the Hopfield model.[2] The eigenvalues of the replicon mode are given by

$$\lambda_\pm = -\frac{1}{2}(u+v) \pm \Big(\frac{1}{4}(u+v)^2 + 1 - uv\Big)^{1/2}, \tag{21}$$

$$u = \alpha\beta^2 \int Dz \langle \cosh^{-4}\Big\{\beta(\sqrt{\alpha r}z + \sum_{\nu\leq 3}\eta^\nu m^\nu - h\eta^{\rm mix}m^{\rm mix})\Big\}\rangle_{\{\eta^1,\eta^2,\eta^3\}}, \tag{22}$$

$$v = \Big(1-\beta(1-q)\Big)^{-2}. \tag{23}$$

The conditions for the AT stability are $\lambda_+ > 0$ and $\lambda_- < 0$. Thus, the AT line is given by

$$uv = 1. \tag{24}$$

2.4 Numerical results of unlearning of mixed state with three patterns for $T > 0$

In this subsection, we compare the numerical and theoretical results for $T > 0$. We performed Markov chain Monte Calro simulations (MCMCs). We updated neurons 500×2 Monte Carlo sweeps (MC sweeps) and took average of order parameters during last 500 MC sweeps. Here, one MC sweeps corresponds to N updates. As initial configurations, we took states near to $\boldsymbol{\xi}^1$, $\boldsymbol{\xi}^{\rm mix}$, and $\boldsymbol{\xi}^4$ depending on solutions we wanted to study. In the region where only the SG state exists, we took random initial configurations and used simulated annealing (SA) method. We took 10 samples. In figures below, we display average values of order parameters together with their standard deviations. We studied the following four cases for $N = 10000$ and $\alpha = 0.02$.

Case 1: $a = 0$ and $h = 0$ (no unlearning), case 2: $a = 0$ and $h = 0.2$ (unlearning), case 3: $a > 0$ and $h = 0$ (no unlearning), case 4: $a > 0$ and $h = 0.2$ (unlearning).

The meaning of curves, symbols, and the conditions of simulations are explained only in the caption of Fig.1 since these are the same in Figs. 2 to 6. Almost all figures, error bars are too small to realize.

2.4.1 Temperature dependences of order parameters

First of all, we study the effect of unlearning when $a = 0$ by comparing the results of cases 1 and 2. As seen from Fig. 1 for case 1, there exist the R state for $\lesssim 0.7$, the M state for $\lesssim 0.1$, and the SG state for $\lesssim 1.1$. As seen from Fig. 2 for case 2, the M state disappears and the temperature region where the retrieval succeeds is reduced. That is, unlearning reduces the stability of the R and M states. This is the same as in the finite loading case of $\alpha = 0$.[14] To see the reason for this, we calculate the local field h_i^1 ($h_i^{\rm mix}$) when the pattern $\boldsymbol{\xi}^1$ ($\boldsymbol{\xi}^{\rm mix}$) is input.

$$h_i^1 = \xi_i^1 + a(\xi_i^2 + \xi_i^3) - \frac{h}{2}(1+a)\xi_i^{\rm mix} + O(\sqrt{\alpha}), \tag{25}$$

$$h_i^{\rm mix} = \frac{1}{2}(1+a)(\xi_i^1 + \xi_i^2 + \xi_i^3) - h\xi_i^{\rm mix} + O(\sqrt{\alpha}). \tag{26}$$

From these equations, we note that when $a = 0$, the ratio of the signal term to the noise term in h_i^1 and $h_i^{\rm mix}$ is smaller for $h > 0$ than for $h = 0$. This is because there is correlations between $\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^{\rm mix}$.

Next, let us study the effect of the correlation a without unlearning by comparing the results of cases 1 and 3. As is shown in Figs. 1 and 3, the correlation enhances the stability of the M state and reduces that of the R state. From eqs. (25) and (26), we can see the reason for this. When $h = 0$, the ratio of the signal term to the noise term in h_i^1 is smaller and is larger in $h_i^{\rm mix}$ for $a > 0$ than for $a = 0$.

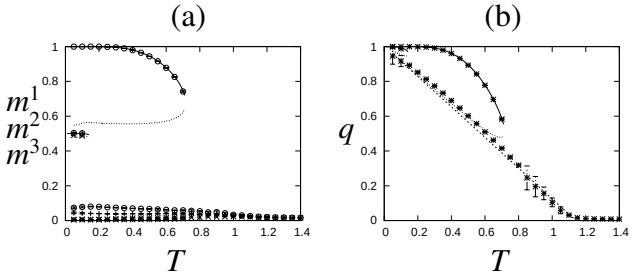

Fig. 1. T dependences of m^μ and q. $a = 0$ and $h = 0$. $\alpha = 0.02$. Curves: RS solution. Solid curve: R state, dashed dotted curve: M state, Dashed curve: SG state. Dotted curves are unstable solutions. Symbols: MCMCs. $\circ : m^2, + : m^2, \times : m^3$, and $*$: SG state. $N = 10000$. Averages are taken from 10 samples. Vertical lines are error bars. (a) m^μ, (b) q.

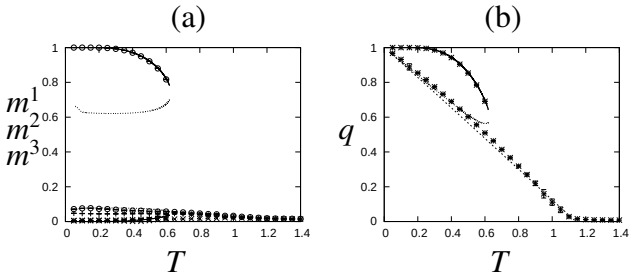

Fig. 2. T dependences of m^μ and q. $a = 0, \alpha = 0.02$, and $h = 0.2$. (a) m^μ, (b) q.

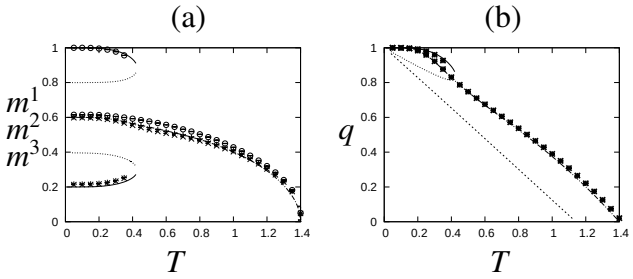

Fig. 3. T dependences of m^μ and q. $a = 0.2$ $h = 0$. $\alpha = 0.02$. (a) m^μ, (b) q.

Now, let us study the effect of unlearning when $a > 0$. In Fig. 4, we show the T dependences of order parameters for case 4. Comparing this figure with Fig. 3, we note that the stable region of the R state is enhanced and that of the M state is reduced. From these figures, we find that at some parameter value where the R state does not exist in case 3, it exists again by unlearning. From these results, it seems that for $a \geq 0$, the effect of unlearning is complicated and difficult to understand by the signal to noise analysis.

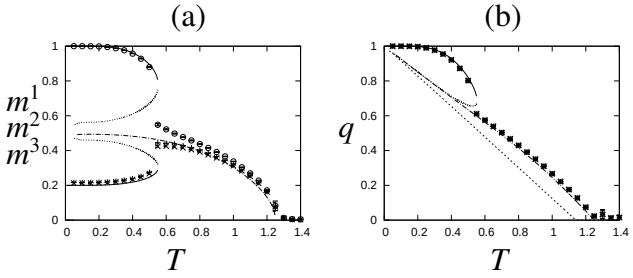

Fig. 4. T dependences of m^μ and q. $a = 0.2$, $\alpha = 0.02$, and $h = 0.2$. (a) m^μ, (b) q.

2.4.2 α dependences of order parameters

We also studied α dependences of order parameters for $a = 0.2$ and $T = 0.5$ with and without unlearning. We found that the R state exists for small α but at some value of α, it disappears and instead the M state appears in both cases with and without unlearning. That is, as α becomes large, the R state tends to disappear and the M state tend to appear. As for the effect of unlearning, we found that the stable α region of the R state becomes wider for $h = 0.2$ than that for $h = 0$, and the stable α region of the M state becomes narrower for $h = 0.2$ than that for $h = 0$. This is consistent with results on the effects by unlearning stated above. We also studied a dependences of order parameters for $\alpha = 0.02$ and $T = 0.5$ with and without unlearning. We found that for small a the R state exists but at some value of a, it disappears and instead the M state appears in both cases with and without unlearning, and this is also consistent with results on the correlation obtained above.

2.4.3 h dependences of order parameters

Here, we study whether the larger h is the better unlearning works. We set values of parameters to $\alpha = 0.02, a = 0.2, T = 0.5$, where only the M state exists for $h = 0$. The P state does not exist. In Fig. 5, we draw the unlearning coefficient h dependences of m and q when $\boldsymbol{\xi}^1$ or $\boldsymbol{\xi}^{\rm mix}$ is retrieved. Numerical and theoretical results on the R and M states agree quite well. The numerical values of q scatter, but a lot of values locate near the theoretical value of the RS solution for the SG state. This shows signs of the breaking of the replica symmetry of

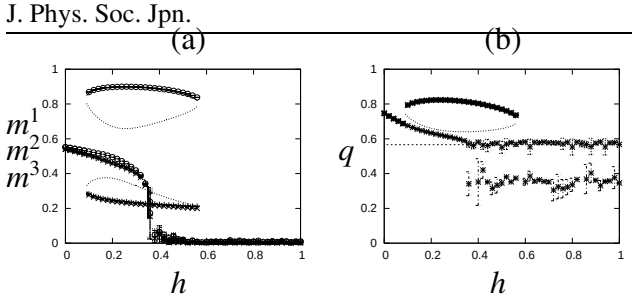

Fig. 5. h dependences of m^μ, q. $\alpha = 0.02, a = 0.2, T = 0.5$. (a) m^μ, (b)q.

the SG states. As is studied later, the eigenvalue of the replicon mode of the Hessian for the SG state is negative, and its absolute value is very small. Thus, we consider even though the replica symmetry is broken for the SG state, q_0 for the replica symmetry breaking solution might have a similar value to that of the RS solution. When we studied the R (M) state, we set initial configurations near to $\boldsymbol{\xi}^1(\boldsymbol{\xi}^{\rm mix})$. We also tried 10 random initial configurations and found that only the M state is retrieved. Thus, it is considered that the basin of the M state is much larger than that of the R state when patterns are correlated. In Fig. 6, we show the result of the case that $\boldsymbol{\xi}^4$ is retrieved. We found that the other embedded states and mixed states than $\boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \boldsymbol{\xi}^3, \boldsymbol{\xi}^{\rm mix}$ are not affected by unlearning of $\boldsymbol{\xi}^{\rm mix}$ as in the finite loading case of $\alpha = 0$.[14)]

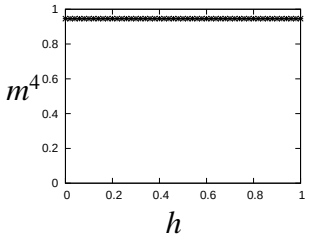

Fig. 6. h dependences of m^4. $\alpha = 0.02, a = 0.2, T = 0.5$.

2.4.4 h dependences of entropy

In Fig. 7, we show the h dependences of entropy for the R, M, and SG solutions. For all solutions, their entropies are positive and the solutions are appropriate.

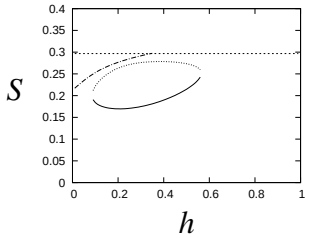

Fig. 7. h dependences of entropy. $\alpha = 0.02, a = 0.2, T = 0.5$. Curves: RS solution.

2.4.5 h dependences of AT stability

In order to study the AT stability of the RS solutions, we studied the h dependences of eigenvalues of the replicon mode for the Hessian of the RS solutions. See Fig. 8.

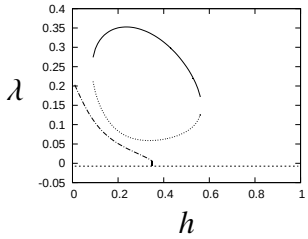

Fig. 8. h dependences of eigenvalues of the replicon mode λ. $\alpha = 0.02, a = 0.2, T = 0.5$. Curves: RS solution.

Eigenvalues are positive for the R and M states and these states are AT stable, but eigenvalues for the SG state take negative values, although absolute values are very small. That is, the SG state is AT unstable and as shown in Fig. 5(b), values of q scatter and this is a sign of the RSB. See Fig. 9.

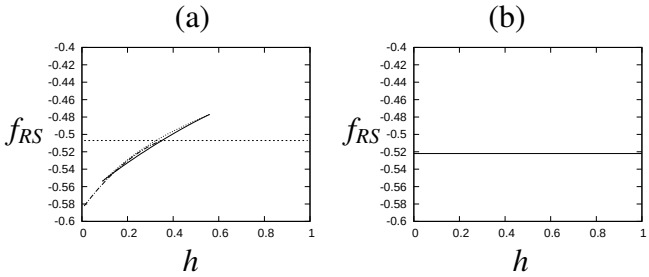

Fig. 9. h dependences of f_{RS}. $\alpha = 0.02, a = 0.2, T = 0.5$. Curves: RS solution. (a) R, M, SG, (b) R_4.

2.4.6 h dependences of free energy

Now, we study the free energy for existing states focusing on the behavior of the R state. Let us denote the free energies of the R, M, SG states, and other embedded pattern R_4 as f_R, f_M, f_{SG}, f_{R_4}, respectively. The SG and R_4 states always exist and f_{SG} and f_{R_4} do not depend on h. We found the following results.

(1) $h < h_1(\sim 0.09)$. All states but the R state exist. $f_M < f_{R_4} < f_{SG}$.

(2) $h_1 < h < h_2(\sim 0.15)$. All states exist and the R state is metastable. $f_M < f_R < f_{R_4} < f_{SG}$.

(3) $h_2 < h < h_3(\sim 0.25)$. All states exist and the R state is stable. $f_R < f_M < f_{R_4} < f_{SG}$.

(4) For $h_3 < h < h_4(\sim 0.35)$. All states exist. The R state is metastable. $f_{R_4} < f_R < f_M < f_{SG}$.

(5) For $h_4 < h < h_5(\simeq 0.56)$. All states but the M state exist. The R state is metastable. $f_{R_4} < f_{SG} < f_R$.

(6) For $h_5 < h$. The R state disappears, and the R_4 and SG states exist. $f_{R_4} < f_{SG}$.
The free energy of the R and M states increase as h increases, and their stabilities reduce. The free energy of the R_4 and SG states are not affected by unlearning.
Thus, to retain the R state and delete the M state, h should be set to $h_4 < h < h_5$.

2.5 1 step replica symmetry breaking solution

We derived the free energy f_{1RSB} and the SPEs. See Appendix B. We tried to find solutions of the SPEs for 1RSB. However, we could not find any of them. Thus, it is considered that more than 1 step replica symmetry is broken.

2.6 Numerical results of unlearning mixed state with 9 patterns for $T = 0$

In this subsection, we study unlearning by the SCSNA. By expressing the state of ith neuron by x_i, the retrieval dynamics is given by

$$\frac{d}{dt}x_i = -x_i + \tanh(\beta \sum_{j\neq i} J_{ij}x_j). \tag{27}$$

By the SCSNA, we obtain the equations which is essentially the same as the SPEs (16), (29),(18), and (19). we compare the numerical and theoretical results for $T = 0$. We study the case of $p = 9$ and investigate the h dependence of m^μ for fixed α. The M state which is unlearned is $\xi_i^{\rm mix} = \mathrm{sgn}(\xi_i^1 + \xi_i^2 + \cdots + \xi_i^9)$. The SPEs are

$$m^\tau = \langle \eta^\tau \mathrm{erf}\Big\{ \frac{1}{\sqrt{2\alpha r}}(\sum_{\nu\leq 9} \eta^\nu m^\nu - h\eta^{\rm mix} m^{\rm mix})\Big\}\rangle_{\{\eta^1,\eta^2,\cdots,\eta^9\}},\ \tau = 1, 2, \cdots, 9. \tag{28}$$

$$m^{\rm mix} = \langle \eta^{\rm mix} \mathrm{erf}\Big\{ \frac{1}{\sqrt{2\alpha r}}(\sum_{\nu\leq 9} \eta^\nu m^\nu - h\eta^{\rm mix} m^{\rm mix})\Big\}\rangle_{\{\eta^1,\eta^2,\cdots,\eta^9\}}, \tag{29}$$

$$r = \frac{1}{(1-U)^2}, \tag{30}$$

where $q = 1, U = \lim_{\beta\to\infty}\beta(1-q)$ and $\mathrm{erf}(x) \equiv \frac{2}{\sqrt{\pi}}\int_0^x Due^{-u^2}$. In Fig. 10, we show the numerical results of unlearning of the M state with initial condition $\boldsymbol{s} = \boldsymbol{\xi}^1$. As shown in Fig. 10(a), for $a = 0.1$ and $\alpha = 0.07$, when $h = 0$, the symmetric M state ($m^1 = m^2 = \cdots = m^9$) is stable and the memory pattern does not exist. For $0.3 < h < 1.3$, the memory pattern m^1 exists and is $\simeq 1$. For $1.3 < h$, it seems that the memory pattern disappears and the system shows complex behavior. Similar result is obtained for $a = 0.1$ and $\alpha = 0.09$. See Fig. 10(b). Next, we study the case of unlearning a M state which is retrieved for $h = 0$. See Fig. 11. Theoretical results are the same as in Fig. 10. We note that region where the memory pattern is stable becomes wider than that in Fig. 10. This implies that unlearning of retrieved M state is more efficient than unlearning of the pure M state, $\xi_i^{\rm mix} = \mathrm{sgn}(\xi_i^1 + \xi_i^2 + \cdots + \xi_i^9)$.

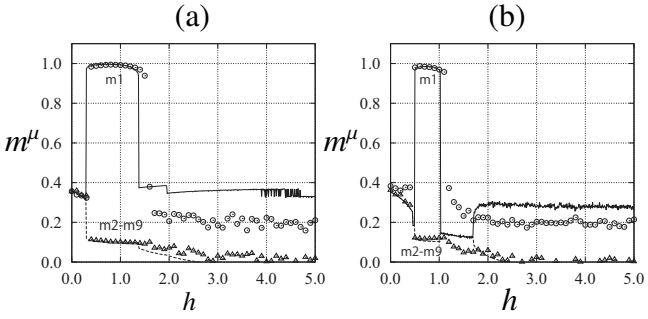

Fig. 10. h dependences of m^μ. $a = 0.1$. Curves: RS solution. Symbols: numerical results by Euler method, time increment $\Delta t = 0.1, N = 10000$. $\circ : m^1, \triangle : m^2 - m^9$. Initial condition is $\boldsymbol{\xi}^1$. Unlearning of the M state. (a) $\alpha = 0.07$, (b) $\alpha = 0.09$.

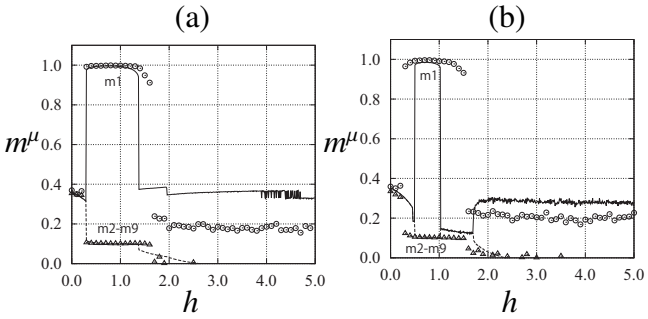

Fig. 11. h dependences of m^μ. $a = 0.1$. Curves: RS solution. Symbols: numerical results by Euler method, time increment $\Delta t = 0.1, N = 10000$. $\circ : m^1, \triangle : m^2 - m^9$. Initial condition is $\boldsymbol{\xi}^1$. Unlearning of a M state which is retrieved for $h = 0$. (a) $\alpha = 0.07$, (b) $\alpha = 0.09$.

3. Case II. Uncorrelated clusters of correlated patterns

Next, we consider the case that patterns in the same cluster are correlated and those in any two different clusters are uncorrelated. For simplicity, we assume that each cluster contains p patterns and the correlation between two of them is a, and the number of clusters is M. We define $\alpha = \frac{Mp}{N}$, N is the total number of neurons. We denote the ν-th pattern in the τ-th cluster as $\boldsymbol{\xi}^{(\tau,\nu)}$, where $\nu = 1, 2, \cdots, p$ and $\tau = 1, 2, \cdots, M$.

$$[\xi_j^{(\tau,\nu)}] = 0, \quad [\xi_j^{(\tau,\nu)}\xi_k^{(\omega,\mu)}] = \left(\delta_{\mu,\nu}(1-a) + a\right)\delta_{j,k}\delta_{\tau,\omega}. \tag{31}$$

We assume that the patterns in the first cluster $\tau = 1$ are retrieved. We denote these patterns as $\boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \cdots, \boldsymbol{\xi}^p$ and the M state $\{\mathrm{sgn}(\xi_i^1 + \xi_i^2 + \cdots \xi_i^p)\}$ as $\{\xi_i^{\mathrm{mix}}\}$. The interaction J_{jk} is given by

$$J_{jk} \quad = \quad \frac{1}{N}\sum_{\tau=1}^{M}\sum_{\nu=1}^{p}\xi_j^{(\tau,\nu)}\xi_k^{(\tau,\nu)} - \frac{h}{N}\xi_j^{\mathrm{mix}}\xi_k^{\mathrm{mix}}. \tag{32}$$

We define the overlap $m^{(\tau,\nu)}$ as

$$m^{(\tau,\nu)} = \frac{1}{N}\sum_{i=1}^{N}\xi_i^{(\tau,\nu)}s_i. \tag{33}$$

The Hamiltonian is given as

$$\begin{aligned} H &= -\sum_{j<k}J_{jk}s_js_k \\ &= -\frac{N}{2}\Big(\sum_{\mu=1}^{p}(m^{(1,\mu)})^2+\sum_{\tau=2}^{M}\sum_{\nu=1}^{p}(m^{(\tau,\nu)})^2\Big)+\frac{hN}{2}(m^{\rm mix})^2+\frac{pM-h}{2}. \end{aligned} \tag{34}$$

By the replica method introducing n replicas $s_i^\rho, (\rho=1,\cdots,n)$, we derive the free energy f as

$$\begin{aligned} nf &= \frac{1}{2}\sum_{\mu=1}^{p}\sum_{\sigma}(m_\sigma^\mu)^2-h\frac{1}{2}\sum_{\sigma}(m_\sigma^{\rm mix})^2+\frac{\alpha}{2p\beta}\,{\rm Tr}\ln K+\frac{\alpha\beta}{2}\sum_{\sigma\neq\rho}r_{\sigma\rho}q_{\sigma\rho} \\ &\quad -\frac{1}{\beta}\langle\ln{\rm Tr}_{\{s^\sigma\}}e^{H_\eta}\rangle_{\{\eta^1,\eta^2,\cdots,\eta^p\}}, \end{aligned} \tag{35}$$

$$H_\eta=\frac{\alpha\beta^2}{2}\sum_{\sigma\neq\rho}r_{\sigma\rho}s^\sigma s^\rho+\beta\sum_{\sigma}\sum_{\mu=1}^{p}\eta^\mu s^\sigma m_\sigma^\mu-\beta h\sum_{\sigma}\eta^{\rm mix}s^\sigma m_\sigma^{\rm mix}, \tag{36}$$

where $m_\rho^{(\tau,\nu)}=\frac{1}{N}\sum_{i=1}^{N}\xi_i^{(\tau,\nu)}s_i^\rho$, $m_\rho^\nu\equiv m_\rho^{(1,\nu)}$, and $q_{\rho\sigma}$ and $r_{\rho\sigma}$ are the same as in case I, K is an $(np)\times(np)$ matrix and given by

$$K = \begin{pmatrix} K_1 & K_2 & \dots & K_2 \\ K_2 & K_1 & \dots & K_2 \\ \vdots & \ddots & \vdots & K_2 \\ K_2 & K_2 & \dots & K_1 \end{pmatrix}, \tag{37}$$

$$K_1, K_2, Q, E_1 :\ n\times n \text{ matrices} \tag{38}$$

$$K_1=(1-\beta)E_1-\beta Q, K_2=-\beta a(E_1+Q), \tag{39}$$

$$Q_{\alpha\beta}=(1-\delta_{\alpha,\beta})q_{\alpha\beta},\ E_1 :\ \text{unit matrix}. \tag{40}$$

$\langle\cdot\rangle_{\{\eta^1,\eta^2,\cdots,\eta^p\}}$ is the same meaning as in I. The free energy of the RS solution $f_{\rm RS}$ becomes

$$\begin{aligned} f_{RS} &= \sum_{\nu_1\leqq p}\frac{(m^{(1,\nu_1)})^2}{2}-h\frac{(m^{\rm mix})^2}{2}+\frac{\alpha\beta}{2}r(1-q) \\ &\quad +\frac{M}{2\beta N}\Bigg[(p-1)\left\{\log\left(1-\beta(1-q)(1-a)\right)-\frac{\beta q(1-a)}{1-\beta(1-q)(1-a)}\right\} \\ &\quad +\log\left(1-\beta(1-q)(1+(p-1)a)\right)-\frac{\beta q(1+(p-1)a)}{1-\beta(1-q)(1+(p-1)a)}\Bigg] \end{aligned}$$

$$-\frac{1}{\beta}\int Dz\langle\log\left\{2\cosh\left(\beta\left(\sqrt{\alpha r}z+\sum_{\nu_1\leqq p}\eta^{\nu_1}m^{(1,\nu_1)}-h\eta^{\rm mix}m^{\rm mix}\right)\right)\right\}\rangle_{\{\eta^1,\eta^2,\cdots,\eta^p\}}. \quad (41)$$

Equations for $m^\tau, m^{\rm mix}$ and q of the RS solution is the same as in case I, and the equation for r becomes

$$r = \frac{q}{3}\Big[\frac{2(1-a)^2}{\Big(1-\beta(1-q)(1-a)\Big)^2}+\frac{(1+2a)^2}{\Big(1-\beta(1-q)(1+2a)\Big)^2}\Big]. \quad (42)$$

See Appendix C for the derivation.

3.1 AT stability

The AT stability is also similar to that in case I. The only difference is the expression of v.

$$\begin{aligned} v = &\Big(1+a^2(p-1)\Big)(l_1-\bar{l}_1)^2+2a(p-1)\Big(2+a(p-2)\Big)(l_1-\bar{l}_1)(l_2-\bar{l}_2)\\ &+(p-1)\Big(1+2a(p-2)+a^2(p^2-3p+3)\Big)(l_2-\bar{l}_2)^2. \end{aligned} \quad (43)$$

$l_1, \bar{l}_1, l_2, \bar{l}_2$ and details of the derivation are given in Appendix D.

3.2 Numerical results

We set $p = 3$. Similar to case I, we numerically study T and h dependences of order parameters etc. We show results for m^1, m^2, m^3, and $m^{\rm mix}$. Results for q are shown in summary and discussion.

We studied the following four cases for $N = 90000$ and $\alpha = 0.02$.

Case 1: $a = 0.15$ and $h = 0$ (no unlearning), case 2: $a = 0.15$ and $h = 0.2$ (unlearning), case 3: $a = 0.2$ and $h = 0$ (no unlearning), case 4: $a = 0.2$ and $h = 0.2$ (unlearning).

The meaning of curves and symbols are the same as in case I.

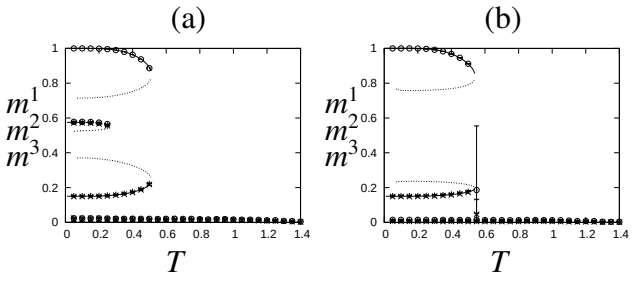

Fig. 12. T dependences of m^μ. $a = 0.15$ and $\alpha = 0.02$. Curves: RS solution, symbols: MCMCs. $N = 90000$. Averages are taken from 10 samples. Vertical lines are error bars. (a) $h = 0$, (b) $h = 0.2$.

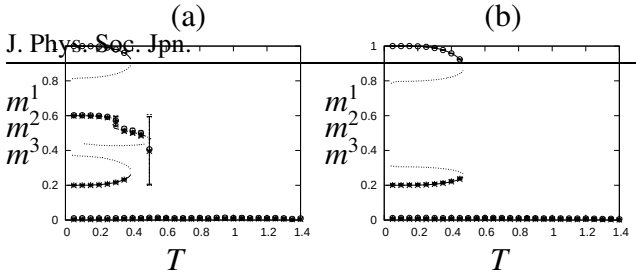

Fig. 13. T dependences of m^μ. $a = 0.2, \alpha = 0.02$. Curves: RS solution, symbols: MCMCs. $N = 90000$. Averages are taken from 10 samples. Vertical lines are error bars. (a) $h = 0$, (b) $h = 0.2$.

3.2.1 Temperature dependences of m^1, m^2 and m^3

Results for cases 1 and 2 with correlation $a = 0.15$ are shown in Fig. 12. As in case I, the M state disappears by unlearning. In cases 3 and 4, the correlation is increased to 0.2 from 0.15 in cases 1 and 2, and results are shown in Fig. 13. We found that the M state disappears by unlearning for $a = 0.2$, as well. Furthermore, by comparing cases 1 and 2 we note that the stability of the M state increases and that of the R state decreases as the correlation increases as in case I.

Now, let us compare the results in case II with those in case I. By comparing Figs. 3 and 13, we found that the existence of clusters reduces the temperature region where the M state is stable. The reason for this seems that in case II, r increases, that is, the effect of non-retrieved patterns is stronger than in case I, because of the existence of other clusters within which the patterns are correlated. Furthermore, the existence of clusters causes more complicated bifurcation structures, e.g., the first-order phase transition between two different mixed states as shown in Fig 14.

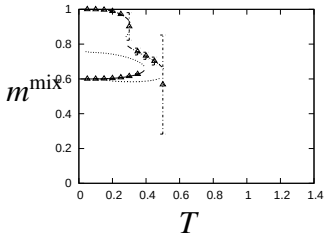

Fig. 14. T dependences of m^{mix}. $\alpha = 0.02, a = 0.2, T = 0.5$. Curves: RS solution.

3.2.2 h dependences of m^1, m^2, and m^3

From Fig. 13, we note that at $T = 0.45$, only the M state exists at $h = 0$, whereas the M state does not exist and the R state exists at $h = 0.2$. To study in more detail, fixing $T = 0.45$ and $a = 0.2$, we studied the h dependences of m^1, m^2, and m^3 both in cases I and II. Comparing Fig. 15 (a) and (b), we note that the M state disappears at a smaller values of h in case II than in case I, and the R state appears at a larger value of h and disappears at a smaller value of

h in case II than in case I. As a result, the existence region of h becomes narrower in case II than in case I. Therefore, it is concluded that the stabilities of the R and M states reduce due to the existence of clusters.

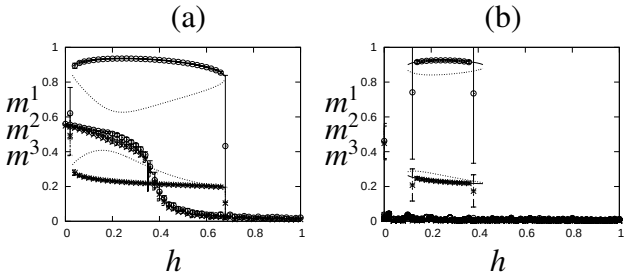

Fig. 15. h dependences of m^μ. $a = 0.2, \alpha = 0.02$, and $T = 0.45$. Curves: RS solution, symbols: MCMCs. $N = 90000$. Averages are taken from 10 samples. Vertical lines are error bars. (a) case I, $N = 100000$, (b) case II, $N = 90000$.

3.2.3 h dependences of entropy

In Fig. 16, we show the h dependences of entropy for the R, M, and SG states. For all states, their entropies are positive and these states are appropriate.

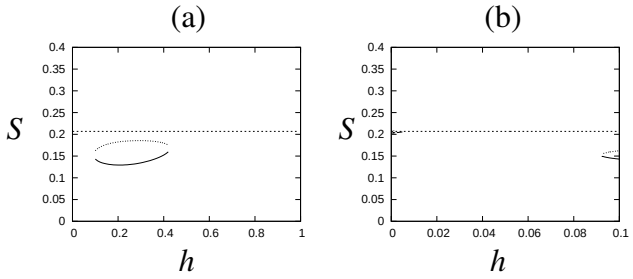

Fig. 16. h dependences of entropy. $\alpha = 0.02, a = 0.2, T = 0.45$. Curves: RS solution. (a) $0 \leqq h \leqq 1$ (b) $0 \leqq h \leqq 0.1$.

3.2.4 h dependences of AT stability

In Fig. 17, we show the h dependences of eigenvalues of the replicon mode for the Hessian of the RS solutions. Similar to case I, the R and M states are AT-stable, but the SG state is AT-unstable.

3.2.5 h dependences of free energy

In Fig. 18, we show the h dependences of free energy for the R, M, and SG states. The free energy of the SG state is lowest among all states for any h, and this result is different from

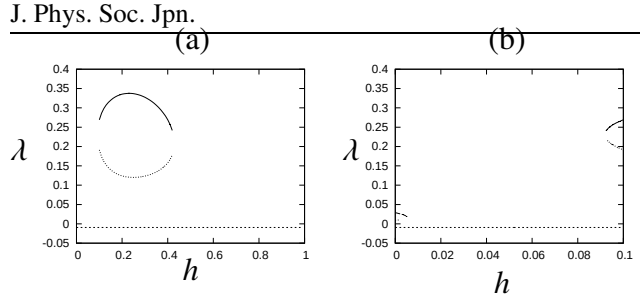

Fig. 17. h dependences of eigenvalues of the replicon mode λ. $\alpha = 0.02, a = 0.2, T = 0.45$. Curves: RS solution. (a) $0 \leqq h \leqq 1$ (b) $0 \leqq h \leqq 0.1$.

the result that the M state has lowest free energy at small values of h in case I. We obtained the following results when h is changed.

(1) $h < h_1 (\sim 0.005)$. The M and SG states exist. $f_{\mathrm{SG}} < f_{\mathrm{M}}$.

(2) $h_1 < h < h_2 (\sim 0.0925)$. Only the SG state exists.

(3) $h_2 < h < h_3 (\sim 0.42)$:.The R state appears. $f_{\mathrm{SG}} < f_{\mathrm{R}}$.

(4) $h_3 < h$. Only the SG state exists.

At $T = 0.45$, the coexistence region of the R and M states does not exist. Thus, the change in stability between the R and M states does not exist in case II different from in case I.

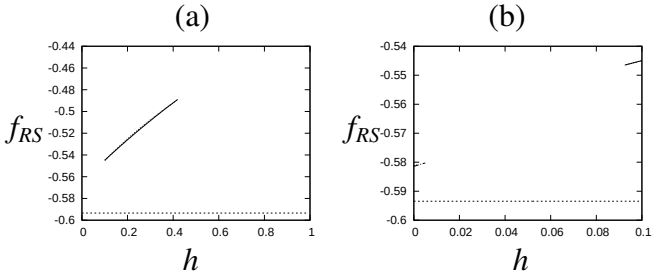

Fig. 18. h dependences of f_{RS}. $\alpha = 0.02, a = 0.2, T = 0.45$. Curves: RS solution. (a) $0 \leqq h \leqq 1$ (b) $0 \leqq h \leqq 0.1$.

4. Summary and Discussion

In conventional Hopfield Model, not only retrieval (R) state but also mixed (M) states are stable. According to previous research, unlearning of mixed states decrease stability of the M states and increase that of the R state for a finite number of embedded patterns. We studied unlearning of a mixed state for the extensive loading case that the number of patterns p is proportional to that of neurons N and for the case that the correlation between patterns exists. We focused on the following two cases; case I: Only three patterns are correlated and the rest of patterns are uncorrelated, case II: Ensemble of clusters composed of correlated patterns such that there is no correlation between the clusters. We used the replica method

and derived the saddle point equations under the ansatz of the replica symmetry. We performed the Markov chain Monte Carlo simulations (MCMCs) and compared the numerical and theoretical results.

First, we summarize the results for case I. We examined the temperature dependence of some order parameters: the overlap between embedded patterns $\{\xi_i^1\} \sim \{\xi_i^3\}$ and the neuron configuration $\{S_i\}$, m^1, m^2, m^3, and the spin glass order parameter, q. We performed MCMCs for the following four cases with the different values of correlation a and unlearning coefficient h: (1) $a = 0, h = 0$, (2) $a = 0, h = 0.2$, (3) $a = 0.2, h = 0$, (4) $a = 0.2, h = 0.2$. Comparing (1) and (2), we noted that the M state is removed by unlearning for $a = 0$. However, the temperature region where the R state is stable is also reduced. From the results of (1) and (3), we found that the stability of the M state increases and that of the R state decreases with the increase of a. For $a = 0.2$ and $T = 0.5$, we also examined the h dependence and found that the R state is removed for too large h. Thus, we should set h appropriately. From the results of h dependences, the retrieval solution R_4, which is the state with the large overlap m_4 and is uncorrelated with the target mixed state for unlearning, is not affected by unlearning. In a spin glass (SG) state, we observed the two values of q, one of which is near to the replica symmetric (RS) solution and the other is different from the RS solution. When we performed the simulations with 10 random initial states, the M state appeared for all of the initial states. Thus, we consider that the basin of the M states are larger than that of the R state.

We examined the h dependence of entropy and confirmed that entropy is positive for the R, M, and SG states. In addition, from the results of the h dependence of the eigenvalue corresponding to the replicon mode λ we found that the RS solution is stable for the R and M states but is unstable for the SG state. Therefore, it is reasonable that there exists disagreement between the theoretical and numerical results for the SG state. Furthermore, we studied the h dependence of the free energy. From the results for the R and M states, we noted that the stability changes as h changes. In ascending order of h, the stability changes as follows; only the M state is stable, the R state is restored but it is metastable, the R state is stable and the M state is metastable, the M state is removed, and the R state is removed. Moreover, the free energy of R_4 is lower than that of the SG state, and they are more stable than the R and M states when h is small but more unstable when h is large.

Next, we summarize the results for case II. We compared the results of the temperature dependence for the following four cases: (1) $a = 0.15, h = 0$, (2) $a = 0.15, h = 0.2$, (3) $a = 0.2, h = 0$, (4) $a = 0.2, h = 0.2$. As is in the case I, the M state is removed by unlearning. Comparing I(3) and II(3), we found that the stable region of the M state for case II become

narrower than that in case I. Furthermore, we compared the results of the h dependence for cases I and II with $a = 0.2, T = 0.45$, and noted that the M state is removed at smaller h for case II than case I. Moreover, the stable region of the R state decreases. Thus, we confirmed that the correlation between embedded patterns decreases the stability of the R and M states.

We studied the h dependence of entropy and λ. We confirmed that entropy is positive for the R, M, SG states, and the RS solution is stable for the R and M states and unstable for the SG states as in case I. For $a = 0.2$ and $T = 0.45$, the R state is restored after the M state is removed. Thus, the coexistence of the R and M states is confirmed in case I, but it isn't observed in case II. In contrast to case I, the SG state is always more stable than the R and M states.

Now, let us discuss the stability reduction of solutions due to the existence of clusters. In both cases I and II, by comparing temperature regions and the range of h where the R and M states exist, we confirmed that the stabilities of the R and M states become more weak in case II than in case I. Furthermore, in case I, by the MCMCs using random initial configurations, we found that the basin of the M state is much larger than that of the R state. Therefore, in case II, it seems that the mixed states with a large basin exist in each cluster, and as a result, each retrieval state and each mixed state are affected and their stabilities reduce.

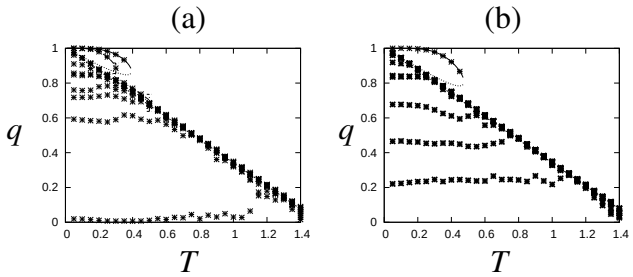

Fig. 19. T dependences of q in case II. $\alpha = 0.02, a = 0.2, h = 0$. Curves: RS solution. , symbols: MCMCs. $N = 90000$. (a) $h = 0$. Among upper two curves existing in low temperature, the upper one is for the R state and the lower one is for the M state. (b) $h = 0.2$. The upper curve existing in low temperature is for the R state.

Next, we discuss the SG states. In case II, we performed simulated annealing (SA) at $a = 0.2$ and $h = 0.2$ for $N = 90000$. We show the temperature dependences of q in Fig. 19. In these figures for the R and M states, the sample average and standard deviation over 10 samples are shown, whereas results of 10 samples are shown for the SG state. From these figures, we confirmed that there are several metastable states for the SG state. Thus, for the SG state, it seems that the replica symmetry breaking (RSB) takes place. In case I, since the stability of the M state is stronger than in case II, when the SA was performed at $a = 0.2$ and

$h = 0.2$, the final states are the M or R state and the SG state does not appear, as shown in Fig. 4(b). By setting to $h = 0.4$ in case I, we could reduce the existence temperature region of the M state and observe the SG state. In Fig. 20 (a) and (b), we show the results by SA for $N = 10000$ and $N = 100000$. From these figures, we confirmed that several metastable states exist in case I, as well. Comparing the result for $N = 10000$ with that for $N = 100000$, we note that the number of metastable states increases as N increases. In both cases I and II, it is found that the SG state is AT-unstable. Furthermore, we examined the h dependence of q in case II for $a = 0.2$ and $T = 0.45$ and found the SG state which has many metastable states. What kind of the RSB takes place is one of future problems.

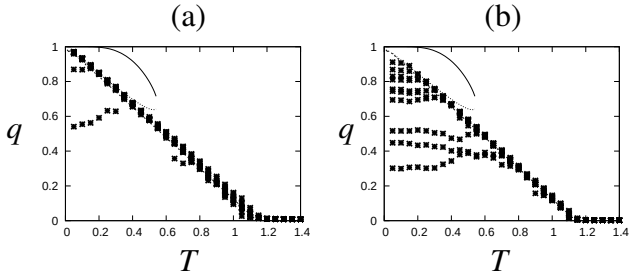

Fig. 20. T dependences of q in case I. $\alpha = 0.02, a = 0.2, h = 0.4$. Curves: RS solution. , symbols: MCMCs. (a) $N = 10000$ (b) $N = 100000$.

Acknoledgement

This work was partially supported by Grant-in-Aid for Scientific Rese arch (C) No. 16K05474 and No. 17K00357 from the Ministry of Education, Culture, Sports, Scienc e and Technology of Japan.

References

1) J. J. Hopfield, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982).
2) D. J. Amit, H. Gutfreund, and H. Somplinski, Phys. Rev. A **32**, 1007 (1985).
3) D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning Internal Representations by Error Propagation, Parallel Distributed Processing, Exploration in the Microstructures of Cognition* (MIT Press, Cambridge, 1986) Vol. 1, Chap. 8, p. 318.
4) J. Hertz, A. Krogh, and R. G.Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City, 1991).
5) F. C. Crick, and G. Mitchison, Nature **304**, 111 (1983).
6) J. J. Hopfield, D. I. Feinstein, and R. G. Palmer, Nature **304**, 158 (1983).
7) S. Wimbauer, N. Klemmer, and J. L. van Hemmen, Neural Networks **7**, 219 (1994).
8) M. C. D. Barrozo, and T. J. P. Penna, Int. J. Mod. Phys. C **5**, 503 (1994).
9) S. Wimbauer, and J. L. van Hemmen, in *Proceedings on Analysis of Dynamical and Cognitive Systems, Advanced Course* (Springer-Verlag, London, 1995) p. 121.
10) S. A. Semenov and I. B. Shuvalova, in *Advances in Neural Information Processing Systems 8 (NIPS 1995)* edited by D. S. Touretzky, M. C. Mozer and M. E. Hasselmo.
11) J. A. Horas, and E. A. Bea, Int. J. Neur. Syst. **12**, 109 (2002).
12) K. Nokura, Phys. Rev. E **54**, 5571 (1996).
13) K. Nokura: J. Phys. A: Math. Gen. **31**, 7447 (1998).
14) H. Ohtani, M. Yoshida, S. Kiyokawa, and T. Uezu, J. Phys. Soc. Jpn., **84**, No. 1, 014002 (2015).
15) J. R. L. de Almeida, and D. J. Thouless, J. Phys. A: Math. Gen. **11**, 983 (1978)
16) G. Parisi, J. Phys. A **13** 1101 (1980); ibid. **13** L115, (1980); ibid. **13** 1887, (1980).

5. Appendix A. Derivation of the free energy in case I

In this appendix, we derive the free energy by the replica method. We start from the replicated partition function,

$$Z^n = e^{-\beta\frac{n(p-h)}{2}}\,\mathrm{Tr}_{\{s_j^\rho\}}\exp\{\frac{N\beta}{2}\sum_{\mu,\rho}(\frac{1}{N}\sum_{j=1}^{N}\xi_j^\mu s_j^\rho)^2-\frac{hN\beta}{2}(\frac{1}{N}\sum_{j=1}^{N}\xi_j^{\mathrm{mix}}s_j^\rho)^2\}, \tag{44}$$

where T is temperature and $\beta = 1/T$. By using the Hubbard-Stratonovich transformation, we obtain

$$Z^n = e^{-\beta\frac{n(p-h)}{2}}\,\mathrm{Tr}_{\{s_j^\rho\}}\int_{-\infty}^{\infty}\Big(\prod_{\mu,\rho}\sqrt{\frac{\beta N}{2\pi}}dm_\rho^\mu\Big)\exp\Big\{\beta N\sum_{\mu,\rho}\Big(-\frac{(m_\rho^\mu)^2}{2}+\frac{1}{N}\sum_{j=1}^{N}\xi_j^\mu s_j^\rho m_\rho^\mu\Big)\Big\}$$

$$\times \int_{-\infty}^{\infty} \Big(\prod_{\rho} \sqrt{\frac{\beta N}{2\pi}} dm_{\rho}^{\rm mix} \Big) \exp\Big\{ \beta N \sum_{\rho} \Big(-\frac{(m_{\rho}^{\rm mix})^2}{2} + \frac{\sqrt{-h}}{N} \sum_{j=1}^{N} \xi_j^{\rm mix} s_j^{\rho} m_{\rho}^{\rm mix} \Big) \Big\}. \quad (45)$$

Below, we study the case that the three patterns are correlated with each other and others are uncorrelated. We assume that m^1, m^2, m^3 and $m^{\rm mix}$ is of the order of $O\Big((\frac{1}{N})^0\Big)$ and others are higher order. We denote the average over $\{\xi^{(m)}\}$ and $\{\xi^{\mu}\}$ by $[\cdot]$. Then, for $\mu \geq 4$, we have

$$[m_{\rho}^{\mu}] = \frac{1}{N} \sum_{j=1}^{N} [\xi_j^{\mu}] s_j^{\rho} = 0, \quad (46)$$

$$[(m_{\rho}^{\mu})^2] = \frac{1}{N^2} \sum_{j,k=1}^{N} [\xi_j^{\mu} \xi_k^{\mu}] s_j^{\rho} s_k^{\rho}. \quad (47)$$

Because of $[\xi_j^{\mu} \xi_k^{\mu}] = \delta_{jk}$, we obtain

$$[(m_{\rho}^{\mu})^2] = \frac{1}{N^2} \sum_{j}^{N} (s_j^{\rho})^2 = \frac{1}{N}. \quad (48)$$

Thus, $m_{\rho}^{\mu} = O(\frac{1}{\sqrt{N}})$ for $\mu \geq 4$. Now, we set $m_{\rho}^{\mu} \to \frac{m_{\rho}^{\mu}}{\sqrt{\beta N}}$ for $\mu \geq 4$. By taking the average over ξ_j^{μ} for $\mu \geq 4$, we obtain,

$$\begin{aligned}
[Z^n] = {} & e^{-\beta \frac{n(p-h)}{2}} \mathrm{Tr}_{\{s_j^{\rho}\}} \int_{-\infty}^{\infty} \Big(\prod_{\rho,\tau \leq 3} \sqrt{\frac{\beta N}{2\pi}} dm_{\rho}^{\tau} \Big) \Big(\prod_{\mu \geq 4, \rho} \frac{1}{\sqrt{2\pi}} dm_{\rho}^{\mu} \Big) \Big(\prod_{\rho} \sqrt{\frac{\beta N}{2\pi}} dm_{\rho}^{\rm mix} \Big) \\
& \times \Big[\exp\Big\{ \beta N \sum_{\tau \leq 3, \rho} \Big(-\frac{(m_{\rho}^{\tau})^2}{2} + \frac{1}{N} \sum_{j=1}^{N} \xi_j^{\tau} s_j^{\rho} m_{\rho}^{\tau} \Big) + \beta N \sum_{\rho} \Big(-\frac{(m_{\rho}^{\rm mix})^2}{2} + \frac{\sqrt{-h}}{N} \sum_{j=1}^{N} \xi_j^{\rm mix} s_j^{\rho} m_{\rho}^{\rm mix} \Big) \\
& + \sum_{\mu \geq 4, \rho} \Big(-\frac{(m_{\rho}^{\mu})^2}{2} \Big) + \frac{\beta}{2N} \sum_{\mu \geq 4} \sum_{j} \sum_{\rho,\sigma} m_{\rho}^{\mu} m_{\sigma}^{\mu} s_j^{\rho} s_j^{\sigma} \Big\} \Big]. \quad (49)
\end{aligned}$$

Now, we take the average over ξ_j^{τ} for $\tau = 1, 2, 3$ and $\xi_j^{(m)}$. Let us define A_j as

$$A_j \equiv \frac{1}{2} \sum_{\xi_j^{(m)} = \pm 1} \langle \exp\Big\{ \beta N \Big(\sum_{\tau \leq 3, \rho} \frac{1}{N} \xi_j^{\tau} s_j^{\rho} m_{\rho}^{\tau} + \sum_{\rho} \frac{\sqrt{-h}}{N} \xi_j^{\rm mix} s_j^{\rho} m_{\rho}^{\rm mix} \Big) \Big\} \rangle_{\xi_j^1, \xi_j^2, \xi_j^3} \exp\Big(\frac{\beta}{2N} \sum_{\mu \geq 4} \sum_{\rho,\sigma} m_{\rho}^{\mu} m_{\sigma}^{\mu} s_j^{\rho} s_j^{\sigma} \Big).$$

We set $\eta_j^{\tau} \equiv \xi_j^{(m)} \xi_j^{\tau}$ ($\tau = 1, 2, 3, \mathrm{mix}$). Then, η_j^{τ} takes 1 with the probability P and -1 with the probability $1 - P$. A_j becomes

$$\begin{aligned}
A_j = {} & \frac{1}{2} \sum_{\xi_j^{(m)} = \pm 1} \langle \exp\Big\{ \beta N \Big(\sum_{\tau \leq 3, \rho} \frac{1}{N} \eta_j^{\tau} \xi_j^{(m)} s_j^{\rho} m_{\rho}^{\tau} + \sum_{\rho} \frac{\sqrt{-h}}{N} \eta_j^{\rm mix} \xi_j^{(m)} s_j^{\rho} m_{\rho}^{\rm mix} \Big) \Big\} \rangle_{\eta_j^1, \eta_j^2, \eta_j^3} \\
& \times \exp\Big(\frac{\beta}{2N} \sum_{\mu \geq 4} \sum_{\rho,\sigma} m_{\rho}^{\mu} m_{\sigma}^{\mu} s_j^{\rho} s_j^{\sigma} \Big).
\end{aligned}$$

Here, $\langle\cdot\rangle_{\{\eta^1,\eta^2,\eta^3\}}$ means the average over η^1,η^2,η^3 where η^μ takes ±1 with the probability $\frac{1\pm\sqrt{a}}{2}$. Furthermore, we take the trace $\mathrm{Tr}_{s_j^\rho}$ and put $s_j^{\rho\prime}=s_j^\rho\xi_j^{(m)}$.

$$\mathrm{Tr}_{s_j^\rho} A_j \quad = \quad \mathrm{Tr}_{s_j^\rho}\langle\exp\Big\{\beta N\Big(\sum_{\tau\le3,\rho}\frac{1}{N}\eta_j^\tau s_j^{\rho\prime}m_\rho^\tau+\sum_\rho\frac{\sqrt{-h}}{N}\eta_j^{\mathrm{mix}}s_j^{\rho\prime}m_\rho^{\mathrm{mix}}\Big)\Big\}\rangle_{\eta_j^1,\eta_j^2,\eta_j^3}\exp\Big(\frac{\beta}{2N}\sum_{\mu\ge4}\sum_{\rho,\sigma}m_\rho^\mu m_\sigma^\mu s_j^{\rho\prime}s_j^{\sigma\prime}\Big).$$

Thus, we obtain

$$\begin{aligned}[Z^n] \quad &= \quad e^{-\beta\frac{n(p-h)}{2}}\mathrm{Tr}_{\{s_j^\rho\}}\int_{-\infty}^{\infty}\Big(\prod_{\rho,\tau\le3}\sqrt{\frac{\beta N}{2\pi}}dm_\rho^\tau\Big)\Big(\prod_{\mu\ge4,\rho}\frac{1}{\sqrt{2\pi}}dm_\rho^\mu\Big)\Big(\prod_\rho\sqrt{\frac{\beta N}{2\pi}}dm_\rho^{\mathrm{mix}}\Big)\\
&\quad\times\exp\Big\{-\beta N\sum_\rho\Big(\sum_{\tau\le3}\frac{(m_\rho^\tau)^2}{2}+\frac{(m_\rho^{\mathrm{mix}})^2}{2}\Big)-\sum_{\mu\ge4,\rho}\frac{(m_\rho^\mu)^2}{2}\Big\}\\
&\quad\times\langle\exp\Big\{\beta N\sum_\rho\Big(\sum_{\tau\le3}\frac{1}{N}\sum_{j=1}^N\eta_j^\tau s_j^\rho m_\rho^\tau+\frac{\sqrt{-h}}{N}\sum_{j=1}^N\eta_j^{\mathrm{mix}}s_j^\rho m_\rho^{\mathrm{mix}}\Big)\Big\}\rangle_{\{\eta_j^1,\eta_j^2,\eta_j^3\}}\exp\Big(\frac{\beta}{2}\sum_{\mu\ge4}\sum_{\rho,\sigma}m_\rho^\mu m_\sigma^\mu q_{\rho\sigma}\Big)\end{aligned} \tag{50}$$

In the expression, we omitted $'$ from s' and define

$$q_{\rho\sigma} \quad \equiv \quad \frac{1}{N}\sum_j s_j^\rho s_j^\sigma. \tag{51}$$

Note that $q_{\rho\rho}=1$. Therefore, we obtain

$$\begin{aligned}[Z^n] \quad &= \quad e^{-\beta\frac{n(p-h)}{2}}\mathrm{Tr}_{\{s_j^\rho\}}\int_{-\infty}^{\infty}\Big(\prod_{\rho,\tau\le3}\sqrt{\frac{\beta N}{2\pi}}dm_\rho^\tau\Big)\Big(\prod_{\mu\ge4,\rho}\frac{1}{\sqrt{2\pi}}dm_\rho^\mu\Big)\Big(\prod_\rho\sqrt{\frac{\beta N}{2\pi}}dm_\rho^{\mathrm{mix}}\Big)\\
&\quad\times\langle\exp\Big\{\beta N\sum_\rho\Big(\sum_{\tau\le3}\Big(-\frac{(m_\rho^\tau)^2}{2}+\frac{1}{N}\sum_{j=1}^N\eta_j^\tau s_j^\rho m_\rho^\tau\Big)-\frac{(m_\rho^{\mathrm{mix}})^2}{2}+\frac{\sqrt{-h}}{N}\sum_{j=1}^N\eta_j^{\mathrm{mix}}s_j^\rho m_\rho^{\mathrm{mix}}\Big)\Big\rangle_{\{\eta_j^1,\eta_j^2,\eta_j^3\}}\\
&\quad\times\exp\Big\{-\sum_{\mu\ge4,\rho}\frac{(m_\rho^\mu)^2}{2}+\frac{\beta}{2}\sum_{\mu\ge4}\sum_{\rho,\sigma}m_\rho^\mu m_\sigma^\mu q_{\rho\sigma}\Big\}.\end{aligned} \tag{52}$$

The argument in the last exponential term is expressed as

$$-\sum_{\mu\ge4,\rho}\frac{(m_\rho^\mu)^2}{2}+\frac{\beta}{2}\sum_{\mu\ge4}\sum_{\rho,\sigma}m_\rho^\mu m_\sigma^\mu q_{\rho\sigma} \quad = \quad -\frac{1}{2}\sum_{\mu\ge4}\sum_{\rho\sigma}m_\rho^\mu K_{\rho\sigma}m_\sigma^\mu, \tag{53}$$

where we define

$$K_{\rho\sigma} \quad \equiv \quad \delta_{\rho\sigma}-\beta q_{\rho\sigma}. \tag{54}$$

Introducing the conjugate variable $r_{\rho\sigma}$ to $q_{\rho\sigma}$, we obtain

$$[Z^n] \quad = \quad e^{-\beta\frac{n(p-h)}{2}}\mathrm{Tr}_{\{s_j^\rho\}}\int\prod_{\rho\ne\sigma}\Big(dq_{\rho\sigma}\Big)\int_{-\mathrm{i}\infty}^{\mathrm{i}\infty}\prod_{\rho\ne\sigma}\Big(\frac{iN\alpha\beta^2}{2}\frac{dr_{\rho\sigma}}{2\pi}\Big)\int_{-\infty}^{\infty}\Big(\prod_{\rho,\tau\le3}\sqrt{\frac{\beta N}{2\pi}}dm_\rho^\tau\Big)$$

$$\times\Big(\prod_{\mu\geq 4,\rho}\frac{1}{\sqrt{2\pi}}dm_\rho^\mu\Big)\Big(\prod_\rho\sqrt{\frac{\beta N}{2\pi}}dm_\rho^{\rm mix}\Big)$$
$$\times\langle\exp\Big\{\beta N\sum_\rho\Big(\sum_{\tau\leq 3}\Big(-\frac{(m_\rho^\tau)^2}{2}+\frac{1}{N}\sum_{j=1}^N\eta_j^\tau s_j^\rho m_\rho^\tau\Big)-\frac{(m_\rho^{\rm mix})^2}{2}+\frac{\sqrt{-h}}{N}\sum_{j=1}^N\eta_j^{\rm mix}s_j^\rho m_\rho^{\rm mix}\Big)\Big\}\rangle_{\{\eta_j^1,\eta_j^2,\eta_j^3\}}$$
$$\times\exp\Big\{-\frac{1}{2}\sum_{\mu\geq 4}\sum_{\rho\sigma}m_\rho^\mu K_{\rho\sigma}m_\sigma^\mu-\sum_{\rho\neq\sigma}\frac{N\alpha\beta^2}{2}r_{\rho\sigma}(q_{\rho\sigma}-\frac{1}{N}\sum_j s_j^\rho s_j^\sigma)\Big\}. \tag{55}$$

The integration with respect to m_ρ^μ ($\mu\geq 4$) is performed and we obtain

$$\int\Big(\prod_{\mu\geq 4,\rho}\frac{1}{\sqrt{2\pi}}dm_\rho^\mu\Big)\exp\Big\{-\frac{1}{2}\sum_{\mu\geq 4}\sum_{\rho\sigma}m_\rho^\mu K_{\rho\sigma}m_\sigma^\mu\Big\} = (\det K)^{-(p-3)/2}=\exp\Big\{-\frac{p-3}{2}\ln\det K\Big\}$$
$$=\exp\Big\{-\frac{p-3}{2}\,{\rm Tr}\ln K\Big\}=\exp\Big\{-\frac{p-3}{2}\,{\rm Tr}\ln\Big((1-\beta)E-\beta Q\Big)\Big\}, \tag{56}$$

where E is the $n\times n$ unit matrix and Q is the $n\times n$ matrix of which diagonal components are zero and off diagonal components are $q_{\rho\sigma}$. In order to take the average over s_j^ρ, we rewrite the relevant part as

$$\langle{\rm Tr}_{\{s_j^\rho\}}\exp\Big\{\beta\sum_{\tau\leq 3,\rho}\sum_{j=1}^N\eta_j^\tau s_j^\rho m_\rho^\tau+\beta\sqrt{-h}\sum_\rho\sum_{j=1}^N\eta_j^{\rm mix}s_j^\rho m_\rho^{\rm mix}+\sum_{\rho\neq\sigma}\frac{\alpha\beta^2}{2}r_{\rho\sigma}\sum_j s_j^\rho s_j^\sigma\Big\}\rangle_{\{\eta_j^1,\eta_j^2,\eta_j^3\}}$$
$$=\exp\Big\{N\langle\ln{\rm Tr}_{\{s^\rho\}}\exp\Big(\beta\sum_{\tau\leq 3,\rho}\eta^\tau s^\rho m_\rho^\tau+\beta\sqrt{-h}\sum_\rho\eta^{\rm mix}s^\rho m_\rho^{\rm mix}+\sum_{\rho\neq\sigma}\frac{\alpha\beta^2}{2}r_{\rho\sigma}s^\rho s^\sigma\Big)\rangle_{\{\eta^1,\eta^2,\eta^3\}}\Big\}. \tag{57}$$

Here, we used the self averaging property and replace $\frac{1}{N}\sum_j g(\{\eta_j^\tau\})$ by $\langle g(\{\eta^\tau\})\rangle_{\eta^\tau}$ since $N\gg 3$. Thus, we obtain

$$[Z^n] = e^{-\beta\frac{n(p-h)}{2}}\int\prod_{\rho\neq\sigma}\Big(dq_{\rho\sigma}\Big)\int_{-{\rm i}\infty}^{{\rm i}\infty}\prod_{\rho\neq\sigma}\Big(\frac{iN\alpha\beta^2}{2}\frac{dr_{\rho\sigma}}{2\pi}\Big)\int_{-\infty}^{\infty}\Big(\prod_{\tau\leq 3,\rho}\sqrt{\frac{\beta N}{2\pi}}dm_\rho^\tau\Big)\Big(\prod_\rho\sqrt{\frac{\beta N}{2\pi}}dm_\rho^{\rm mix}\Big)$$
$$\times\exp\Big\{N\Big\{-\beta\sum_{\tau\leq 3,\rho}\frac{(m_\rho^\tau)^2}{2}-\beta\sum_\rho\frac{(m_\rho^{\rm mix})^2}{2}-\frac{\alpha}{2}{\rm Tr}\ln\Big((1-\beta)E-\beta Q\Big)-\frac{\alpha\beta^2}{2}\sum_{\rho\neq\sigma}r_{\rho\sigma}q_{\rho\sigma}$$
$$+\langle\ln{\rm Tr}_{\{s^\rho\}}\exp\Big(\sum_{\rho\neq\sigma}\frac{\alpha\beta^2}{2}r_{\rho\sigma}s^\rho s^\sigma+\beta\sum_{\tau\leq 3,\rho}\eta^\tau s^\rho m_\rho^\tau+\beta\sqrt{-h}\sum_\rho\eta^{\rm mix}s^\rho m_\rho^{\rm mix}\Big)\rangle_{\{\eta^1,\eta^2,\eta^3\}}\Big\}\Big\}, \tag{58}$$

where $\alpha\equiv\frac{p}{N}$. We study the case $N\gg 1$ and thus the integration can be estimated at the saddle point. Assuming the self-averaging property, the free energy per neuron $f=[F]/N=-T[\ln Z]/N$ is expressed as

$$f = -T[\ln Z]/N=-\frac{1}{\beta}\frac{[Z^n]-1}{nN}. \tag{59}$$

Thus, we obtain

$$f = \frac{1}{n}\sum_{\tau\le 3,\rho}\frac{(m_\rho^\tau)^2}{2}+\frac{1}{n}\sum_{\rho}\frac{(m_\rho^{\rm mix})^2}{2}+\frac{\alpha}{2n\beta}\,{\rm Tr}\ln\Big((1-\beta)E-\beta Q\Big)+\frac{\alpha\beta}{2n}\sum_{\rho\neq\sigma}r_{\rho\sigma}q_{\rho\sigma}$$
$$-\frac{1}{n\beta}\langle\ln {\rm Tr}_{\{s^\rho\}}\exp(\beta H_\eta)\rangle_{\{\eta^1,\eta^2,\eta^3\}}, \tag{60}$$

$$\beta H_\eta = \sum_{\rho\neq\sigma}\frac{\alpha\beta^2}{2}r_{\rho\sigma}s^\rho s^\sigma+\beta\sum_{\tau\le 3,\rho}\eta^\tau s^\rho m_\rho^\tau+\beta\sqrt{-h}\sum_\rho\eta^{\rm mix}s^\rho m_\rho^{\rm mix}. \tag{61}$$

By making the variable transformation $m_\rho^{\rm mix}\to\sqrt{-h}m_\rho^{\rm mix}$, we obtain

$$f = \frac{1}{n}\sum_{\tau\le 3,\rho}\frac{(m_\rho^\tau)^2}{2}-h\frac{1}{n}\sum_{\rho}\frac{(m_\rho^{\rm mix})^2}{2}+\frac{\alpha}{2n\beta}\,{\rm Tr}\ln\Big((1-\beta)E-\beta Q\Big)+\frac{\alpha\beta}{2n}\sum_{\rho\neq\sigma}r_{\rho\sigma}q_{\rho\sigma}$$
$$-\frac{1}{n\beta}\langle\ln {\rm Tr}_{\{s^\rho\}}\exp(\beta H_\eta)\rangle_{\{\eta^1,\eta^2,\eta^3\}}, \tag{62}$$

$$\beta H_\eta = \sum_{\rho\neq\sigma}\frac{\alpha\beta^2}{2}r_{\rho\sigma}s^\rho s^\sigma+\beta\sum_{\tau\le 3,\rho}\eta^\tau s^\rho m_\rho^\tau-\beta h\sum_\rho\eta^{\rm mix}s^\rho m_\rho^{\rm mix}. \tag{63}$$

Here, $\langle\cdot\rangle_{\{\eta^1,\eta^2,\eta^3\}}$ means the average over η^1,η^2,η^3 where η^μ takes ±1 with the probability $\frac{1\pm\sqrt{a}}{2}$. Note that $q_{\rho\rho}=1$ in the above expression. The following relations hold.

$$m_\rho^\mu = \frac{1}{N}\sum_j\xi_j^\mu s_j^\rho, \tag{64}$$

$$r_{\rho\sigma} = \frac{1}{\alpha}\sum_{\mu\ge 4}m_\rho^\mu m_\sigma^\mu \tag{65}$$

6. Appendix B. 1RSB solution in case I

By the standard recipe, the 1RSB solution is obtained and is expressed as

$$f_{\rm 1RSB} = \sum_{\tau\le 3}\frac{1}{2}(m^\tau)^2-\frac{h}{2}(m^{\rm mix})^2$$
$$+\frac{\alpha}{2\beta}\Big[\frac{m_1-1}{m_1}\ln\Big(1-\beta(1-q_1)\Big)+\frac{1}{m_1}\ln\Big\{1-\beta\Big(1+(m_1-1)q_1-m_1q_0\Big)\Big\}\Big]+\frac{\alpha\beta}{2}r_1$$
$$-\frac{\alpha}{2}\frac{q_0}{1-\beta\Big(1+(m_1-1)q_1-m_1q_0\Big)}+\frac{\alpha\beta}{2}\Big((m_1-1)r_1q_1-m_1r_0q_0\Big)$$
$$-\frac{1}{\beta m_1}\langle\int Du\ln\int Dv\cosh^{m_1}\Xi\rangle_{\{\eta^1,\eta^2,\eta^3\}}-\frac{1}{\beta}\ln 2, \tag{66}$$

$$\Xi=\beta\Big(\sqrt{\alpha r_0}u+\sqrt{\alpha(r_1-r_0)}v+\sum_{\tau\le 3}\eta^\tau m^\tau-h\eta^{\rm mix}m^{\rm mix}\Big), \tag{67}$$

$$m^\tau = \langle\eta^\tau\int Du\frac{\int Dv(\cosh\Xi)^{m_1}\tanh\Xi}{\int Dv(\cosh\Xi)^{m_1}}\rangle_{\{\eta^1,\eta^2,\eta^3\}},\quad \tau=1,2,3, \tag{68}$$

$$m^{\rm mix} = \langle \eta^{\rm mix} \int Du \frac{\int Dv(\cosh\Xi)^{m_1}\tanh\Xi}{\int Dv(\cosh\Xi)^{m_1}} \rangle_{\{\eta^1,\eta^2,\eta^3\}}, \tag{69}$$

$$q_0 = \langle \int Du \Big(\frac{\int Dv(\cosh\Xi)^{m_1}\tanh\Xi}{\int Dv(\cosh\Xi)^{m_1}} \Big)^2 \rangle_{\{\eta^1,\eta^2,\eta^3\}}, \tag{70}$$

$$q_1 = \langle \int Du \frac{\int Dv(\cosh\Xi)^{m_1}\tanh^2\Xi}{\int Dv(\cosh\Xi)^{m_1}} \rangle_{\{\eta^1,\eta^2,\eta^3\}}, \tag{71}$$

$$r_0 = \frac{q_0}{\Big\{1-\beta\Big(1+(m_1-1)q_1-m_1q_0\Big)\Big\}^2}, \tag{72}$$

$$r_1 = r_0 + \frac{q_1-q_0}{\Big(1-\beta(1-q_1)\Big)\Big\{1-\beta\Big(1+(m_1-1)q_1-m_1q_0\Big)\Big\}}, \tag{73}$$

$$\frac{\alpha}{2\beta}\ln\Big\{\frac{1-\beta(1-q_1)}{1-\beta\Big(1+(m_1-1)q_1-m_1q_0\Big)}\Big\} + \frac{1}{\beta}\langle\int Du \ln\int Dv(\cosh\Xi)^{m_1}\rangle_{\{\eta^1,\eta^2,\eta^3\}}$$

$$= \frac{\alpha}{2}m_1(q_1-q_0)\frac{1-\beta\Big(1+(m_1-1)q_1\Big)}{\Big\{1-\beta\Big(1+(m_1-1)q_1-m_1q_0\Big)\Big\}\Big(1-\beta(1-q_1)\Big)}$$

$$+\frac{m_1}{\beta}\langle\int Du\frac{\int Dv(\cosh\Xi)^{m_1}\ln(\cosh\Xi)}{\int Dv(\cosh\Xi)^{m_1}}\rangle_{\{\eta^1,\eta^2,\eta^3\}}. \tag{74}$$

7. Appendix C. Derivation of the free energy in case II

In this appendix, we formulate case II in a rather general situation, that is, there are M clusters, and in the ωth cluster, the number of patterns is p_ω. The ν_ωth pattern in the ωth cluster is represented by $\{\xi_j^{(\omega,\nu_\omega)}\}$. We assume that patterns in the cluster $\omega = 1$ and mixed states composed of these patterns are retrieved. We study unlearning the following mixed state $\xi^{\rm mix}$,

$$\xi_i^{\rm mix} = {\rm sgn}(\xi_i^{(1,1)} + \xi_i^{(1,2)} \cdots + \xi_i^{(1,p_1)}). \tag{75}$$

Now, we derive the free energy. From the Hamiltonian defined in eq. (34), and introducing n replicas $s_i^\rho, (\rho = 1, \cdots, n)$, we obtain

$$[Z^n] = e^{-\beta n\frac{p-h}{2}}{\rm Tr}_{\{S_j^\rho\}}\int_{-\infty}^{\infty}\Big(\prod_{\rho,\nu_1\leqq p}\sqrt{\frac{\beta N}{2\pi}}dm_\rho^{(1,\nu_1)}\Big)\int_{-\infty}^{\infty}\Big(\prod_\rho\sqrt{\frac{\beta N}{2\pi}}dm_\rho^{\rm mix}\Big)\int_{-\infty}^{\infty}\Big(\prod_{\rho,\omega\geqq 2,\nu_\omega}\frac{1}{\sqrt{2\pi}}dm_\rho^{(\omega,\nu_\omega)}\Big)$$

$$\times\exp\Big\{-\beta N\Big(\sum_{\rho,\nu_1\leqq 3}\frac{(m_\rho^{(1,\nu_1)})^2}{2}+\sum_\rho\frac{(m_\rho^{\rm mix})^2}{2}\Big)\Big\}$$

$$\times\left[\exp\left\{\beta N\left(\sum_{\rho,\nu_1\leqq p}\frac{1}{N}\sum_j \xi_j^{(1,\nu_1)}S_j^\rho m_\rho^{(1,\nu_1)}+\sum_\rho\frac{\sqrt{-h}}{N}\sum_j\xi_j^{\mathrm{mix}}S_j^\rho m_\rho^{\mathrm{mix}}\right)\right\}\right]$$
$$\times\exp\left\{-\sum_{\rho,\omega\geqq 2,\nu_\omega}\frac{(m_\rho^{(\omega,\nu_\omega)})^2}{2}\right\}\times\left[\exp\left\{\sqrt{\frac{\beta}{N}}\sum_{\rho,\omega\geqq 2,\nu_\omega}\sum_j\xi_j^{(\omega,\nu_\omega)}S_j^\rho m_\rho^{(\omega,\nu_\omega)}\right\}\right], \tag{76}$$

where we define $m_\rho^{(\omega,\nu_\omega)}=\frac{1}{N}\sum_i\xi^{(\omega,\nu_\omega)}s_i^\rho$, and $m_\rho^{\mathrm{mix}}=\frac{1}{N}\sum_i\xi^{\mathrm{mix}}s_i^\rho$. Here, $m_\rho^{(\omega,\nu_\omega)}=O(1/\sqrt{N})$ and we replace $m_\rho^{(\omega,\nu_\omega)}$ by $m_\rho^{(\omega,\nu_\omega)}/\sqrt{\beta N}$ for $\omega\geqq 2$. The term for $\omega\geqq 2$ is calculated as follows.

$$\left[\exp\left\{\sqrt{\frac{\beta}{N}}\sum_{\rho,\omega\geqq 2,\nu_\omega}\sum_j\xi_j^{(\omega,\nu_\omega)}S_j^\rho m_\rho^{(\omega,\nu_\omega)}\right\}\right]$$
$$=\prod_{\omega\geqq 2}\exp\left\{\frac{\beta}{2}\left(\sum_{\nu_\omega}\sum_{\rho,\sigma}m_\rho^{(\omega,\nu_\omega)}m_\sigma^{(\omega,\nu_\omega)}q_{\rho\sigma}+\sum_{\nu_\omega\neq\nu'_\omega}\sum_{\rho,\sigma}am_\rho^{(\omega,\nu_\omega)}m_\sigma^{(\omega,\nu'_\omega)}q_{\rho\sigma}\right)\right\}, \tag{77}$$

where $q_{\rho\sigma}=\frac{1}{N}\sum_i s_i^\rho s_i^\sigma$. Thus, we obtain

$$\prod_{\omega\geqq 2}\exp\left\{-\sum_{\rho,\nu_\omega}\frac{(m_\rho^{(\omega,\nu_\omega)})^2}{2}+\frac{\beta}{2}\left(\sum_{\nu_\omega}\sum_{\rho,\sigma}m_\rho^{(\omega,\nu_\omega)}m_\sigma^{(\omega,\nu_\omega)}q_{\rho\sigma}+\sum_{\nu_\omega\neq\nu'_\omega}\sum_{\rho,\sigma}am_\rho^{(\omega,\nu_\omega)}m_\sigma^{(\omega,\nu'_\omega)}q_{\rho\sigma}\right)\right\}$$
$$=\prod_{\omega\geqq 2}\exp\left\{-\frac{1}{2}\sum_{\nu_\omega,\nu'_\omega}\sum_{\rho,\sigma}m_\rho^{(\omega,\nu_\omega)}K_{\rho\sigma}^{\nu_\omega\nu'_\omega}m_\sigma^{(\omega,\nu'_\omega)}\right\}, \tag{78}$$

where we define

$$K_{\rho\sigma}^{\nu_\omega\nu'_\omega}=\delta_{\nu_\omega\nu'_\omega}(\delta_{\rho\sigma}-\beta q_{\rho\sigma})+(1-\delta_{\nu_\omega\nu'_\omega})(-\beta a q_{\rho\sigma}). \tag{79}$$

Let us define $K^{(\omega)}$ for the ωth cluster as follows.

$$K^{(\omega)}=\begin{pmatrix}K_1 & K_2 & \cdots & K_2\\ K_2 & K_1 & & \vdots\\ \vdots & & \ddots & K_2\\ K_2 & \cdots & K_2 & K_1\end{pmatrix},\ K_1=\begin{pmatrix}1-\beta & & -\beta q_{\rho\sigma}\\ & \ddots & \\ -\beta q_{\rho\sigma} & & 1-\beta\end{pmatrix},\ K_2=\begin{pmatrix}-\beta a & & -\beta a q_{\rho\sigma}\\ & \ddots & \\ -\beta a q_{\rho\sigma} & & -\beta a\end{pmatrix}, \tag{80}$$

where $K^{(\omega)}$ is the $np_\omega\times np_\omega$ matrix and K_1 and K_2 are the $n\times n$ matrices. The integration with respect to $m_\rho^{(\omega,\nu_\omega)}(\omega\geqq 2)$ is performed and we obtain

$$\int_{-\infty}^{\infty}\left(\prod_{\omega\geqq 2,\nu_\omega,\rho}\frac{1}{\sqrt{2\pi}}dm_\rho^{(\omega,\nu_\omega)}\right)\exp\left\{-\frac{1}{2}\sum_{\omega\geqq 2}\sum_{\nu_\omega,\nu'_\omega}\sum_{\rho,\sigma}m_\rho^{(\omega,\nu_\omega)}K_{\rho\sigma}^{\nu_\omega\nu'_\omega}m_\sigma^{(\omega,\nu'_\omega)}\right\}$$
$$=\exp\left\{-\frac{M-1}{2}\log(\det K^{(\omega)})\right\}. \tag{81}$$

Now, we study the case that p_ω is constant and is set to p, and thus we put $K^{(\omega)}=K$. As is in the case I, we set $\eta_j^{\nu_1}\equiv\xi_j^{(1,\nu_1)}\xi_j^{(m)}$, $\eta_j^{\mathrm{mix}}\equiv\xi_j^{\mathrm{mix}}\xi_j^{(m)}$, and $(S_j^\rho)'\equiv S_j^\rho\xi_j^{(m)}$. The integration can be

estimated at the saddle point and we obtain

$$[Z^n] \simeq e^{-\beta n \frac{p-h}{2}} \exp\left\{ N \left\{ -\beta \sum_{\rho,\nu_1 \leqq p} \frac{(m_\rho^{\nu_1})^2}{2} + \beta h \sum_\rho \frac{(m_\rho^{\rm mix})^2}{2} - \frac{M}{2N} \log(\det K) - \frac{\alpha\beta^2}{2} \sum_{\rho\neq\sigma} r_{\rho\sigma} q_{\rho\sigma} \right.\right.$$
$$\left.\left. + \langle \log {\rm Tr}_{\{S^\rho\}} \exp\left(\beta \sum_{\rho,\nu_1 \leqq p} \eta^{\nu_1} S^\rho m_\rho^{\nu_1} - \beta h \sum_\rho \eta^{\rm mix} S^\rho m_\rho^{\rm mix} + \sum_{\rho\neq\sigma} \frac{\alpha\beta^2}{2} r_{\rho\sigma} S^\rho S^\sigma \right) \rangle_{\{\eta^1,\cdots,\eta^p\}} \right\} \right\}, \tag{82}$$

where we define $m_\rho^{\nu_1} \equiv m_\rho^{(1,\nu_1)}$ and replace $m_\rho^{\rm mix}$ by $\sqrt{-h} m_\rho^{\rm mix}$ and omit the superscript $'$ from $S^\rho{}'$. $\langle\cdot\rangle_{\eta^1,\cdots,\eta^p}$ and $r_{\rho\sigma}$ are the same as in case I.

We assume the replica symmetry and K in Eq.(80) is expressed as

$$K = \begin{pmatrix} K_1 & K_2 & \cdots & K_2 \\ K_2 & K_1 & & \vdots \\ \vdots & & \ddots & K_2 \\ K_2 & \cdots & K_2 & K_1 \end{pmatrix}, \; K_1 = \begin{pmatrix} 1-\beta & & -\beta q \\ & \ddots & \\ -\beta q & & 1-\beta \end{pmatrix}, \; K_2 = \begin{pmatrix} -\beta a & & -\beta a q \\ & \ddots & \\ -\beta a q & & -\beta a \end{pmatrix}. \tag{83}$$

K_1 and K_2 can be diagonalized by an orthogonal matrix U_1. We define the eigenvalues of K_1 and K_2 as $\lambda_1, \cdots, \lambda_n$、 $\lambda'_1, \cdots, \lambda'_n$, respectively.

$$\Lambda_1 \equiv \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} = U_1^T K_1 U_1, \; \Lambda_2 \equiv \begin{pmatrix} \lambda'_1 & & 0 \\ & \ddots & \\ 0 & & \lambda'_n \end{pmatrix} = U_1^T K_2 U_1. \tag{84}$$

Let us define the matrix U which diagonal blocks are U_1 and off diagonal blocks are zero. Thus we obtain

$$\Lambda \equiv \begin{pmatrix} \Lambda_1 & \Lambda_2 & \cdots & \Lambda_2 \\ \Lambda_2 & \Lambda_1 & & \vdots \\ \vdots & & \ddots & \Lambda_2 \\ \Lambda_2 & \cdots & \Lambda_2 & \Lambda_1 \end{pmatrix} = U^T K U. \tag{85}$$

$$\det \Lambda = \prod_{\rho=1}^{n} (\lambda_\rho - \lambda'_\rho)^{(p-1)} \{(p-1)\lambda'_\rho + \lambda_\rho\}. \tag{86}$$

λ_ρ and λ'_ρ ($\rho = 1, \cdots, n$) are expressed as follows.

$$\lambda_1 = 1 - \beta + \beta q - n\beta q \, , \; \lambda_2 = \cdots = \lambda_n = 1 - \beta + \beta q,$$

$$\lambda'_1 = -\beta a(1 - q + nq) \, , \; \lambda'_2 = \cdots = \lambda'_n = -\beta a(1 - q).$$

Thus we obtain

$$\log(\det K) = n\left[(p-1)\left\{\log\left(1-\beta(1-q)(1-a)\right)-\frac{\beta q(1-a)}{1-\beta(1-q)(1-a)}\right\}\right.$$
$$\left.+\log\left(1-\beta(1-q)(1+(p-1)a)\right)-\frac{\beta q(1+(p-1)a)}{1-\beta(1-q)(1+(p-1)a)}\right]. \quad (87)$$

Therefore, the replica symmetric free energy f_{RS} is obtained as

$$f_{RS} = \sum_{\nu_1 \leqq p}\frac{(m^{\nu_1})^2}{2} - h\frac{(m^{\rm mix})^2}{2} + \frac{\alpha\beta}{2}r(1-q)$$
$$+\frac{M}{2\beta N}\left[(p-1)\left\{\log\left(1-\beta(1-q)(1-a)\right)-\frac{\beta q(1-a)}{1-\beta(1-q)(1-a)}\right\}\right.$$
$$\left.+\log\left(1-\beta(1-q)(1+(p-1)a)\right)-\frac{\beta q(1+(p-1)a)}{1-\beta(1-q)(1+(p-1)a)}\right]$$
$$-\frac{1}{\beta}\int Dz\langle\log\left\{2\cosh\left(\beta\left(\sqrt{\alpha r}z+\sum_{\nu_1\leqq p}\eta^{\nu_1}m^{\nu_1}-h\eta^{\rm mix}m^{\rm mix}\right)\right)\right\}\rangle_{\{\eta^1,\cdots,\eta^p\}}. \quad (88)$$

8. Appendix D. Derivation of the AT stability in case II

Here, we derive the AT stability in case II. We calculate the following $A_{(\rho\sigma)(\gamma\delta)}$ for K in eq. (80).

$$A_{(\rho\sigma)(\gamma\delta)} = \frac{M}{2\beta N}\frac{\partial}{\partial q_{\gamma\delta}}\left(\frac{1}{\det K}\frac{\partial}{\partial q_{\rho\sigma}}\det K\right). \quad (89)$$

The matrices K_1' and K_2' are defined

$$K_1' : (K_1')_{\alpha\sigma} = -\beta\delta_{\alpha\rho},\ \text{for}\ \alpha = 1,\cdots,n,\ \ (K_1')_{\alpha\beta} = (K_1)_{\alpha\beta}\ \text{for}\ \alpha,\beta = 1,\cdots,n,(\beta\neq\sigma),$$

$$K_2' : (K_2')_{\alpha\sigma} = -\beta a\delta_{\alpha\rho},\ \text{for}\ \alpha = 1,\cdots,n,\ \ (K_2')_{\alpha\beta} = (K_2)_{\alpha\beta}\ \text{for}\ \alpha,\beta = 1,\cdots,n,(\beta\neq\sigma),$$

where ρ and σ are fixed indices with $\rho<\sigma$.

$$\frac{\partial}{\partial q_{\rho\sigma}}\det K = \frac{\partial}{\partial q_{\rho\sigma}}\begin{vmatrix} K_1 & K_2 & \cdots & K_2\\ K_2 & K_1 & & \vdots\\ \vdots & & \ddots & K_2\\ K_2 & \cdots & K_2 & K_1\end{vmatrix}$$
$$= 2p\left\{\begin{vmatrix} K_1' & K_2 & \cdots & K_2\\ 0 & K_1 & & \vdots\\ \vdots & & \ddots & K_2\\ 0 & K_2 & \cdots & K_1\end{vmatrix}+\begin{vmatrix} 0 & K_2 & \cdots & K_2\\ K_2' & K_1 & & \vdots\\ \vdots & & \ddots & K_2\\ 0 & K_2 & \cdots & K_1\end{vmatrix}+\cdots+\begin{vmatrix} 0 & K_2 & \cdots & K_2\\ 0 & K_1 & & \vdots\\ \vdots & & \ddots & K_2\\ K_2' & K_2 & \cdots & K_1\end{vmatrix}\right\}$$
$$= 2p\{-\beta\tilde{K}_{\sigma\rho}-\beta a\tilde{K}_{\sigma(n+\rho)}-\cdots-\beta a\tilde{K}_{\sigma((p-1)n+\rho)}\}. \quad (90)$$

where $\tilde{K}_{\sigma\rho}$ is a cofactor of the respective (σ,ρ) components. Thus we obtain

$$\begin{aligned} A_{(\rho\sigma)(\gamma\delta)} &= -\frac{pM}{N}\frac{\partial}{\partial q_{\gamma\delta}}\left\{\frac{1}{\det K}\left(\tilde{K}_{\sigma\rho}+a\tilde{K}_{\sigma(n+\rho)}+\cdots+a\tilde{K}_{\sigma((p-1)n+\rho)}\right)\right\} \\ &= -\alpha\frac{\partial}{\partial q_{\gamma\delta}}\left\{(K^{-1})_{\rho\sigma}+a\left((K^{-1})_{(n+\rho)\sigma}+\cdots+(K^{-1})_{((p-1)n+\rho)\sigma}\right)\right\}. \end{aligned} \tag{91}$$

We calculate $\partial(K^{-1})_{xy}/\partial q_{\gamma\delta}$ as in case I. Let us introduce the variables x, y, α, and β which take values from one to np as distinguished from ρ, σ. We define the label of the block which α and β belong to as l and l', respectively. Then we obtain

$$\frac{\partial K_{\alpha\beta}}{\partial q_{\gamma\delta}} = -\beta\sum_{l=0}^{p-1}(\delta_{\alpha(nl+\gamma)}\delta_{\beta(nl+\delta)}+\delta_{\alpha(nl+\delta)}\delta_{\beta(nl+\gamma)})-\beta a\sum_{l\neq l'}(\delta_{\alpha(nl+\gamma)}\delta_{\beta(nl'+\delta)}+\delta_{\alpha(nl+\delta)}\delta_{\beta(nl'+\gamma)}). \tag{92}$$

Therefore,

$$\begin{aligned} \frac{\partial(K^{-1})_{xy}}{\partial q_{\gamma\delta}} &= \beta\sum_{l=0}^{p-1}\left\{(K^{-1})_{x(nl+\gamma)}(K^{-1})_{(nl+\delta)y}+(K^{-1})_{x(nl+\delta)}(K^{-1})_{(nl+\gamma)y}\right\} \\ &\quad +\beta a\sum_{l\neq l'}\left\{(K^{-1})_{x(nl+\gamma)}(K^{-1})_{(nl'+\delta)y}+(K^{-1})_{x(nl+\delta)}(K^{-1})_{(nl'+\gamma)y}\right\}. \end{aligned} \tag{93}$$

Let us define K^{-1} as

$$K^{-1}=\begin{pmatrix} L_1 & L_2 & \cdots & L_2 \\ L_2 & L_1 & & \vdots \\ \vdots & & \ddots & L_2 \\ L_2 & \cdots & L_2 & L_1 \end{pmatrix}. \tag{94}$$

Thus, $A_{(\rho\sigma)(\gamma\delta)}$ is expressed as

$$\begin{aligned} A_{(\rho\sigma)(\gamma\delta)} = -\alpha\beta\Big\{&(1+(p-1)a^2)\left(L_1^{\rho\gamma}L_1^{\delta\sigma}+L_1^{\rho\delta}L_1^{\gamma\sigma}\right) \\ &+a(p-1)(2+(p-2)a)\left(L_1^{\rho\gamma}L_2^{\delta\sigma}+L_1^{\rho\delta}L_2^{\gamma\sigma}+L_2^{\rho\gamma}L_1^{\delta\sigma}+L_2^{\rho\delta}L_1^{\gamma\sigma}\right) \\ &+(p-1)(1+2(p-2)a+(p^2-3p+3)a^2)\left(L_2^{\rho\gamma}L_2^{\delta\sigma}+L_2^{\rho\delta}L_2^{\gamma\sigma}\right)\Big\}. \end{aligned}$$

Let us derive the concrete formula of L_1 and L_2. From the relation $KK^{-1}=E$, we obtain

$$K_1L_1+K_2L_2+(p-2)K_2L_2=E_n, \tag{95}$$

$$K_1L_2+K_2L_1+(p-2)K_2L_2=0. \tag{96}$$

Then, we obtain

$$L_1 = L_2+L_4, \tag{97}$$

$$L_2 = -L_3K_2L_4, \tag{98}$$

$$L_3 = (K_1 + (p-1)K_2)^{-1}, \tag{99}$$

$$L_4 = (K_1 - K_2)^{-1}. \tag{100}$$

By taking the limit of $n \to 0$, the diagonal component l_i and the off diagonal component $\overline{l_i}$ of L_i are expressed as follows.

$$\begin{aligned}
l_1 &= l_2 + l_4, \\
\overline{l_1} &= \overline{l_2} + \overline{l_4}, \\
l_2 &= -\beta a\{-l_3 l_4 + (1-2q)\overline{l_3}\,\overline{l_4} + q(l_3\overline{l_4} + \overline{l_3} l_4)\}, \\
\overline{l_2} &= -\beta a\{-(1-2q)(l_3\overline{l_4} + \overline{l_3} l_4) + (2-3q)\overline{l_3}\,\overline{l_4} - q l_3 l_4\}, \\
l_3 &= \frac{1-\beta(1-2q)(1+(p-1)a)}{\{1-\beta(1-q)(1+(p-1)a)\}^2}, \\
\overline{l_3} &= \frac{\beta q(1+(pa-1)a)}{\{1-\beta(1-q)(1+(p-1)a)\}^2}, \\
l_4 &= \frac{1-\beta(1-2q)(1-a)}{\{1-\beta(1-q)(1-a)\}^2}, \\
\overline{l_4} &= \frac{\beta q(1-a)}{\{1-\beta(1-q)(1-a)\}^2}.
\end{aligned} \tag{101}$$