Learning of non-monotonic target functions by Ising perceptrons — learning curve, perfect learning and perfect anti-learning —

Kanako Tooriyama and Tatsuya Uezu*

Graduate School of Human Culture, Nara Women's University, Nara 630-8506

(Received February 3, 2003)

We study supervised learning by Ising perceptrons for a class of non-monotonic teacher input-output relations employing the replica method with the Gibbs algorithm. We consider up to one-step replica symmetry breaking (1RSB) solutions. We find that the behaviour of the learning curve depends on the shape of the teacher input-output relation. In particular, for the class of teacher input-output relations, which represents an extension of the reversed-wedge problem, a peculiar type of learning, perfect anti-learning(PAL), exists, whereas perfect learning (PL) does not exist. Further, we find that the students' vectors cannot tend to the teacher vector, that is, the students cannot learn the teacher completely. The theoretical results are confirmed by the results of Monte Carlo simulations.

KEYWORDS: Ising perceptron, Perfect anti learning, Replica method

§1. Introduction

There have been many studies on the problem of supervised learning by perceptrons[1)]-.[4)] From these studies, it has come to be known that the generalization error ϵ_g is a function of the ratio α of the number of examples p to the number of synaptic weights N, $\alpha \equiv \frac{p}{N}$, when N is sufficiently large. From studies of batch learning and on-line learning, it has been found that the generalization errors $\epsilon_g(\alpha)$ (which are also sometimes referred to as 'learning curves') decay according to power laws and exponential laws. In addition, it is known that in some situations there exists perfect learning (PL), in which the students' weight vectors coincide with the teacher's weight vector at a finite value of α.[5,6)] For example, this phenomenon exists in the case of Ising weights, whose components take the discrete values ± 1.

In recent years, we have been studying supervised learning by perceptrons in the presence of external noise.[7–10)] We believe that such studies are important because they account for the effect of external disturbances, which are unavoidable in any real system, and they allow us to investigate the robustness of learning.

We consider a system consisting of a teacher and students. The teacher's role is to take a finite number of questions and to each give an answer. The students' role is to learn from these examples. The teacher is represented by a fixed 'weight vector' $\boldsymbol{w}^o$ and a probability function $\mathcal{P}$, according to which its answers are generated. $\mathcal{P}$ represents the effect of external 'noise'. Each student is

* uezu@ki-rin.phys.nara-wu.ac.jp

characterized by only a weight vector $\boldsymbol{w}$. Each question is represented by an N-dimensional vector $\boldsymbol{x}$, and to each, the teacher's answer is either +1 or -1. In any given case, the answer that a teacher gives is determined stochastically according to the quantity $\mathcal{P}(u^o)$, which is the probability that the teacher answers +1 to a question $\boldsymbol{x}$, where $u^o = (\boldsymbol{w}^o \cdot \boldsymbol{x})/\sqrt{N}$. To each question, there is a 'correct' answer, which is defined by $\mathrm{sgn}(\boldsymbol{w}^o \cdot \boldsymbol{x})$, but due to the nature of $\mathcal{P}$, the teacher does not always give correct answers. Here, $\mathrm{sgn}(x) = 1$ for $x \geq 0$ and $\mathrm{sgn}(x) = -1$ for $x < 0$. A student's answer to a given question is obtained deterministically from its weight vector [explicitly, as $\mathrm{sng}(\boldsymbol{w} \cdot \boldsymbol{x})$]. A student attempts to learn from the teacher's examples by, in some sense (in a manner made clear below), minimizing the discrepancy between her answers and the teacher's answers to a given set of questions. The goal of the student is to realize the teacher's weight vector $\boldsymbol{w}^o$, in which case, all of her answers would be correct. The precise details of the model are given in the next section. Using the Gibbs algorithm as the learning algorithm, we studied batch learning in systems both with and without noise. In the cases that the conditional probabilites $\mathcal{P}(u^o)$ are monotonic functions withe respect to u^o, we found that for Ising weights, PL is realized in certain situations for both noisy and noiseless systems, even in the thermodynamic limit (i.e. the limit of large N), and obtained the necessary and sufficient conditions for the existence of PL in the general situation.[10] Contrastingly, it has been found that PL does not exist in the thermodynamic limit for a system with spherical weights whose components can take any real number values and are normalized to $\sqrt{N}$.[7,11] PL does not exist in a system with spherical weights because in such a system, students' weight vectors vary continuously, and the probability of a student vector being identical to the teacher's weight vector is zero. In order to elucidate the general conditions under which PL can be realized, it is necessary to consider cases in which the conditional probabilities are given by non-monotonic functions with respect to u^o. The purpose of the present study is to investigate such cases.

In this paper, we consider a class of non-monotonic functions of u^o as the teacher input-output relation (see Eq.(1) and Fig.1). The reversed-wedge problem is one special case of the present class ($k = 0$), and in that case, the input-output relation is realized by a two-layer perceptron. This problem was studied by Kabashima and Inoue in the case of spherical weights,[12] and several types of interesting behaviour of the learning curve were found. For example, they found that there is a discontinuous transition from the 'poor generalization phase' to the 'good generalization phase', and that the power characterizing the decay of the asymptotic learning curve depends on a certain system parameter. Further, they found that asymptotically realized solutions correspond to the minimum generalization error. However, their model does not exhibit PL, because the weight vectors are spherical.

In the present study, using a statistical mechanical method, we investigate the learning curve for a system with Ising weights. As the learning algorithm, we adopt the Gibbs algorithm, introducing

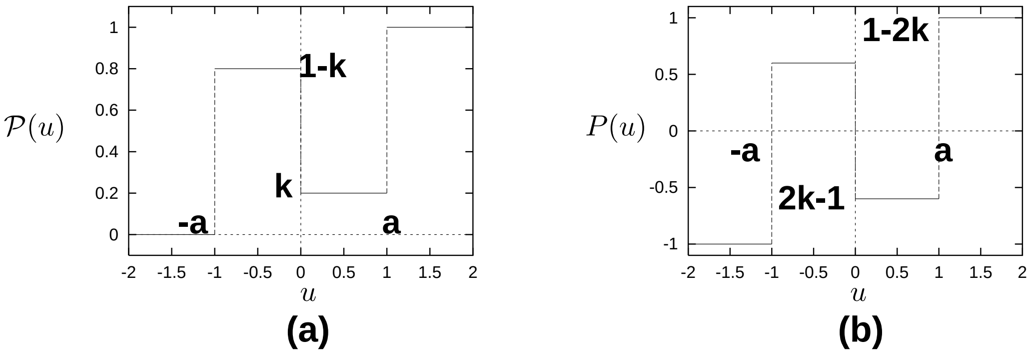

Fig. 1. Nonmonotonic stochastic teacher input-output relation: (a) $\mathcal{P}(u)$;(b) $P(u)$.

an 'energy' characterizing each student vector and a 'temperature' T. In the system we study, PL does not exist, but there does exist a different, somewhat unusual type of learning, perfect anti-learning (PAL), in which at a finite value of α, the students' weight vectors coincide with the opposite of the teacher's weight vector.

In the next section, we give the formulation of the model. In §3, the theoretical results and numerical results are presented and compared first to those for the reversed-wedge problem and then to those for the more general case. In §4, a summary and discussion are given. Some mathematical details are clarified in Appendices A and B.

§2. Formulation

2.1 General formulation

We study supervised learning by perceptrons in the presence of external noise. We consider a stochastic input-output relation, i.e. that of the teacher, between an N-dimensional input vector $\boldsymbol{x}$ and a binary output $r^o \in \{1, -1\}$, which is represented by a conditional probability $p_r(r^o|\boldsymbol{x})$. Input vectors $\boldsymbol{x}$ are normalized as $|\boldsymbol{x}| = \sqrt{N}$, and $p_r(r^o|\boldsymbol{x})$ is defined as a function of the inner product of the input $\boldsymbol{x}$ and the optimal weight $\boldsymbol{w^o}$ (the teacher's weight vector) as

$$
\begin{aligned}
p_r(+1|\boldsymbol{x}) \quad &= \quad \mathcal{P}(u^o) = \frac{1+P(u^o)}{2}, \\
& \qquad u^o = (\boldsymbol{x}\cdot\boldsymbol{w^o})/\sqrt{N}.
\end{aligned}
\tag{1}
$$

Here, $\mathcal{P}(u^o) \in [0,1]$ represents the probability that u^o will yield an output of $+1$. The form of $\mathcal{P}(u^o)$ used in this work, characterized by two parameters, is dipicted in Fig.1 and defiend explicitely in Eq.(3) below. Unlike the teacher, whose output is stochastic, the students are pure perceptrons;

that is, an output r by a student with weight vector $\boldsymbol{w}$ resulting from an input $\boldsymbol{x}$ is given by

$$r = \mathrm{sgn}(u),$$

where $u \equiv (\boldsymbol{x} \cdot \boldsymbol{w})/\sqrt{N}$. In this paper, as the teacher and students, we consider Ising perceptrons, meaning that the components of their weight vectors $\boldsymbol{w^o}$ and $\boldsymbol{w}$ take the discrete values ± 1. Further, $P(-u) = -P(u)$ is stipulated for simplicity.

The learning procedure begins with p input vectors $\boldsymbol{x}^1, \boldsymbol{x}^2, \cdots, \boldsymbol{x}^p$ chosen randomly and independently from a uniform distribution on a hypersphere of radius $\sqrt{N}$ centered at the origin in an N-dimensional Euclidean space. From each such $\boldsymbol{x}^\mu$, an output $r^o_\mu (= 1 \text{ or } = -1)$ is obtained with the conditional probability $p_r(r^o_\mu | \boldsymbol{x}^\mu)$. In this way, a set of p examples of input and teacher-generated output, $\xi_p = \{(\boldsymbol{x}^1, r^o_1), (\boldsymbol{x}^2, r^o_2), ..., (\boldsymbol{x}^p, r^o_p)\}$, which are to act as the lessons from which the students learn, are obtained. For a given realization of examples ξ_p, we define the 'energy' $E[w, \xi_p]$ of a student with weight vector $\boldsymbol{w}$ as the number of false predictions,

$$E[w, \xi_p] = \sum_{\mu=1}^{p} \Theta(-r^o_\mu u_\mu), \quad u_\mu \equiv (\boldsymbol{x}^\mu \cdot \boldsymbol{w})/\sqrt{N}, \tag{2}$$

where $\Theta(x) = 1$ for $x \geq 0$ and $\Theta(x) = 0$ for $x < 0$. The learning performance of a given student is represented by the generalization error ϵ_g, which is a false prediction of the student, and is calculated as

$$\epsilon_g \equiv \langle \Theta(-r^o r) \rangle = \frac{1}{2} - \int_0^\infty DyP(y) + 2\int_0^\infty DyP(y)H(\frac{Ry}{\sqrt{1-R^2}}).$$

Here, $\langle \cdots \rangle$ represents the average over inputs, which are the same for the teacher and the student, and over outputs r^o. $Dy = \exp(-y^2/2)dy/\sqrt{2\pi}$, $H(x) = \int_x^\infty Dy$, and R is the overlap of the optimal weight vector and the weight vector of the student, $R = (\boldsymbol{w^o} \cdot \boldsymbol{w})/N$. In this study, we adopt the Gibbs algorithm with temperature T as the learning algorithm. Then, from the 'energy' defined by (2), the partition function Z with inverse temperature β is given by

$$Z = \mathrm{Tr}_w e^{-\beta E[w,\xi_p]} = \mathrm{Tr}_w \Pi^p_{\mu=1}[e^{-\beta} + (1 - e^{-\beta})\Theta(r^o_\mu u_\mu)],$$

where Tr_w represents the summation over all possible $\boldsymbol{w}$. The average free energy f per synaptic weight is calculated using the replica method as

$$-\beta N f = \langle \ln Z \rangle_{\xi_p, w^o} = \lim_{n\to 0} \frac{1}{n}(\langle Z^n \rangle_{\xi_p, w^o} - 1),$$

where $\langle \cdots \rangle_{\xi_p, w^o}$ denotes the average over quenched variables.

The quantity $\langle Z^n \rangle_{\xi_p, w^o}$ is a function of several replica order parameters, namely, the overlap of student weight vectors, $q^{ab} = \frac{(\boldsymbol{w}^a \cdot \boldsymbol{w}^b)}{N}$, its conjugate $\hat{q}^{ab}$, the overlap of the student weight vector and the optimal weight vector, $R^a = \frac{(\boldsymbol{w^o} \cdot \boldsymbol{w}^a)}{N}$, and its conjugate $\hat{R}^a$. Here, a and b are replica indices. (See Ref.[10] for details.)

2.2 Non-monotonic functions

In this paper, we consider mainly the case in which the optimal weight $\boldsymbol{w^o}$, the teacher's weight vector, yields an incorrect output for an input vector $\boldsymbol{x}$ with probability $1-k$ when $|u^o| \le a$, where $0 \le k < 1/2$.[13] More precisely, in the present work, the probability $\mathcal{P}(u^o)$ that the teacher answers $+1$ is given by the following (see Fig.1):

$$\mathcal{P}(u^o) = \begin{cases} 0 & (\text{ for } u^o < -a) \\ 1-k & (\text{ for } -a \le u^o < 0), \\ k & (\text{ for } 0 \le u^o < a) \\ 1 & (\text{ for } a \le u^o) \end{cases}$$

where the value a defines the interval of values of u^o, $[-a, a]$, over which there is a finite probability that the teacher will return the wrong answer. (The case $k = 0$ corresponds to the reversed-wedge problem.) With $\mathcal{P}$ as given by (3), the generalization error becomes

$$\epsilon_g(R) = 1 - k - 2(1-k)H(a) + \frac{1}{\pi}\cos^{-1} R - 4(1-k)\int_0^a DyH(\frac{Ry}{\sqrt{1-R^2}}), \quad (3)$$

which satisfied $\epsilon_g(-R) = 1 - \epsilon_g(R)$. In this study, first we consider the replica symmetric (RS) solution. When T is sufficiently small and α is sufficiently large, the entropy of the RS solution becomes negative and the RS solution becomes invalid. Then, next, we consider the one-step replica symmetry breaking (1RSB) solution.

2.3 RS solution

The replica symmetry ansatz is that no quantity depends on replica indices. Threfore, we replace $q^{ab}, \hat{q}^{ab}, R^a$ and $\hat{R}^a$ by the new quantities $q, \hat{q}, R$ and $\hat{R}$. Thus, the free energy under the replica symmetry ansatz, f_{RS}, is calculated as

$$-\beta f_{RS}(q, \hat{q}, R, \hat{R}, \beta) = -\frac{\hat{q}}{2}(1-q) - R\hat{R} + \alpha K + I, \quad (4)$$

where

$$K \equiv \int Dy 2\mathcal{P}(y) \int Du \ln \tilde{H}(\frac{\sqrt{q-R^2}u - Ry}{\sqrt{1-q}}) = \int Du \ln \tilde{H}(u/Q)E(u/Q), \quad (5)$$

$$I \equiv \int Dt \ln[2\cosh(\sqrt{\hat{q}}t + \hat{R})], \quad (6)$$

with

$$E(u) = 2[1 - (2k-1)H(v) + (k-1)\{H(v + \frac{a}{\xi}) + H(v - \frac{a}{\xi})\}], \quad (7)$$

$$\tilde{H}(u) \equiv e^{-\beta} + (1 - e^{-\beta})H(u)$$

$$\xi = \sqrt{1 - \frac{R^2}{q}}, \quad Q = \sqrt{\frac{1-q}{q}}, v = -\frac{R}{\sqrt{q}\chi}u, \quad \chi = \frac{\xi}{Q}.$$

$q, \hat{q}, R$ and $\hat{R}$ are determined as the saddle points of f_{RS}. Then, the saddle point equations (SPE) are

$$
\begin{aligned}
q &= \int Du \tanh^2(\sqrt{\hat{q}}u + \hat{R}), && (8)\\
R &= \int Du \tanh(\sqrt{\hat{q}}u + \hat{R}), && (9)\\
\hat{q} &= \frac{\alpha}{1-q} \int Du(\tilde{\varphi}(u/Q))^2 E(u/Q), && (10)\\
\hat{R} &= -\frac{\alpha}{Q\sqrt{q-R^2}} \int Du \tilde{\varphi}(u/Q) w(u/Q), && (11)
\end{aligned}
$$

where

$$
\begin{aligned}
& w(u) = 2[(1-k)\{h(v+\frac{a}{\xi}) + h(v-\frac{a}{\xi})\} + (2k-1)h(v)] = w(-u),\\
& \tilde{\varphi}(u) = \frac{\tilde{H}'(u)}{\tilde{H}(u)},\ h(u) = e^{-u^2/2}/\sqrt{2\pi}.
\end{aligned}
$$

The entropy s_{RS} per synapse and the energy (training error) e_t per synapse are given by

$$
\begin{aligned}
s_{RS} &= -\beta f_{RS} - \alpha\beta e^{-\beta} J, && (12)\\
e_t &= -\alpha e^{-\beta} J, && (13)
\end{aligned}
$$

where

$$
J = \int Dy 2\mathcal{P}(y) \int Du \frac{H(\frac{\sqrt{q-R^2}u-Ry}{\sqrt{1-q}}) - 1}{\tilde{H}(\frac{\sqrt{q-R^2}u-Ry}{\sqrt{1-q}})}.
$$

2.4 1RSB solution

When T is sufficiently small and α is sufficiently large, the entropy of the RS solution becomes negative. This implies that the replica symmetry ansatz is no longer valid in this situation, therefore, we have to consider the breaking of the replica symmetry. Presently, the only successful ansatz in the case of replica symmetry breaking is that given by Parisi.[14] We adopt the one-step version of this replica symmetry breaking ansatz of Parisi, because it is the simplest.

For the 1RSB solution, the matrix q^{ab} is divided into $(n/m)^2$ small matrices of dimension $m \times m$. All components of each of the off-diagonal such matrices are q_0, and all components of each of the diagonal matrices are q_1, except for the diagonal components, which are 0. The structure of the matrix $\hat{q}^{ab}$ is idential, with the values $\hat{q}_0$ and $\hat{q}_1$ replacing q_0 and q_1. Further, we stipulate

$R^a = R$ and $\hat{R}^a = \hat{R}$. Then, the free energy under the 1RSB ansatz, f_{1RSB}, is given by

$$\begin{aligned}
&-\beta f_{1RSB}(q_0, \hat{q_0}, q_1, \hat{q_1}, R, \hat{R}, m, \beta) \\
&= -\frac{\hat{q_1}}{2}(1-q_1) + \frac{m}{2}(\hat{q_0}q_0 - \hat{q_1}q_1) - R\hat{R} \\
&\quad + \frac{\alpha}{m}\int Dy 2\mathcal{P}(y)\int Dz_0 \ln \int Dz_1[\tilde{H}(\frac{\sqrt{q_0 - R^2}z_0 + \sqrt{q_1 - q_0}z_1 - Ry}{\sqrt{1-q_1}})]^m \\
&\quad + \frac{1}{m}\int Dz_0 \ln \int Dz_1[2\cosh(\sqrt{\hat{q_0}}z_0 + \sqrt{\hat{q_1} - \hat{q_0}}z_1 + \hat{R})]^m.
\end{aligned} \tag{14}$$

Next, following Krauth and Mézard,[15)] we take the limits $q_1 \to 1$ and $\hat{q_1} \to \infty$. Then we obtain

$$f_{1RSB}(q_0, \hat{q_0}, q_1 = 1, \hat{q_1} = \infty, R, \hat{R}, m, \beta) = f_{RS}(q_0, m^2\hat{q_0}, R, m\hat{R}, m\beta). \tag{15}$$

By differentiating f_{1RSB} with respect to $q_0, \hat{q_0}, R$ and $\hat{R}$, we obtain a set of saddle point equations identical to (9)-(12), with q_0 replacing q, $m^2\hat{q_0}$ replacing $\hat{q}$, R remaining as R, $m\hat{R}$ replacing $\hat{R}$ and $m\beta$ replacing β. Further, by differentiating f_{1RSB} with respect to m, we obtain $s_{RS}(q_0, m^2\hat{q_0}, R, m\hat{R}, m\beta) = 0$, where s_{RS} is the entropy for the RS solution. Let us denote the solutions of these coupled equations by $q_0 = q_c, m^2\hat{q_0} = \hat{q}_c, R = R_c, m\hat{R} = \hat{R}_c$ and $m\beta = \beta_c$. That is, $q_c, \hat{q}_c, R_c$ and $\hat{R}_c$ are RS solutions at the temperature $T_c = \frac{1}{\beta_c}$, at which the RS entropy s_{RS} is 0. Then, the 1RSB solutions are given by $q_0 = q_c, \hat{q_0} = (\frac{\beta}{\beta_c})^2\hat{q}_c,\ \ R = R_c,\ \hat{R} = \frac{\beta}{\beta_c}\hat{R}_c$ and $m = \frac{\beta_c}{\beta}$. Since $m \le 1$, these solutions are valid for $T < T_c$. Thus, the 1RSB solutions for $T < T_c$ are expressed in terms of the RS solutions at $T = T_c$.

In the next section, we first consider the case $k = 0$, i.e. the reversed-wedge problem, and then we consider the case $k > 0$.

§3. Theoretical and numerical analyses of the case with a non-monotonic teacher input-output relation

As proved in Appendix A, for the class of non-monotonic teacher input-output relations, perfect learning (PL) does not exist, but perfect anti-learning (PAL) always exists.

3.1 The reversed-wedge problem ($k = 0$)

In this subsection, we give theoretical results and compare them with the results of Monte Carlo simulations(MCS) for the reversed-wedge problem. First, we explain the numerical method and discuss its accuracy.

We carried out Monte Carlo similations using the simulated annealing method. First, in order to check the accuracy of the MCS, we compare the overlap $R_{MCS}(N)$ obtained using the MCS with the overlap $R_{enn}(N)$ obtained using the enumeration method for a system size N, for several values of α, the largest of which being 6. We found that the ratio $|(R_{MCS}(10) - R_{enn}(10))/R_{enn}(10)|$ is small, $\simeq 0.05$. Then, we carried out MCS for several values of the system size N, N=15, 25, 35 and

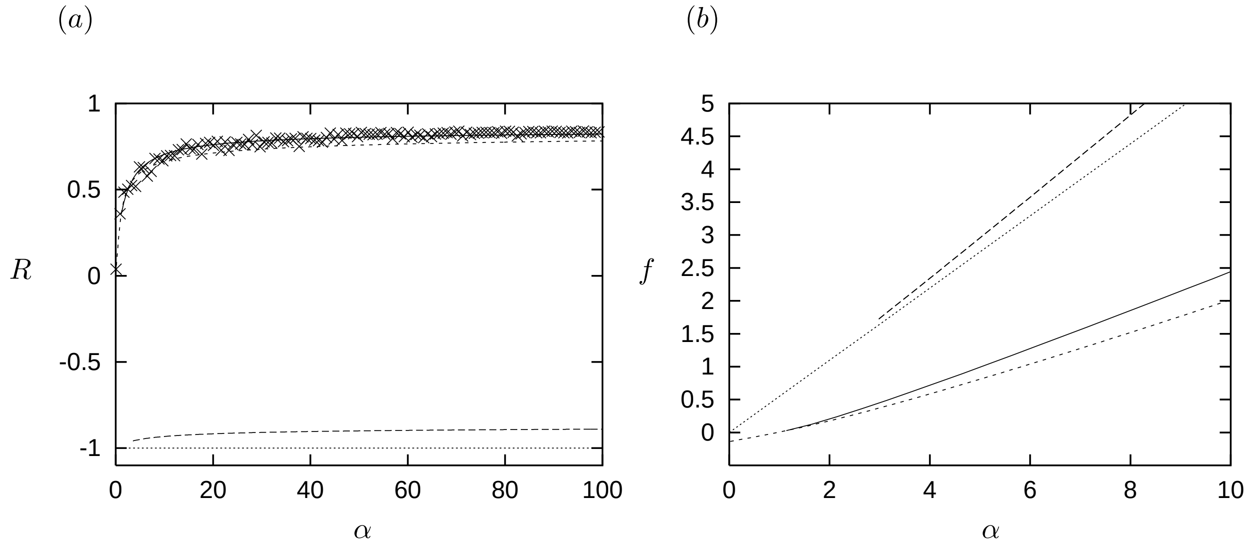

Fig. 2. Region (A). The α dependence of R and the free energy f for $k = 0, a = 0.6$ and $T = 0.2$. The solid curve represents the 1RSB solution in branch I, the long-dashed curve represents the 1RSB solution in branch II, the short-dashed curve represents the RS solution in branch I, the dotted curve represents the PAL solution, and $\times$s represent the numerical results for $N = 15$. We omit the RS solution in branch II because that this solution is irrelevant. Since the solid curve agrees with numerical data and is covered by $\times$s, it is difficult to identify it.

45, using various values for a and T and for several values of α, the largest of which being 6. From these calculations, we found that for $N, N' \geq 25$, the ratio $|(R_{MCS}(N) - R_{MCS}(N'))/R_{MCS}(N)|$ is about $0.02 - 0.2$, and that this ratio is a decreasing function of T. Hence, it seems that results obtained for $N = 25$ are sufficient to surmise the $N = \infty$ behaviour, at least for $\alpha \leq 6$ and for large T.

Next, we give our theoretical results and compare them with the numerical results. There are three regions of values of a in which the behaviour of the learning curve differs qualitatively: (A) $0 < a < a_{sp}(\simeq 0.81)$, (B) $a_{sp} \leq a < a_c(= \sqrt{2 \ln 2} \simeq 1.18)$, and (C) $a_c < a$. This classification does not depend on temperature. a_{sp} and a_c are defined in terms of $\epsilon_g(R)$ as follows. a_{sp} is defined by the equation $\epsilon_g(R_+) = \epsilon_g(-1)$, and a_c is defined by the equations $\frac{\partial \epsilon_g}{\partial R}|_{R=0} = 0$ and $\frac{\partial^2 \epsilon_g}{\partial R^2}|_{R=0} = 0$, which give $a_c = \sqrt{2 \ln 2}$.

Regions (A) and (B):

First, we describe the behaviour common to regions (A) and (B). When α is small, there exist two RS solutions and the PAL solution of the SPE. Two 1RSB solutions appear when α exceeds some value, say α_c, at which the RS entropy becomes 0. For both the RS and 1RSB cases, the two solutions correpond to two branches, which we refer to as I and II. As α tends to infinity, the 1RSB solution in branch I converges to a positive constant $R_\infty^{(I)}(< 1)$, while the 1RSB solution in branch II converges to a negative constant $R_\infty^{(II)}(> -1)$ (see Fig.2(a)). The free energy of the RS solution in branch I is smaller than that of the PAL solution for $\alpha \leq \alpha_c$, and thus it is expected that in the MCS, for $\alpha \leq \alpha_c$, the RS solution in branch I will appear. The actual results of our MCS agree

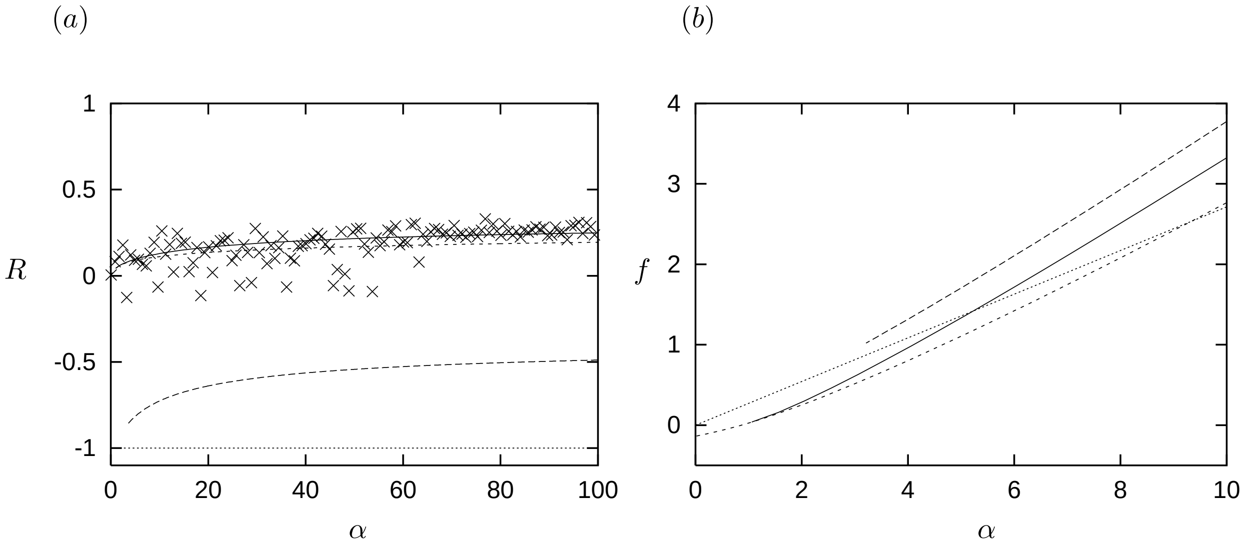

Fig. 3. Region (B). The α dependence of R and the free energy f for $k = 0, a = 1.1$ and $T = 0.2$. The solid curve represents the 1RSB solution in branch I, the long-dashed curve represents the 1RSB solution in branch II, the short-dashed curve represents the RS solution in branch I, the dotted curve represents the PAL solution, and $\times$ s represent the numerical results for $N = 25$. We omit the RS solution in branch II because that this solution is irrelevant.

with the RS solution in branch I for $\alpha \leq \alpha_c$.

The behaviour for $\alpha_c < \alpha$ is different in the two regions (A) and (B). We now describe the behaviour in each. First, we consider region (A). For $\alpha_c < \alpha$, the free energy of the 1RSB solution in branch I, f^I_{1RSB}, is smallest among those of the existing solutions (see Fig.2(b)). Therefore, it is expected that when MCS are carried out, for $\alpha > \alpha_c$ the 1RSB solution in branch I will appear. The actual results of our MCS show that this is indeed the case (see Fig.2(a)). Next, we consider region (B). For $\alpha > \alpha_c$, f^I_{1RSB} is smallest up to some value of α, α_{cr}, while above α_{cr}, the free energy of the PAL solution, f_{PAL}, becomes smallest (see Fig.3(b)). Thus it is expected that the 1RSB solution will appear for $\alpha \in [\alpha_c, \alpha_{cr}]$ and the PAL solution will appear for $\alpha > \alpha_{cr}$ in the MCS. The actual result obtained from the MCS is displayed in Fig.3(a). From this, we find that even after α exceeds α_{cr}, the system remains in the 1RSB state in branch I. This seems to result from the fact that we employed a simulated annealing method to obtain the equilibrium states, and for this reason, the system might become trapped in a local minimum of the free energy. To check this conjecture, we carried out further MCS and found that if the initial state is near the PAL state, the system asymptotically approaches the PAL state. Therefore, there exist two local minima, the 1RSB state in branch I and the PAL state.

The types of behaviour observed in the regions (A) and (B) can be understood as follows. From numerical calculations using the SPE, we found that the temperature $T_c(\alpha)$ at which the 1RSB solution appears tends to infinity as $\alpha \to \infty$. In the case $T \gg 1$ ($\beta \ll 1$), we can derive the relation $f_{1RSB} \simeq \alpha\epsilon_g$ (see Appendix B). Therefore, in the asymptotic region of α (i.e., $\alpha \gg 1$), we only have

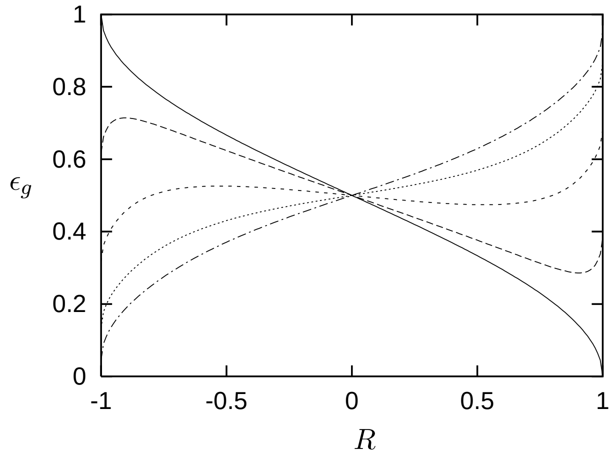

Fig. 4. The R dependence of the generalization error $\epsilon_g(R)$ for several values of a in the reversed-wedge problem. Solid curve corresponds to $a = 0$. The other curves correspond to, in order from the bottom on the right-hand side of the figure, $a = 0, 0.5, 1.0, 1.5, 2.0$. Note that $0.5 < a_{sp}(\simeq 0.81) < 1.0 < a_c(\simeq 1.18) < 1.5 < 2.0$.

to study ϵ_g to investigate the stable states of the system. As shown in Fig.4, for $0 < a < a_c$, ϵ_g has two local minima, one is at a positive value of R, say $R_+(0 < R_+ < 1)$, and the other is at $R = -1$ (PAL). For $0 < a < a_{sp}$, $R = R_+$ is the global minimum, and for $a_{sp} < a \leq a_c$, $R = -1$ is the global minimum. We have also confirmed that R_+ and $R_\infty^{(I)}$ coincide. Thus, the results obtained from our MCS can be understood completely in terms of the free energy. That is, the final states obtained from the MCS in the regions (A) and (B) are the solutions with $R = R_+$, and the state in the region (A) corresponds to the global minimum of f_{1RSB}, while that in the region (B) corresponds to the local minimum of f_{1RSB}.

Region (C):

In this region, $\epsilon_g(R)$ is a decreasing function of R and $\epsilon_g(-1)$ is the minimum. Thus, it is expected that at some value of α, α_{max}, the RS solutions in branches I and II merge and disappear. [See Fig.5(a).]

First, let us consider the case that T is sufficently large that the RS entropy in branch I does not become negative and the 1RSB solutions do not exist. Among the existing solutions, the free energy of the RS solution in branch I, f_{RS}^I, takes the smallest value below some value of α, α_{tr}, and that of the PAL solution becomes smallest above α_{tr}. Therefore, for $\alpha < \alpha_{tr}$, we predict that the RS solution in branch I will appear, and for $\alpha_{tr} < \alpha$, the PAL solution will appear in the MCS. In Fig. 5(a), both the numerical result obtained from the MCS for $N = 25$ and the theoretical results are displayed. It is seen that the numerical data deviate from the theoretical curve at some value of α near 2.0, less than α_{tr} ($\simeq 3$). We found that as N is increased, the length of the interval of

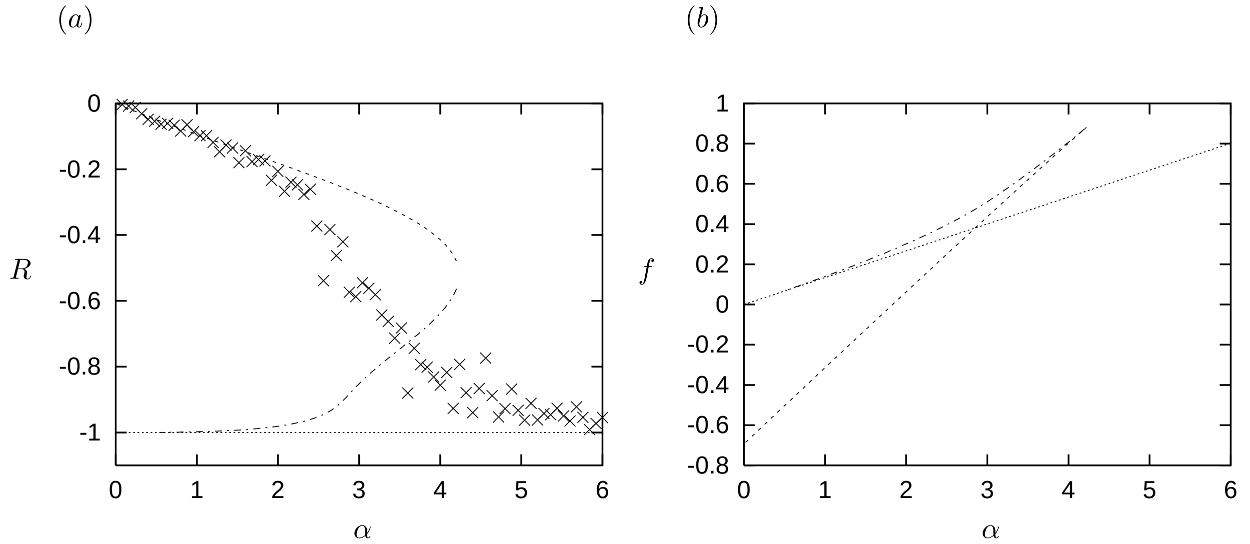

Fig. 5. Region (C). The α dependence of R and the free energy f for $k = 0,\ a = 1.5$, and $T = 1.0$. The dashed curve represents the RS solution in branch I, the dashed-dotted curve represents the RS solution in branch II, the dotted curve represents the PAL solution, and $\times$s represent the numerical results for $N = 25$. There is no RSB solution in this case.

α values over which the MCS results agree with the theoretical results increases. This leads us to believe that the deviation of the numerical results from the theoretical results is a finite-size effect. In the case discussed next, we investigate this effect in detail.

Next, let us consider the case that T is not sufficiently large for the RS entropy in branch I to become negative. Let α_{max}^{1RSB} be the value of α where 1RSB solutions in the branches I and II merge and disappear. Then the 1RSB solutions exist in the interval $\alpha \in [\alpha_c, \alpha_{max}^{1RSB}]$. [See Fig.6(a).] As α is increased, the solution with the smallest free energy changes as follows: the RS solution in branch I$\rightarrow$ the 1RSB solution in branch I$\rightarrow$ the PAL solution. Let α_{tr}^{1RSB} be the value of α where $f_{1RSB}^{I} = f_{PAL}$. Thus, the theoretical result predicts that a transition from the 1RSB solution in branch I to the PAL solution will take place at $\alpha = \alpha_{tr}^{1RSB}$.

In Fig.6, for $a = 1.5$, the theoretical result and numerical result obtained from the MCS for $N = 25$ are displayed for the case of small T $(= 0.2)$. We see that the numerical result deviates from the theoretical prediction at some value of α, α^*. (Of course, in practice, α^* can only be determined approximately.) For the parameter values used in Fig.6, $\alpha^* \simeq 1.3$, while $\alpha_{tr}^{1RSB} \simeq 2.2$. To see the system size dependence of α^*, we carried out MCS for larger values of N. As seen from Fig.7, the length of the interval $[0, \alpha^*]$ over which the results from the MCS agrees with the theoretical curve increases as N is increased. It is thus reasonable to conclude that the observed disagreement between the numerical results and the theoretical results above α^* is a finite-size effect.

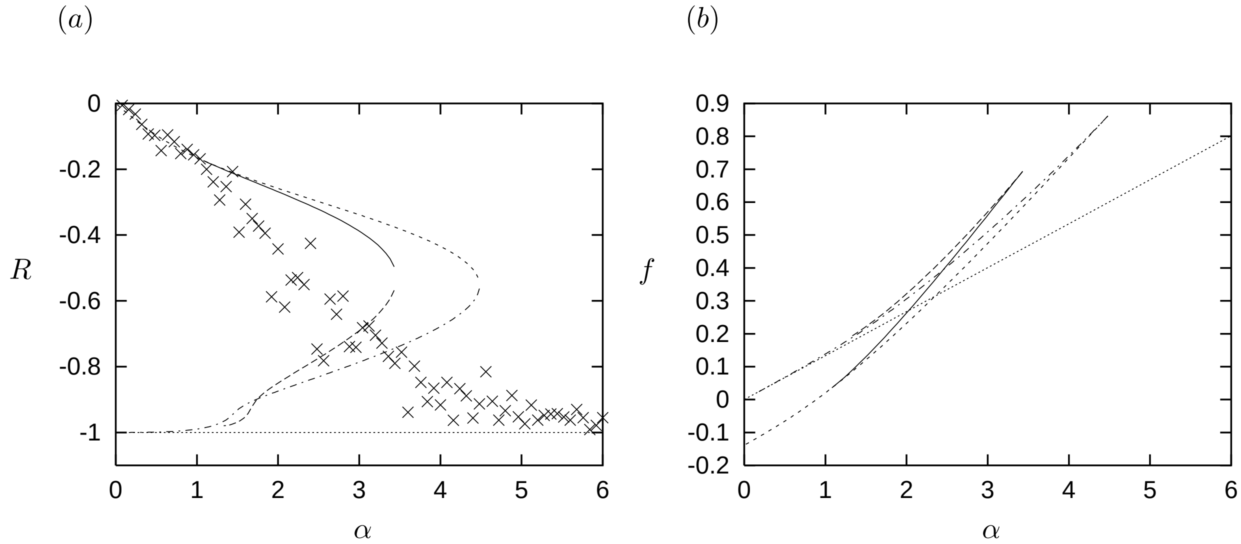

Fig. 6. Region (C). The α dependence of R and the free energy f for $k = 0, a = 1.5$ and $T = 0.2$. The solid curve represents the 1RSB solution in branch I, the long-dashed curve represents the 1RSB solution in branch II, the short-dashed curve represents the RS solution in branch I, the dashed-dotted curve represents the RS solution in branch II, the dotted curve represents the PAL solution, and $\times$s represent the numerical results for $N = 25$.

3.2 The case $0 < k < 1/2$

In this subsection, we treat the non-monotonic functions with $0 < k < 1/2$. We study learning behaviour exhibited by the system for various values of the two parameters a and k, and T fixed to 0.2. By numerically solving the SPE, we found that, as in the case $k = 0$, if the RS solutions exist for any α, their entropies s_{RS} become negative for large α. Then, the 1RSB solutions bifurcate from the RS solutions at the value of α where s_{RS} become 0. The critical temperature $T_c(\alpha)$ at which the 1RSB solution appears tends to infinity as $\alpha \to \infty$. Thus, as in the reversed-wedge problem, in the present case we can derive the relation $f_{1RSB} \simeq \alpha\epsilon_g$ (see Appendix B). For this reason, if solutions exist for any value of α, to investigate stable solutions, we only have to study the behaviour of the generalization error. We now proceed with this.

First, we consider the case that k is fixed. In this case, for $a < a_c(k)$, there exist two extrema of ϵ_g, $R_\pm = \pm\sqrt{1 - (\frac{a}{a_c(k)})^2}$, where $a_c(k) \equiv \sqrt{2(\ln(\frac{2(1-k)}{1-2k}))}$. For $a_c(k) \leq a$, ϵ_g is a strictly increasing function of R. There exists another value of a at which the behaviour of the system changes, $a_{sp}(k)$. This is the value at which $\epsilon_g(R_+) = \epsilon_g(-1)$, and it satisfies the relation $0 < a_{sp}(k) < a_c(k)$. Then, the minimum of ϵ_g depends on the value of a as follows.

(1) $\underline{0 < a < a_{sp}(k)}$

$\epsilon_g(R_+)$ is the global minimum, and $\epsilon_g(-1)$ is a local minimum.

(2) $\underline{a_{sp}(k) < a < a_c(k)}$

$\epsilon_g(-1)$ is the global minimum, and $\epsilon_g(R_+)$ is a local minimum.

(3) $a_c(k) \leq a$

$\epsilon_g(R)$ is a strictly increasing function of R, and thus $\epsilon_g(-1)$ is the global minimum.

Next, we allow k to change with a fixed. We define $k_c(a)$ as the inverse function of $a_c(k)$, $k_c(a) = \frac{e^{a^2/2}-2}{2(e^{a^2/2}-1)}$.

In this case, the minimum of ϵ_g depends on the values of a and k as follows.

(1) $\sqrt{2\ln 2} < a$

a. $0 \leq k \leq k_c$: $\epsilon_g(R)$ is a strictly increasing function of R, and thus $\epsilon_g(-1)$ is the global minimum.

b. $k_c < k < k_{sp}(a)$: $\epsilon_g(-1)$ is the global minimum, and $\epsilon_g(R_+)$ is a local minimum.

c. $k_{sp}(a) < k < 1/2$: $\epsilon_g(R_+)$ is the global minimum, and $\epsilon_g(-1)$ is a local minimum.

For $a = 1.5$, $k_{sp} \simeq 0.394$ and $k_c \simeq 0.260$.

(2) $a < \sqrt{2\ln 2}$.

$0 \leq k < 1/2$: $\epsilon_g(R)$ has two minimal, at $R = R_+$ and $R = -1$, and two maximal, at $R = R_-$ and $R = 1$. Whether or not the spinodal point k_{sp} exists depends on a.

We have carried out numerical calculations using the SPE in each of the above cases, and in each case, we obtained results similar to those for the reversed-wedge problem. For this reason, we forgo detailed descriptions of the transitions. As an example, we simply give the final solutions obtained from the MCS as α is increased when k is changed for $\sqrt{2\ln 2} < a$.

(1) For $0 < k \leq k_c$, the solution realized is that of PAL.

(2) For $k_c < k \leq k_{sp}$, $R \to R_+$ as $\alpha \to \infty$. This would seem to imply that $R = R_+$ corresponds to the global minimum, but we have found that when the initial state is near the PAL state, the system approaches this state as time increases. Thus, there are two local minima. From theoretical consideration, we know that the solution with $R = R_+$ corresponds to a local minimum of f_{1RSB}. It thus appears that the $R \to R_+$ result is due to a finite size effect and that the PAL state corresponds to the global minimum.

(3) For $k_{sp} < k < 1/2, R \to R_+$ as $\alpha \to \infty$. In this case, this is a 'true' result, not due to a finite size effect.

Now, we compare the results for $k > 0$ with those for $k = 0$. For this purpose, we carried out MCS with $k = 0.2, T = 0.2$ and $N = 15$ and considered several values of a. We found that when $a = 0.2, 0.6$ and 0.9, as $\alpha \to \infty$, R tends to R_+, but when $a = 1.5$, PAL is realized. Considering the cases $k = 0$ and $k > 0$ with equal values of a and T, if $R \to R_+$ as $\alpha \to \infty$ in both cases, then the value of R_+ is greater than for the case $k > 0$ than the case $k = 0$. Next, consider the case that the system jumps into the PAL state as α increases. Then, let α^* represent the value of α at which the data generated by the MCS begin to deviate from the theoretical curve. We find that the value of α^* for $k > 0$ is larger than that for $k = 0$. These results are quite reasonable, because they imply that the teacher gives correct outputs more frequently and therefore the students learn more quickly for $k > 0$ than for $k = 0$.

§4. Summary and discussion

We have studied supervised learning by Ising perceptrons governed by a class of non-monotonic teacher input-output relations using a statistical mechanical method. The teacher input-output relation is a generalization of that employed in the reversed-wedge problem. This relation is characterized by two parameters, a and k: The teacher gives a wrong output to an input $\boldsymbol{x}$ with probability $1-k$ when u $(= (\boldsymbol{w^o} \cdot \boldsymbol{x})/\sqrt{N})$ is in $(-a, a)$.

As the learning algorithm, we used the Gibbs algorithm, introducing an 'energy' of each student and temperature T. Then, employing the replica method, we obtained the following results for $0 \leq k < 1/2$.

(1) In the system we study, PL does not exist, but PAL does exist.

(2) For the case $k = 0$, the behaviour of the learning curve can be classified into three cases, depending on the value of parameter a:

(A) $0 < a < a_{sp}$

In this range, for large α, the entropy of the RS solution in branch I becomes negative, and a 1RSB solution in branch I appears. As $\alpha \to \infty$, $R(\alpha) \to R_+$ for the 1RSB solution, where in this case, R_+ (which always satisfies $0 < R_+ < 1$) is the value of R at which ϵ_g and the free energy are minimal.

(B) $a_{sp} \leq a < a_c$

In this range, for large α again, the entropy of the RS solution in branch I becomes negative, and a 1RSB solution in branch I appears. Also, again here, as $\alpha \to \infty$, $R(\alpha) \to R_+$ for the 1RSB solution, where again in this case R_+ corresponds to a local minimum of ϵ_g and the free energy. In this case, the state corresponding to the minimum of the free energy is that of PAL.

(C) $a_c < a$.

In this region, all solutions other than the PAL solution are continued to the finite range of α values. When T is large, only RS solutions exist, and when T is small, 1RSB solutions exist. There is a first-order transition from the RS solution in branch I or the 1RSB solution in branch I to the PAL solution as α is increased.

(3) For the case $0 < k < \frac{1}{2}$, results similar to those of the case $k = 0$ are obtained. Further, we found that as k increases, in the case that $R \to R_+$, the value of R_+ increases, and when the system jumps into the PAL state as α increases, the theoretical value of α_{max}, at which the transition takes place, and the value $\alpha^*(< \alpha_{max})$, at which the results obtained from the MCS begin to deviate from the theoretical curve, both increase. The increase of these values reflects very natural behaviour of the model: As the number of mistakes made by the teacher decreases, it becomes easier for the students to learn.

From these results, we note that the asymptotic behaviour of learning curves depends on a and k as $\alpha \to \infty$. That is, for small a, $R \to R_+$ and for large a, $R \to -1$(PAL) when k is fixed. On the other hand, for small k, $R \to -1$ and for large k, $R \to R_+$ when a is fixed. Therefore, we conclude that, unlike in the case of monotonic teacher input-output relations, the nature of the learning exhibited by the students is not determined by the properties of $\mathcal{P}(u)$ around the decision boundary $u = 0$ but, rather, depends on the shape of $\mathcal{P}(u)$ over the entire range of values of u.

Kabashima and Inoue studied the reversed-wedge problem in the case of spherical weights using the replica method within the RS ansatz.[12)] They studied the learning of non-monotonic rule by simple perceptrons. The rule they treated is

$$T_a(u) = \mathrm{sgn}(-u - a)\mathrm{sgn}(u)\mathrm{sgn}(u - a), \tag{16}$$

which is related to the function $P(u)$ with $k = 0$ in this paper as

$$T_a(u) = -P(u). \tag{17}$$

They obtained the following asymptotic behaviour as $\alpha \to \infty$. For large a, $R \to 1$ and for small a, $R \to R_-(-1 < R_- < 0)$. They also obtained a discontinuous transition from the 'poor generalization phase' to the 'good generalization phase' for intermediate values of a. Now, let us comper their results with ours.

First, in our case, R does not tend to 1, that is, the students cannot learn the teacher completely. This is because the probability that the teacher returns the correct answer is 0 around the decision boundary ($\mathcal{P}(u) = 0$ for $|u| < a$). On the other hand, in their case, for large a, $R \to 1$ as $\alpha \to \infty$, that is, the students can learn the teacher completly. This is because $\mathcal{P}(u) = \frac{1+T_a(u)}{2}$, which is interpreted as the probability that the teacher returns the correct answer, behaves as $\mathcal{P}(u) = 1$ for $|u| < a$. Second, as for the asymptotic behaviour as $\alpha \to \infty$, asymptotic solutions are determined by the minimum of the generalization error in the both cases. Let us denote the generalization error for $T_a(u)$ and that for $P(u)$ with $k = 0$ treated here by $\epsilon_g^{(KI)}(a, R)$ and $\epsilon_g(a, R)$, respectively. Then, from the relation (17), we derive the relation,

$$\epsilon_g^{(KI)}(a, R) = 1 - \epsilon_g(a, R) = \epsilon_g(a, -R). \tag{18}$$

Since in the interval $(-a, a)$, $\mathcal{P}(u) = 0$ for our case and $\mathcal{P}(u) = 1$ for their case, large(small) a in our case correponds to small(large) a in their case. From this fact and the relation (18), we find that the results of asymptotic behaviour in both cases are not contradictory but are considered to be similar. Third, in our case, we have the PAL state, but in their case there exists no such state. This is due to the very nature of the descreteness of the Ising weights in our case.

Finally, let us discuss the validity of the ansatz we consided in this paper, RS and 1RSB. With regard to the AT stability of the RS solution, according to our results, as α increases, the entropy of the RS solution in branch I becomes 0 before its AT-stability is lost. The numerical results indicate that the RS solution is valid at least as long as its entropy is positive. We did not calculate the AT-stability for the 1RSB solution, but we judged the validity of the 1RSB ansatz by comparing the theoretical results and the numerical results obtained from the MCS. Although there exist finite-size effects in the MCS, we found that the numerical results support our ansatz.

Appendix A: The conditions for the existence of PL and PAL

PL is characterized by $q = R = 1$ and PAL is characterized by $q = 1$ and $R = -1$. Thus, from the SPE (8) and (9) we have the following necessary and sufficient conditions for PL and PAL,

$$\hat{R} \to \infty \quad \text{and } \tau = \frac{\hat{R}}{\sqrt{\hat{q}}} \to \infty \text{ for PL,}$$

$$\hat{R} \to -\infty \quad \text{and } \tau = \frac{\hat{R}}{\sqrt{\hat{q}}} \to -\infty \text{ for PAL.}$$

First, we estimate $\hat{q}$ and $\hat{R}$. Substituting $v = -\frac{Ru}{\sqrt{q}\chi}$ into Eqs.(10) and (11), we obtain

$$\begin{aligned}
\hat{q} &= \frac{2\alpha Q}{1-q}\int \tilde{Du}(\tilde{\varphi}(u))^2 \\
&\times[1-(2k-1)H(-\frac{Ru}{\sqrt{q}\chi}) + (k-1)H(-\frac{Ru}{\sqrt{q}\chi}+\frac{a}{\xi}) + (k-1)H(-\frac{Ru}{\sqrt{q}\chi}-\frac{a}{\xi})], \\
\hat{R} &= \frac{-2\alpha}{q\xi}\int \tilde{Du}\tilde{\varphi}(u) \\
&\times[(1-k)h(-\frac{Ru}{\sqrt{q}\chi}+\frac{a}{\xi}) + (1-k)h(-\frac{Ru}{\sqrt{q}\chi}-\frac{a}{\xi}) + (2k-1)h(-\frac{Ru}{\sqrt{q}\chi})].
\end{aligned}$$

In the cases of PL and PAL, $q = 1$ and $|R| = 1$. Then, if we assume $q = |R|$ and $0 < \chi < \infty$ for finite α and take the limit $q \to 1$, $\xi = +0$ follows, and we obtain

$$\begin{aligned}
\hat{q} &\simeq \frac{2\alpha}{\sqrt{1-q}}(1-k)\int \frac{du}{\sqrt{2\pi}}(\tilde{\varphi}(u))^2 + \frac{2\alpha}{\sqrt{1-q}}(2k-1)\int \frac{du}{\sqrt{2\pi}}(\tilde{\varphi}(u))^2 H(\frac{Ru}{\chi}) \\
&= \frac{\alpha}{\sqrt{1-q}}\{(1-k)\hat{q}_1 + k\hat{q}_2\}, \\
\hat{R} &\simeq \frac{-2\alpha}{\xi}(2k-1)\int \frac{du}{\sqrt{2\pi}}\tilde{\varphi}(u)h(\frac{u}{\chi}) = -\frac{2\alpha}{Q\sqrt{2\pi}}(2k-1)\int Du\tilde{\varphi}(\chi u) \\
&= -\frac{\alpha}{\sqrt{1-q}}(1-2k)\hat{R}_1,
\end{aligned}$$

with

$$\begin{aligned}
\hat{q_1} &= \frac{2}{\sqrt{2\pi}}\int du(\tilde{\varphi}(u))^2 H(-\frac{Ru}{\chi}), \ \ \hat{q_2} = \frac{2}{\sqrt{2\pi}}\int du(\tilde{\varphi}(u))^2 H(\frac{Ru}{\chi}), \\
\hat{R_1} &= \frac{2}{\sqrt{2\pi}}\int Du|\tilde{\varphi}(\chi u)|.
\end{aligned}$$

Thus, if $q = |R|$ and $0 < \chi < \infty$ and the limit $q \to 1$ is taken, then $\hat{q}_1, \hat{q}_2$ and $\hat{R}_1$ are all finite. Also, in this case, $\hat{q}$ is infinite and for $0 \le k < \frac{1}{2}$, $\hat{R}$ is negatively infinite. Therefore, we have

$$\tau = \frac{\hat{R}}{\sqrt{\hat{q}}} = -\alpha^{\frac{1}{2}}(1-q)^{-\frac{1}{4}}\{\frac{(1-2k)\hat{R}_1}{\sqrt{(1-k)\hat{q}_1 + k\hat{q}_2}}\} \to -\infty.$$

Thus, for $0 \le k < \frac{1}{2}$, PL does not exist, but PAL does exist.

Appendix B: Derivation of the relation $f_{1RSB} \simeq \alpha\epsilon_g$ for $\beta \ll 1$

As found from our numerical calculations of the SPE, for all cases studied in this paper, in the case that solutions for which $|R| < 1$ exist for arbitrarily large values of α, as $\alpha \to \infty$, $\beta_c (= \frac{1}{T_c}) \to 0$. In this appendix, assuming $\beta \ll 1$ and $P(-u) = -P(u)$, we derive the relation $f_{RSB} \simeq \alpha\epsilon_g$.

Because the entropy s per synapse can be expressed as $s = -\beta f - \alpha\beta e^{-\beta} J$, we first estimate J. For $\beta \ll 1$, $(\tilde{H}(u))^{-1} \simeq 1 + \beta H(-u)$ holds. Then, we obtain

$$\begin{aligned} J &= \int Dy 2\mathcal{P}(y) \int Du \frac{H(Y)-1}{\tilde{H}(Y)} = \int Du \frac{H(\frac{u}{Q})-1}{\tilde{H}(\frac{u}{Q})} E(\frac{u}{Q}) \\ &= -\int Du \frac{H(-\frac{u}{Q})}{\tilde{H}(\frac{u}{Q})} E(\frac{u}{Q}) \simeq -\int Du H(-\frac{u}{Q})\{1 + \beta H(-\frac{u}{Q})\} E(\frac{u}{Q}) \\ &\equiv J_0 + \beta J_1, \end{aligned} \tag{B·1}$$

with

$$\begin{aligned} J_0 &= -\int Du H(-\frac{u}{Q}) E(\frac{u}{Q}) = -K_a, \\ J_1 &= \int Du H(\frac{u}{Q}) H(-\frac{u}{Q}) E(\frac{u}{Q}) - \int Du H(-\frac{u}{Q}) E(\frac{u}{Q}) = K_b - K_a, \\ K_a &\equiv \int Du H(-\frac{u}{Q}) E(\frac{u}{Q}), \qquad \text{(B·2)} \\ K_b &\equiv \int Du H(\frac{u}{Q}) H(-\frac{u}{Q}) E(\frac{u}{Q}) = \int Du H(\frac{u}{Q}) H(-\frac{u}{Q}). \qquad \text{(B·3)} \end{aligned}$$

Here, we have used the following relations, which hold when $P(u)$ is an odd function:

$$E(u) = 1 - E_1(u), \quad E_1(u) \equiv \int Dy P(\xi y + \frac{R}{\sqrt{q}} Q u), \tag{B·4}$$

$$E_1(-u) = -E_1(u). \tag{B·5}$$

Then, we have

$$J \simeq -K_a + \beta(K_b - K_a). \tag{B·6}$$

Using the condition $s = 0$ for the 1RSB solution, we therefore obtain

$$f_{1RSB} = -\alpha e^{-\beta} J \simeq -\alpha\{-K_a + \beta K_b\} = \alpha\{K_a - \beta K_b\}, \tag{B·7}$$

where K_a is given by

$$K_a = \int DuH(-\frac{u}{Q})E(\frac{u}{Q}) = \int Dy2\mathcal{P}(y)\int DuH(-Y)$$
$$= \int Dy2\mathcal{P}(y)H(\frac{Ru}{\sqrt{1-R^2}}),$$

with $Y = \frac{\sqrt{q-R^2}u - Ry}{\sqrt{1-q}}$. Next, note that ϵ_g can be calculated from its definition as

$$\epsilon_g = \frac{1}{2} - \int DyH(-\frac{Ru}{\sqrt{1-R^2}})[\mathcal{P}(u) - \mathcal{P}(-u)].$$

Using the relations $\mathcal{P}(u) = \frac{1}{2}(1+P(u))$ and $P(-u) = -P(u)$, this becomes

$$\epsilon_g = \frac{1}{2} - \int DuH(-\frac{Ru}{\sqrt{1-R^2}})P(u) = \int DuH(\frac{Ru}{\sqrt{1-R^2}})(1+P(u))$$
$$= \int Du2\mathcal{P}(u)H(\frac{Ru}{\sqrt{1-R^2}}).$$

Thus, we obtain $K_a = \epsilon_g$, and hence

$$f_{1RSB} \simeq \alpha(\epsilon_g - \beta K_b). \tag{B·8}$$

If the minimum of ϵ_g is positive, because K_b (>0) is bounded from above, we obtain, for $\beta \ll 1$,

$$f_{1RSB} \simeq \alpha\epsilon_g. \tag{B·9}$$

---

1) See for example, T. L. H. Watkin, R. Rau and M. Biehl: Rev. Mod. Phys.**65**(1993)499.
2) A. Engel and C. V. d. Broeck: *Statistical Mechanics of Learning*, (Cambridge,2001).
3) *On-line Learning in Neural Networks* edited by David Saad, (Cambridge, 1998).
4) H. S. Seung, H. Sompolinsky and N. Tishby N: Phys. Rev. A **45**(1992)6056.
5) G. Györgyi G: *Phys. Rev. A* **41**(1990) 7097.
6) H. Sompolinsky, N. Tishby and H. S. Seung: Phys. Rev. Lett. **65**(1990) 1683.
7) T. Uezu, Y. Kabashima, K. Nokura and N. Nakamura: J. Phys. Soc. Japan **65**(1996) 3797
8) T. Uezu and Y. Kabashima: J. Phys. A: Math. Gen. **29**(1996) L55.
9) T. Uezu and Y. Kabashima: J. Phys. A: Math. Gen. **29**(1996) L439.
10) T. Uezu: J. Phys. Soc. Japan **71**, No.8(2002) 1882.
11) T. Uezu : J. of Phys. A: Math. Gen. **30**(1997) L777.
12) Y. Kabashima and J. Inoue, J. Phys. A: Math. Gen. **31**(1998) 123.
13) If k satisfied $1/2 \leq k \leq 1$, $\mathcal{P}(u^o)$ would be
14) G. Parisi: *J. Phys. A* **13**(1980) 1101; **13**(1980) L115; **13**(1980) 1887.
15) W. Krauth and M. Mézard: *J. Phys. France* **50**(1989) 3057.

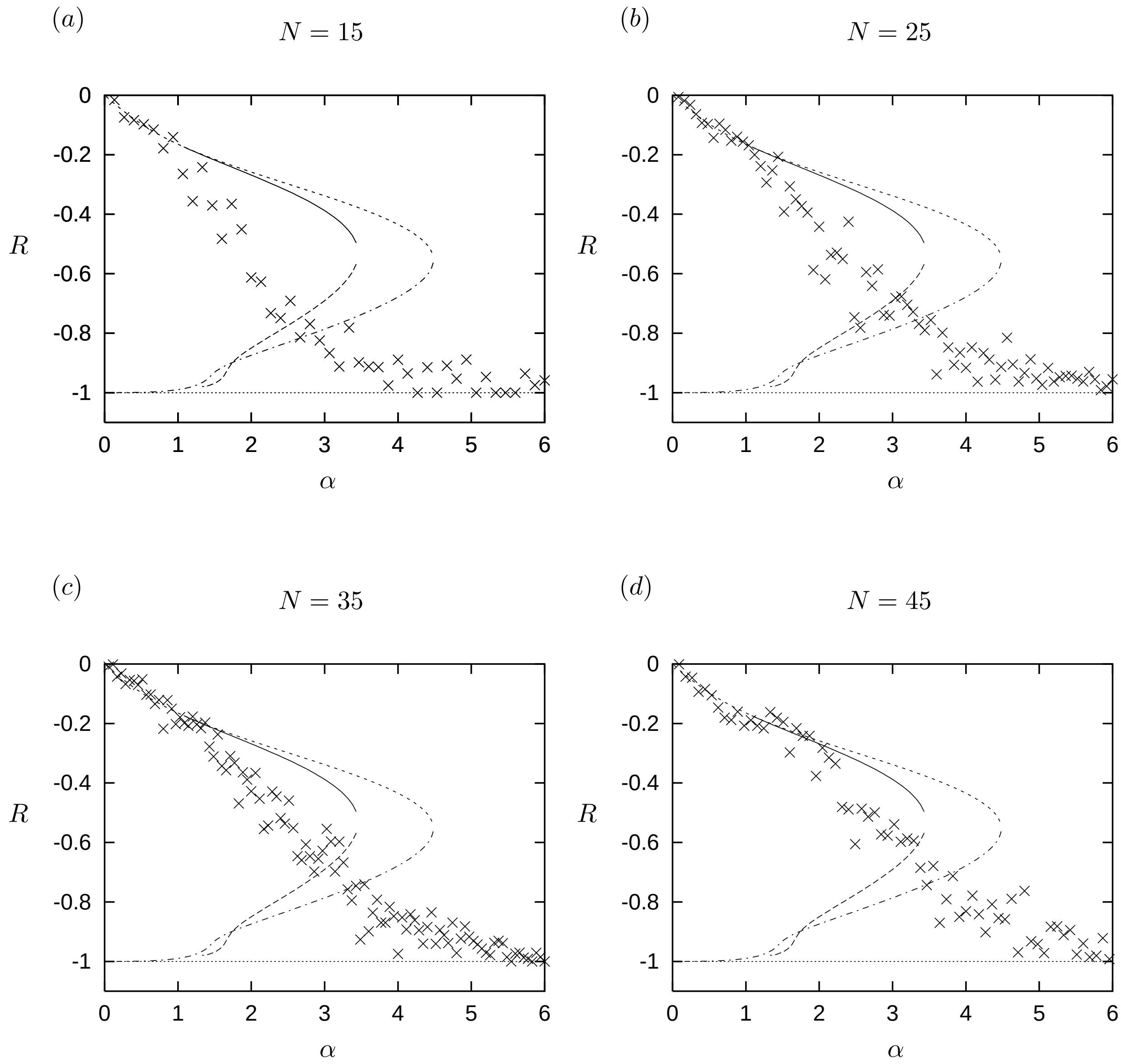

Fig. 7. Region (C). The α dependence of R for several values of N, with $k = 0, a = 1.5$ and $T = 0.2$. The solid curve represents the 1RSB solution in branch I, the long-dashed curve represents the 1RSB solution in branch II, the short-dashed curve represents the RS solution in branch I, the dashed-dotted curve represents the RS solution in branch II, the dotted curve represents the PAL solution, and $\times$s represent the numerical results.