INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **38** (2005) 5627–5646 doi:10.1088/0305-4470/38/25/003

Anti-pairing in learning of a neural network with redundant hidden units

Chulan Kwon and Hyong Kyun Kim

Department of Physics, Myongji University, Namdong San 38-2, Yongin, Kyonggi-Do 449-728, Korea

E-mail: ckwon@mju.ac.kr

Received 11 January 2005, in final form 12 April 2005
Published 8 June 2005
Online at stacks.iop.org/JPhysA/38/5627

Abstract

We study the statistical mechanics of learning from examples between the two-layered committee machines with different numbers of hidden units using the replica theory. The number M of hidden units of the student network is larger than the number M_{T} of those of the target network called the teacher. We choose the networks to have binary synaptic weights, ± 1, which makes it possible to compare the calculation with the Monte Carlo simulation. We propose an effective teacher as a virtual target network which has the same M hidden units as the student and gives identical outputs with those of the original teacher. This is a way of making a conjecture for a ground state of a thermodynamic system, given by the weights of the effective teacher in our study. We suppose that the weights on M_{T} hidden units of the effective teacher are the same as those of the original teacher while those on $M - M_{\mathrm{T}}$ redundant hidden units are composed of anti-pairs, $\{1, -1\}$, with probability $1 - p$ in the limit $p \to 0$. For $p = 0$ exact, there are no terms related to the effective teacher in the calculation, for the contributions of anti-pairs to outputs are exactly cancelled. In the limit $p \to 0$, however, we find that the learnt weights of the student are actually equivalent to those of the suggested effective teacher, which is not possible from the calculation for $p = 0$. p plays the role of a symmetry breaking parameter for anti-pairing ordering, which is analogous to the magnetic field for the Ising model. A first-order phase transition is found to be signalled by breaking of symmetry in permuting hidden units. Above a critical number of examples, the student is shown to learn perfectly the effective teacher. Anti-pairing can be measured by a set of order parameters; zero in the permutation-symmetric phase and nonzero in the permutation symmetry breaking phase. Results from the Monte Carlo simulation are shown to be in good agreement with those from the replica calculation.

PACS numbers: 87.10.+e, 05.50.+q, 64.60.Cn

1. Introduction

The learning of a feed-forward neural network from examples given by a set of input–output rules is considered as a typical model for the supervised learning used in many applications. Statistical mechanics has been proven useful for the study of this subject. The replica theory [1, 2] has been used in theoretical studies in order to perform the average over quenched disorder. The statistical mechanical approach to the learning from examples was developed [3, 4] by utilizing Gardner's theory of the storage capacity [5–8], which is defined as the maximum number of patterns that can be stored in a neural network. Early studies were done for a simple perceptron with a single layer. Interesting results were found such as gradual or sudden learning depending on continuous or discrete weights and the enhancement of learning via phase transition. In some cases, the spin-glass nature was observed with replica symmetry breaking [3, 4, 9, 10].

There has been remarkable progress in studies for more realistic two-layer neural networks, in particular the parity and the committee machine. The connectivity between input units and hidden units is categorized into the tree structure where each hidden unit is connected to a part of input units without overlapping and the fully-connected structure where each hidden unit is connected to all input units.

In studies of the storage capacity, replica symmetry was found to break when the number of patterns is close to the storage capacity. The parity machine with the tree structure was studied successfully within the one-step replica symmetry breaking scheme [11]. For the committee machine, the earlier works were done only within the replica-symmetric scheme [12, 13] and the breaking of permutation symmetry was found to be characteristic of fully-connected machines. Later, the much-needed values of the storage capacity of the committee machine for both the tree structure and the fully-connected structure were obtained within the one-step replica symmetry breaking scheme [14]. Monasson and Zecchina developed a new theory that needs no replica symmetry breaking scheme [15]. Using this new approach, an equivalent result was found for the fully-connected committee machine [16, 17] and a new result for the fully-connected parity machine [18].

The breaking of permutation symmetry was also found to play an important role in the learning by the fully-connected committee machine. It was found that there is a first-order phase transition from the poor to the good learning phase, which is signalled by permutation symmetry breaking [19, 20]. Unlike the case of the storage capacity, the breaking of replica symmetry is not found [21, 22]. This is probably because the weights of the given target network serve for a recognizable attractor in the weight space.

In most previous studies of learning, the target network, called the teacher, and the student network have the same architecture. In real applications, however, one does not know the architecture of the teacher which is usually a human. We can reasonably assume that the two networks have the same number of input units that is for example the number of pixels of the screen in which the input image is captured. We can also assume that the two networks have the same number of output units which determines the number of classes in patterns. Consequently, a crucial difference may come from the hidden-layer structure. We consider the fully-connected committee machines which have realistic two-layer structures and complexity, due to the fully-connected structure, as can be handled by analytical calculation. We consider the situation where the number M of hidden units of the student is larger than the number M_{T} of those of the target network. Previous work in this direction was done for the networks having finite numbers of hidden units ($M = 3, M_{\mathrm{T}} = 1$) and binary weights [23], and also for those having large numbers of hidden units with a scaling $M_{\mathrm{T}}/M \to 0$ and continuous weights with spherical constraint [24]. In this paper we will investigate the situation in which

the networks have large numbers of hidden units with a finite ratio M_{T}/M which seems more realistic than the extreme scaling $M_{\mathrm{T}}/M \to 0$. For the opposite case with $M < M_{\mathrm{T}}$ the learning mechanism is completely different, for which we will present our result in a separate paper.

The student is to utilize all possible weights on its hidden units, not knowing the architecture of the teacher. We therefore suppose that the student learns an effective teacher, rather than the original teacher, which has the same M hidden units. The weights of the effective teacher are supposed to correspond to a ground state. We consider the networks to have binary weights which can be regarded as the Ising spins. Then we have an advantage to accompany the Monte Carlo simulation. The calculation might depend on the conjecture of the ground state. It is indeed quite valuable to use the simulation to confirm the reliability of our findings from the conjecture.

We summarize the following sections. In section 2, the statistical mechanical formalism via the replica theory is developed. The effective teacher having a partial anti-pairing of redundant weights is introduced. In section 3 infinite degeneracy in ground states and the resultant phase transition are discussed. In section 4, the permutation-symmetric state is investigated. In section 5, the permutation symmetry breaking state is found and the resultant first-order transition from the permutation-symmetric to permutation symmetry breaking phase is discussed. Results from the Monte Carlo simulation are shown to be in a good agreement with the theoretical expectation. In section 6, we summarize our study and comment about our future work for the case with $M < M_{\mathrm{T}}$.

2. Statistical mechanical formalism

We consider the teacher network that is a fully-connected committee machine with one hidden layer and one output unit. In the fully-connected structure, all the hidden units are connected to every input unit. It has N input units and M_{T} hidden units. By definition of the committee machine, the weights connecting hidden units to output unit are set to unity. The student network is also a fully-connected committee machine with the same architecture, but has more M hidden units. Let W^0_{ji} and W_{ji} be the weights of the teacher and student respectively which connect an input unit i to a hidden unit j. We consider binary weights having a value either 1 or -1. The teacher provides a training set consisting of P examples. An example μ is given by an input–output rule: $\{\xi^\mu_i; i = 1, \ldots, N\} \to o^\mu$. Input variable ξ^μ_i on an input unit i is independently and randomly distributed with variance unity, which may describe the brightness of an image at a pixel i of the screen. The output of the teacher is then given by $o^\mu = \mathrm{sgn}\big(M_{\mathrm{T}}^{-1/2} \sum_j^{M_{\mathrm{T}}} \mathrm{sgn}\big(N^{-1/2} \sum_i^N W^0_{ji}\xi^\mu_i\big)\big)$. Given input variables, the student produces its own output $\sigma^\mu = \mathrm{sgn}\big(M^{-1/2} \sum_j^M \mathrm{sgn}\big(N^{-1/2} \sum_i^N W_{ji}\xi^\mu_i\big)\big)$. The error function is defined by $1/4 \sum_\mu (o^\mu - \sigma^\mu)^2$. The learning algorithm is that the student learns the teacher by adjusting its weights so as to minimize this error function.

The statistical mechanical approach can be applied by identifying the error function as the energy E, and taking the thermodynamic limit $N \to \infty$. Weights are treated as thermodynamic variables. The temperature β^{-1} can be interpreted as a stochastic or noise parameter inherent in the network or environment. $\{\xi^\mu_i\}$ and $\{W^0_{ji}\}$, which are fixed during the learning process, can be interpreted as quenched disorder. We can then apply the replica theory to find averaged properties over disorder. We consider the teacher to have uncorrelated weights: $N^{-1} \sum_i W^0_{ji} W^0_{ki} = \delta_{jk}$. The question may arise whether there exist M orthogonal vectors $\boldsymbol{W}_j$ with N components having a value either +1 or -1 which meet this constraint. The constraint is met for $N = 2^l$ for integer l, and also approximately so for $N \gg M$.

We consider the latter case, which seems to apply to networks with many input units used in real applications.

The student will try to adjust all possible weights on its M hidden units without knowing the architecture of the teacher. We propose an effective teacher as a virtual target network which has the same M hidden units as those of the student and produces identical outputs with those of the original teacher. This is in fact a conjecture for a ground state of the thermodynamic system on which the weights of the student will converge in the weight space as learning makes progress. In solving the statistical mechanical problems, it is sometimes very crucial to know a ground state and find a relevant order parameter. A good example is the anti-ferromagnetic system. In this problem the relevant order parameter is the staggered magnetization which is the thermal average of the overlap of spins with the staggered spins of the ground state. In a similar way we will suppose the weights of the effective teacher as a ground state and introduce various order parameters from overlaps between the weights of the student and those of the effective teacher. There also appear different types of order parameters due to introducing replicas.

Let us divide the hidden units of the effective teacher into three blocks: B_1 for $1 \leqslant j \leqslant M_{\mathrm{T}}$, B_2 for $M_{\mathrm{T}}+1 \leqslant j \leqslant M_{\mathrm{T}}+L$, and B_3 for $M_{\mathrm{T}}+L+1 \leqslant j \leqslant M$, where $L=(M-M_{\mathrm{T}})/2$. We choose the weights for the hidden units in B_1 to be identical to those of the original teacher. An anti-pair is composed of W^0_{ki} for $k \in B_2$ and $W_{k+L,i}$ for $k+L \in B_3$ with the condition $W_{ki} = -W_{k+L,i}$. If we choose the whole weights for redundant hidden units in B_2 to make anti-pairs with those for B_3, the contributions of anti-pairs to outputs will be cancelled, making the original and the effective teacher yield the same outputs. However, there are no places involved with redundant weights of the effective teacher, since the energy depends only on outputs. While this effective teacher is a candidate for a ground state, it is not possible to examine from the calculation whether or not the student may indeed learn the effective teacher.

To resolve this difficulty, we consider the effective teacher which has a partial anti-pairing of redundant weights:

$$W^0_{ki} = \begin{cases} -W^0_{k+L,i} & \text{for} \quad N(1-p) \text{ input units} \\ W^0_{k+L,i} & \text{for} \quad Np \text{ input units.} \end{cases} \tag{1}$$

In this way, the effective teacher has a probability $1-p$ of redundant weights being made up of anti-pairs. Ultimately, we will take the limit $p \to 0$. The supposed weights of the effective teacher will be given explicitly in the calculation, but this does not mean that the student has *a priori* information about the teacher. The student only learns through outputs of which the effective teacher produces the same values as the original teacher in the limit $p \to 0$. This is a hypothetical scheme designed for the purpose of theoretical calculation. The introduction of the effective teacher might seem costly, making the problem unnecessarily complicated. In section 5, however, we will see that the calculation without introducing the effective teacher, that is equivalent to the case for $p=0$ exact, leads to an undesirable result, while the calculation based on the effective teacher in the limit $p \to 0$ gives rise to a consistent result with what is found from the simulation. One of the order parameters describing anti-pairing is the thermal average of the overlap $r_- = N^{-1}\sum_i W^0_{ki} W_{k+L,i}$. In the limit $p \to 0$ we will verify our anti-pairing learning scenario by finding $r_- \to -1$ in the learnt state, appearing for sufficient examples, while $r_- = 0$ in the unlearning state. For $p=0$ exact, however, we will find r_- to be always zero, independent of the number of examples. In this sense p can be regarded as a symmetry breaking parameter for anti-pairing ordering, which is analogous

to the magnetic field for the Ising model. We are now to solve the statistical mechanical problem by making a conjecture for a ground state with a symmetry breaking parameter and constructing order parameters associated with it. In the computer simulation by the Monte Carlo algorithm we will not use the effective teacher, but can check whether the learnt weights of the student are equivalent to those of the effective teacher.

Using the replica theory, the free energy F can be found as

$$-\beta F = \lim_{n\to 0} \frac{\partial}{\partial n} \overline{Z^n} = N(G_0 + \alpha G_r), \tag{2}$$

where we define $\alpha \equiv P/N$ as the number of examples per input unit and Z is the partition function. The over-bar denotes the average over disorder: $\{\xi_i^\mu\}, \{W_{ji}^0\}$. We take the limit in which $M, M_{\mathrm{T}}, M - M_{\mathrm{T}}$ go to ∞ with a finite ratio $\kappa = M_{\mathrm{T}}/M$, which makes analytic calculation possible. Various order parameters can be written as matrices:

$$\begin{aligned} Q_{jk}^{\sigma} &= \overline{\left\langle \frac{1}{N} \sum_i W_{ji}^{\sigma} W_{ki}^{\sigma} \right\rangle}, \\ C_{jk}^{\sigma\rho} &= \overline{\left\langle \frac{1}{N} \sum_i W_{ji}^{\sigma} W_{ki}^{\rho} \right\rangle} \qquad (\sigma \neq \rho), \\ R_{jk}^{\sigma} &= \overline{\left\langle \frac{1}{N} \sum_i W_{ji}^{\sigma} W_{ki}^{0} \right\rangle}. \end{aligned} \tag{3}$$

Here $\sigma, \rho = 1, \ldots, n$ are replica indices. The bracket denotes the thermal average. We use the replica-symmetric ansatz that the order parameters above do not depend on specific replica indices and specific pairs of replica indices. Dropping replica indices, we can then rewrite the above order parameter matrices as Q_{jk}, C_{jk} and R_{jk}, respectively.

The first two order parameter matrices in equation (3) are symmetric and the matrix elements are written as

$$Q_{jk} = \begin{cases} 1 & \text{for} \quad j = k \\ Q & \text{for} \quad j, k \in B_1 \ (j \neq k) \\ w & \text{for} \quad j \in B_2, k = j + L \in B_3 \\ Q_{1+} & \text{for} \quad j \in B_1, k \in B_2 \\ Q_{1-} & \text{for} \quad j \in B_1, k \in B_3 \\ Q_{2+} & \text{for} \quad j, k \in B_2 \text{ or } B_3 \ (j \neq k) \\ Q_{2-} & \text{for} \quad j \in B_2, k \in B_3 \ (k \neq j + L) \end{cases} \tag{4}$$

$$C_{jk} = \begin{cases} q & \text{for} \quad j = k \text{ and } j, k \in B_1 \\ q_+ & \text{for} \quad j = k \text{ and } j, k \in B_2 \text{ or } B_3 \\ q_- & \text{for} \quad j \in B_2, k = j + L \in B_3 \\ C & \text{for} \quad j, k \in B_1 \ (j \neq k) \\ C_{1+} & \text{for} \quad j \in B_1, k \in B_2 \\ C_{1-} & \text{for} \quad j \in B_1, k \in B_3 \\ C_{2+} & \text{for} \quad j, k \in B_2 \text{ or } B_3 \ (j \neq k) \\ C_{2-} & \text{for} \quad j \in B_2, k \in B_3 \ (k \neq j + L). \end{cases} \tag{5}$$

The elements of the asymmetric order parameter matrix R_{jk} are also written as

$$R_{jk} = \begin{cases} r & \text{for} \quad j = k \text{ and } j, k \in B_1 \\ r_+ & \text{for} \quad j = k \text{ and } j, k \in B_2 \text{ or } B_3 \\ r_- & \text{for} \quad j \in B_2, k = j + L \in B_3 \text{ or } k \in B_2, j = k + L \\ R & \text{for} \quad j, k \in B_1 \ (j \neq k) \\ R^*_{1+} & \text{for} \quad j \in B_1, k \in B_2 \\ R^*_{1-} & \text{for} \quad j \in B_1, k \in B_3 \\ R_{1+} & \text{for} \quad j \in B_2, k \in B_1 \\ R_{1-} & \text{for} \quad j \in B_3, k \in B_1 \\ R_{2+} & \text{for} \quad j, k \in B_2 \text{ or } B_3 \ (j \neq k) \\ R_{2-} & \text{for} \quad j \in B_2, k \in B_3 \text{ or } j \in B_3, k \in B_2 \ (k \neq j + L). \end{cases} \tag{6}$$

Weights on different hidden units that connect the same input unit, for example W_{ji} and W_{ki}, could be correlated because they share common input patterns, which is characteristic of the fully-connected structure. In fact, this is true in the permutation-symmetric state, explained later, which is stable for an insufficient number of examples less than a critical value. The order parameters between different hidden units are written in uppercase in the above equations, e.g., Q, C, R etc. They are found to be of $\mathcal{O}(M^{-1})$. However, they give a non-vanishing contribution to the free energy via rescaling given by multiplying a number of $\mathcal{O}(M)$. We rescale them in the following way:

$$\begin{aligned} \bar{Q} &= (M_{\mathrm{T}} - 1)Q, & \bar{Q}_{1\pm} &= \sqrt{M_{\mathrm{T}}L}\,Q_{1\pm}, & \bar{Q}_{2\pm} &= (L-1)Q_{2\pm}, \\ \bar{C} &= (M_{\mathrm{T}} - 1)C, & \bar{C}_{1\pm} &= \sqrt{M_{\mathrm{T}}L}\,C_{1\pm}, & \bar{C}_{2\pm} &= (L-1)C_{2\pm}, \\ \bar{R} &= (M_{\mathrm{T}} - 1)R, & \bar{R}^{(*)}_{1\pm} &= \sqrt{M_{\mathrm{T}}L}\,R^{(*)}_{1\pm}, & \bar{R}_{2\pm} &= (L-1)R_{2\pm}. \end{aligned} \tag{7}$$

G_r in equation (2) can be found from $\overline{Z^n} = \sum_{\{W^\sigma_{ji}\}} \mathrm{e}^{nPG_r}$ where the disorder average over input patterns ξ^μ_i is carried out, but the average over weights of the teacher is left. We find

$$G_r(A_1, A_2, A_3) = 2\int_{-\infty}^{\infty} \mathrm{D}t\, H\left(\frac{A_3}{\sqrt{A_2 - A_3^2}}t\right) \ln\left(\mathrm{e}^{-\beta} + (1 - \mathrm{e}^{-\beta})H\left(\sqrt{\frac{A_2}{A_1 - A_2}}t\right)\right), \tag{8}$$

where

$$\begin{aligned} A_1 &= \frac{2}{\pi}\frac{1}{M}\sum_{j,k} \sin^{-1} Q_{jk} \\ A_2 &= \frac{2}{\pi}\frac{1}{M}\sum_{j,k} \sin^{-1} C_{jk} \\ A_3 &= \frac{2}{\pi}\frac{1}{\sqrt{M M_{\mathrm{T}}}}\sum_{j,k} \sin^{-1} R_{jk}. \end{aligned} \tag{9}$$

We use $\mathrm{D}t = \frac{\mathrm{d}t}{\sqrt{2\pi}}\mathrm{e}^{-t^2/2}$ and $H(t) = \int_t^\infty \mathrm{D}x$. Using equations (4)–(7), we rewrite A_1, A_2, A_3 as

$$
\begin{aligned}
A_1 &= 1+(1-\kappa)\frac{2}{\pi}\sin^{-1}w+\frac{2}{\pi}\{\kappa\bar{Q}+(1-\kappa)(\bar{Q}_{2+}+\bar{Q}_{2-})+\sqrt{2\kappa(1-\kappa)}(\bar{Q}_{1+}+\bar{Q}_{1-})\}\\
A_2 &= \frac{2}{\pi}\{\kappa\sin^{-1}q+(1-\kappa)(\sin^{-1}q_{+}+\sin^{-1}q_{-})\\
&\qquad +\kappa\bar{C}+(1-\kappa)(\bar{C}_{2+}+\bar{C}_{2-})+\sqrt{2\kappa(1-\kappa)}(\bar{Q}_{1+}+\bar{Q}_{1-})\}\\
A_3 &= \frac{2}{\pi}(\kappa\gamma)^{-\frac{1}{2}}\Bigg\{\kappa\sin^{-1}r+(1-\kappa)(\sin^{-1}r_{+}+\sin^{-1}r_{-})\\
&\qquad +\kappa\bar{R}+(1-\kappa)(\bar{R}_{2+}+\bar{R}_{2-})+\sqrt{\frac{\kappa(1-\kappa)}{2}}(\bar{R}_{1+}+\bar{R}_{1-}+\bar{R}^{*}_{1+}+\bar{R}^{*}_{1-})\Bigg\}
\end{aligned}
\tag{10}
$$

where we define $\gamma\equiv 1+(\kappa^{-1}-1)(1-(2/\pi)\sin^{-1}(1-2p))$, going to 1 in the limit $p\to 0$.

Before the order parameters are introduced above in G_r, we first encounter a multiple integral for various variables. For example, as a part of the integral we have $\int \mathrm{d}Q^{\sigma}_{jk}\delta\big(Q^{\sigma}_{jk}-a\big)$ where $a=N^{-1}\sum_i W^{\sigma}_{ji}W^{\sigma}_{ki}$. From the integral representation of the delta function, this leads to $\int \mathrm{d}Q^{\sigma}_{jk}\int \mathrm{d}K^{\sigma}_{jk}(N/2\pi)\exp\big[\mathrm{i}NK^{\sigma}_{jk}\big(Q^{\sigma}_{jk}-a\big)\big]$. From the saddle-point approximation for this integral in the limit $N\to\infty$, Q^{σ}_{jk} and $-\mathrm{i}K^{\sigma}_{jk}$ take the saddle-point values, each corresponding to an order parameter Q_{jk} and its conjugate order parameter $\hat{Q}_{jk}$, respectively, assuming that the saddle point is replica-symmetric. In this manner, we define the hatted order parameters conjugate to the order parameters given in equations (4)–(6). The conjugate order parameters do not have direct physical meanings, so we can redefine them for convenience as

$$
\begin{aligned}
&M_{\mathrm{T}}^{-1}(\hat{q}-\hat{C})\to\hat{q}, && 2L^{-1}(\hat{q}_{\pm}-\hat{C}_{2\pm})\to\hat{q}_{\pm}, && M_{\mathrm{T}}^{-1}(\hat{r}-\hat{R})\to\hat{r},\\
&L^{-1}(\hat{r}_{\pm}-2\hat{R}_{2\pm})\to\hat{r}_{\pm}, && L^{-1}(\hat{w}-2\hat{Q}_{2-}-2\hat{q}_{-}+2\hat{C}_{2-})\to\hat{w}, &&\\
&\hat{Q}-\hat{C}\to\hat{Q}, && \hat{Q}_{1\pm}-\hat{C}_{1\pm}\to\hat{Q}_{1\pm}, && \hat{Q}_{2\pm}-\hat{C}_{2\pm}\to\hat{Q}_{2\pm}.
\end{aligned}
\tag{11}
$$

Summing over $\{W^{\sigma}_{ji}\}$ for the replicated partition function $\overline{Z^n}$, we get $\overline{Z^n}=\mathrm{e}^{nN(G_0+\alpha G_r)}$. Then G_0 can be found as

$$
G_0=-\frac{1}{2}\operatorname{Tr}\boldsymbol{Q}\cdot\hat{\boldsymbol{Q}}-\frac{1}{2}\operatorname{Tr}\boldsymbol{C}\cdot\hat{\boldsymbol{C}}-\operatorname{Tr}\boldsymbol{R}^{\mathrm{t}}\cdot\hat{\boldsymbol{R}}-\frac{L}{2}w\hat{w}+\frac{1}{2}\boldsymbol{q}\cdot\hat{\boldsymbol{q}}-\boldsymbol{r}\cdot\hat{\boldsymbol{r}}+\frac{1}{n}\overline{\ln Z_{\mathrm{eff}}}.
\tag{12}
$$

The superscript 't' denotes the transpose of the matrix. In this equation, we define the matrices:

$$
\begin{aligned}
\boldsymbol{Q}&=\begin{pmatrix}1+\bar{Q} & \bar{Q}_{1+} & \bar{Q}_{1-}\\ \bar{Q}_{1+} & 1+\bar{Q}_{2+} & w+\bar{Q}_{2-}\\ \bar{Q}_{1-} & w+\bar{Q}_{2-} & 1+\bar{Q}_{2+}\end{pmatrix},\qquad
\hat{\boldsymbol{Q}}=\begin{pmatrix}\hat{Q} & \hat{Q}_{1+} & \hat{Q}_{1-}\\ \hat{Q}_{1+} & \hat{Q}_{2+} & \hat{Q}_{2-}\\ \hat{Q}_{1-} & \hat{Q}_{2-} & \hat{Q}_{2+}\end{pmatrix},\\
\boldsymbol{C}&=\begin{pmatrix}1+\bar{Q}-q-\bar{C} & \bar{Q}_{1+}-\bar{C}_{1+} & \bar{Q}_{1-}-\bar{C}_{1-}\\ \bar{Q}_{1+}-\bar{C}_{1+} & 1+\bar{Q}_{2+}-q_{+}-\bar{C}_{2+} & w+\bar{Q}_{2-}-q_{-}-\bar{C}_{2-}\\ \bar{Q}_{1-}-\bar{C}_{1-} & w+\bar{Q}_{2-}-q_{-}-\bar{C}_{2-} & 1+\bar{Q}_{2+}-q_{+}-\bar{C}_{2+}\end{pmatrix},\\
\hat{\boldsymbol{C}}&=\begin{pmatrix}\hat{C} & \hat{C}_{1+} & \hat{C}_{1-}\\ \hat{C}_{1+} & \hat{C}_{2+} & \hat{C}_{2-}\\ \hat{C}_{1-} & \hat{C}_{2-} & \hat{C}_{2+}\end{pmatrix},\\
\boldsymbol{R}&=\begin{pmatrix}r+\bar{R} & \bar{R}^{*}_{1+} & \bar{R}^{*}_{1-}\\ \bar{R}_{1+} & r_{+}+\bar{R}_{2+} & r_{-}+\bar{R}_{2-}\\ \bar{R}_{1-} & r_{-}+\bar{R}_{2-} & r_{+}+\bar{R}_{2+}\end{pmatrix},\qquad
\hat{\boldsymbol{R}}=\begin{pmatrix}\hat{R} & \hat{R}^{*}_{1+} & \hat{R}^{*}_{1-}\\ \hat{R}_{1+} & \hat{R}_{2+} & \hat{R}_{2-}\\ \hat{R}_{1-} & \hat{R}_{2-} & \hat{R}_{2+}\end{pmatrix}.
\end{aligned}
\tag{13}
$$

We also define the vectors:

$$\boldsymbol{q} = \begin{pmatrix} q-1 \\ q_+ - 1 \\ q_- - w \end{pmatrix}, \qquad \hat{\boldsymbol{q}} = \begin{pmatrix} M_{\mathrm{T}}\hat{q} \\ L\hat{q}_+ \\ L\hat{q}_- \end{pmatrix}, \qquad \boldsymbol{r} = \begin{pmatrix} r \\ r_+ \\ r_- \end{pmatrix}, \qquad \hat{\boldsymbol{r}} = \begin{pmatrix} M_{\mathrm{T}}\hat{r} \\ L\hat{r}_+ \\ L\hat{r}_- \end{pmatrix}. \tag{14}$$

Z_{eff} is the effective partition function per input unit for which the summation is carried out over the weights branching from one input unit, $\{W_j^\sigma\}$, where the input index is dropped. The over-bar on $\ln Z_{\mathrm{eff}}$ denotes the disorder average over the weights of the effective teacher, which is the remaining disorder. Note that, in finding G_r, the disorder average over $\{\xi_i^\mu\}$ has already been carried out. In the appendix we include the calculation in detail to obtain Z_{eff}.

The training error ϵ_{t} is defined as the average error per example,

$$\epsilon_{\mathrm{t}} = \frac{1}{P}\overline{\langle E \rangle} = -\frac{\partial G_r}{\partial \beta}. \tag{15}$$

The generalization error ϵ_{g} is defined as the average error for a general example not trained. ϵ_{g} can be found from

$$\epsilon_{\mathrm{g}} = \overline{\left\langle \int \mathrm{D}^N\xi \frac{1}{4}\left\{o\left(\{W_{ki}^0\};\{\xi_i\}\right) - \sigma\left(\{W_{ji}\};\{\xi_i\}\right)\right\}^2 \right\rangle} = \frac{1}{\pi}\cos^{-1}\left(\frac{A_3}{\sqrt{A_1}}\right), \tag{16}$$

where $\int \mathrm{D}^N\xi$ denotes the N-dimensional Gaussian integration over $\{\xi_i\}$, equivalent to the average over arbitrary examples. ϵ_{t} and ϵ_{g} are found to decay as α increases, implying that learning improves as the number of examples increases.

3. Infinite degeneracies

The set of weights, $\{W_{ij}\}$, can be mapped to a NM-dimensional vector $\boldsymbol{W}$. The energy $E(\boldsymbol{W})$ gives the ragged energy surface with many valleys in this weight vector space. The bottom of a valley, separated by an infinite energy barrier from other valleys, corresponds to a ground state. There are two sources for degeneracy in ground states: (i) permutation symmetry and (ii) symmetry in anti-pairing. The former is characteristic of every network with the fully-connected structure. The latter is due to redundancy of hidden units of the student, not observed for $M = M_{\mathrm{T}}$.

Permutation symmetry is the invariance of the output under permutation of hidden units. Given a ground state, successive permutations of hidden units will generate $M!$ equivalent ground states. In order to examine degeneracy solely due to symmetry in anti-pairing, let us consider a set of ground-state weight vectors, all having the same weights for the hidden units in B_1 as the original teacher. The number of degeneracies is then equal to 2^{NL} which is the number of ways of choosing random numbers for the weights for B_2. The weights for B_3 are determined automatically from the condition of an anti-pair. Thus the total number of degeneracies in ground states is $M!2^{NL}$.

The learnt weight vector $\boldsymbol{W}$ of the student after being trained through sufficient examples will be one of $M!2^{NL}$ equivalent ground states that may not be $\boldsymbol{W}^0$ of the effective teacher. One can choose any of those ground-state weight vectors for the hypothetical effective teacher with no loss of generality. The current choice seems the most convenient for notation.

The Hamming distance, $(\boldsymbol{W} - \boldsymbol{W}')^2/4$, measures how far two weight vectors are apart in the weight vector space. The Hamming distance between ground states for usual thermodynamic systems such as the Ising model is extensive with the size of the system. An interesting feature in this problem is that the Hamming distance between two anti-paired ground states may not be extensive. In fact the shortest Hamming distance is equal to 2 for

the case in which two ground states differ only by a single anti-pair. It is hard to imagine the energy landscape where adjacent valleys a finite distance apart are separated by infinite energy barrier. This might be explained by the argument that it is probabilistically, though not energetically, hard to flip one weight of an anti-pair and exactly the other successively in order to move from one valley to the adjacent one. There are two possible ways to show this feature. In simulation it should be shown that the relaxation time from one valley to another increases with the size of the system, NL. A rigorous numerical study for this, that is extremely time-consuming, is not done in this paper, though we have some supportive data from the simulation for systems of rather small size. Theoretically we should show that the saddle-point solution extremizing the free energy manifests the anti-pairing ordering formed by single anti-paired ground states, not by a mixture of them. We will concentrate ourselves on the theoretical investigation in later sections.

For an insufficient number of examples, up to $\mathcal{O}(NM^a)$ for $a < 1$, the dominant probability distribution spreads over a wide region which is singly connected and located approximately equidistantly from valleys in the weight vector space. This means that there is only one thermodynamic state, called the permutation-symmetric state. As a result permutation symmetry is preserved and no specific anti-pairing is built up. As the number of examples goes to $\mathcal{O}(NM)$, the region of the dominant probability distribution is disconnected into many subregions, each of which is confined within a valley and corresponds to a single thermodynamic state called the permutation symmetry breaking state. There still exists the subregion with preserved permutation symmetry. The free energy of a subregion can be defined as minus temperature times the logarithm of the restricted partition function for which the summation over weights is done within the subregion. Let $F_{\rm PS}$ be the free energy of the subregion with permutation symmetry. Let $F_{\rm PSB}$ be the free energy of a single valley with anti-pairing and broken permutation symmetry, having the same value for all valleys. The permutation symmetry breaking states have the so-called extensive configuration entropy equal to $NL\ln 2 + \ln M!$. Therefore the total free energy of the valleys is equal to $F_{\rm PSB} - \beta^{-1}NL\ln 2$ in the limit $N \gg M$.

One can expect that the first-order phase transition from the permutation-symmetric to the permutation symmetry breaking state occurs at a critical number of examples, $P_{\rm c} = MN\alpha'_{\rm c}$. However, the determination of $\alpha'_{\rm c}$ might depend on the dynamics. More specifically, it might depend on the observation time (learning time) for which the time average, supposed to be the same as the thermal average, is carried out. Depending on the observation time, the student may visit a different number of thermodynamic states. If the student is allowed to learn for a sufficiently long time, enough to visit everywhere including all valleys in the weight space, the configuration entropy should be considered. In this case for the longest observation time, $\alpha'_{\rm c}$ is determined from the condition:

$$F_{\rm PS} = F_{\rm PSB} - \beta^{-1}NL\ln 2. \tag{17}$$

For minimal observation time which is the relaxation time from the permutation-symmetric region to one of the valleys, the transition might be signalled by the condition:

$$F_{\rm PS} = F_{\rm PSB}. \tag{18}$$

4. Permutation-symmetric phase

We can obtain the set of equations for order parameters from the saddle-point condition for the free energy. When the number of examples is of order of NM^a for $a < 1$, it can be found that there exists only a solution with permutation symmetry for the saddle-point equations. Due

to permutation symmetry, the order parameters do not depend on hidden unit indices, so we have $w = Q_{2-}, q = C, q_\pm = C_{2\pm}, r = R, r_\pm = R_{2\pm}$. Note that each of the order parameters in lowercase is diagonal in a block matrix given in equations (4)–(6) while the corresponding one in uppercase is off-diagonal. The order parameters written in lowercase can be found to be of the order of M^{-1}, so they are neglected in the large-M limit. The other order parameters written in uppercase are of the same order, but give non-negligible contribution through the rescaling given in equation (7). The hatted order parameters conjugate with the negligible order parameters, such as $\hat{w}, \hat{q}, \hat{q}_\pm, \hat{r}$, and $\hat{r}_\pm$, can also be shown to vanish. For example, the saddle-point condition yields

$$\hat{q} \propto \frac{\partial G_r}{\partial q} \propto \frac{1}{\sqrt{1-q^2}} - 1, \tag{19}$$

where we use the fact that G_r depends on $\sin^{-1} q - q$. One can easily see $\hat{q} \to 0$ as $q \to 0$.

We consider only the case for $p = 0$, since p plays an important role only in the permutation symmetry breaking phase. Using the above results, we can obtain the contribution of the effective partition function to the free energy in equation (A.14) in the appendix. Then equation (12) leads to

$$G_0 = M \ln 2 - \tfrac{1}{2} \operatorname{Tr} \boldsymbol{Q} \cdot \hat{\boldsymbol{Q}} - \tfrac{1}{2} \operatorname{Tr} \boldsymbol{C} \cdot \hat{\boldsymbol{C}} - \operatorname{Tr} \boldsymbol{R}^{\mathrm{t}} \cdot \hat{\boldsymbol{R}} - \tfrac{1}{2}(\ln \det(\boldsymbol{I} - \hat{\boldsymbol{Q}}) + \operatorname{Tr}(\boldsymbol{I} - \hat{\boldsymbol{Q}})^{-1} \cdot \hat{\boldsymbol{C}} + \operatorname{Tr} \boldsymbol{\Gamma} \cdot \hat{\boldsymbol{R}}^{\mathrm{t}} \cdot (\boldsymbol{I} - \hat{\boldsymbol{Q}})^{-1} \cdot \hat{\boldsymbol{R}}). \tag{20}$$

Note that the hatted order parameters appear only in G_0. Therefore, from the saddle-point condition, the derivatives of G_0 with respect to $\hat{\boldsymbol{Q}}, \hat{\boldsymbol{C}}$ and $\hat{\boldsymbol{R}}$ should vanish. Therefore we find

$$\boldsymbol{C} = (\boldsymbol{I} - \hat{\boldsymbol{Q}})^{-1}, \tag{21}$$

$$\boldsymbol{R} = (\boldsymbol{I} - \hat{\boldsymbol{Q}})^{-1} \cdot \hat{\boldsymbol{R}} \cdot \boldsymbol{\Gamma}, \tag{22}$$

$$\boldsymbol{Q} = (\boldsymbol{I} - \hat{\boldsymbol{Q}})^{-1} + (\boldsymbol{I} - \hat{\boldsymbol{Q}})^{-1} \cdot (\hat{\boldsymbol{C}} + \hat{\boldsymbol{R}} \cdot \boldsymbol{\Gamma} \cdot \hat{\boldsymbol{R}}^{t}) \cdot (\boldsymbol{I} - \hat{\boldsymbol{Q}})^{-1}, \tag{23}$$

where $\boldsymbol{\Gamma}$ is defined in equation (A.15). Using these equations, we can eliminate the hatted order parameters from G_0.

From the saddle-point condition, each hatted order parameter is given by the derivative of G_r with respect to its corresponding conjugate order parameter, which can be found from the expression for G_0 given in equation (20). In G_r, $\bar{R}_{1\pm}$, $\bar{R}^*_{1\pm}$ have the same multiplicative factor $\sqrt{\kappa(1-\kappa)/2}$ and $\bar{R}_{2+}$, $\bar{R}_{2-}$ have the same factor $1 - \kappa$. This gives

$$\hat{R}_{1+} = \hat{R}_{1-} = \hat{R}^*_{1+} = \hat{R}^*_{1-}, \qquad \hat{R}_{2+} = \hat{R}_{2-}. \tag{24}$$

Then we can show that equation (22) yields

$$\bar{R}^*_{1+} = \bar{R}^*_{1-} = \bar{R}_{2+} = \bar{R}_{2-} = 0. \tag{25}$$

These order parameters involve the weights for redundant hidden units of the effective teacher that are missing in the original teacher. As expected, the introduction of the effective teacher does not have any effect in the permutation-symmetric phase. We will show in the next section that in the permutation symmetry breaking phase the effective teacher plays a crucial role, even for $p \to 0$.

We also find that A_1, A_2 and A_3 in G_r are written as linear combinations of the order parameters. As a result, the elements of each hatted order parameter matrix differ only by constant factors, so that there is only one independent element for each matrix. Therefore there are three independent hatted order parameters for three matrices: $\hat{\boldsymbol{Q}}, \hat{\boldsymbol{C}}$ and $\hat{\boldsymbol{R}}$. From

equations (21)–(23), we can see that this is also true for order parameters. Thus we can express all order parameters in terms of three independent order parameters, e.g., $\bar{Q}$, $\bar{C}$ and $\bar{R}$. Through a bit lengthy calculation, we can eliminate the hatted order parameters. Interestingly, the free energy can be shown to be form-invariant, independent of M_{T}, for three redefined quantities given by

$$\begin{aligned} Q_{\mathrm{inv}} &= \kappa(1+\bar{Q}-(1-\kappa)), \\ C_{\mathrm{inv}} &= \kappa(1+\bar{Q}-\bar{C}-(1-\kappa)), \\ R_{\mathrm{inv}} &= \sqrt{\kappa}\,\bar{R}. \end{aligned} \tag{26}$$

M_{T} appears only implicitly in these quantities, as well as M. The form-invariant free energy in the permutation-symmetric phase can be found as

$$\begin{aligned} -\frac{\beta F_{\mathrm{PS}}}{N} &= M\ln 2 - \frac{Q_{\mathrm{inv}}}{2} + \frac{Q_{\mathrm{inv}}-R_{\mathrm{inv}}^2}{2C_{\mathrm{inv}}} + \frac{1}{2}\ln C_{\mathrm{inv}} \\ &\quad + 2\alpha\int \mathrm{D}t\, H(k_1 t)\ln(\mathrm{e}^{-\beta}+(1-\mathrm{e}^{-\beta})H(k_2 t)), \end{aligned} \tag{27}$$

where

$$k_1 = \frac{\frac{2}{\pi}R_{\mathrm{inv}}}{\sqrt{\frac{2}{\pi}(Q_{\mathrm{inv}}-C_{\mathrm{inv}})-\left(\frac{2}{\pi}R_{\mathrm{inv}}\right)^2}}, \qquad k_2 = \sqrt{\frac{\frac{2}{\pi}(Q_{\mathrm{inv}}-C_{\mathrm{inv}})}{1-\frac{2}{\pi}+\frac{2}{\pi}C_{\mathrm{inv}}}}. \tag{28}$$

Imposing the saddle-point condition on F_{PS} with respect to Q_{inv}, C_{inv} and R_{inv} yields three self-consistent equations that can be solved numerically. All the properties obtained from the solution of the self-consistent equations are independent of M_{T} as well as M, and therefore are the same as for $M = M_{\mathrm{T}}$. ϵ_{g} and ϵ_{t} are given by

$$\epsilon_{\mathrm{g}} = \frac{1}{\pi}\cos^{-1}\left(\frac{\frac{2}{\pi}R_{\mathrm{inv}}}{\sqrt{1-\frac{2}{\pi}+\frac{2}{\pi}Q_{\mathrm{inv}}}}\right), \qquad \epsilon_{\mathrm{t}} = \frac{2}{\pi}\int \mathrm{D}t\,\frac{H(k_1 t)(1-H(k_2 t))}{H(k_2 t)+(\mathrm{e}^{\beta}-1)^{-1}}. \tag{29}$$

The learning curve given by ϵ_{g} and ϵ_{t} versus α in the permutation-symmetric phase is independent of M_{T}.

As α increases, ϵ_{g} and ϵ_{t} decrease to reach plateaus. Taking the limit $\alpha \to \infty$, we can show

$$C_{\mathrm{inv}},\, Q_{\mathrm{inv}} - R_{\mathrm{inv}}^2 \sim \mathcal{O}(\alpha^{-1}) \to 0. \tag{30}$$

R_{inv} is found self-consistently from

$$R_{\mathrm{inv}} = \sqrt{1-\frac{2}{\pi}}\frac{A}{B}, \tag{31}$$

where

$$\begin{aligned} A &= \int \mathrm{D}x\,\frac{\mathrm{e}^{-R_{\mathrm{inv}}^2 x^2/\pi}}{H\left(\frac{2}{\pi}R_{\mathrm{inv}}x\right)+(\mathrm{e}^{\beta}-1)^{-1}}, \\ B &= \int \mathrm{D}x\, H(k_1 x)\left(\frac{\mathrm{e}^{-k_2^2 x^2}}{(H(k_2 x)+(\mathrm{e}^{\beta}-1)^{-1})^2} - \frac{\sqrt{2\pi}k_2 x}{H(k_2 x)+(\mathrm{e}^{\beta}-1)^{-1}}\right). \end{aligned} \tag{32}$$

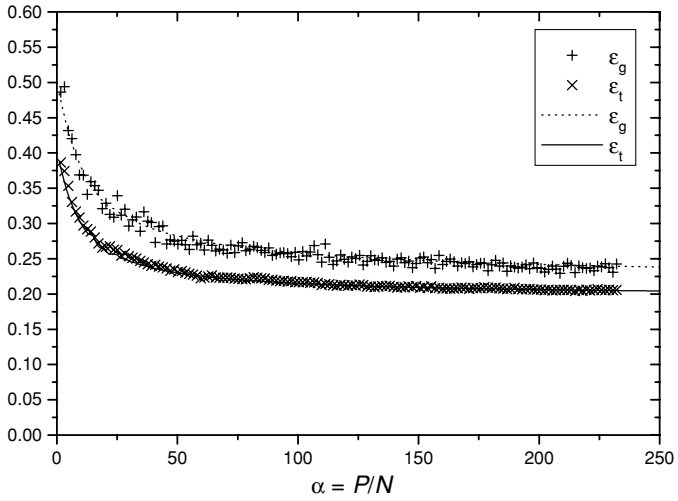

Figure 1. ϵ_g (+) and ϵ_t (×) versus $\alpha = P/N$ at $T = 2.5$ are plotted from the Monte Carlo simulation for $N = 51$, $M = 41$, $M_T = 21$. The solid line and the broken line denote the theoretical plots for ϵ_g and ϵ_t respectively, obtained by solving self-consistent equations numerically, showing a good agreement with those from the simulation.

Using equation (30), we find that k_1 and k_2 simply become $\sqrt{\frac{2}{\pi-2}}$ and $\sqrt{\frac{2}{\pi-2}}R_{\rm inv}$ respectively. Then we can obtain the limiting values of ϵ_g and ϵ_t at plateaus from equation (29). We can also obtain the free energy for large α,

$$-\frac{\beta F_{\rm PS}}{N} = M\ln 2 + 2\alpha\int_{-\infty}^{\infty} {\rm D}t\, H\left(\sqrt{\frac{2}{\pi-2}}t\right)\ln\left(e^{-\beta} + (1-e^{-\beta})H\left(\sqrt{\frac{2}{\pi-2}}R_{\rm inv}t\right)\right). \tag{33}$$

This result can also hold for $\alpha \sim \mathcal{O}(M)$ where the term of $\mathcal{O}(\ln M)$ is neglected.

Figure 1 shows the result from the Monte Carlo simulation at $T = 2.5$. From the theoretical calculation, $\epsilon_t \to 0.20$ and $\epsilon_g \to 0.21$ as $\alpha \to \infty$, which agrees well with the values from the simulation. As the temperature T goes down, ϵ_t decreases to 0 while ϵ_g increases. This implies that at low T, where the tendency to minimize the energy is strong, weight vectors producing small training error for given examples are not quite suitable for a general, not-trained example. This is a kind of over-fitting caused by a strict minimization of the energy. The stochastic or noisy learning turns out to be helpful in overcoming this kind of over-fitting.

5. Permutation symmetry breaking phase

The permutation symmetry breaking state is expected to appear for sufficiently many examples. $F_{\rm PS}$ will increase to be comparable to $F_{\rm PSB}$, that will be found in this section, as α grows. As seen in equation (33), $R_{\rm inv}$ is fixed in the limit $\alpha \to \infty$ for given β, so only change will be made by α. Therefore we expect $\mathcal{O}(\alpha) \geqslant M$; otherwise α would yield an inconsiderable change in $F_{\rm PS} \simeq -NM\beta^{-1}\ln 2$. In fact the transition from the permutation-symmetric to the permutation symmetry breaking state will be found to occur for $\alpha \sim \mathcal{O}(\mathcal{M})$.

Let us define α' by α/M. We expect that there exists a new solution for the saddle-point equations for order parameters such that

$$q \to 1, \qquad q_+ \to 1, \qquad r \to 1, \qquad r_+ \to 1. \tag{34}$$

The limit $q \to 1, r \to 1$ means that the weights of the student and the effective teacher for hidden units in B_1 become identical. Similarly, the limit $q_+ \to 1, r_+ \to 1$ means that the weights of the two networks for hidden units in B_2, B_3 get close. As $p \to 0$, the student will have anti-pairing as perfectly as the effective teacher. So we also expect the new permutation symmetry breaking solution to give

$$q_- \to -1, \qquad r_- \to -1, \qquad w \to -1 \qquad \text{as} \quad p \to 0. \tag{35}$$

At an early step, we tried to find the permutation symmetry breaking solution without introducing the effective teacher. In this approach, the matrix R_{jk} in equation (6) is $M \times M_{\mathrm{T}}$. This gives the correct permutation-symmetric solution. Note that the matrix elements associated with redundant hidden units of the effective teacher vanish, as can be seen in equation (25). There are also missing order parameters, r_+ and r_-. These order parameters measure the overlap of weights between the student and the effective teacher for redundant hidden units. Technically, this approach corresponds to the case where $r_\pm, \hat{r}_\pm = 0$ in the formalism based on the effective teacher. However, we cannot find a desired solution with $q, q_+ \to 1$ and $q_-, w \to -1$ in this approach. It is very crucial to introduce the effective teacher as a ground state and the overlaps with it as the order parameters. A similar scheme can be seen in the anti-ferromagnetic system where the stagger magnetization is introduced as a crucial order parameter.

There are two kinds of divergence for the hatted order parameters. One is due to permutation symmetry breaking yielding equation (34). It makes $\hat{q}, \hat{q}_+, \hat{r}$ and $\hat{r}_+$ divergent. For example,

$$\hat{q} \propto \alpha' \frac{\partial G_r}{\partial q} \propto \left(\frac{1}{\sqrt{1-q^2}} - 1 \right) \times (\cdots) \sim \frac{1}{\sqrt{1-q}} \times (\cdots), \tag{36}$$

where '$\cdots$' comes from the differentiation of G_r. In the same way, $\hat{q}_+ \sim (1-q_+)^{-1/2}, \hat{r} \sim (1-r)^{-1/2}, \hat{r}_+ \sim (1-r_+)^{-1/2}$. From the Monte Carlo simulation for finite M and large N, we can still observe that the weight vector of the student collapses to a single ground-state weight vector. Presumably this divergence is due to the thermodynamic limit $N \to \infty$. We can then use

$$\hat{q}, \hat{q}_+, \hat{r}, \hat{r}_+ \to \infty. \tag{37}$$

The exact scaling behaviour on this divergence is beyond the saddle-point approximation. We can also use a strong condition,

$$\frac{\hat{r}}{\sqrt{\hat{q}}} \to \infty, \qquad \frac{\hat{r}_+}{\sqrt{\hat{q}_+}} \to \infty. \tag{38}$$

This guarantees self-consistently that $r \to 1, r_+ \to 1$ as well as $q \to 1, q_+ \to 1$. There is a similar divergence due to equation (35) in the limit $p \to 0$. As a result, $\hat{q}_- \sim (1+q_-)^{-1/2}, \hat{r}_- \sim (1+r_-)^{-1/2}$ and $\hat{w} \sim (1+w)^{-1/2}$. We consider that p^{-1} is large, but $N \gg p^{-1}$. Therefore, we use

$$\hat{q}, \hat{q}_+, \hat{r}, \hat{r}_+ \gg \hat{q}_-, \hat{r}_-, \hat{w}. \tag{39}$$

In the appendix, using this together with equation (38), we can simplify many involved terms.

The other divergence is due to the large-M limit. For example,

$$\hat{Q} \propto \alpha \frac{\partial G_r}{\partial \bar{Q}} \simeq M\alpha' \times (\cdots), \tag{40}$$

where $\alpha \sim \mathcal{O}(M)$ is used and '$\cdots$' comes from the differentiation of G_r. This divergence is applied to the matrix elements of $\hat{\boldsymbol{Q}}, \hat{\boldsymbol{C}}$ and $\hat{\boldsymbol{R}}$. We are considering $N \gg M$, so the divergence in equations (37) and (38) is stronger than that due to the large-M limit. This fact is also used in the appendix.

Taking divergence into account, we can simplify G_0 in equation (12) (see the appendix):

$$G_0 = -\frac{1}{2}\mathrm{Tr}(\boldsymbol{Q}-\boldsymbol{H})\cdot\hat{\boldsymbol{Q}} - \frac{1}{2}\mathrm{Tr}(\boldsymbol{C}-\boldsymbol{A})\cdot\hat{\boldsymbol{C}} - \mathrm{Tr}(\boldsymbol{R}^{\mathrm{t}}-\boldsymbol{H})\cdot\hat{\boldsymbol{R}}$$
$$-\frac{L}{2}(w+1-2p)\hat{w} + \frac{M_{\mathrm{T}}}{2}(q-1)\hat{r} - M_{\mathrm{T}}(r-1)\hat{r} + \frac{L}{2}((q_+-1)\hat{q}_+$$
$$+(q_-+1-2p)\hat{q}_-) - L((r_+-1)\hat{r}_+ + (r_-+1-2p)\hat{r}_-). \tag{41}$$

The matrix $\boldsymbol{H}$ is defined in equation (A.22) in the appendix. Since the hatted order parameters appear only in G_0, the derivatives of G_0 with respect to them should vanish. Therefore, we find

$$q = q_+ = r = r_+ = 1, \qquad w = q_- = r_- = -1+2p. \tag{42}$$

This implies that in the permutation symmetry breaking state, the weight vector of the student is frozen to one of the ground-state weight vectors. We can also find

$$\boldsymbol{Q} = \boldsymbol{R} = \boldsymbol{H}, \qquad \boldsymbol{C} = \boldsymbol{A} = 0. \tag{43}$$

The structure of $\boldsymbol{H}$ manifests anti-pairing for $p \to 0$. Therefore, we get $G_0 = 0$.

The three terms in equation (10) for G_r can be found as

$$A_1 = A_2 = A_3^2 = \kappa + \frac{4}{\pi}(1-\kappa)\sqrt{p}. \tag{44}$$

This leads to $G_r = 0$. Therefore the free energy of a single valley for the permutation symmetry breaking solution vanishes,

$$F_{\mathrm{PSB}} = 0. \tag{45}$$

This means that the entropy and the energy are equal to zero. The zero entropy confirms that the weight vector of the student is frozen to a single ground state, not a mixture. The zero energy means that the student becomes equivalent to the effective teacher. The training error ϵ_{t} and generalization error ϵ_{g} can certainly be shown to vanish.

The two extreme conditions in equations (17) and (18) might yield the lower and the upper bound respectively for the critical value α_{c}'. We can find

$$\alpha_{\mathrm{c}}' = -\frac{\gamma \ln 2}{2}\left[\int_{-\infty}^{\infty} \mathrm{D}t\, H\left(\sqrt{\frac{2}{\pi-2}}t\right)\ln\left(\mathrm{e}^{-\beta} + (1-\mathrm{e}^{-\beta})H\left(\sqrt{\frac{2}{\pi-2}}R_{\mathrm{inv}}t\right)\right)\right]^{-1}, \tag{46}$$

where R_{inv} is the value for $\alpha \to \infty$ in equation (31). The parameter γ is defined by

$$\gamma = \begin{cases} 1 & \text{(upper bound)} \\ \frac{1}{2}(1+M_{\mathrm{T}}/M) & \text{(lower bound).} \end{cases} \tag{47}$$

Note that the upper bound is independent of M_{T}, the same as the critical value for $M = M_{\mathrm{T}}$.

In the computer simulation, the target network is chosen to have M_{T} hidden units, so is not the effective teacher in the theoretical scheme. Figure 2 shows the result from the Monte Carlo simulation at $T = 2.5$. The simulation was run so optimally as to barely detect the transition, which might correspond to the regime of the shortest observation time. In fact α_{c}' measured from the simulation is close to the theoretical estimate which is equal to the critical value for $M = M_{\mathrm{T}}$.

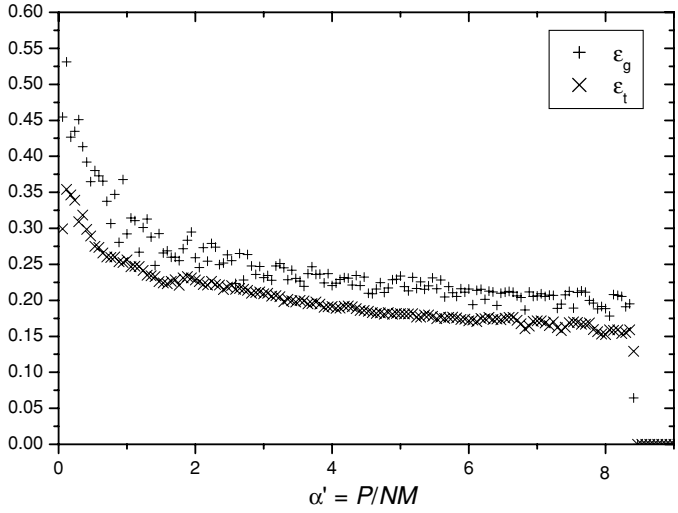

Figure 2. ϵ_g (+) and ϵ_t (×) versus $\alpha' = P/NM$ at $T = 2.5$ are plotted from the Monte Carlo simulation for $N = 25, M = 17, M_T = 11$. We use smaller values of N, M, M_T than those in figure 1 in order to observe the phase transition. The upper bound of α'_c is expected to be independent of M_T. It is in fact close to the theoretical estimate, which is about 8.3.

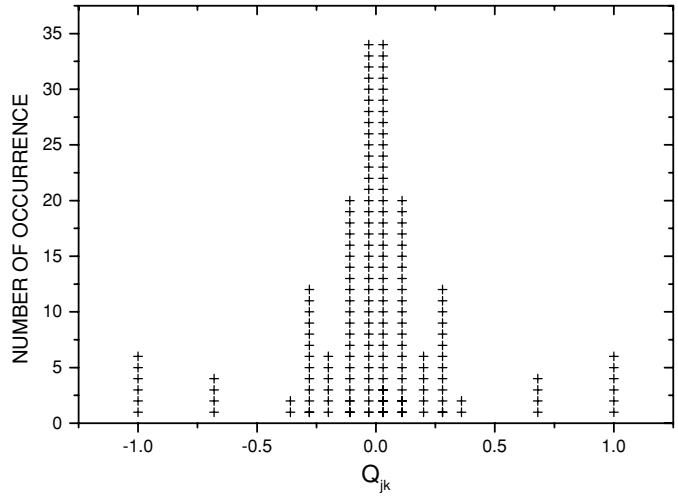

Figure 3. The distribution of the overlaps Q_{jk} outside the diagonal block for $j, k \in B_1$ for permutation symmetry breaking state is plotted. The ordinate denotes the number of occurrences of matrix elements on the abscissa. Data are obtained from the Monte Carlo simulation at $T = 2.5$ and for $N = 25, M = 17, M_T = 11$. The distribution is symmetric about zero, which manifests anti-pairing ordering.

Figure 3 shows the distribution of overlaps $N^{-1}\sum_{i=1}^{N} W_{ji}W_{ki}$ for $j, k = 1, \ldots, M$ for $\alpha' > \alpha'_c$ from which the distribution of overlaps for the teacher itself $N^{-1}\sum_{i=1}^{N} W^0_{ji}W^0_{ki}$ for $j, k = 1, \ldots, M_T$ is extracted. It gives the distribution of Q_{jk} outside the diagonal block ($j, k \in B_1$) in the theoretical scheme, given in equation (4). Because of anti-pairing, this distribution is expected to be symmetric about zero. In fact, the figure shows perfect symmetry, which directly verifies our scenario based on anti-pairing. Calculation is done

under the condition: $N \gg M \gg 1$ and finite M_T/M. We still observe a good agreement between the results from the theory and the simulation even under a looser condition.

6. Discussions and future

We practised the simulation for networks with smaller size than that used for figure 2. We observed that relaxation took place from valley to valley in a very complicated manner, where valleys are identified as having zero energy. We measured the anti-pairing ordering and the Hamming distance of each valley from an initial valley. We found that each valley has perfect anti-pairing and that the Hamming distance remains constant inside a valley but varies from valley to valley. The latter implies that each valley corresponds to a single weight vector, not a mixture. We also observed that the relaxation time to escape from a valley tends to increase with the network size. This observation may support the idea that each thermodynamic state, represented by a valley, is given by single anti-paired weight vector. For more rigorous analysis, we should examine the scaling behaviour of the relaxation time in system size NM, which is not done in this paper.

Interpretation of the two limits for the critical value α_c' in equation (46) is not yet clearly made. Presumably the lower bound corresponds to the case for the longest observation time and the upper bound to that for the shortest observation time, both of which are infinite in the thermodynamic limit. Recent work on structural glasses has given us rich and novel concepts about glassy states with extensive configuration entropy [25–28]. We do not attempt to investigate the dynamical aspect for our problem in this direction, which is beyond our present scope.

Our study presents a particular way for a generic situation in which a more complicated student adjusts its redundant structure, though not recognizing it in the process of learning, to learn a simpler teacher. It is anti-pairing of redundant weights in our case. The redundant structure may provide the student with more diversity in solving the examples than the teacher. This is manifested in our study by the appearance of infinitely many ground states due to anti-pairing. The student may learn a simpler teacher from a smaller number of examples, which can be seen from our finding that there exists a lower bound of α_c'.

Learning in application areas is usually imperfect because of the limitation in constructing a network. One does not know the proper size of the hidden layer, which cannot be enlarged indefinitely. One happens to encounter the situation in which the student has insufficient hidden units. This situation can be studied in our model for $M < M_T$ to which we are now extending our investigation. The learning mechanism is quite different from that for $M > M_T$. In this case, there are no ideal attractors yielding the desired output in the weight vector space of the student. When $M_T - M$ is small, a possible candidate for an attractor might be a weight vector composed of the weights which are the same as those on partial M hidden units of the teacher. This weight vector will give a better generalization error because it is very similar to the weight vector of the teacher. However, it always yields a nonzero training error for given examples. Such partial learning may not be probable at low temperatures where the tendency to minimize the energy is dominant. Therefore, partial learning, if possible, will be found at finite temperatures. Noise is expected to play a more drastic role, through a phase transition, in escaping from over-fitting at low temperatures than in this study. However, it is not clear whether such partial learning might be possible even for large $M_T - M$, resulting in a relatively large generalization error. Limit $M_T \gg M$ will lead to $\epsilon_g = 1/2$, definitely larger than the value given from the permutation-symmetric state, which is not desirable. Presumably there will be a lower bound of M below which there is no partial learning. There is also the possibility of replica symmetry breaking, which was found to be diagnostic of

imperfect learning in some cases [3, 4, 9, 10]. We are now investigating this problem carefully by using the one-step replica symmetry breaking scheme.

Acknowledgments

This research was funded by the Brain Science & Engineering Research Program (the Ministry of Science and Technology) in Korea.

Appendix. Effective partition function

The effective partition function in equation (12) is written as

$$
\begin{aligned}
Z_{\mathrm{eff}} = \sum_{\{W_j^\sigma\}} \int \prod_{\sigma=1}^{n} \prod_{a=1}^{3} \frac{\mathrm{d}\Lambda_a^\sigma \, \mathrm{d}\hat{\Lambda}_a^\sigma}{2\pi} \int \prod_{a=1}^{3} \mathrm{D}_{\hat{C}} p_a \\
\times \exp\left(\mathrm{i} \sum_\sigma \boldsymbol{\Lambda}^\sigma \cdot \hat{\boldsymbol{\Lambda}}^\sigma + \sum_\sigma \boldsymbol{p} \cdot \boldsymbol{\Lambda}^\sigma + \frac{1}{2} \sum_\sigma \boldsymbol{\Lambda}^\sigma \cdot \hat{\boldsymbol{Q}} \cdot \boldsymbol{\Lambda}^\sigma + \sum_\sigma \boldsymbol{\Lambda}^\sigma \cdot \hat{\boldsymbol{R}} \cdot \boldsymbol{\Lambda}^0 \right) \\
\times \exp\left[-\mathrm{i} \sum_\sigma \left\{ \frac{\hat{\Lambda}_1^\sigma}{\sqrt{M_{\mathrm{T}}}} \sum_{j \in B_1} W_j^\sigma + \frac{\hat{\Lambda}_2^\sigma}{\sqrt{L}} \sum_{j \in B_2} W_j^\sigma + \frac{\hat{\Lambda}_3^\sigma}{\sqrt{L}} \sum_{j \in B_3} W_j^\sigma \right\} \right. \\
+ \frac{1}{2} \hat{q} \sum_{j \in B_1} \left(\sum_\sigma W_j^\sigma \right)^2 + \hat{r} \sum_{j \in B_1} \sum_\sigma W_j^\sigma W_j^0 + \frac{1}{2} \hat{w} \sum_{j \in B_2} \sum_\sigma W_j^\sigma W_{j+L}^\sigma \\
+ \frac{1}{8} (\hat{q}_+ + \hat{q}_-) \sum_{j \in B_2} \left(\sum_\sigma \left(W_j^\sigma + W_{j+L}^\sigma \right) \right)^2 + \frac{1}{8} (\hat{q}_+ - \hat{q}_-) \\
\times \sum_{j \in B_2} \left(\sum_\sigma \left(W_j^\sigma - W_{j+L}^\sigma \right) \right)^2 + \frac{1}{4} (\hat{r}_+ + \hat{r}_-) \sum_{j \in B_2} \sum_\sigma \left(W_j^\sigma + W_{j+L}^\sigma \right) \left(W_j^0 + W_{j+L}^0 \right) \\
\left. + \frac{1}{4} (\hat{r}_+ - \hat{r}_-) \sum_{j \in B_2} \sum_\sigma \left(W_j^\sigma - W_{j+L}^\sigma \right) \left(W_j^0 - W_{j+L}^0 \right) \right]. \qquad \text{(A.1)}
\end{aligned}
$$

In this equation $\boldsymbol{p}$ is a three-dimensional vector with components p_a for $a = 1, 2, 3$ and $\int \prod_a \mathrm{D}_{\hat{C}} p_a$ denotes the multi-variable Gaussian integration with the variance $\overline{p_a p_b} = \hat{C}_{ab}$. $\boldsymbol{\Lambda}^\sigma$ and $\hat{\boldsymbol{\Lambda}}^\sigma$ are three-dimensional vectors with the components Λ_a^σ and $\hat{\Lambda}_a^\sigma$, respectively. Also a vector $\boldsymbol{\Lambda}^0$ is defined by

$$
\boldsymbol{\Lambda}^0 = \left(\Lambda_1^0, \Lambda_2^0, \Lambda_3^0 \right) = \left(\sum_{j \in B_1} \frac{W_j^0}{\sqrt{M_{\mathrm{T}}}}, \sum_{j \in B_2} \frac{W_j^0}{\sqrt{L}}, \sum_{j \in B_3} \frac{W_j^0}{\sqrt{L}} \right). \qquad \text{(A.2)}
$$

Λ_a^0 are random variables due to the random distribution of weights of the effective teacher. In the large-M limit we only need moments up to the second order,

$$
\overline{\Lambda_a^0} = 0, \qquad \overline{\left(\Lambda_a^0 \right)^2} = 1, \qquad \overline{\Lambda_2^0 \Lambda_3^0} = r_0, \qquad \overline{\Lambda_1^0 \Lambda_2^0} = \overline{\Lambda_1^0 \Lambda_3^0} = 0. \qquad \text{(A.3)}
$$

Higher order moments, at most $\mathcal{O}(M^{-1})$, are neglected. The matrices $\hat{\boldsymbol{C}}$, $\hat{\boldsymbol{Q}}$ and $\hat{\boldsymbol{R}}$ are defined in equation (13).

The effective partition function contains the quadratic terms in which weights are coupled between replicas. They can be decoupled by the Gaussian transformation $\exp(h^2/2) = \int \mathrm{D}x \exp(hx)$. Introducing Gaussian integral variables, z_j, z_{+j}, z_{-j}, these quadratic terms become linear as follows:

$$\sum_{\sigma,j\in B_1} \sqrt{\hat{q}} z_j W_j^\sigma + \sum_{\sigma,j\in B_2} \left(\sqrt{\hat{q}_+ + \hat{q}_-} z_{+j} \frac{W_j^\sigma + W_{j+L}^\sigma}{2} + \sqrt{\hat{q}_+ - \hat{q}_-} z_{-j} \frac{W_j^\sigma - W_{j+L}^\sigma}{2} \right). \tag{A.4}$$

Now we can easily carry out the sum over $\{W_j^\sigma\}$. In this summation we use the cumulant expansion up to the second order in $\hat{\Lambda}_a^\sigma$ which is multiplied by $M_{\mathrm{T}}^{-1/2}$ or $L^{-1/2}$, giving a non-vanishing contribution to the integration over $\hat{\Lambda}_a^\sigma$. We also rename weights by $W_j^\sigma W_j^0 \to W_j^\sigma$. As a result, we can find the effective partition function as being independent of site index j. In this procedure we introduce the effective fields given as

$$h(z) = \sqrt{\hat{q}} z + \hat{r}, \qquad h_{\pm k}(z_\pm) = \sqrt{\hat{q}_+ \pm \hat{q}_-} z_\pm + \tfrac{1}{2}(\hat{r}_+ \pm \hat{r}_-)\big(1 \pm W_{+k}^0 W_{-k}^0\big), \tag{A.5}$$

where z, z_+, z_- are the same Gaussian variables as in equation (A.4) with site index j dropped. W_{+k}^0 and W_{-k}^0 denote the partially anti-paired weights of the effective teacher, W_k^0 and W_{k+L}^0 respectively for $k \in B_2$. Then we can impose the probability of anti-pairing for the overlap $r_0 = W_{+k}^0 W_{-k}^0$ defined in equation (A.3), given as

$$P(r_0) = (1-p)\delta(r_0+1) + p\delta(r_0-1). \tag{A.6}$$

Let us define for $k \in B_2$

$$\begin{aligned} c_k &= \mathrm{e}^{\hat{w}/2} \cosh h_{+k} + \mathrm{e}^{-\hat{w}/2} \cosh h_{-k}, \\ s_{\pm k} &= \mathrm{e}^{\hat{w}/2} \sinh h_{+k} \pm \mathrm{e}^{-\hat{w}/2} \sinh h_{-k}. \end{aligned} \tag{A.7}$$

Then we abbreviate many involved terms:

$$\begin{aligned} &a = \overline{\tanh^2 h}, \qquad a_\pm = \frac{1}{L}\sum_k \overline{\left(\frac{s_{\pm k}}{c_k}\right)^2}, \qquad a' = \frac{1}{L}\sum_k \overline{\left(\frac{2\sinh \hat{w}}{c_k^2}\right)}, \\ &b = \overline{\tanh h}, \qquad b_\pm = \frac{1}{\sqrt{L}}\sum_k \overline{\left(\frac{W_{+k}^0 s_{\pm k}}{c_k}\right)}, \\ &b_1 = \frac{\mathrm{e}^{\hat{w}/2}}{\sqrt{L}}\sum_k \overline{\frac{W_{+k}^0 \sinh h_{+k}}{c_k}}, \qquad b_2 = \frac{\mathrm{e}^{-\hat{w}/2}}{\sqrt{L}}\sum_k \overline{\frac{W_{+k}^0 \sinh h_{-k}}{c_k}}, \\ &d = a - \big(\Lambda_1^0\big)^2 b^2, \qquad d_+ = \overline{b_1^2} - \overline{b_1}^2, \qquad d_- = \overline{b_2^2} - \overline{b_2}^2. \end{aligned} \tag{A.8}$$

In this expression, the over-bar denotes the average over z, z_+, z_-. The quadratic terms in $\hat{\Lambda}_a^\sigma$, yielded by the cumulant expansion, can also be decoupled by the Gaussian transformation:

$$\begin{aligned} &-\frac{1}{2}d\left(\sum_\sigma \hat{\Lambda}_1^\sigma\right)^2 - \frac{1}{2}d_+\left(\sum_\sigma \big(\hat{\Lambda}_2^\sigma + \hat{\Lambda}_3^\sigma\big)\right)^2 - \frac{1}{2}d_-\left(\sum_\sigma \big(\hat{\Lambda}_2^\sigma - \hat{\Lambda}_3^\sigma\big)\right)^2 \\ &\qquad \to -\mathrm{i}t d\sum_\sigma \hat{\Lambda}_1^\sigma - \mathrm{i}t_+ d_+ \sum_\sigma \big(\hat{\Lambda}_2^\sigma + \hat{\Lambda}_3^\sigma\big) - \mathrm{i}t_- d_- \sum_\sigma \big(\hat{\Lambda}_2^\sigma - \hat{\Lambda}_3^\sigma\big), \end{aligned} \tag{A.9}$$

where t, t_+, t_- are a new set of Gaussian integral variables with variance 1.

We define a vector $\boldsymbol{t}$ by

$$\boldsymbol{t} = \begin{pmatrix} \Lambda_1^0 b + \sqrt{d}t \\ b_+ + \sqrt{d_+}t_+ + \sqrt{d_-}t_- \\ b_- + \sqrt{d_+}t_+ - \sqrt{d_-}t_- \end{pmatrix}. \tag{A.10}$$

We also define a matrix $\boldsymbol{A}$ by

$$\boldsymbol{A} = \begin{pmatrix} 1-a & 0 & 0 \\ 0 & 1-a_+ & a' \\ 0 & a' & 1-a_- \end{pmatrix}. \tag{A.11}$$

Using $\boldsymbol{A}$ and $\boldsymbol{t}$, we can write the contribution of the effective partition function to the free energy in a compact manner:

$$\begin{aligned} \frac{1}{n}\overline{\ln Z_{\rm eff}} = {} & M_{\rm T}\,\overline{\ln 2\cosh h(z)} + L\,\overline{\ln 2(\mathrm{e}^{\hat{w}/2}\cosh h_+(z_+) + \mathrm{e}^{-\hat{w}/2}\cosh h_-(z_-))} \\ & - \frac{1}{2}\overline{\ln\det(\boldsymbol{I}-\boldsymbol{A}\cdot\hat{\boldsymbol{Q}})} + \frac{1}{2}\,\mathrm{Tr}(\boldsymbol{A}^{-1}-\hat{\boldsymbol{Q}})^{-1}\cdot\hat{\boldsymbol{C}} - \frac{1}{2}\overline{\boldsymbol{t}\cdot\boldsymbol{A}^{-1}\cdot\boldsymbol{t}} \\ & + \frac{1}{2}\overline{\boldsymbol{t}\cdot\boldsymbol{A}^{-1}\cdot(\boldsymbol{I}-\boldsymbol{A}\cdot\hat{\boldsymbol{Q}})^{-1}\cdot\boldsymbol{t}} + \frac{1}{2}\overline{\boldsymbol{\Lambda}^0\cdot\hat{\boldsymbol{R}}^{\rm t}\cdot(\boldsymbol{A}^{-1}-\hat{\boldsymbol{Q}})^{-1}\cdot\hat{\boldsymbol{R}}\cdot\boldsymbol{\Lambda}^0} \\ & + \overline{\boldsymbol{t}\cdot(\boldsymbol{I}-\boldsymbol{A}\cdot\hat{\boldsymbol{Q}})^{-1}\cdot\hat{\boldsymbol{R}}\cdot\boldsymbol{\Lambda}^0}, \end{aligned} \tag{A.12}$$

where the over-bar denotes the average over all possible random variables $z, z_\pm, r_0, t, t_\pm, \Lambda_a^0$, and $\boldsymbol{I}$ is the 3×3 identity matrix. This seems to be very complicated, but can be simplified by the useful limits found for the permutation-symmetric and the permutation symmetry breaking states.

In the permutation-symmetric state, $q, q_\pm, r, r_\pm, w$ and the corresponding hatted order parameters, $\hat{q}, \hat{q}_\pm, \hat{r}, \hat{r}_\pm, \hat{w}$ vanish. Therefore

$$\boldsymbol{A}\to\boldsymbol{I}, \qquad \boldsymbol{t}\to 0. \tag{A.13}$$

Using this, we obtain

$$\frac{1}{n}\overline{\ln Z_{\rm eff}} = M\ln 2 - \frac{1}{2}[\ln\det(\boldsymbol{I}-\hat{\boldsymbol{Q}}) + \mathrm{Tr}(\boldsymbol{I}-\hat{\boldsymbol{Q}})^{-1}\cdot\hat{\boldsymbol{C}} + \mathrm{Tr}\,\boldsymbol{\Gamma}\cdot\hat{\boldsymbol{R}}^t\cdot(\boldsymbol{I}-\hat{\boldsymbol{Q}})^{-1}\cdot\hat{\boldsymbol{R}}], \tag{A.14}$$

where

$$\boldsymbol{\Gamma} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix}. \tag{A.15}$$

Therefore we find equation (20).

In the permutation symmetry breaking state, the strong condition in equation (38) together with equation (39), gives

$$a = a_\pm \to 1, \qquad a'\to 0, \qquad \text{i.e., } \boldsymbol{A}\to 0. \tag{A.16}$$

Also, the limit $N \gg M$ gives

$$\boldsymbol{A}\cdot\hat{\boldsymbol{Q}}\to 0. \tag{A.17}$$

Then we can find

$$\begin{aligned} &(\boldsymbol{I}-\boldsymbol{A}\cdot\hat{\boldsymbol{Q}})^{-1} = \boldsymbol{I} + \boldsymbol{A}\cdot\hat{\boldsymbol{Q}} + \cdots, \\ &\ln\det(\boldsymbol{I}-\boldsymbol{A}\cdot\hat{\boldsymbol{Q}}) = -\mathrm{Tr}\,\boldsymbol{A}\cdot\hat{\boldsymbol{Q}} + \cdots. \end{aligned} \tag{A.18}$$

Then the contribution of the effective partition function to the free energy in the permutation symmetry breaking state is found as

$$\begin{aligned} \frac{1}{n}\overline{\ln Z_{\rm eff}} = {} & M_{\rm T}\,\overline{\ln 2\cosh h(z)} + L\,\overline{\ln 2(\mathrm{e}^{\hat{w}/2}\cosh h_+(z_+) + \mathrm{e}^{-\hat{w}/2}\cosh h_-(z_-))} \\ & + \frac{1}{2}\,\mathrm{Tr}\,\boldsymbol{A}\cdot\hat{\boldsymbol{C}} + \frac{1}{2}\overline{\boldsymbol{t}\cdot\hat{\boldsymbol{Q}}\cdot\boldsymbol{t}} + \overline{\boldsymbol{t}\cdot\hat{\boldsymbol{R}}\cdot\boldsymbol{\Lambda}^0}, \end{aligned} \tag{A.19}$$

where the higher order terms in $\boldsymbol{A}\cdot\hat{\boldsymbol{Q}}$ are neglected. The first two terms can be found as

$$\overline{\ln 2\cosh h(z)} \to \hat{r},$$

$$\overline{\ln 2(\mathrm{e}^{\hat{w}/2}\cosh h_+(z_+) + \mathrm{e}^{-\hat{w}/2}\cosh h_-(z_-))} \to \hat{r}_+ - (1-2p)\left(\hat{r}_- + \tfrac{1}{2}\hat{w}\right). \tag{A.20}$$

The two terms containing $\boldsymbol{t}$ can be simplified as

$$\overline{\boldsymbol{t}\cdot\hat{\boldsymbol{Q}}\cdot\boldsymbol{t}} \to \mathrm{Tr}\,\boldsymbol{H}\cdot\hat{\boldsymbol{Q}}, \qquad \overline{\boldsymbol{t}\cdot\hat{\boldsymbol{R}}\cdot\boldsymbol{\Lambda}^0} \to \mathrm{Tr}\,\boldsymbol{H}\cdot\hat{\boldsymbol{R}} \tag{A.21}$$

where the matrix $\boldsymbol{H}$ is given by

$$\boldsymbol{H} = (1-p)\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & -1\\ 0 & -1 & 1\end{pmatrix} + p\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 1\\ 0 & 1 & 1\end{pmatrix}. \tag{A.22}$$

Therefore, we obtain G_0 in equation (41).

References

[1] Sherrington D and Kirkpatrick S 1975 *Phys. Rev. Lett.* **35** 1792
[2] Parisi G 1980 *J. Phys. A: Math. Gen.* **13** L115
Parisi G 1980 *J. Phys. A: Math. Gen.* **13** 1101
Parisi G 1980 *J. Phys. A: Math. Gen.* **13** 1807
[3] Sompolinsky H, Tishby N and Seung H S 1990 *Phys. Rev. Lett.* **65** 1683
[4] Seung H S, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 6056
[5] Gardner E 1987 *Europhys. Lett.* **4** 481
[6] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
[7] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[8] Krauth W and Mézard M 1989 *J. Phys. France* **50** 3057
[9] Kwon C, Park Y and Oh J-H 1993 *Phys. Rev.* E **47** 3707
[10] Park K and Kwon C 1996 *J. Phys. A: Math. Gen.* **29** 1397
[11] Barkai E, Hansel D and Kanter I 1990 *Phys. Rev. Lett.* **65** 2312
[12] Barkai E, Hansel D and Sompolinsky H 1992 *Phys. Rev.* A **45** 4146
[13] Engel A, Köhler H M, Tschepke F, Vollmayr H and Zippelius A 1992 *Phys. Rev.* A **45** 7590
[14] Kwon C and Oh J-H 1997 *J. Phys. A: Math. Gen.* **30** 6273
[15] Monasson R and Zecchina R 1995 *Phys. Rev. Lett.* **75** 2432
[16] Xiong Y S, Oh J-H and Kwon C 1997 *Phys. Rev.* E **56** 4540
[17] Urbanczik R 1995 *J. Phys. A: Math. Gen.* **28** 7907
[18] Xiong Y S, Kwon C and Oh J-H 1998 *J. Phys. A: Math. Gen.* **31** 7043
[19] Schwarze H and Hertz J 1992 *Europhys. Lett.* **20** 375
[20] Kang K, Oh J-H, Kwon C and Park Y 1993 *Phys. Rev.* E **48** 4805
[21] Kang K, Oh J-H, Kwon C and Park Y 1995 *Neural Networks: The Statistical Mechanics Perspective* ed J-H Oh *et al* (Singapore: World Scientific) p 18
[22] Kang K, Oh J-H, Kwon C and Park Y 1996 *Phys. Rev.* E **54** 1811
[23] Schwarze H, Opper M and Kinzel W 1992 *Phys. Rev.* A **46** R6185
[24] Urbanczik R 1998 *Phys. Rev.* E **58** 2298
[25] Mézard M and Parisi G 1998 *Phys. Rev. Lett.* **82** 747
[26] Mézard M and Parisi G 1999 *J. Chem. Phys.* **111** 1076
[27] Leoenard R D, Angelani L, Parisi G and Ruocco G 2000 *Phys. Rev. Lett.* **84** 6054
[28] Griegera T S, Martin-Mayor V, Parisi G and Verrocchio P 2004 *Phys. Rev.* B **70** 014202