Learning of non-monotonic rules by simple perceptrons

1998 J. Phys. A: Math. Gen. 31 123

(http://iopscience.iop.org/0305-4470/31/1/015)

J. Phys. A: Math. Gen. **31** (1998) 123–144. Printed in the UK PII: S0305-4470(98)85416-4

Learning of non-monotonic rules by simple perceptrons

Yoshiyuki Kabashima†‖ and Jun-ichi Inoue‡§

† Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama 226, Japan

‡ Department of Physics, Tokyo Institute of Technology, Ohokayama, Meguro-ku, Tokyo 152, Japan

§ Laboratory for Information Representation, RIKEN, Hirosawa 2-1, Wako-shi, Saitama 351-01, Japan

Received 27 June 1997, in final form 3 October 1997

Abstract. In this paper, we study the generalization ability of a simple perceptron which learns an unrealizable Boolean function represented by a perceptron with a non-monotonic transfer function of reversed-wedge type. This type of non-monotonic perceptron is considered as a variant of multilayer perceptron and is parametrized by a single 'wedge' parameter a. Reflecting the non-monotonic nature of the target function, a discontinuous transition from the poor generalization phase to the good generalization phase is observed in the learning curve for intermediate values of a. We also find that asymptotic learning curves are classified into the following two categories depending on a. For large a, the learning curve obeys a power law with exponent 1. On the other hand, a power law with exponent $\frac{2}{3}$ is obtained for small a. Although these two exponents are obtained from unstable replica symmetric solutions by using the replica method, they are consistent with the results obtainable without using the replica method in a low-dimensional version of this learning problem. This suggests that our results are good approximations even if they are not exact.

1. Introduction

Recently, the problem of learning from examples has been an attractive topic in statistical mechanics [1]. In order to investigate how well a generalization ability can be acquired by learning, learning curves of generalization error ε, which is a probability of making a false prediction for a novel example, were calculated for various types of networks by using the replica method. These studies revealed the following feature of learning. When the number of examples P is small relative to that of weight parameters N, learning curves exhibit rich behaviours depending on the architectures of networks. In contrast, there are some universal properties in the asymptotic region where $\alpha = P/N$ is large. For example, in the case where a teacher's relation is realizable and there is no noise, learning curves of Boolean networks all obey the universal scaling law

$$\varepsilon \sim \alpha^{-1}. \tag{1.1}$$

It is an interesting and important question whether a similar feature of learning holds as well in more realistic cases where examples are corrupted by noise or the teacher's rule is unrealizable. Recently, learning of a simple perceptron from noisy examples was studied precisely and the following answer was given to this question [2, 3]. When learning is

‖ E-mail address: kaba@dis.titech.ac.jp

0305-4470/98/010123+22$19.50

disrupted by noise, the learning curve does not obey equation (1.1) and the scaling law depends on the type of noise. For example, when the teacher is a simple perceptron the sign of whose output is reverse to the opposite with a fixed probability λ, the learning curve decays as

$$\varepsilon - \varepsilon_{\min} \sim \alpha^{-1} \tag{1.2}$$

where $\varepsilon_{\min}$ is the minimum value of generalization error which is attained by a unique optimal weight. On the other hand, when each input vector $\boldsymbol{x}$ is disrupted by noise vector $\boldsymbol{\eta}$, the decay of learning curve is rather slow as

$$\varepsilon - \varepsilon_{\min} \sim \alpha^{-2/3} \tag{1.3}$$

within a logarithmic precision.

When the teacher's rule is unrealizable by the student, the input–output relation seems rather noisy to the student. Therefore, it is expected that similar features that are obtained in *learning from noisy examples* are also observed in *learning of unrealizable rules*. Learning of a simple perceptron which learns a multilayer one, such as a committee machine and a parity machine, is a typical example of learning of unrealizable rules. However, detailed analysis of such problems is much involved and only a few established conclusions on the learning curves are obtained so far [4, 5].

In this paper, we study a simple perceptron which learns a perceptron with a non-monotonic transfer function of reversed-wedge type in order to clarify what type of learning curve appears when target rule is unrealizable by the student. The input–output relation of our non-monotonic perceptron is defined as follows. For an N-dimensional input vector $\boldsymbol{x}$, this machine returns $+1$ if $u_0 \in (-a, 0)$ or $u_0 \in (a, \infty)$ and -1 otherwise, where u_0 is the normalized inner product of its synaptic weight $\boldsymbol{w}_0$ and $\boldsymbol{x}$. The properties of such neural networks with non-monotonic transfer function have been studied in the context of the associative memory [6–9] and the storage capacity [10–12]. The authors of these studies reported that these non-monotonic neural networks can store more patterns than conventional monotonic neural networks. This kind of non-monotonic perceptron can be regarded as a variant of parity machines with three hidden units of which outputs are $\mathrm{sign}(u-a)$, $\mathrm{sign}(u)$ and $\mathrm{sign}(-u-a)$, respectively. The product of these three outputs is the final output of this parity machine [11]. This enhanced structure of the non-monotonic networks may partly explain its greater ability than that of monotonic networks. In addition, the calculation necessary for analysis is much easier than that for parity machines and committee machines. For this reason, this type of network has been investigated as a toy model of the multilayer network [13].

A similar learning model to ours was once investigated by Watkin and Rau [5]. They studied learning curves of two conventional learning algorithms, 'high-temperature learning' and 'maximum stability algorithm' by solving the saddle point (SP) equations numerically. It should be remarked that their investigation was limited to the region in which the number of examples is relatively small compared with that of the synaptic weights and no analytical conclusion on the asymptotic property is obtained. In contrast, we will investigate learning by the 'minimum error algorithm', namely 'zero-temperature Gibbs learning' with Gardner–Derrida [14] cost function and give analytical conclusions on its asymptotic behaviour.

The results obtained in this paper are summarized as follows. It is clear that our non-monotonic perceptron is realizable for the two limiting values of a, $a = 0$ and $+\infty$. In these two special cases, the learning curve obeys the scaling law (1.1). Except for these values of a, the behaviour of learning is found to be classified into the following four categories depending on a: for $a > a_{c0} \sim 1.53$, the learning curve smoothly decays to its minimum and

its asymptote obeys relation (1.2); for $a_{c0} > a > a_{c1} = 1.17$, a discontinuous transition from the poor generalization phase to the good generalization phase takes place at some value of $\alpha = \alpha_{\rm th} \sim {\rm O}(1)$ and the quasistable solution disappears at the spinodal point $\alpha = \alpha_{\rm sp} > \alpha_{\rm th}$. The asymptotic learning curve has the form of equation (1.2); for $a_{c1} > a > a_{c2} = 0.8$, the discontinuous transition from the poor generalization phase to the good generalization phase also takes place at some value of $\alpha = \alpha_{\rm th} \sim {\rm O}(1)$. However, the spinodal point $\alpha_{\rm sp}$ becomes infinity and the quasistable solution persists even in the limit $\alpha \to \infty$. This quasistable solution exhibits the slow convergence (1.3) in the asymptotic region $\alpha \gg 1$, although the asymptotic form of the globally stable solution obeys equation (1.2); for $a_{c2} > a > 0$, the discontinuous transition disappears and the learning curve decays to its minimum smoothly exhibiting the slow convergence (1.3) in the asymptotic region. These results suggest that the scaling relations obtained in the problems of learning from noisy examples generally appear in the problem of learning unrealizable rules as well. We should also address that the globally stable solution obtained by the minimum error algorithm realizes the optimal generalization error in the limit $\alpha \to \infty$ for an arbitrary a.

The above results are obtained by using the replica method under the replica-symmetric (RS) ansatz. Unfortunately, it is known that the RS solution of zero-temperature learning with the Gardner–Derrida cost function becomes thermodynamically unstable when the teacher's rule is unrealizable [15, 5, 16]. Furthermore, it is conjectured that any finite step of replica symmetry breaking (RSB) is not sufficient to obtain a thermodynamically stable solution [17]. Nevertheless, we have a conjecture that our results offer a good approximation at least qualitatively because the same exponents of asymptotic learning curves, 1 and $\frac{2}{3}$, are also obtainable without using the replica method in a low-dimensional version of our learning model.

This paper is organized as follows. In section 2, the problem is formulated and the general properties of the generalization error are investigated. In section 3, the learning curves are calculated in the framework of statistical mechanics. In particular, the asymptotic behaviours of the solutions are investigated analytically. In section 4, we discuss the validity of our RS solution. Section 5 is devoted to a summary.

2. Model

Hereafter, we assume that an arbitrary N-dimensional vector $\boldsymbol{a}$ is normalized as $|\boldsymbol{a}| = \sqrt{N}$. We consider a teacher perceptron with synaptic weight $\boldsymbol{w}_0$ which has a non-monotonic transfer function of reversed-wedge type parametrized by a non-negative number a

$$T_a(x) = {\rm sign}(-x - a)\,{\rm sign}(x)\,{\rm sign}(x - a) \tag{2.1}$$

where ${\rm sign}(x)$ is the function that returns the sign of argument x. For an N-dimensional input vector $\boldsymbol{x}$, this machine returns the output y as

$$y = T_a(u_0) \tag{2.2}$$

where $u_0 \equiv \boldsymbol{w}_0 \cdot \boldsymbol{x}/\sqrt{N}$.

On the other hand, the student in this problem is a simple perceptron with synaptic weight $\boldsymbol{w}$. Following to a given set of examples $\xi^P \equiv \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_P, y_P)\}$ which are independently and uniformly drawn from the N-dimensional sphere S^N centred at the origin, this student adjusts $\boldsymbol{w}$ in order to acquire the generalization ability. This ability is measured by the generalization error $\varepsilon(\boldsymbol{w})$ which is the probability of making a false prediction on a future example. Due to the assumption that the distribution of inputs is uniform on S^N, $\varepsilon(\boldsymbol{w})$ becomes a function of overlap between the two weights $\boldsymbol{w}_0$ and

$\boldsymbol{w}$, $R = \boldsymbol{w}_0 \cdot \boldsymbol{w}/N$. Note that in the limit $N \gg 1$, $u_0 = \boldsymbol{w}_0 \cdot \boldsymbol{x}/\sqrt{N}$ and $u = \boldsymbol{w} \cdot \boldsymbol{x}/\sqrt{N}$ obeys a joint Gaussian distribution

$$P_R(u_0, u) = \frac{1}{2\pi\sqrt{1-R^2}} \exp\left[-\frac{u_0^2 - 2Ru_0u + u^2}{1-R^2}\right] \tag{2.3}$$

when $\boldsymbol{x}$ is uniformly drawn from S^N. This enables us to calculate the generalization error as

$$\begin{aligned}\varepsilon(\boldsymbol{w}) \equiv \varepsilon(R) &= \langle\Theta(-T_a(u_0)\,\mathrm{sign}(u))\rangle_{\boldsymbol{x}} \\ &= \int_{-\infty}^{+\infty} \mathrm{d}u_0 \int_{-\infty}^{+\infty} \mathrm{d}u\; P_R(u_0, u)[\Theta(T_a(u_0))\Theta(-u) + \Theta(-T_a(u_0))\Theta(u)] \\ &= 2\int_{-\infty}^{0} \mathrm{D}t\; \Omega(R; t)\end{aligned} \tag{2.4}$$

where $\langle\cdots\rangle_{\boldsymbol{x}}$ represents the average over input vector $\boldsymbol{x}$, $\Theta(x)$ is the Heaviside's step function which returns $+1$ for $x > 0$ and 0 otherwise, $\mathrm{D}t$ is the Gaussian measure $\exp[-t^2/2]/\sqrt{2\pi}$, and $\Omega(R; t)$ is a function defined as

$$\begin{aligned}\Omega(R : t) \equiv \int_{-\infty}^{+\infty} \mathrm{D}z \Big[&\Theta\left(-\sqrt{1-R^2}z - Rt - a\right) + \Theta\left(\sqrt{1-R^2}z + Rt\right) \\ &- \Theta\left(\sqrt{1-R^2}z + Rt - a\right)\Big].\end{aligned} \tag{2.5}$$

In figure 1, we plot $\varepsilon(R)$ for several values of parameter a. This figure shows that for $a = \infty$, $\varepsilon(R)$ goes to zero when R approaches 1. In contrast, for $a = 0$, $\varepsilon(R)$ goes to zero when R approaches -1. This is easily understood because the teacher's input–output relation of $a = 0$ is completely opposite to that of $a = \infty$. Between these two limiting values $a = 0$ and $a = \infty$, $\varepsilon(R)$ exhibits a highly non-trivial behaviour. For

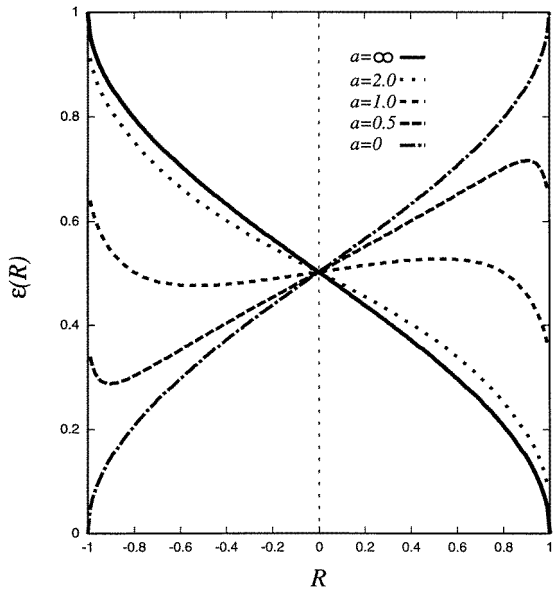

Figure 1. Generalization error as a function of overlap R, $\varepsilon(R)$, for several values of a.

$a > a_{c1} = \sqrt{2\log 2} = 1.17$, $\varepsilon(R)$ is a monotonically decreasing function of R which takes the non-zero minimum value at $R = +1$. However, for $a < a_{c1}$, ε is locally minimized at

$$R_{-}(a) \equiv -\sqrt{\frac{2\log 2 - a^2}{2\log 2}} \tag{2.6}$$

and locally maximized at

$$R_{+}(a) \equiv +\sqrt{\frac{2\log 2 - a^2}{2\log 2}} = -R_{-}(a). \tag{2.7}$$

Further, for $0 < a < a_{c2} = 0.8$, $\varepsilon(R_{-}(a)) < \varepsilon(+1)$. Namely, $\varepsilon(R)$ is globally minimized at $R = R_{-}(a)$. In figures 2(*a*) and (*b*), we plot the global minimum value of $\varepsilon(R)$ and the value of R which gives the global minimum as functions of a, respectively. From these figures, we find that for $a > a_{c2}$, it is desirable for the student to find the teacher's weight $\boldsymbol{w}_0$. In contrast, for $0 < a < a_{c2}$, it is more desirable for the student to find a weight $\boldsymbol{w}_*$ which satisfies the condition $\boldsymbol{w}_* \cdot \boldsymbol{w}_0/N = R_{-}(a)$. This is a very interesting situation because most of the previous works have mainly focused on the problem of how fast the student finds a *unique* optimal weight.

3. Statistical mechanics

For the purpose of acquiring a good generalization ability, it is a natural learning strategy to minimize the number of false predictions on the given set of examples ξ^P

$$E(\boldsymbol{w}|\xi^P) = \sum_{\mu=1}^{P} \Theta(-y_\mu \cdot u_\mu) \tag{3.1}$$

where $u_\mu \equiv \boldsymbol{w} \cdot \boldsymbol{x}_\mu/\sqrt{N}$. We call the learning algorithm following this strategy the 'minimum error algorithm'. The cost function (3.1) is identical to that of Gardner and Derrida [14] and the learning process of the minimum error algorithm is investigated through their analytical method as follows.

From the 'energy' defined by equation (3.1), the 'partition function' with the inverse temperature β is given by

$$\begin{aligned} Z(\beta) &= \int \mathrm{d}\boldsymbol{w}\, \delta(|\boldsymbol{w}|^2 - N) \exp[-\beta E(\boldsymbol{w}|\xi^P)] \\ &= \int \mathrm{d}\boldsymbol{w}\, \delta(|\boldsymbol{w}|^2 - N) \prod_{\mu=1}^{P} [\mathrm{e}^{-\beta} + (1 - \mathrm{e}^{-\beta})\Theta(y_\mu \cdot u_\mu)]. \end{aligned} \tag{3.2}$$

Minimization of $E(\boldsymbol{w}|\xi^P)$ corresponds to the limit $\beta \to \infty$ in Gibbs–Boltzmann distribution and we focus on this limit hereafter.

3.1. Below α_c

Due to the storage ability of perceptrons, there remain some weights which completely reproduce the input–output relations among ξ^P until the ratio $\alpha = P/N$ increases up to some critical capacity α_c even if the teacher's relation is unrealizable. This enables us to calculate the learning curve below α_c by evaluating the logarithm of 'Gardner–Derrida volume' $V_{\mathrm{GD}} = Z(\infty)$ through the formula

$$\frac{\ln V_{\mathrm{GD}}}{N} = \frac{\langle\langle \ln Z(\infty) \rangle\rangle_{\xi^P}}{N} = \frac{1}{N} \lim_{n \to 0} \frac{\langle\langle Z^n(\infty) \rangle\rangle_{\xi^P} - 1}{n} \tag{3.3}$$

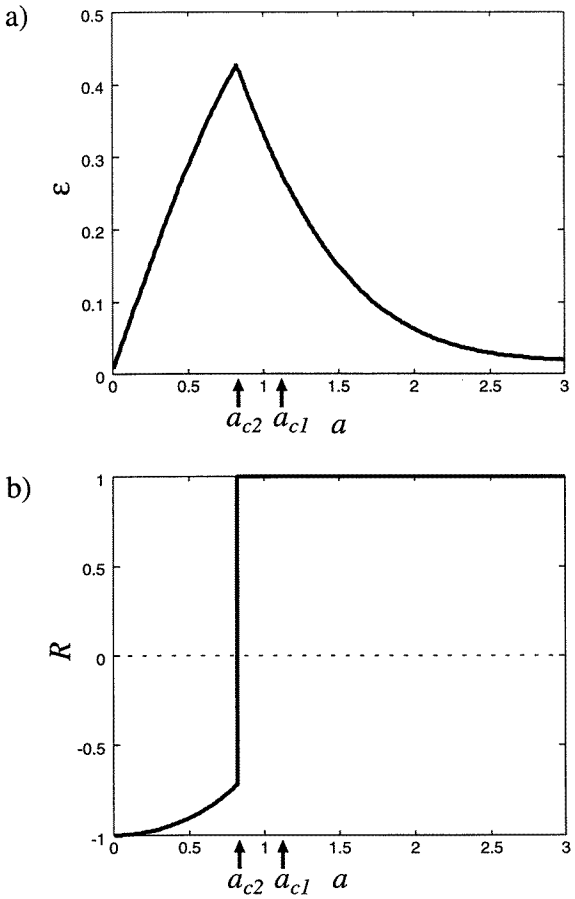

Figure 2. (*a*) The global minimum value of $\varepsilon(R)$ as a function of a. (*b*) R which globally minimize $\varepsilon(R)$ as a function of a.

where $\langle\langle \cdots \rangle\rangle_{\xi^P}$ represents the average over the example set ξ^P. $Z^n(\infty)$ is the simultaneous partition function of n-replicated systems sharing the same random variables ξ^P and becomes a function of order parameters

$$R_a = \frac{\boldsymbol{w}_0 \cdot \boldsymbol{w}_a}{N} \tag{3.4}$$

$$q_{ab} = \frac{\boldsymbol{w}_a \cdot \boldsymbol{w}_b}{N} \tag{3.5}$$

where $a = 1, \ldots, n$ and $b = 1, \ldots, n$ are the indices which represent replicated systems.

Under the RS ansatz

$$R_a = R \tag{3.6}$$

$$q_{ab} = q \tag{3.7}$$

we can show (see appendix A) that the equation (3.3) is evaluated as

$$\mathrm{ext}_{\{R,q\}} \left\{ 2\alpha \int_{-\infty}^{+\infty} \mathrm{D}t\, \Omega\left(\frac{R}{\sqrt{q}} : t\right) \ln \Xi(q : t) + \frac{1}{2}\ln(1-q) + \frac{q - R^2}{2(1-q)} \right\} \tag{3.8}$$

where $\Omega(R:t)$ is defined as equation (2.5) and

$$\Xi(q:t) \equiv \int_{-\infty}^{+\infty} \mathrm{D}z\, \Theta(\sqrt{1-q}z + \sqrt{q}t). \tag{3.9}$$

From the identities

$$\frac{\partial}{\partial R}\int_{-\infty}^{+\infty} \mathrm{D}t\, \Omega\left(\frac{R}{\sqrt{q}}:t\right) F(t) = \frac{1}{R}\int_{-\infty}^{+\infty} \mathrm{D}t\, \frac{\partial \Omega}{\partial t}\left(\frac{R}{\sqrt{q}}:t\right)\frac{\partial F(t)}{\partial t} \tag{3.10}$$

$$\frac{\partial}{\partial q}\int_{-\infty}^{+\infty} \mathrm{D}t\, \Xi(q:t)F(t) = \frac{1}{2q}\int_{-\infty}^{+\infty} \mathrm{D}t\, \frac{\partial \Xi}{\partial t}(q:t)\frac{\partial F(t)}{\partial t} \tag{3.11}$$

we find that equation (3.8) yields the following set of SP equations

$$2\alpha\int_{-\infty}^{+\infty} \mathrm{D}t\, \Omega \times \left(\frac{\Omega_t}{\Omega}\right) \times \left(\frac{\Xi_t}{\Xi}\right) = \frac{R^2}{1-q} \tag{3.12}$$

$$2\alpha\int_{-\infty}^{+\infty} \mathrm{D}t\, \Omega \times \left(\frac{\Xi_t}{\Xi}\right)^2 = \frac{q(q-R^2)}{(1-q)^2} \tag{3.13}$$

where F_t represents the abbreviation of the partial derivative of a function F with respect to t. By solving the SP equations (3.12) and (3.13), we can investigate the learning process below α_c.

The critical capacity α_c is defined as a ratio $\alpha = P/N$ at which overlap q becomes 1. Taking the limit $q \to 1$ in the SP equations (3.12) and (3.13), we obtain a couple of equations which determines α_c as

$$-2\alpha_c\int_{-\infty}^{0} \mathrm{D}t\, t\Omega_t(R_c:t) = {R_c}^2 \tag{3.14}$$

$$2\alpha_c\int_{-\infty}^{0} \mathrm{D}t\, t^2\Omega(R_c:t) = 1 - {R_c}^2 \tag{3.15}$$

where R_c is the value of R at the critical capacity α_c. Here, we have used the relation

$$\frac{\Xi_t}{\Xi} \sim -\frac{q}{1-q}t\Theta(-t) \tag{3.16}$$

which is valid in the limit $q \to 1$. α_c and R_c which are obtained from these equations are plotted as functions of a in figures 3(a) and (b).

3.2. Beyond α_c

The solution of equations (3.12) and (3.13) disappears for $\alpha > \alpha_c$. This reflects the fact that beyond α_c there is no weight which completely reproduces the input–output relations among ξ^P. This makes the Gardner–Derrida volume V_{GD} shrink to zero. Therefore, we can not investigate the learning process by evaluating equation (3.3). Instead, the 'free energy'

$$-f = \lim_{\beta\to\infty}\frac{\langle\!\langle \ln Z(\beta)\rangle\!\rangle_{\xi^P}}{N\beta} = \lim_{\beta\to\infty}\lim_{n\to 0}\frac{\langle\!\langle Z^n(\beta)\rangle\!\rangle_{\xi^P} - 1}{nN\beta} \tag{3.17}$$

gives us a solution for $\alpha > \alpha_c$.

$\langle\!\langle Z^n(\beta)\rangle\!\rangle_{\xi^P}$ also becomes a function of order parameters R_a and q_{ab}. Under the RS ansatz equations (3.6) and (3.7), it can be shown (see appendix A) that equation (3.17) with finite β is evaluated as

$$\mathrm{ext}_{\{R,q\}}\left\{\frac{2\alpha}{\beta}\int_{-\infty}^{+\infty} \mathrm{D}t\, \Omega\left(\frac{R}{\sqrt{q}}:t\right)\ln \Xi_\beta(q:t) + \frac{1}{2\beta}\ln(1-q) + \frac{q-R^2}{2\beta(1-q)}\right\} \tag{3.18}$$

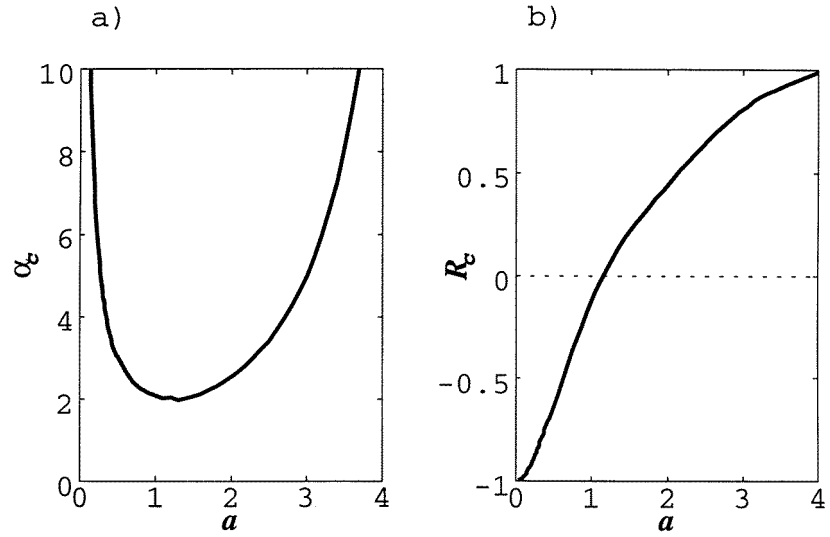

Figure 3. (a) α_c versus a. (b) R_c versus a.

where

$$\Xi_\beta(q:t) \equiv \mathrm{e}^{-\beta} + (1-\mathrm{e}^{-\beta})\Xi(q:t). \tag{3.19}$$

In the limit $\beta \to \infty$, a non-trivial result is obtained only when $q \to 1$ keeping $x \equiv \beta(1-q)$ finite. Then, equation (3.18) becomes

$$\mathrm{ext}_{\{R,x\}} \left\{ -2\alpha \left[\int_{-\infty}^{0} \mathrm{D}t\, \Omega(R:t) \left\{ \Theta(-t-\sqrt{2x}) + \frac{t^2}{2x}\Theta(t+\sqrt{2x}) \right\} \right] + \frac{1-R^2}{2x} \right\} \tag{3.20}$$

which yields the following SP equations

$$-2\alpha \int_{-\sqrt{2x}}^{0} \mathrm{D}t\, t\Omega_t(R:t) = R^2 \tag{3.21}$$

$$2\alpha \int_{-\sqrt{2x}}^{0} \mathrm{D}t\, t^2\Omega(R:t) = 1-R^2. \tag{3.22}$$

In the derivation of equation (3.20), we have used the relation

$$\Xi_\beta(q:t) \sim \begin{cases} \mathrm{e}^{-\beta} & \text{for } t < -\sqrt{\dfrac{2\beta(1-q)}{q}} \\ -\dfrac{\sqrt{1-q}}{\sqrt{2\pi q}\,t}\mathrm{e}^{-\frac{q}{2(1-q)}t^2} & \text{for } -\sqrt{\dfrac{2\beta(1-q)}{q}} < t < 0 \\ 1 & \text{for } 0 < t \end{cases} \tag{3.23}$$

which is valid in the limits $\beta \to \infty$ and $q \to 1$.

Before proceeding further, we mention the stability of the RS solution obtained from the SP equations (3.21) and (3.22). Unfortunately, our RS solution becomes thermodynamically unstable for $\alpha > \alpha_c$ (see appendix B). This results from a similar reason to that which Bouten (1994) pointed out for a problem of storing random patterns in a perceptron [16]. Therefore, we have to take the RSB into account in order to obtain a stable solution. However, it is much involved to compute RSB solutions and such computation is beyond the purpose of

this paper. Hence, here we only present unstable RS solutions hoping that they are still good approximations and discuss their validity by comparing them with the results obtainable in a low-dimensional version of the present problem in the next section.

By solving the SP equations (3.21) and (3.22), we found that the feature of the learning is classified into the following five types depending on a.

3.2.1. $a = \infty$, 0 (realizable cases). The teacher becomes realizable for $a = \infty$ because the teacher is identical to the student with $R = 1$ ($\boldsymbol{w} = \boldsymbol{w}_0$). In addition, the teacher is also realizable for $a = 0$. This is because for $a = 0$ its input–output relation is completely opposite to that of $a = \infty$, which means the student with $R = -1$ ($\boldsymbol{w} = -\boldsymbol{w}_0$) exactly mimics the teacher. For these special values of a, α_c becomes infinity and the learning is described by equations (3.12) and (3.13) even in the limit $\alpha \to \infty$. The solution of these equations is thermodynamically stable and the learning curve is identical to that obtained in a realizable problem [15, 18] which has the asymptote

$$\varepsilon \sim 0.624\alpha^{-1}. \tag{3.24}$$

This is consistent with the universal scaling (1.1) observed in general realizable problems.

3.2.2. $a > a_{c0} \sim 1.53$. In this parameter region, we found that the order parameter R monotonically increases to $+1$ as $\alpha \to \infty$ (figure 4(a)). On the other hand, x decreases from $+\infty$ to some value, and after that, approaches up to $a^2/2$ in the limit $\alpha \to \infty$ (figure 4(b)).

In order to investigate how fast R and x converge to these limiting values, we expand equations (3.21) and (3.22) with small parameters $\Delta R = 1 - R$ and $\Delta x = a^2/2 - x$. This yields the following equations

$$\alpha \left[\frac{a\mathrm{e}^{-a^2/2}}{\sqrt{2\pi}} H\left(\frac{\Delta x}{a\sqrt{2\Delta R}} \right) + \frac{(1 - \mathrm{e}^{-a^2/2})}{2\pi} \sqrt{2\Delta R} \right] \sim 1 \tag{3.25}$$

$$\alpha \left[\frac{\Delta R^{3/2}}{\sqrt{2\pi}} + \frac{a^2\mathrm{e}^{-a^2/2}}{\sqrt{2\pi}} H\left(\frac{\Delta x}{a\sqrt{2\Delta R}} \right) \right] \sim 2\Delta R \tag{3.26}$$

where $H(x) \equiv \int_x^{+\infty} \mathrm{d}t \, \exp[-t^2/2]/\sqrt{2\pi}$ and these imply the following scalings

$$\Delta R \sim \alpha^{-2} \tag{3.27}$$

$$\Delta x \sim (\ln \alpha)^{1/2} \alpha^{-1}. \tag{3.28}$$

From equations (2.4) and (2.5), we found the relation

$$\varepsilon(1 - \Delta R) - \varepsilon(+1) \sim \mathrm{O}(\Delta R^{1/2}). \tag{3.29}$$

Note that $\varepsilon(+1)$ is the minimum value of $\varepsilon(R)$ for this parameter region. Substituting equation (3.27) into equation (3.29), we obtain the learning curve

$$\varepsilon - \varepsilon_{\min} \sim \alpha^{-1} \tag{3.30}$$

which is identical to equation (1.2) discovered in the problem of learning disrupted by output noise (figure 4(c)).

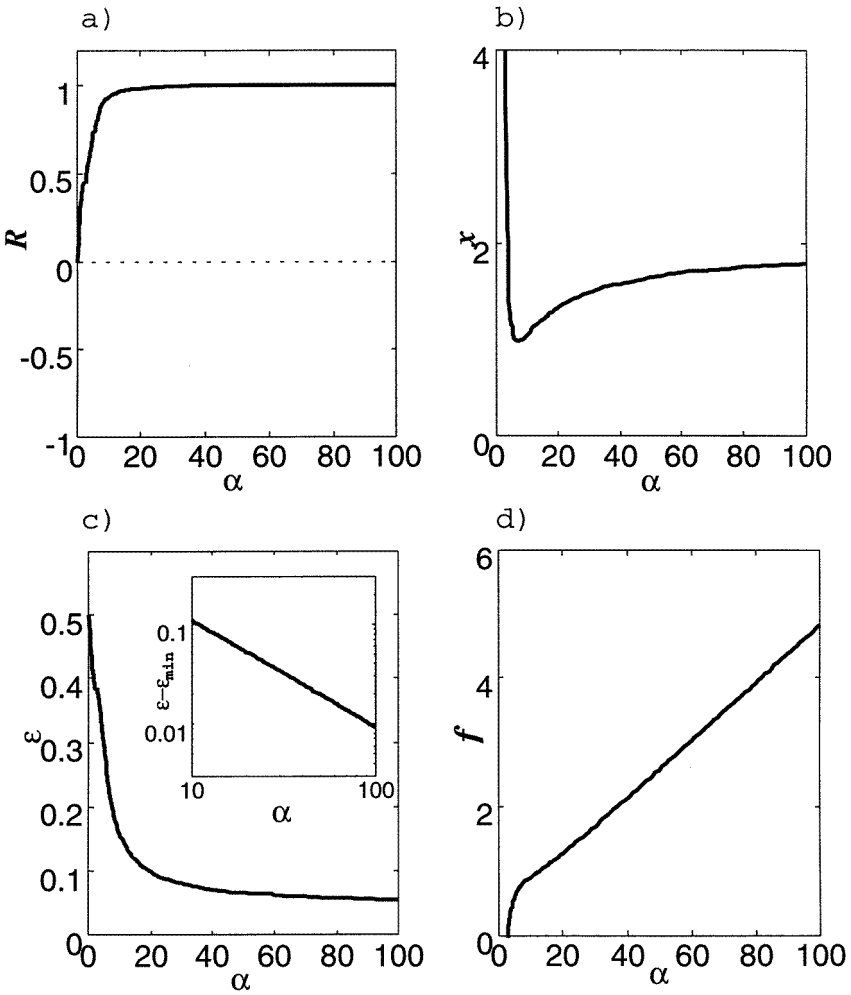

Figure 4. Solution for $a = 2.0$. R, x, ε and f are plotted as functions of α in (a), (b), (c) and (d), respectively. The inner graph of (c) is consistent with the analytically obtained scaling $\varepsilon - \varepsilon_{\min} \sim \alpha^{-1}$.

3.2.3. $a_{c0} > a > a_{c1}$. A discontinuous transition from the poor generalization phase to the good generalization phase is observed at $\alpha \sim \mathrm{O}(1)$ in this parameter region. In figures 5(a)–(d), we plot R, x, f and ε for $a = 1.3$ as functions of α, respectively. We can observe that there are three solutions for $12.5 < \alpha < \alpha_{\mathrm{sp}} \sim 24.2$. For $\alpha < \alpha_{\mathrm{th}} \sim 14.7$, the solution which has the smallest R among the three has the lowest free energy and therefore is the globally stable solution. As α is increased beyond α_{th}, the solution which has the largest R becomes the global minimum of free energy. Namely, a thermodynamic phase transition takes place at $\alpha = \alpha_{\mathrm{th}}$. Nevertheless, the solution with the smallest R persists until the spinodal point α_{sp} is reached. The solution with the middle R is the local maximum of free energy and represents a unstable solution. A similar transition was also reported by Engel and Reimer [13] in a problem where a non-monotonic perceptron learns the same type of non-monotonic perceptron, although teacher's rule is realizable in their problem.

In the limit $\alpha \to \infty$, R approaches $+1$ which achieves the global minimum of the

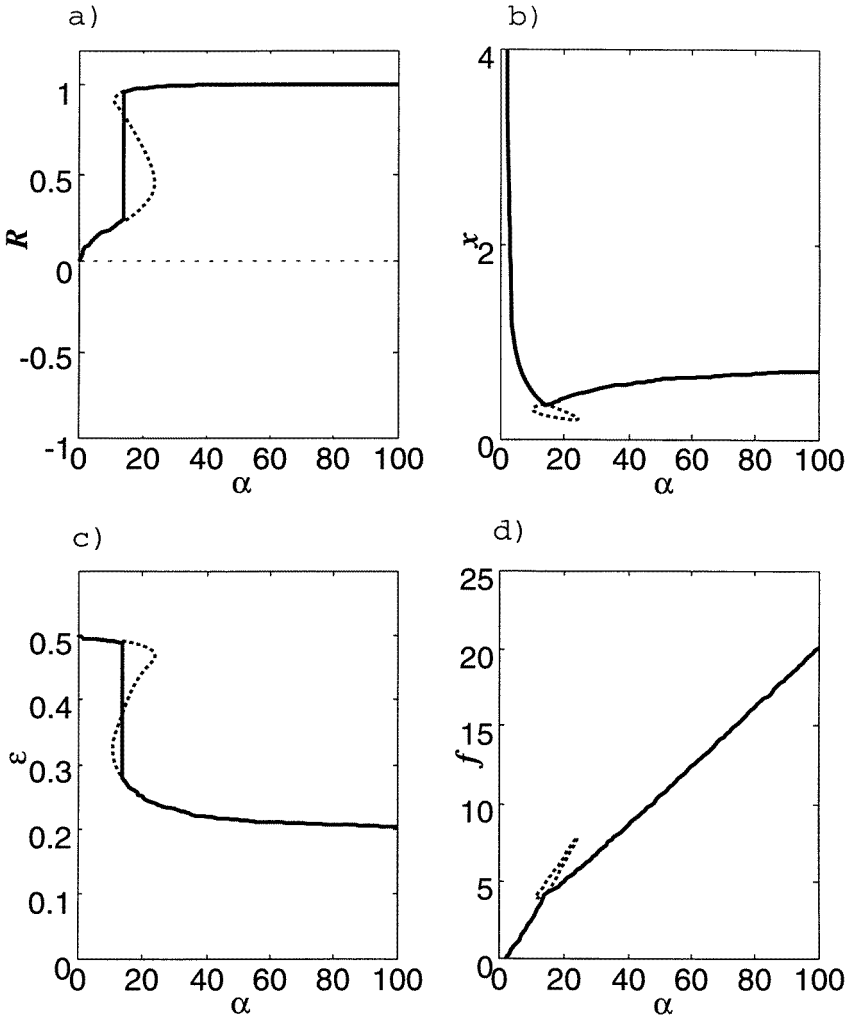

Figure 5. Solution for $a = 1.3$. R, x, ε and f are plotted as functions of a in (a), (b), (c) and (d), respectively. The globally stable solution is plotted by heavy curves.

generalization error in this parameter region. The asymptotic behaviour of the learning curve is identical to equation (3.30).

3.2.4. $a_{c1} > a > a_{c2}$. The discontinuous transition from the poor generalization phase to the good generalization phase is observed as well as in the previous subsection. However, the spinodal point $\alpha_{\rm sp}$ becomes infinity $a < a_{c1}$, which means that the quasistable solution beyond $\alpha_{\rm th}$ persists even in the limit $\alpha \to \infty$.

This is easily understood by the following consideration. In thermodynamical systems, physical quantities correspond to the minimum point of free energy which consists of 'energy' and 'entropy'. In our system, the energy (3.1) increases with α, although the effect of entropy in free energy is not proportional to α. Therefore, in the limit $\alpha \to \infty$, it is expected that properties of the system are determined almost only by energy. As a result of the central limit theorem, energy is nearly proportional to the generalization error

$\varepsilon(R)$ for $\alpha \gg 1$. This implies that R obtained from the SP equations (3.21) and (3.22) in the limit $\alpha \to \infty$ are identical to the extreme points of $\varepsilon(R)$. For $a > a_{c1}$, $R = 1$ is the unique minimum point of $\varepsilon(R)$. Hence, other solutions of the SP equations should disappear as $\alpha \to \infty$ even if they exist for $\alpha \sim \mathrm{O}(1)$, which explains why α_{sp} is finite for $a_{c1} < a < a_{c0}$. On the other hand, for $a_{c2} < a < a_{c1}$, $\varepsilon(R)$ has two extreme points $R = R_{-}(a)$ and $R = R_{+}(a)$ besides $R = 1$ which remains the global minimum of $\varepsilon(R)$. This suggests that the SP equations have three solutions in the limit $\alpha \to \infty$ corresponding to the three extreme points of $\varepsilon(R)$, i.e. $R = 1$, $R_{-}(a)$ and $R_{+}(a)$, which means that α_{sp} is infinity.

The solutions for $a = 1.0$ are plotted in figures 6(a)–(d). In these figures, we find three solutions which all persist in the limit $\alpha \to \infty$. One solution (solution (I)) starts from

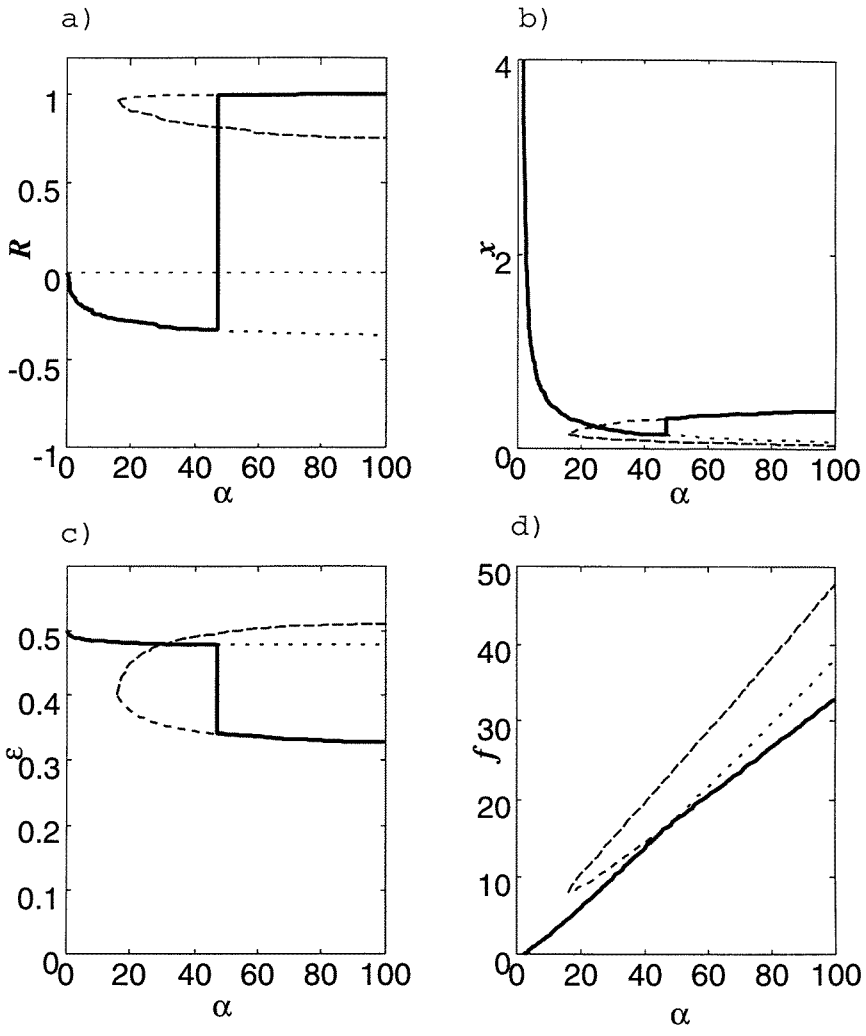

Figure 6. Solution for $a = 1.0$. R, x, ε and f are plotted as functions of α in (a), (b), (c) and (d), respectively. Globally stable solution is plotted by heavy curves.

$\alpha_c = 2.05$ and reaches the local minimum of $\varepsilon(R)$ as

$$\text{solution (I)} \qquad \begin{cases} R \to R_-(a) \\ x \to 0 \end{cases} \tag{3.31}$$

in the limit $\alpha \to \infty$.

On the other hand, two other solutions emerge at $\alpha \sim 16$. One of them (solution (II)) approaches to the local maximum of $\varepsilon(R)$ as

$$\text{solution (II)} \qquad \begin{cases} R \to R_+(a) \\ x \to 0 \end{cases} \tag{3.32}$$

in the limit $\alpha \to \infty$. This solution corresponds to the local maximum of free energy and therefore unstable. The last one (solution (III)) is another (local) minimum of free energy approaching to the global minimum of $\varepsilon(R)$, $R = 1$ as

$$\text{solution (III)} \qquad \begin{cases} R \to 1 \\ x \to a^2/2 \end{cases} \tag{3.33}$$

in the limit $\alpha \to \infty$.

For $\alpha < \alpha_{\text{th}} \sim 47$, solution (I) is the global minimum of free energy. As α increases beyond α_c, solution (III) becomes the global minimum of free energy, which means that a thermodynamical transition from solution (I) to solution (III) takes place at α_{th}. Hence, we obtain $R \to 1$ as $\alpha \to \infty$, which achieves the global minimum of $\varepsilon(R)$ for this parameter region of a. The asymptotic learning curve of this solution obeys the same power law as equation (3.30).

In addition to the globally stable solution (III), we now have a locally stable solution (I) in the limit $\alpha \to \infty$. R of this solution approaches $R_-(a)$ which locally minimizes $\varepsilon(R)$. In order to investigate how fast R and x converge as equation (3.31), we expand equations (3.21) and (3.22) with small parameters $\Delta R = R - R_-(a)$ and x. After some algebra, we obtain the following equations

$$2\alpha \frac{\Omega_{ttt}(R_-(a) : 0)}{\sqrt{2\pi}\, R_-(a)^2} \Delta R x \sim R_-(a)^2 \tag{3.34}$$

$$2\alpha \frac{x^{3/2}}{3\sqrt{\pi}} \sim 1 - R_-(a)^2 \tag{3.35}$$

which suggest the scalings

$$\Delta R \sim \alpha^{-1/3} \tag{3.36}$$

$$x \sim \alpha^{-2/3}. \tag{3.37}$$

In figures 7(a) and (b), we plot asymptotic behaviours of ΔR and x as functions of α, respectively, which are consistent with the scalings of equations (3.36) and (3.37).

From equations (2.4) and (2.5), it is found that the relation

$$\varepsilon(R_-(a) + \Delta R) - \varepsilon(R_-(a)) \sim \mathrm{O}(\Delta R^2) \tag{3.38}$$

holds for small ΔR. Substituting equation (3.36) into equation (3.38), we obtain the scaling

$$\varepsilon - \varepsilon_{l.\min} \sim \alpha^{-2/3} \tag{3.39}$$

where $\varepsilon_{l.\min} = \varepsilon(R_-(a))$. It should be remarked that this scaling form is identical to equation (1.3) discovered in the problem of learning disrupted by input noise, although $\varepsilon_{l.\min}$ is not the global but the local minimum of generalization error (figure 6(c)).

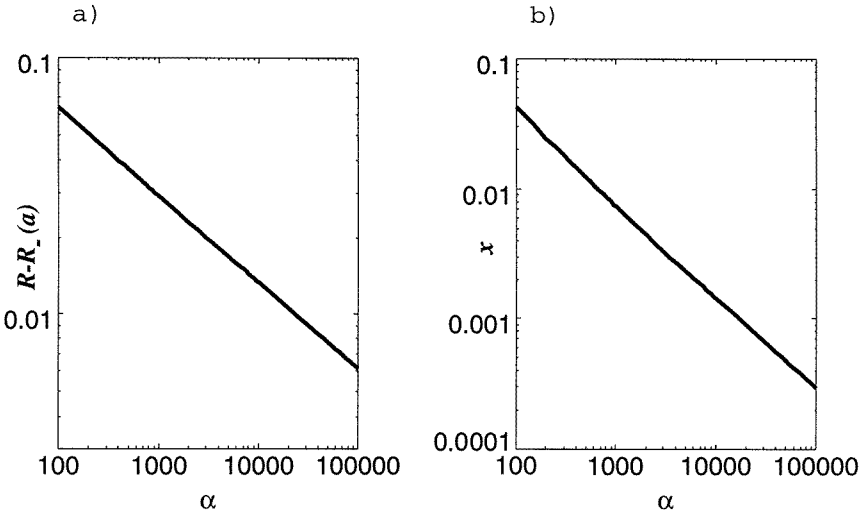

Figure 7. Asymptotic behaviour of the quasistable solution for $a = 1$. $\Delta R = R - R_{-}(a)$ and x are plotted as functions of α in (a) and (b), respectively. These graphs are consistent with the analytically derived scalings $\Delta R \sim \alpha^{-1/3}$ and $x \sim \alpha^{-2/3}$.

3.2.5. $a_{c2} > a > 0$. In this parameter region, $\varepsilon(R)$ is minimized not at $R = 1$ but at $R = R_{-}(a)$. As a result, solution (I) obtained in section 3.2.4 remains the global minimum of free energy until $\alpha \to \infty$. Namely, the thermodynamic transition from solution (I) to solution (III) disappears and the learning curve decays smoothly to its minimum as

$$\varepsilon - \varepsilon_{\min} \sim \alpha^{-2/3} \tag{3.40}$$

where $\varepsilon_{\min} = \varepsilon(R_{-}(a))$ is the minimum value of $\varepsilon(R)$ in this parameter region (figures 8(a)–(d)).

Remark that equation (3.40) suggests that the exponent of learning curve in the limit $a \to 0$ is different from that of $a = 0$ (realizable case) in equation (1.1). In contrast, as for the other realizable case $a = \infty$, the exponents of learning curves for $a \to \infty$ are the same as that of $a = \infty$. This implies that non-monotonic teacher with small a is more difficult for a simple perceptron to learn than that with large a.

4. Discussion

In this section, we discuss the validity of the results obtained under the RS ansatz in the previous section. First, we comment about the critical values of a, i.e. a_{c0}, a_{c1} and a_{c2}. $a_{c0} \sim 1.53$ is the point below which a discontinuous transition appears in the learning curve. This value is intrinsic of the RS ansatz and therefore will be changed if we proceed to RSB calculations. However, we conjecture that a_{c1}, which is defined as the point below which $\alpha_{\rm sp}$ becomes infinity, and a_{c2}, which is defined as the point below which $\alpha_{\rm th}$ becomes infinity, will be unchanged by RSB calculations because they result from changes in the shape of $\varepsilon(R)$ which is independent of the ansatz on the replica calculations.

Secondly, we mention the critical values of α, i.e. α_c, $\alpha_{\rm th}$ and $\alpha_{\rm sp}$. α_c is the point at which $q \to 1$. In our case, this value is identical to α_{AT} beyond which the RS solution becomes unstable (see appendix B). Therefore, this is invariant if we take the RSB into

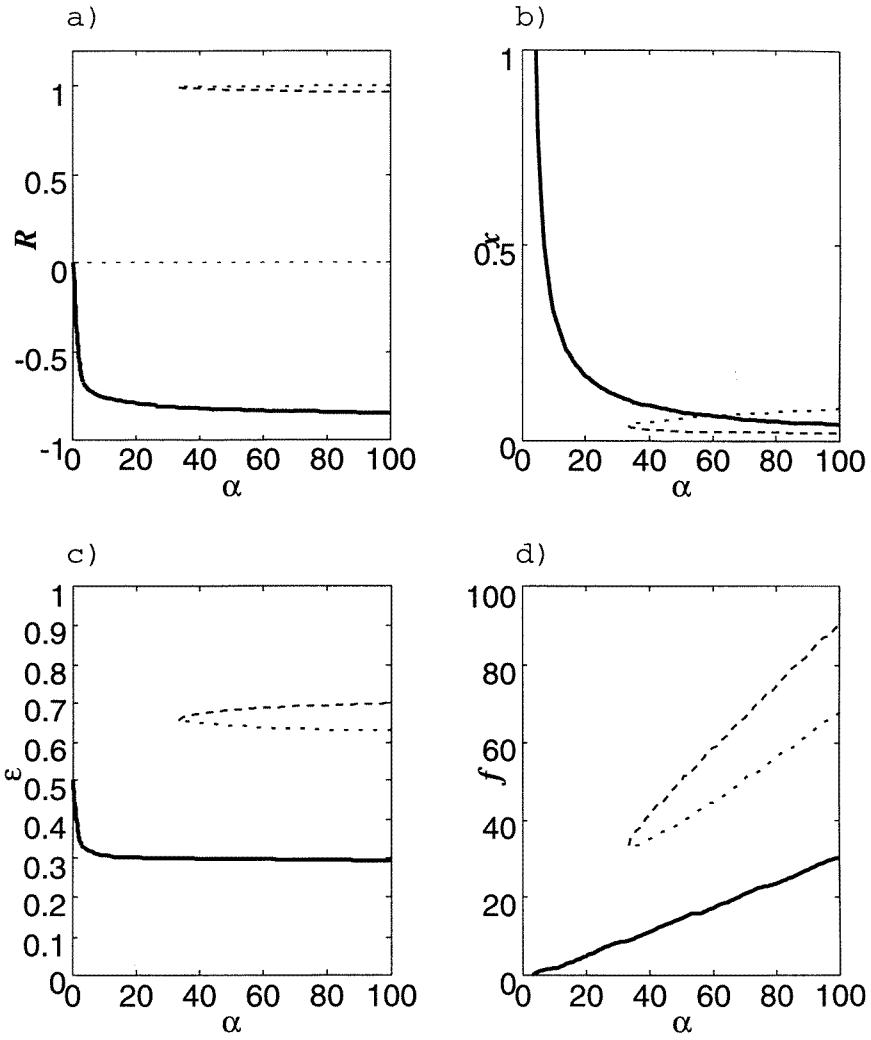

Figure 8. Solution for $a = 0.5$. R, x, ε and f are plotted as functions of a in (a), (b), (c) and (d), respectively. Globally stable solution is plotted by heavy curves.

account. However, $\alpha_{\rm th}$ and $\alpha_{\rm sp}$ will be changed by RSB calculations because they are intrinsic of the RS solution.

Finally, we discuss the asymptotic behaviours of learning curves. In the previous section, we found two types of asymptotic learning curves with exponents 1 and $\frac{2}{3}$ for unrealizable cases. Although they are unstable RS solutions, the two exponents 1 and $\frac{2}{3}$ are consistent with those obtainable without using the replica method in a two-dimensional version of the present problem as follows. This suggests that our results are good approximations even if they are not exact.

Let us consider the following two-dimensional learning problem. In this problem, the teacher is a two-dimensional non-monotonic perceptrons with the weight $\boldsymbol{w}_0 = (w^0{}_1, w^0{}_2)$ which returns output

$$y = T_a(\boldsymbol{w}_0 \cdot \boldsymbol{x}) \tag{4.1}$$

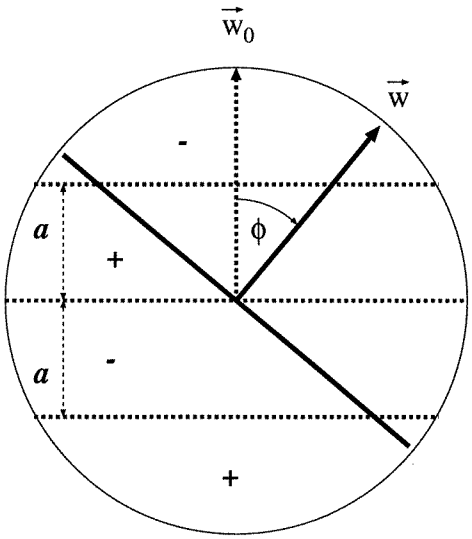

Figure 9. The two-dimensional learning problem. From the two-dimensional nature of the problem, the system is specified by a single parameter ϕ.

for two-dimensional input $\boldsymbol{x}$. Here, $T_a(x)$ is defined as equation (2.2) and it is assumed that $|\boldsymbol{w}_0| = 1$. On the other hand, the student is a two-dimensional simple perceptron with weight $\boldsymbol{w} = (w_1, w_2)$. In order to acquire a good generalization ability, this student learns from a given set of examples $\xi^P = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots (\boldsymbol{x}_P, y_P)\}$ which are assumed to be independently and identically drawn from the two-dimensional Gaussian distribution $\exp[-({x_1}^2 + {x_2}^2)/2]/(2\pi)$, following the error minimum algorithm (figure 9).

From the two-dimensional nature of the problem, the system can be specified by a single parameter ϕ which is the angle between $\boldsymbol{w}_0$ and $\boldsymbol{w}$. In figure 10, we plot the number of false predictions on a realization of ξ^P, E_P, versus ϕ together with its expectation $\langle E_P(\phi)\rangle = P \times \varepsilon(\phi)$ for $a = 0.5$. Here, $\varepsilon(\phi)$ is the generalization error as a function of ϕ. In the figure, we only plot the graphs for positive ϕ because these graphs are statistically symmetric under the reverse operation $\phi \leftrightarrow -\phi$.

From this figure, it is found that $E_P(\phi)$ is minimized around $\phi = 0$, which globally minimizes $\varepsilon(\phi)$. At the same time, $E_P(\phi)$ is locally minimized around $\phi = \phi_* \sim 2.3$[rad], which is the local minimum point of $\varepsilon(\phi)$. Let us estimate how these (local) minimum points fluctuate around $\phi = 0$ or $\phi = \phi_*$ by the following consideration. A similar method was once applied to explain the asymptotic learning curve of a stochastic learning problem [19].

$E_P(\phi)$ is the number of examples which satisfy the condition that $y_\mu = 1$ and $\boldsymbol{w}_0 \cdot \boldsymbol{x}_\mu < 0$ or $y_\mu = -1$ and $\boldsymbol{w}_0 \cdot \boldsymbol{x}_\mu > 0$ ($\mu = 1, 2, \ldots, P$). First, we evaluate how the expectation of $E_P(\phi)$ increases around $\phi = 0$, and $\phi = \phi_*$. From figure 10, we find that this increases as

$$\langle E_P(\phi) - E_P(0)\rangle \sim P \times |\phi| \tag{4.2}$$

around $\phi = 0$. On the other hand, $\langle E_P(\phi)\rangle$ quadratically increases around the local minimum $\phi = \phi_*$, as

$$\langle E_P(\phi) - E_P(\phi_*)\rangle \sim P \times (\phi - \phi_*)^2. \tag{4.3}$$

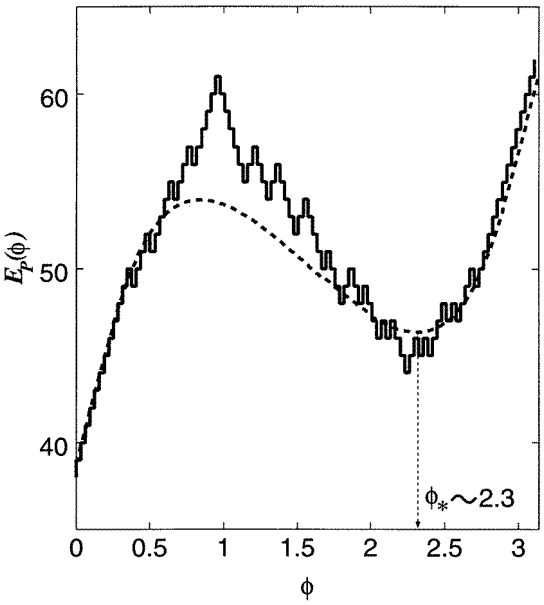

Figure 10. $E_P(\phi)$ for a realization of an example set $\xi^P = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_P, y_P)\}$. This graph is for $a = 0.5$ and $P = 100$. The broken curve represents the expectation $\langle E_P(\phi)\rangle = P \times \varepsilon(\phi)$, where $\varepsilon(\phi)$ is the generalization error for ϕ.

Next, we estimate the fluctuation of equations (4.2) and (4.3). Suppose ϕ moves from $\phi = 0$ to $\phi = \pi$[rad]. Every time the decision boundary of $\boldsymbol{w}$ comes across an input $\boldsymbol{x}_\mu$ $(1 < \mu < P)$, $E_P(\phi)$ increases or decreases discontinuously by 1. Around $\phi = 0$, $E_P(\phi)$ almost always increases because positive and negative examples are clearly separated around the boundary $\boldsymbol{w}_0 \cdot \boldsymbol{x} = 0$. This means that the fluctuation of equation (4.2) is very small and the minimum point fluctuates of the order of P^{-1} which is a rough estimate of the width between two neighbouring examples. Therefore, we obtain

$$\varepsilon - \varepsilon_{\min} \sim \varepsilon(\phi) - \varepsilon(0) \sim |\phi| \sim P^{-1} \tag{4.4}$$

which has the same exponent 1 as that of equation (3.30).

In contrast, $E_P(\phi)$ increases or decreases almost randomly around $\phi = \phi_*$. This motion of $E_P(\phi)$ is analogous to that of a 'random walk' if we regard ϕ as 'time'. From this analogy, we obtain the following relation with respect to the fluctuation of equation (4.3)

$$\Delta(E_P(\phi) - E_P(\phi_*)) \sim \sqrt{P \times |\phi - \phi_*|}. \tag{4.5}$$

The balance between equations (4.3) and (4.5) determines the fluctuation of the (local) minimum point of $E_P(\phi)$ around $\phi = \phi_*$. This gives the scaling $|\phi - \phi_*| \sim P^{-1/3}$, which yields the learning curve

$$\varepsilon - \varepsilon_{l.\min} \sim \varepsilon(\phi) - \varepsilon(\phi_*) \sim (\phi - \phi_*)^2 \sim P^{-2/3} \tag{4.6}$$

which shares the same exponent $\frac{2}{3}$ with equations (3.39) and (3.40).

In order to confirm the above discussion, we performed numerical experiments for $a = 1.0$ (large a) where $\phi = 0$ is the global minimum of $\varepsilon(\phi)$, and for $a = 0.1$ (small a) where $\phi = \phi_* = 2.99$[rad] globally minimizes $\varepsilon(\phi)$. The numerically obtained data exhibits the following behaviours. As the number of examples P increases, the parameter obtained by learning converges to the global minimum of $\varepsilon(\phi)$, namely, to 0 for $a = 1.0$ and to ϕ_*

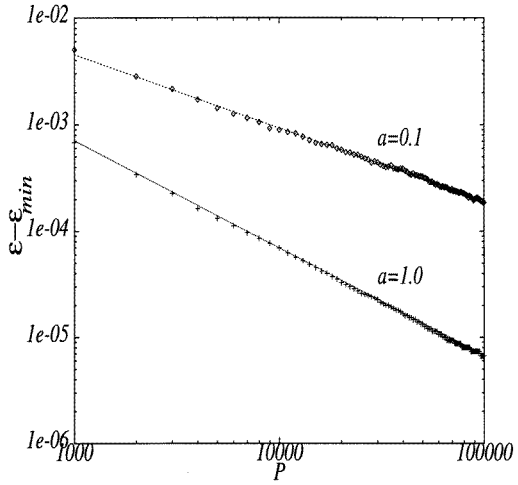

Figure 11. Average of $\varepsilon - \varepsilon_{\min}$ versus the number of examples P for $a = 1.0$ and 0.1. The lines were drawn by the least-square fit under the assumption $\varepsilon - \varepsilon_{\min} \sim P^{-\gamma}$. The obtained exponents are $\gamma = 1.01 \pm 0.01$ for $a = 1.0$ and $\gamma = 0.68 \pm 0.02$ for $a = 0.1$, which are in good agreement with our theoretical predictions $\gamma = 1$ and $\frac{2}{3}$.

for $a = 0.1$. The average generalization error ε taken over 1000 sets of examples is plotted in figure 11. This figure indicates scaling relations $\varepsilon - \varepsilon_{\min} \sim \mathrm{O}(P^{-\gamma})$ for both cases. The exponents obtained from the least-square method are $\gamma = 1.01 \pm 0.01$ for $a = 1.0$ and $\gamma = 0.68 \pm 0.02$ for $a = 0.1$, which are highly consistent with our theoretical predictions $\gamma = 1$ for large a and $\frac{2}{3}$ for small a.

5. Summary

In this paper, we have studied the learning process of a simple perceptron which learns an unrealizable Boolean rule represented by a perceptron with a non-monotonic transfer function of reversed-wedge type. This type of non-monotonic perceptron is considered as a variant of multilayer perceptron and is parametrized by a single parameter a. Reflecting the non-monotonic nature of the target rule, it was found that a discontinuous transition from the poor generalization phase to the good generalization phase takes place when the number of examples P is relatively small compared with that of synaptic weight N for intermediate values of a. We also found that asymptotic learning curves are classified into the following two categories depending on a. For large a, the learning curve obeys the scaling relation with exponent 1. On the other hand, the learning curve with exponent $\frac{2}{3}$ is obtained for small a. These exponents are the same as those found in learning from noisy examples [2, 3]. Therefore, we conjecture that these types of learning curves generally appear in learning of unrealizable Boolean functions independent of the cause of unrealizability.

Acknowledgments

The authours are grateful to H Nishimori for valuable comments and discussions. YK is partially supported by the programme 'Research for the Future (RFTF)' of the Japan Society for the Promotion of Science. He also thanks D Saad and other members of the Neural

Computing Research Group (NCRG) in Aston University for valuable discussions and their hospitality, while he was staying there. JI is partially supported by the Junior Research Associate programme of RIKEN. He also acknowledges C Van den Broeck for stimulus discussions in TANC'97.

Appendix A. Calculation of $\langle\langle \ln Z(\beta)\rangle\rangle_{\xi^P}$

In this appendix, we derive equations (3.3) and (3.18) by the replica method. Under the RS ansatz (3.6) and (3.7), $\langle\langle Z^n(\beta)\rangle\rangle_{\xi^P}$ is evaluated by the SP method with regard to R and q as

$$\begin{aligned}\langle\langle Z^n(\beta)\rangle\rangle_{\xi^P} &= \mathrm{ext}_{\{R,q\}}\left\{\int\prod_{a=0}^{n}\mathrm{d}\boldsymbol{w}_a\prod_{a=1}^{n}\delta(\boldsymbol{w}_a\cdot\boldsymbol{w}_a-N)\prod_{a=1}^{n}\delta(\boldsymbol{w}_0\cdot\boldsymbol{w}_a-NR)\right.\\ &\qquad\left.\times\prod_{a>b}\delta(\boldsymbol{w}_a\cdot\boldsymbol{w}_b-Nq)\left\langle\!\!\left\langle\prod_{a=1}^{n}\prod_{\mu=1}^{P}[\mathrm{e}^{-\beta}+(1-\mathrm{e}^{-\beta})\Theta(y_\mu\cdot u_{a\mu})]\right\rangle\!\!\right\rangle_{\xi^P}\right\}\\ &= \mathrm{ext}_{\{R,q\}}\{A_0(R,q:n)\times A_1(R,q,\beta:n)\} \qquad (A.1)\end{aligned}$$

$$u_{a\mu}\equiv\frac{\boldsymbol{w}_a\cdot\boldsymbol{x}_\mu}{\sqrt{N}}. \qquad (A.2)$$

In the last line of equation (A.1), we defined $A_0(R,q:n)$ and $A_1(R,q,\beta:n)$ as

$$A_0(R,q:n)\equiv\int\prod_{a=0}^{n}\mathrm{d}\boldsymbol{w}_a\prod_{a=1}^{n}\delta(\boldsymbol{w}_a\cdot\boldsymbol{w}_a-N)\prod_{a=1}^{n}\delta(\boldsymbol{w}_0\cdot\boldsymbol{w}_a-NR)\prod_{a>b}\delta(\boldsymbol{w}_a\cdot\boldsymbol{w}_b-Nq) \qquad (A.3)$$

and

$$A_1(R,q,\beta:n)\equiv\left\langle\!\!\left\langle\prod_{a=1}^{n}\prod_{\mu=1}^{P}[\mathrm{e}^{-\beta}+(1-\mathrm{e}^{-\beta})\Theta(y_\mu\cdot u_{a\mu})]\right\rangle\!\!\right\rangle_{\xi^P} \qquad (A.4)$$

respectively.

By using the usual SP method, $A_0(R,q:n)$ is evaluated as

$$A_0(R,q:n)=\left[\left(1+n\frac{q-R^2}{1-q}\right)\times(1-q)^n\right]^{N/2} \qquad (A.5)$$

except for a numerical factor [20]. Next, we evaluate $A_1(R,q,\beta:n)$. Since it is assumed that each $\boldsymbol{x}_\mu$ ($\mu=1,2\ldots P$) is drawn independently and iteratively from an identical distribution, the average with respect to ξ^P in equation (A.4) is replaced by the product of the averages

$$\left\langle\!\!\left\langle\prod_{a=1}^{n}[\mathrm{e}^{-\beta}+(1-\mathrm{e}^{-\beta})\Theta(y_\mu\cdot u_{a\mu})]\right\rangle\!\!\right\rangle_{(\boldsymbol{x}_\mu,y_\mu)} \qquad (A.6)$$

where $\mu=1,\ldots P$. These averages are independent of index μ and therefore we drop μ in the evaluation of equation (A.6) hereafter. For an input $\boldsymbol{x}$, the 'probability' that $y=1$ is returned by the non-monotonic teacher is

$$\begin{aligned}P(y=+1|\boldsymbol{x}) &= \Theta\left(-\frac{\boldsymbol{w}_0\cdot\boldsymbol{x}}{\sqrt{N}}-a\right)+\Theta\left(\frac{\boldsymbol{w}_0\cdot\boldsymbol{x}}{\sqrt{N}}\right)-\Theta\left(\frac{\boldsymbol{w}_0\cdot\boldsymbol{x}}{\sqrt{N}}-a\right)\\ &= \Theta(-u_0-a)+\Theta(u_0)-\Theta(u_0-a)\\ &= 1-P(y=-1|\boldsymbol{x})=P(y=-1|-\boldsymbol{x}). \qquad (A.7)\end{aligned}$$

By taking the average with respect to y first in equation (A.6) using equation (A.7) and taking the symmetry between $\boldsymbol{x}$ and $-\boldsymbol{x}$ into account, we obtain

$$\left\langle\!\!\left\langle 2[\Theta(-u_0-a)+\Theta(u_0)-\Theta(u_0-a)]\prod_{a=1}^{n}[\mathrm{e}^{-\beta}+(1-\mathrm{e}^{-\beta})\Theta(u_a)]\right\rangle\!\!\right\rangle_{\boldsymbol{x}}. \tag{A.8}$$

It should be remarked that under the RS ansatz (3.6) and (3.7), $u_0, u_1, \ldots u_n$ become a set of joint Gaussian random variables which satisfy the condition

$$\langle u_a \cdot u_b\rangle = (1-q)\delta_{ab}+q \qquad \text{for } a,b=1,\ldots,n \tag{A.9}$$

$$\langle u_0 \cdot u_a\rangle = R \qquad \text{for } a=1,\ldots,n \tag{A.10}$$

$$\langle {u_0}^2\rangle = 1 \tag{A.11}$$

when $\boldsymbol{x}$ is uniformly drawn from S^N. Here, $\langle\cdots\rangle$ stands for the statistical average of $\cdots$. These joint Gaussian random variables are represented explicitly by $n+2$ independent Gaussian random variables z_a $(a=0,1,\ldots,n)$ and t which satisfy the condition

$$\langle z_a \cdot z_b\rangle = \delta_{ab} \qquad \text{for } a,b=0,1,\ldots,n \tag{A.12}$$

$$\langle t \cdot z_a\rangle = 0 \qquad \text{for } a=0,1,\ldots,n \tag{A.13}$$

$$\langle t^2\rangle = 1 \tag{A.14}$$

as

$$u_a = \sqrt{1-q}\,z_a+\sqrt{q}\,t \qquad \text{for } a=1,\ldots,n \tag{A.15}$$

$$u_0 = \sqrt{1-\frac{R^2}{q}}\,z_0+\frac{R}{\sqrt{q}}\,t. \tag{A.16}$$

Substituting equations (A.15) and (A.16) into equation (A.8) and taking the average with respect to z_a $(a=0,1,\ldots,n)$ and t instead of $\boldsymbol{x}$, we obtain

$$\begin{aligned}
&\left\langle\!\!\left\langle \prod_{a=1}^{n}[\mathrm{e}^{-\beta}+(1-\mathrm{e}^{-\beta})\Theta(y\cdot u_a)]\right\rangle\!\!\right\rangle_{(\boldsymbol{x},y)} \\
&\quad= 2\int \mathrm{D}t\left[\int \mathrm{D}z_0\left[\Theta\left(-\sqrt{1-\frac{R^2}{q}}\,z_0-\frac{R}{\sqrt{q}}\,t-a\right)\right.\right. \\
&\quad\quad\left.\left.+\Theta\left(\sqrt{1-\frac{R^2}{q}}\,z_0+\frac{R}{\sqrt{q}}\,t\right)-\Theta\left(\sqrt{1-\frac{R^2}{q}}\,z_0+\frac{R}{\sqrt{q}}\,t-a\right)\right]\right] \\
&\quad\quad\times\prod_{a=1}^{n}\int \mathrm{D}z_a\left[\mathrm{e}^{-\beta}+(1-\mathrm{e}^{-\beta})\Theta\left(\sqrt{1-q}\,z_a+\sqrt{q}\,t\right)\right] \\
&\quad= 2\int \mathrm{D}t\;\Omega\left(\frac{R}{\sqrt{q}}:t\right)\{\Xi_\beta(q:t)\}^n
\end{aligned} \tag{A.17}$$

which means

$$A_1(R,q,\beta:n)=\left[2\int \mathrm{D}t\;\Omega\left(\frac{R}{\sqrt{q}}:t\right)\{\Xi_\beta(q:t)\}^n\right]^P. \tag{A.18}$$

For small n, equations (A.5) and (A.18) are expanded with respect to n as

$$A_0(q,R:n)\sim 1+n\times N\times\left[\frac{1}{2}\ln(1-q)+\frac{q-R^2}{2(1-q)}\right]+\mathrm{O}(n^2) \tag{A.19}$$

and

$$A_1(q, R, \beta : n) \sim 1 + n \times 2P \times \left[\int \mathrm{D}t\, \Omega \left(\frac{R}{\sqrt{q}} : t \right) \ln \Xi_\beta(q : t) \} \right] + \mathrm{O}(n^2) \tag{A.20}$$

respectively. From these, we obtain

$$\frac{\langle\langle \ln Z(\beta) \rangle\rangle_{\xi^P}}{N} = \mathrm{ext}_{\{R,q\}} \left\{ 2\alpha \int \mathrm{D}t\, \Omega \left(\frac{R}{\sqrt{q}} : t \right) \ln \Xi_\beta(q : t) + \frac{1}{2} \ln(1-q) + \frac{q - R^2}{2(1-q)} \right\} \tag{A.21}$$

which yields equations (3.3) and (3.18).

Appendix B. Stability of the RS solution

Here, we show that our RS solution is unstable for $\alpha > \alpha_c$. The Almeida–Thouless (AT) stability for the RS solution is judged by the following quantity for any temperature [21, 16],

$$\Lambda_3 = \alpha \lambda_3 \tilde{\lambda_3} - 1 \tag{B.1}$$

where

$$\lambda_3 \equiv \frac{2}{q^2} \int \mathrm{D}t\, \Omega \left(\frac{R}{\sqrt{q}} : t \right) \left[\frac{\partial^2}{\partial t^2} \ln \Xi_\beta(q : t) \right]^2 \tag{B.2}$$

$$\tilde{\lambda_3} \equiv (1-q)^2. \tag{B.3}$$

From equation (3.23), we obtain the following relation for the RS solution

$$\ln \Xi_\beta(q : t) \sim \begin{cases} -\beta & \text{for } t < -\sqrt{2x} \\ -\dfrac{\beta}{2x} t^2 & \text{for } -\sqrt{2x} < t < 0 \\ 0 & \text{for } 0 < t \end{cases} \tag{B.4}$$

where $x = \beta(1-q)$, when β is large and $1-q$ is small. This yields the relation

$$\frac{\partial^2}{\partial t^2} \ln \Xi_\beta(q : t) \sim \beta \left[\sqrt{\frac{2}{x}} \delta(t + \sqrt{2x}) - \frac{1}{x} \left(\Theta \left(t + \sqrt{2x} \right) - \Theta(t) \right) \right] \tag{B.5}$$

which means that Λ_3 is calculated as

$$\Lambda_3 = 2\alpha x^2 \int \Omega(R : t) \left[\sqrt{\frac{2}{x}} \delta \left(t + \sqrt{2x} \right) - \frac{1}{x} \left(\Theta \left(t + \sqrt{2x} \right) - \Theta(t) \right) \right]^2 - 1 \tag{B.6}$$

in the limit $\beta \to \infty$. This diverges to infinity unless x is infinite because the right-hand side of equation (B.6) includes a term such as $\delta^2(t + \sqrt{2x})$ in the integral. For $\alpha > \alpha_c$, x of our RS solution is finite, which means that this solution is thermodynamically unstable.

We should mention that Λ_3 becomes 0 at $\alpha = \alpha_c$. Namely, the RS solution loses the AT stability *just at* α_c. This is explained as follows. By partially integrating the left-hand side of equation (3.15), we obtain

$$2\alpha_c \int_{-\infty}^{0} \mathrm{D}t\, [\Omega(R_c : t) + t\Omega_t(R_c : t)] = 1 - {R_c}^2. \tag{B.7}$$

Adding equation (3.14) to this equation, we obtain the following relation at $\alpha = \alpha_c$

$$2\alpha_c \int_{-\infty}^{0} \mathrm{D}t\, \Omega(R_c : t) = 1. \tag{B.8}$$

Note that $x = \infty$ at $\alpha = \alpha_c$. Then, the right-hand side of equation (B.6) becomes

$$2\alpha_c \int_{-\infty}^{0} \mathrm{D}t\, \Omega(R_c : t) - 1. \tag{B.9}$$

From equation (B.8), this means that $\Lambda_3 = 0$ at $\alpha = \alpha_c$.

References

[1] Watkin T H L, Rau A and Biehl M 1993 *Rev. Mod. Phys.* **65** 499
[2] Uezu T and Kabashima Y 1996 *J. Phys. A: Math. Gen.* **29** L55
[3] Uezu T, Kabashima Y, Nokura K and Nakamura N 1996 *J. Phys. Soc. Japan* **65** 3797
[4] Seung H S, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 6056
[5] Watkin T L H and Rau A 1992 *Phys. Rev.* A **45** 4111
[6] Morita M, Yoshizawa S and Nakano K 1990 *Trans. IEICE* **J73-D-II** 242
[7] Shiino M and Fukai T 1993 *J. Phys. A: Math. Gen.* **26** L831
[8] Nishimori H and Opris I 1993 *Neural Networks* **6** 1061
[9] Inoue J 1996 *J. Phys. A: Math. Gen.* **29** 4815
[10] Boffetta G, Monasson R and Zecchina R 1993 *J. Phys. A: Math. Gen.* **26** L507
[11] Monasson R and O'Kane D 1994 *Europhys. Lett.* **27** 85
[12] Bex G J and Van den Broeck C 1997 *Phys. Rev.* E **56** 870
[13] Engel A and Reimers L 1994 *Europhys. Lett.* **28** 531
[14] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[15] Györgyi G and Tishby N 1990 *Neural Networks and Spin Glasses* (Singapore: World Scientific)
[16] Bouten M 1994 *J. Phys. A: Math. Gen.* **27** 6021
[17] Whyte W and Sherrington D 1996 *J. Phys. A: Math. Gen.* **29** 3036
[18] Opper M and Haussler D 1991 *Proc. 4th ACM Workshop on Computational Learning Theory* (San Mateo: Morgan Kaufmann)
[19] Kabashima Y and Shinomoto S 1992 *Neural Comput.* **4** 712
[20] Opper M and Kinzel W 1996 *Models of Neural Networks* vol III (New York: Springer)
[21] de Almeida J R and Thouless D J 1978 *J. Phys. A: Math. Gen.* **11** 983