Journal of the Physical Society of Japan
Vol. 65, No. 12, December, 1996, pp. 3797-3809

Analysis of Learning from Stochastic Rules in the Framework of Replica Symmetry Breaking

Tatsuya UEZU*, Yoshiyuki KABASHIMA**,
Kazuo NOKURA[1] and Naomi NAKAMURA***

Department of Physics, Nara Women's University, Nara 630
[1]*Shonan Institute of Technology, Fujisawa 251*

(Received April 25, 1996)

We study learning from examples by a perceptron when a return to an example is given by a stochastic relation which is represented by the conditional probability distribution $\mathcal{P}(u)$, where u is proportional to the inner product between the optimal synaptic weight vector and the example vector. The problem is analyzed by replica method in the case of spherical weights. Since the replica symmetric(RS) solution turns out to be unstable for stochastic cases, we consider the one-step replica symmetry breaking (RSB) solution. We investigate the asymptotic behavior of the learning curve as $\alpha \equiv \frac{p}{N} \to \infty$, where p is the number of samples and N is the dimension of the synaptic weights. The average generalization error $\epsilon_{\rm g}$ is expressed as $(\epsilon_{\rm g} - \epsilon_{\rm min}) \propto \alpha^{-\gamma}$ in the asymptotic region. For the minimum-error algorithm, we find $\gamma = \frac{1+\delta}{1+3\delta}$ for the RS solution and $\gamma = \frac{1+\delta}{1+2\delta}$ for the one-step RSB solution in the case of the function $\mathcal{P}(u)$ which is monotonic and expressed as $\mathcal{P}(u) \simeq$const.$+a$ sgn$(u)|u|^\delta$ around $u = 0$. That is, the exponent γ is determined by the local property of $\mathcal{P}(u)$ around $u = 0$. In particular, for the case of $\delta = 1$, the one-step RSB ansatz, unlike the RS ansatz, gives the same result up to a logarithmic correction as previous results obtained by non-replica methods in the one-dimensional model.

KEYWORDS: neural network, learning, perceptron, replica method, one-step RSB ansatz

§1. Introduction

Recently, study on learning in neural networks has attracted many researchers. Since there are many methods of learning, it is important to investigate the efficiency of the learning rules. Several results have been already obtained on such problems. By using the replica method, learning curves or generalization error, which are the probability of a false prediction on a novel example, were calculated for various types of learning machines (for details, see ref. 1)). These studies revealed a rich behavior of learning curves *depending* on the architecture of machines, when the number of examples p is small relative to the number of adjustable machine parameters N.[2-4] However, the most surprising result is the universality of asymptote of learning curves when p/N is large.[5,6] For the cases that the machine parameters are continuous and the target rules are deterministic and realizable by the learning machines, the generalization error $\epsilon_{\rm g}$ obeys the universal scaling relation

$$\epsilon_{\rm g} \sim (p/N)^{-1} \tag{1}$$

independent of the architecture of machines. It should be remarked that the relation (1) can also be derived by methods other than the replica trick, such as the uniform convergence technique in computational learning theory[7,8] and the asymptotic analysis in statistics.[9]

However, it was pointed out by several authors that the scaling relation (1) does not hold when the target rule is unrealizable or seems stochastic to the learner.[2,5,10-13] Among other studies, the model studied by Györgyi and Tishby[5] is worth noting because it is simple and shows a non-trivial behavior of power law different from (1). They considered learning from examples generated by a perceptron. A learner is a perceptron with coupling weight $\boldsymbol{J}$. Its output R for an input $\boldsymbol{\xi}$ is,

$$R = \mathrm{sgn}\left[\frac{1}{\sqrt{N}}(\boldsymbol{J}\cdot\boldsymbol{\xi})\right],$$

where $\boldsymbol{J}$ and $\boldsymbol{\xi}$ are N-dimensional vectors. The target rule is also a perceptron with coupling weight $\boldsymbol{J}'$, whose input $\boldsymbol{\xi}$ is subject to an external noise $\boldsymbol{\eta}$. The output R' is,

$$R' = \mathrm{sgn}\left[\frac{1}{\sqrt{N}}(\boldsymbol{J}'\cdot(\boldsymbol{\xi}+\boldsymbol{\eta}))\right]. \tag{2}$$

$\boldsymbol{\eta}$ is a Gaussian random variable of zero mean. When the training set

$$\{\boldsymbol{\xi}^\mu, R'^\mu\}, \mu = 1, \cdots, p,$$

is given, learning occurs by comparing the output sequence R^μ with the reference output R'^μ. Treating this problem by the replica method, they obtained as asymptote

$$\epsilon_{\rm g} - \epsilon_{\rm min} \sim (p/N)^{-1/2}, \tag{3}$$

* E-mail: uezu@cc.nara-wu.ac.jp
** Present address: Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama 226. E-mail: kaba@dis.titech.ac.jp
*** Present address: Yokohama Works, Sumitomo Electric Industries, Ltd., Yokohama 244.

where $\epsilon_{\min}$ is the minimum value of the generalization error.

On the other hand, Kabashima and Shinomoto[12] studied a one-dimensional problem which is considered to be a one-dimensional version of Györgyi and Tishby's problem. In their model, the target rule is a stochastic relation $p_r(y|x)$ between a real number input $x \in [0,1]$ and a binary output $y \in \{-1,+1\}$. For a novel input x, a learner's return is $y = \mathrm{sgn}(x-\theta)$, and the learner's task is to decide the threshold θ so as to minimize the generalization error for a given set of examples. They obtained the following asymptotic behavior of the learning curve by transforming the problem into a one-dimensional random walk,

$$\epsilon_{\mathrm{g}} - \epsilon_{\min} \sim p^{-2/3}. \tag{4}$$

Kim and Pollard[13] obtained the same result as (4) in the field of statistics.

Thus, we have two different scalings, (3) and (4), for similar problems. Now, let us examine the above results a little more closely. As for Györgyi-Tishby's result, it concerns the RS solution which is shown to be unstable, although they conjectured that the RS ansatz is a good approximation because their result is identical with that of the worst case analysis by Haussler *et al.*[10] As for Kabashima-Shinomoto's approach, it can not be applied to more than one dimension because the analysis depends strongly on the one-dimensional nature of the problem. Therefore, it is not clear what the scaling of the generalization error in higher dimensional space is.

In this paper, we study this problem by investigating the direct extension of Kabashima-Shinomoto's problem to higher dimensions and derive the asymptotic behavior of the learning curve. The model of Györgyi and Tishby is included in our model.

Let us describe our problem. We consider a stochastic relation between an N-dimensional input vector $\boldsymbol{\xi}$ and its output $R \in \{1,-1\}$ which is represented by a conditional probability distribution $p_r(+1|\boldsymbol{\xi}) = 1 - p_r(-1|\boldsymbol{\xi})$. We assume the following form of the probability

$$p_r(+1|\boldsymbol{\xi}) = \mathcal{P}(u_0) = \frac{1+P(u_0)}{2},$$
$$u_0 \equiv (\boldsymbol{\xi}\cdot\boldsymbol{J}^0)/\sqrt{N}.$$

Furthermore, we assume that $\mathcal{P}(u)$ is an increasing function of u. Then, the prediction $R = \mathrm{sgn}(u_0)$, which is the return of the perceptron with the weight vector $\boldsymbol{J}^0$, minimizes the generalization error. That is, $\boldsymbol{J}^0$ is the optimal weight. We consider the learner to be a perceptron with a synaptic weight $\boldsymbol{J}$, and the learning strategy to be the minimum-error algorithm, which minimizes the training error, i.e. the number of false predictions on training examples. Under these assumptions, we study the asymptotic behavior of the leaning curve of the minimum-error algorithm using the replica method. Since the RS solution turns out to be unstable when $\alpha \equiv \frac{p}{N} \to \infty$, we proceed to the one-step RSB calculation.

In this paper, we treat the function $P(u)$ which behaves as $P(u) \simeq a\ \mathrm{sgn}(u)|u|^\delta$ around $u=0$ in order to clarify what determines the power of the decay of the learning curve. We denote this family as $P_\delta(u)$. In the model studied by Györgyi and Tishby,[5] an input $\boldsymbol{\xi}$ is subject to an external noise $\boldsymbol{\eta}$ and the conditional probability is represented as $p_r(+1|\boldsymbol{\xi}) = H[-(\boldsymbol{\xi}\cdot\boldsymbol{J}^0)/\sqrt{\langle(\boldsymbol{\eta}\cdot\boldsymbol{J}^0)^2\rangle}]$ where $H(u) = \int_u^\infty \mathrm{d}x \exp[-x^2/2]/\sqrt{2\pi}$ and $\langle\cdots\rangle$ represents the average over $\boldsymbol{\eta}$. Since an input is subject to a noise, this model is called an input noise model and corresponds to the case of $\delta = 1$ in our model. Another typical model is the output noise model studied by Opper and Haussler.[6] In their model, the target rule is a perceptron the sign of whose output is reversed by noise with a probability λ. Such a rule is represented by the conditional probability $p_r(+1|\boldsymbol{\xi}) = \lambda + (1-2\lambda)\Theta[(\boldsymbol{\xi}\cdot\boldsymbol{J}^0)/\sqrt{N}]$, where Θ is the Heaviside function. The probability is rewritten as $p_r(+1|\boldsymbol{\xi}) = (1 + k\ \mathrm{sgn}[(\boldsymbol{\xi}\cdot\boldsymbol{J}^0)/\sqrt{N}])/2$ with $k = 1-2\lambda$. This corresponds to the case of $\delta = 0$ in our model. In particular, the case of k=1 corresponds to the existence of a deterministic relation between input and output.

Thus, with appropriate choice of the parameter δ our model includes the above typical models of learning disturbed by noise.

The results obtained in this paper are as follows. We calculate the learning curve for $P_\delta(u)$ where $\delta \geq 0$. For the RS solution, the learning curve scales as

$$\epsilon_{\mathrm{g}} - \epsilon_{\min} \sim (p/N)^{-\frac{1+\delta}{1+3\delta}}, \tag{5}$$

for $\delta \geq 0$. We show that this solution becomes unstable at a finite value of $\alpha \equiv \frac{p}{N}$, α_{AT}, and that α_{AT} coincides with the critical capacity α_c. The results obtained for the one-step RSB solution are for $\delta > 0$

$$\epsilon_{\mathrm{g}} - \epsilon_{\min} \sim (p/N)^{-\frac{1+\delta}{1+2\delta}}\left(\ln\frac{p}{N}\right)^{\frac{1+\delta}{2(1+2\delta)}}, \tag{6}$$

and for $\delta = 0$

$$\epsilon_{\mathrm{g}} - \epsilon_{\min} \sim (p/N)^{-1}. \tag{7}$$

Expressing the average generalization error for large α as

$$\epsilon_{\mathrm{g}} - \epsilon_{\min} \sim \alpha^{-\gamma},$$

in both the RS and the RSB solutions, the exponent γ is determined by the index δ which characterizes the function $P_\delta(u)$ around $u = 0$.

The asymptotic forms (3) and (4) obtained previously are for the case of $\delta = 1$. The scaling (3) is for the infinite dimensional model and (4) is for the one-dimensional model. Substituting $\delta = 1$ into the eqs. (5) and (6) of our results, we obtain $\gamma = 1/2$ for the RS solution and $\gamma = 2/3$ for the RSB solution. Thus, our result of $\gamma = 2/3$ for the RSB solution, unlike $\gamma = 1/2$ for the unstable RS solution, agrees with that of Kabashima-Shinomoto and Kim-Pollard up to a logarithmic correction. That is, the scaling (4) is also valid for the infinite dimensional model. In the case of $\delta = 0$, we obtain the exponent $\gamma = 1$ for both the RS and the RSB solutions. This result is the same as Opper and Haussler's result which was obtained in their model for a special temperature within the RS ansatz.

The construction of the paper is as follows. In §2, we formulate our problem and derive the expression of the generalization error for general P. In §3, the RS solution and their asymptotic behavior are studied and then the generalization errors are estimated. The one-step RSB solution is studied in §4 and the asymptotic expression of the generalization error is derived. Section 5 is devoted to the summary and discussions.

§2. Formulation

The stochastic relation we consider is the following. Examples $\boldsymbol{\xi}$s are drawn independently from a uniform distribution on an N-dimensional sphere with radius $\sqrt{N}$. Then an output R^μ for the μ-th example $\boldsymbol{\xi}^\mu$ is $+1$ or -1 and is given according to the following conditional probability. The probability of $R^\mu = 1$ is

$$p_r(+1|\boldsymbol{\xi}^\mu) = \mathcal{P}(u_0^\mu) = \frac{1+P(u_0^\mu)}{2}, \tag{8}$$
$$u_0^\mu \equiv (\boldsymbol{\xi}^\mu \cdot \boldsymbol{J}^0)/\sqrt{N}.$$

The output of a learner for $\boldsymbol{\xi}^\mu$ is

$$\mathrm{sgn}(u^\mu), \quad u^\mu \equiv (\boldsymbol{\xi}^\mu \cdot \boldsymbol{J})/\sqrt{N}.$$

The optimal weight $\boldsymbol{J}^0$ and the weight of the learner $\boldsymbol{J}$ are N-dimensional vectors normalized to $\sqrt{N}$.

As the learning strategy, in this paper we adopt the minimum-error algorithm. When the training set

$$\{\boldsymbol{\xi}^\mu, R^\mu\}, \quad \mu = 1, \cdots, p,$$

is given, the number of errors for $\boldsymbol{J}$ is the number of different outputs between $\mathrm{sgn}(u^\mu)$ and R^μ. In the minimum-error algorithm, $\boldsymbol{J}$ is determined so as to minimize the number of errors. Therefore, we investigate the weights of learners whose numbers of errors are minimum.

When the parameter $\alpha \equiv p/N$ is changed, the learning behavior becomes different above and below some critical value α_c. Above α_c the average number of errors changes from zero to non-zero.

For $\alpha < \alpha_c$, we consider the volume of weights of learners whose errors are zero. Let V be the ratio of this volume to the total volume of weights.[15)]

$$\begin{aligned} V &= \int \mathcal{D}(\boldsymbol{J}) \prod_{\mu=1}^{p} \Theta(R^\mu u^\mu) \Big/ \int \mathcal{D}(\boldsymbol{J}), \\ &= \int \mathcal{D}(\boldsymbol{J}) \prod_{\mu=1}^{p} \int_0^\infty \mathrm{d}\lambda^\mu \int_{-\infty}^{\infty} \frac{\mathrm{d}x^\mu}{2\pi} \\ &\quad \times \exp\left(-\mathrm{i}x^\mu (R^\mu u^\mu - \lambda^\mu)\right)/S_N, \end{aligned} \tag{9}$$

$$\mathcal{D}(\boldsymbol{J}) \equiv \left(\prod_{j=1}^{N} \mathrm{d}J_j\right) \delta\left(\sum_{j=1}^{N} J_j^2 - N\right), \quad S_N = \int \mathcal{D}(\boldsymbol{J}),$$

where we used the following formula,

$$\Theta(u) = \int_0^\infty \mathrm{d}\lambda \int_{-\infty}^{\infty} \frac{\mathrm{d}x}{2\pi} \exp\left(-\mathrm{i}x(u-\lambda)\right).$$

Using the standard replica method and assuming the self-averaging, $\ln V$ is calculated as follows.

$$\ln V = \langle \ln V \rangle_{R,\xi,J^0} = \lim_{n\to 0} \frac{\langle V^n \rangle_{R,\xi,J^0} - 1}{n}.$$

Introducing n replicas and taking sample averages over R, $\boldsymbol{\xi}$, and $\boldsymbol{J}^0$, we obtain the expression for $\langle V^n \rangle_{R,\xi,J^0}$ as the function of $q^{\alpha\beta}$, M^α, etc., where $q^{\alpha\beta} = \frac{1}{N}\sum_{j=1}^{N} J_j^\alpha J_j^\beta$ is the overlap between weights of learners and $M^\alpha = \frac{1}{N}\sum_{j=1}^{N} J_j^\alpha J_j^0$ is the overlap between the weight of a learner and the optimal weight. (See Appendix A.)

For $\alpha > \alpha_c$, we have to choose weights with minimum errors, since we adopt the minimum-error algorithm. On that account we introduce an energy as the number of errors, and introduce the temperature T and then take the limit $T \to 0$. That is, the energy of the learner is,

$$E_p[J,\xi] = \sum_{\mu=1}^{p} \Theta(-R^\mu u^\mu). \tag{10}$$

Then the partition function Z is,

$$\begin{aligned} Z &= \int \mathcal{D}(\boldsymbol{J}) \mathrm{e}^{-E_p[J,\xi]/T} = \int \mathcal{D}(\boldsymbol{J}) \prod_{\mu=1}^{p} \mathrm{e}^{-\beta\Theta(-R^\mu u^\mu)}, \\ &= \int \mathcal{D}(\boldsymbol{J}) \prod_{\mu=1}^{p} \left[\mathrm{e}^{-\beta}\int_{-\infty}^{0} \mathrm{d}\lambda^\mu + \int_0^\infty \mathrm{d}\lambda^\mu\right] \int_{-\infty}^{\infty} \frac{\mathrm{d}x^\mu}{2\pi} \exp\left(-\mathrm{i}x^\mu (R^\mu u^\mu - \lambda^\mu)\right), \end{aligned} \tag{11}$$

where $\beta = 1/T$. For $\alpha < \alpha_c$, $Z(T \to 0)/S_N$ coincides to V. For $\alpha > \alpha_c$ to estimate the fraction of volume of weights with minimum energy, we have to calculate the minimum error fraction f,[16)] which is defined as

$$f = -\lim_{\beta\to\infty} \frac{\langle \ln Z \rangle_{R,\xi,J^0}}{N\alpha\beta}. \tag{12}$$

Thus, in order to obtain f we calculate $\langle \ln Z \rangle_{R,\xi,J^0}$ using the replica method. Comparing the eqs. (9) and (11) we note that the expression of $\langle Z^n \rangle_{R,\xi,J^0}$ is obtained by the following replacement in that of $\langle V^n \rangle_{R,\xi,J^0}$,

$$\int_0^\infty \mathrm{d}\lambda^\alpha \longrightarrow \mathrm{e}^{-\beta}\int_{-\infty}^{0} \mathrm{d}\lambda^\alpha + \int_0^\infty \mathrm{d}\lambda^\alpha.$$

The free energy per synapse is expressed as $F = \alpha f$.

We evaluate the integrand of the expression of G by the saddle-point method assuming some symmetries of solutions.

Now, we define the generalization error ϵ_g. Let us consider the probability of false prediction on a novel example. ϵ_g is the expectation value of this probability,

$$\epsilon_g = \langle\!\langle \mathcal{P}(u_0)(1-\Theta(u)) + (1-\mathcal{P}(u_0))\Theta(u) \rangle\!\rangle. \tag{13}$$

$\langle\!\langle \cdot \rangle\!\rangle$ denotes the average over samples. Since u_0 and u

are sums of N independent random variables and the averages of u_0^2, u^2 and $u_0 u$ are 1, 1, and M respectively, the distribution function of (u_0, u), $Prob(u_0, u)$, becomes the Gaussian distribution,

$$Prob(u_0, u) = \frac{1}{2\pi\sqrt{1-M^2}} \times \exp[-(u_0^2 - 2Mu_0u + u^2)/2(1-M^2)],$$

where M is the overlap between the optimal weight and the weight of the learner. Thus, we obtain

$$\begin{aligned}\epsilon_g &= \int du_0 \int du Prob(u_0, u)\{\mathcal{P}(u_0)(1-\Theta(u)) \\ &\quad + (1-\mathcal{P}(u_0))\Theta(u)\}, \\ &= \epsilon_{\min} + \int_0^\infty \mathrm{D}y H(y/\zeta)[P(y) - P(-y)], \quad (14)\end{aligned}$$

$$\begin{aligned}\epsilon_{\min} &= \frac{1}{2} - \frac{1}{2}\int_0^\infty \mathrm{D}y[P(y) - P(-y)], \\ &= \int_0^\infty \mathrm{D}y(1-\mathcal{P}(y)) + \int_{-\infty}^0 \mathrm{D}y\mathcal{P}(y), \quad (15)\end{aligned}$$

where $\mathrm{D}y = \exp(-y^2/2)\mathrm{d}y/\sqrt{2\pi}$, $\zeta = \frac{\sqrt{1-M^2}}{M}$. These are exact expressions. $\epsilon_{\min}$ is the minimum value of the generalization error obtained by the optimal weight. Hereafter, we assume the symmetry $P(-y) = -P(y)$ for simplicity. For the family $P_\delta(y)$, which behaves as $P_\delta(y) \simeq a\ \mathrm{sgn}(y)|y|^\delta$ around $y = 0$, and for $M \sim 1 (\zeta \sim 0)$, we get the approximate formula,

$$\epsilon_g \simeq \epsilon_{\min} + \frac{2s}{\sqrt{2\pi}(1+\delta)}\zeta^{1+\delta}, \quad (16)$$

$$s \equiv a\int_0^\infty \mathrm{D}y y^{1+\delta}.$$

$\epsilon_g - \epsilon_{\min}$ is expected to become smaller as the number of examples increases. We study the behavior of $\epsilon_g - \epsilon_{\min}$ as α becomes large. Concretely, we put for $\alpha \gg 1$,

$$\epsilon_g = \epsilon_{\min} + \epsilon_0 \alpha^{-\gamma}. \quad (17)$$

In the subsequent sections, we calculate the the exponent γ and the coefficient ϵ_0 using the statistical mechanical method.

§3. RS Solution

In this section, we study the RS solution which has the highest symmetry in the replica space. We solve saddle-point equations in the large α limit.

Assuming that all quantities are independent of replica indices, especially putting $q^{\alpha\beta} = q$ and $M^\alpha = M$, the minimum error fraction f is obtained as follows. (See Appendix A.)

$$f(\alpha, \zeta, z) = -\frac{1}{\alpha z^2} - \frac{1}{2\alpha\beta}\ln 2\pi(1-q) + f_1, \quad (18)$$

$$f_1 = \int \mathrm{D}y 2\mathcal{P}(y)\left[H(y/\zeta + z) + \frac{1}{z^2}\int_{y/\zeta}^{y/\zeta+z} \times \mathrm{D}u(u - y/\zeta)^2\right], \quad (19)$$

where $z = \frac{x}{\sqrt{1-M^2}}$ and $\zeta = \frac{\sqrt{1-M^2}}{M}$. Here, we took the limits $\beta \to \infty$ and $q \to 1$ fixing $x \equiv \sqrt{2\beta(1-q)}$ to obtain non-trivial results. Then the saddle-point equations for ζ and z are

$$\int \mathrm{D}y\ 2\mathcal{P}(y)y\int_{y/\zeta}^{y/\zeta+z} \mathrm{D}u(u - y/\zeta) = 0, \quad (20)$$

$$\int \mathrm{D}y\ 2\mathcal{P}(y)\int_{y/\zeta}^{y/\zeta+z} \mathrm{D}u(u - y/\zeta)^2 = \alpha^{-1}. \quad (21)$$

We solved numerically these eqs. (20) and (21) for several functions $\mathcal{P}(y)$. In all cases, we found $M \to 1$ as $\alpha \to \infty$, that is the learning succeeds. Therefore, hereafter we assume this behavior. Furthermore, we found that the behavior of z is different in the two cases of $\delta > 0$ and $\delta = 0$. In the former case z tends to zero while in the latter case it tends to a non-zero value. A similar situation is met in the analysis of the RSB solutions. Thus, we have to make different treatments in the two cases and will consider them separately.

Case 1. $P(y)$ is continuous $(\delta > 0)$.

Case 2. $P(y)$ is discontinuous at $y = 0$ $(\delta = 0)$.

For later use, we rewrite eqs. (20) and (21) as follows.

$$\frac{\zeta^2}{1+\zeta^2}\int_0^z \mathrm{D}t\ t\int \mathrm{D}y 2\mathcal{P}\left(\frac{\zeta y}{\sqrt{1+\zeta^2}}\right) y\exp\left(-\frac{ty}{\sqrt{1+\zeta^2}}\right) = 0, \quad (22)$$

$$\frac{\zeta}{\sqrt{1+\zeta^2}}\int_0^z \mathrm{D}t\ t^2\int \mathrm{D}y 2\mathcal{P}\left(\frac{\zeta y}{\sqrt{1+\zeta^2}}\right) \exp\left(-\frac{ty}{\sqrt{1+\zeta^2}}\right) = \alpha^{-1}. \quad (23)$$

Case 1. $P(y)$ is continuous $(\delta > 0)$.

We treat the family $P_\delta(y)$ for $\delta > 0$. Then, taking the lowest orders in eqs. (22) and (23), we obtain

$$\frac{z^3}{3\sqrt{2\pi}} \simeq a\zeta^\delta L(z), \quad (24)$$

$$\zeta\frac{z^3}{3\sqrt{2\pi}} \simeq \alpha^{-1}, \quad (25)$$

$$L(z) \equiv \int_0^z \mathrm{D}t\ t\int_0^\infty \mathrm{D}y\ y^{1+\delta}(\mathrm{e}^{-ty} + \mathrm{e}^{ty}).$$

From eq. (24), $z \sim 0$ follows, so, expanding $L(z)$ by z, this equation becomes

$$z \simeq 3s\zeta^\delta. \quad (26)$$

Thus, we finally have

$$z \simeq (3\sqrt{2\pi}\eta)^{\frac{1}{3}}\alpha^{-\frac{\delta}{1+3\delta}} \propto \alpha^{-\frac{\delta}{1+3\delta}}, \quad \zeta \simeq \eta^{-1}\alpha^{-\frac{1}{1+3\delta}}, \quad (27)$$

$$\eta \equiv \left(\frac{9s^3}{\sqrt{2\pi}}\right)^{\frac{1}{1+3\delta}}.$$

Case 2. $P(y)$ is discontinuous at $y=0$ $(\delta=0)$.

Since $\zeta \to 0$ as $\alpha \to \infty$, we may replace $\mathcal{P}(\frac{\zeta y}{\sqrt{1+\zeta^2}})$ by $\mathcal{P}(0+)$ or $\mathcal{P}(0-)$ in integrations in eqs. (22) and (23). Then, for the lowest order we have

$$\int_0^z \mathrm{D}t\, t \int_0^\infty \mathrm{D}y\, y[(1+k)\mathrm{e}^{-ty} - (1-k)\mathrm{e}^{ty}] \simeq 0, \quad (28)$$

$$\zeta \int_0^z \mathrm{D}t\, t^2 \int_0^\infty \mathrm{D}y[(1+k)\mathrm{e}^{-ty} + (1-k)\mathrm{e}^{ty}] \simeq \alpha^{-1}, \quad (29)$$

where $k \equiv \lim_{y\searrow 0} P(y)$. We assume $0<k<1$ to exclude the deterministic case of $k=1$. Executing integrations we obtain the following equations,

$$\mathcal{F}(z) \simeq 0, \quad (30)$$

$$\zeta \frac{k}{\pi}\left(1-\exp\left[-\frac{z^2}{2}\right]\right) \simeq \alpha^{-1}, \quad (31)$$

where $\mathcal{F}(z)$ is given by

$$\mathcal{F}(z) \equiv \frac{k}{\pi} - \frac{z^3}{\sqrt{2\pi}}(1-k) + \frac{k}{\pi}\mathrm{e}^{-\frac{z^2}{2}}(z^2-1) - \frac{2k}{\sqrt{2\pi}}z^3 H(z).$$

Thus, asymptotic solutions are

$$\zeta \simeq \frac{\pi}{k}\left(1-\exp\left[-\frac{z^{*2}}{2}\right]\right)^{-1}\alpha^{-1} \propto \alpha^{-1}, \quad (32)$$

$$z \simeq z^*, \quad (33)$$

where $z^*(>0)$ is the solution of the equation $\mathcal{F}(z^*)=0$.

Here, let us calculate the generalization error.

Case 1.

For $P_\delta(y)$, from eq. (16) in §2

$$\epsilon_{\mathrm{g}} - \epsilon_{\min} \simeq \frac{2s}{\sqrt{2\pi}(1+\delta)}\zeta^{1+\delta} \simeq \epsilon_0 \alpha^{-\frac{1+\delta}{1+3\delta}}, \quad (34)$$

$$\epsilon_0 = \frac{2s}{\sqrt{2\pi}(1+\delta)}\eta^{-(1+\delta)}.$$

Case 2.

From eq. (14) in §2

$$\epsilon_{\mathrm{g}} - \epsilon_{\min} \simeq \frac{k}{\pi}\zeta \simeq \left(1-\exp\left[-\frac{z^{*2}}{2}\right]\right)^{-1}\alpha^{-1} \propto \alpha^{-1}. \quad (35)$$

Substituting $\delta=1$ into the relation (34), $\epsilon_{\mathrm{g}} - \epsilon_{\min} \propto \alpha^{-\frac{1}{2}}$ follows. This result agrees with that of Györgyi and Tishby for the unstable RS solution. However, this is different from the result $\gamma = \frac{2}{3}$ by Kabashima-Shinomoto and Kim-Pollard. This suggests that the RS solution is inadequate for the large α limit. In fact, we can show by a similar analysis to Bouten[17] that α_{AT}, where the AT instability takes place, is finite and equal to the critical capacity α_c. (See Appendix B.) Thus, it turns out that the RS solution is unstable in the limit of $\alpha \to \infty$ in both cases 1 and 2 unlike the deterministic case of $P(y) = \mathrm{sgn}(y)$ in which the RS solution is stable for any α. Then, we have to consider the RSB solution.

§4. One-Step RSB Solution

In this section, we consider the following one-step replica symmetry breaking solution.

$$\left.\begin{array}{l} q^{\alpha\beta} = q_0 \\ F^{\alpha\beta} = \hat{q}_0 \end{array}\right\} \text{ for } I\left(\frac{\alpha}{m}\right) \neq I\left(\frac{\beta}{m}\right),$$

$$\left.\begin{array}{l} q^{\alpha\beta} = q_1 \\ F^{\alpha\beta} = \hat{q}_1 \end{array}\right\} \text{ for } I\left(\frac{\alpha}{m}\right) = I\left(\frac{\beta}{m}\right) \text{ and } \alpha \neq \beta,$$

where $I(x)$ is the Gauss function. Furthermore, we put

$$M^\alpha = M,$$
$$\overline{M}^\alpha = \overline{M},$$
$$E^\alpha = E, \text{ etc.}$$

Eliminating other variables, we obtain G as the function of q_0, q_1, M, m and β.

$$\frac{G_{\mathrm{RSB}}}{n} = \frac{1}{2} + \frac{\varphi}{2}(q_0 - M^2) + \frac{1}{2}\ln 2\pi(1-q_1) - \frac{1}{2m}\ln\varphi(1-q_1) + \frac{\alpha}{n}G_1, \quad (36)$$

$$\varphi \equiv [m(1-q_0) + (1-m)(1-q_1)]^{-1},$$

$$\frac{1}{n}G_1 = \frac{1}{m}\int \mathrm{D}y 2\mathcal{P}(y) \int \mathrm{D}z_0 \ln\{\int \mathrm{D}z_1 \tilde{H}(Y_1)^m\}, \quad (37)$$

$$Y_1 = \frac{\sqrt{q_0 - M^2}z_0 + \sqrt{q_1 - q_0}z_1 - My}{\sqrt{1-q_1}}.$$

To obtain the minimum error fraction f, the limit $\beta \to \infty$ is taken. Further, we take the limits $q_1 \to 1$ and $m \to 0$ fixing $x \equiv \sqrt{2\beta(1-q_1)}$ and $c \equiv \frac{m}{1-q_1}$ to obtain non-trivial results. Then, f is obtained as follows.

$$f(Q,r,\xi,z') = -\frac{1}{z'^2}\left[\alpha^{-1}\left(\frac{\xi^2}{Q^2} - \frac{2\ln r}{1-r^2}\right) + \frac{2K}{1-r^2}\right], \quad (38)$$

$$K = Q\int \tilde{\mathrm{D}}u \ln \hat{H}(u) E(u),$$

$$\tilde{\mathrm{D}}u = \frac{\mathrm{d}u}{\sqrt{2\pi}}\exp\left(-\frac{Q^2u^2}{2}\right),$$

$$E(u) = \int \mathrm{D}y 2\mathcal{P}(-\Lambda) = 1 - \int \mathrm{D}y P(\Lambda),$$

$$\Lambda = \xi y + \sqrt{1-\xi^2}Qu,$$

$$Q \equiv \sqrt{\frac{1-q_0}{q_0}}, \quad r \equiv \frac{1}{\sqrt{1+c(1-q_0)}},$$

$$\xi \equiv \sqrt{1-\frac{M^2}{q_0}}, \quad z' \equiv \frac{x}{r\sqrt{1-q_0}},$$

$$\hat{H}(u) = H(u) + \mathrm{e}^{-\frac{1-r^2}{2}z'^2}H(-u+rz') + r\mathrm{e}^{-\frac{1-r^2}{2}u^2}\{H(ru-z') - H(ru)\}.$$

In the above expressions, we transformed variables from x, c, q_0 and M to Q, ξ, r and z' since they are easy to handle. We used the following formula for any functions $\mathcal{P}$ and F and any real constants a and b,

$$\int \mathrm{D}y\, \mathcal{P}(y) \int \mathrm{D}t\, F(ay+bt) = \int \mathrm{D}u\, F(\sqrt{a^2+b^2}\, u) \times \int \mathrm{D}y\, \mathcal{P}\left(\frac{by+au}{\sqrt{a^2+b^2}}\right).$$

Here, for later use, let us define χ as $\frac{\xi}{Q}$,

$$\chi \equiv \frac{\xi}{Q}.$$

After a little algebra, the following saddle-point equations for Q, ξ, r and z' are derived.

$$\alpha^{-1}\chi^2 = -Q\int \tilde{\mathrm{D}}u u E(u)\frac{\mathcal{L}(u)}{\hat{H}(u)}, \tag{39}$$

$$\alpha^{-1}\chi^2 = -\frac{\xi}{\sqrt{1-\xi^2}}\int \tilde{\mathrm{D}}u w(u)\frac{\mathcal{L}(u)}{\hat{H}(u)}, \tag{40}$$

$$\alpha^{-1}\left(2r\ln r + \frac{1-r^2}{r}\right) = Q\int \tilde{\mathrm{D}}u E(u) g(u), \tag{41}$$

$$\alpha^{-1}\{\chi^2(1-r^2) - 2\ln r\} = -Q\int \tilde{\mathrm{D}}u E(u)\psi(u), \tag{42}$$

$$w(u) = \int \mathrm{D}y y P(\Lambda),$$

$$\mathcal{L}(u) = \frac{\hat{H}'(u)}{1-r^2} = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(z'^2-2rz'u+u^2)} - \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}u^2}$$
$$- ru\mathrm{e}^{-\frac{1-r^2}{2}u^2}\{H(ru-z') - H(ru)\},$$

$$g(u) = \left\{2r + (1-r^2)\frac{\partial}{\partial r}\right\} \ln \hat{H}(u),$$

$$\psi(u) = \left(2 - z'\frac{\partial}{\partial z'}\right) \ln \hat{H}(u),$$

$$= 2\ln \hat{H}(u) + (1-r^2)z'^2\mathrm{e}^{-\frac{1-r^2}{2}z'^2}\frac{H(rz'-u)}{\hat{H}(u)}.$$

Next, we study the asymptotic behavior as $\alpha \to \infty$.

Similarly to the case of the RS solution, we assume $M \to 1$ as $\alpha \to \infty$ because this behavior was observed for all functions we treated numerically. Since $M \leq q_0$, $Q \to 0$ and $\xi \to 0$ follow as $M \to 1$. Now, we discuss the cases 1 and 2 separately.

Case 1.

For several functions $P(y)$, we solved the saddle-point equations (39)–(42) numerically and obtained only one solution. In Fig. 1, as an example, we show a bifurcation diagram for the RS and the RSB solutions as functions of α. Since numerical results for several $P(y)$ show that z', r, Q and ξ tend to 0 and $\chi = \frac{\xi}{Q}$ tends to ∞ as $\alpha \to \infty$, we assume these results. That is, we treat the family of continuous functions for which z', r, Q and ξ tend to 0 and χ tends to ∞ as $\alpha \to \infty$.

Expanding f in terms of z', r, ξ and Q, we obtain the following asymptotic form of f. (See Appendix C.)

$$f \simeq \epsilon_{\min} - \frac{\alpha^{-1}}{z'^2}\left(\frac{\xi^2}{Q^2} - 2\ln r\right) - \frac{2rz'Q}{3\sqrt{2\pi}} - \frac{Qz'^2}{4\sqrt{2\pi}}$$
$$+ \frac{2s\left(\dfrac{Q^2+\xi^2}{1-\xi^2}\right)^{(1+\delta)/2}}{\sqrt{2\pi}(1+\delta)}. \tag{43}$$

Then, the saddle-point equations are

$$\alpha^{-1}\chi^2 \simeq \frac{1}{\sqrt{2\pi}}Qz'^2\left[\frac{rz'}{3} + \frac{z'^2}{8\sqrt{\pi}} - sQ\xi^{\delta-1}\right], \tag{44}$$

$$\alpha^{-1}\chi^2 \simeq \frac{s}{\sqrt{2\pi}}z'^2\xi^{\delta+1}, \tag{45}$$

$$\alpha^{-1} \simeq \frac{1}{3\sqrt{2\pi}}Qrz'^3, \tag{46}$$

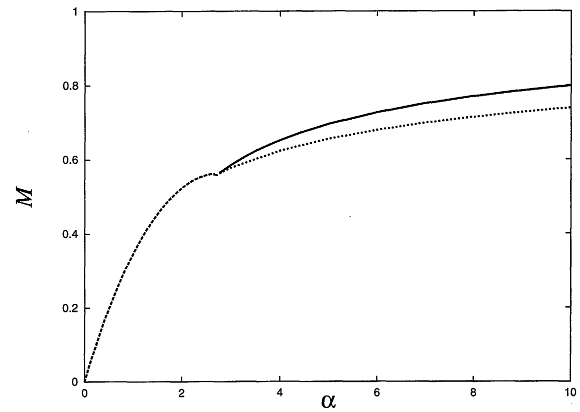

Fig. 1. Bifurcation diagram for $\mathcal{P}(y) = H(-y)$. Dashed and dotted curves denote the stable RS solution and the unstable RS solution, respectively. The solid curve is the RSB solution.

$$\alpha^{-1}(\chi^2 - 2\ln r) \simeq Qz'^3\left[\frac{r}{3\sqrt{2\pi}} + \frac{z'}{4\pi\sqrt{2}}\right]. \tag{47}$$

Solving these equations, the asymptotic forms of variables are given by

$$z' \simeq z_0\alpha^{-\frac{\delta}{2(1+2\delta)}}(\ln\alpha)^{\frac{3\delta+1}{4(1+2\delta)}}, \tag{48}$$

$$r \simeq r_0\alpha^{-\frac{\delta}{2(1+2\delta)}}(\ln\alpha)^{-\frac{5\delta+3}{4(1+2\delta)}}, \tag{49}$$

$$\xi \simeq \xi_0\alpha^{-\frac{1}{1+2\delta}}(\ln\alpha)^{\frac{1}{2(1+2\delta)}}, \tag{50}$$

$$Q \simeq Q_0\alpha^{-\frac{1}{1+2\delta}}(\ln\alpha)^{-\frac{\delta}{1+2\delta}}, \tag{51}$$

where z_0, r_0, ξ_0 and Q_0 depend on δ but not on α. For original variables, we have

$$x \simeq Q_0r_0z_0\alpha^{-\frac{1+\delta}{1+2\delta}}(\ln\alpha)^{-\frac{1+3\delta}{1+2\delta}}, \tag{52}$$

$$c \simeq \frac{1}{r_0^2Q_0^2}\alpha^{-\frac{2+\delta}{1+2\delta}}(\ln\alpha)^{\frac{3(1+3\delta)}{2(1+2\delta)}}, \tag{53}$$

$$M \simeq 1 - \frac{\xi_0^2}{2}\alpha^{-\frac{2}{1+2\delta}}(\ln\alpha)^{\frac{1}{1+2\delta}}, \tag{54}$$

$$q_0 \simeq 1 - Q_0^2\alpha^{-\frac{2}{1+2\delta}}(\ln\alpha)^{-\frac{2\delta}{1+2\delta}}. \tag{55}$$

Substituting $\delta \simeq 0$ into eqs. (48)–(51), we obtain

$$z' \propto \delta^{\frac{1}{4}}\alpha^{-\frac{\delta}{2}}(\ln\alpha)^{\frac{1}{4}},\ \ r \propto \delta^{-\frac{3}{4}}\alpha^{-\frac{\delta}{2}}(\ln\alpha)^{-\frac{3}{4}},$$
$$\xi \propto \delta^{\frac{1}{2}}\alpha^{-1}(\ln\alpha)^{\frac{1}{2}},\ \ Q \propto \alpha^{-1}(\ln\alpha)^{-\delta}. \tag{56}$$

Note that at $\delta = 0$, the variables z' and r do not converge to 0, and $\chi = \frac{\xi}{Q}$ does not tend to ∞ as $\alpha \to \infty$. These results are not consistent with our assumptions. Thus, we have to treat the case of $\delta = 0$ separately.

Case 2.

Solving eqs. (39)–(42) numerically for several functions $P(y)$, we found there is only one solution. In Fig. 2, for $P = k\ \mathrm{sgn}(y)$ we show a bifurcation diagram for the RS and the RSB solutions as functions of α. For the RSB solution, $z' \to z'_0$, $r \to r_0$, $Q \to 0$ and $\xi \to 0$ as $\alpha \to \infty$. Here, $z'_0 > 1$ and $0 < r_0 < 1$ are constants. These behaviors were observed for all functions we treated. Thus, we assume these asymptotic behaviors and study saddle-point equations analytically.

Transforming $u \to \chi u$ and substituting $Q = 0$, $\xi = 0$,

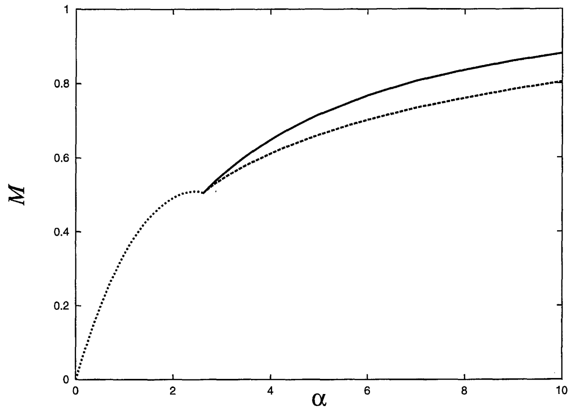

Fig. 2. Bifurcation diagram for $\mathcal{P}(y) = \frac{1}{2}(1 + k\ \mathrm{sgn}(y))$ with $k = 0.7$. Dashed and dotted curves denote the stable RS solution and the unstable RS solution, respectively. The solid curve is the RSB solution.

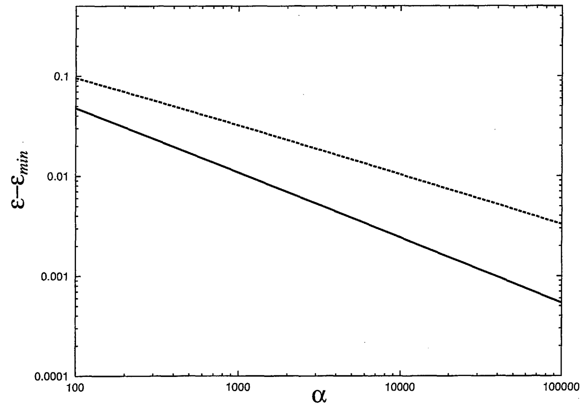

Fig. 3. log-log plot of learning curves for $\mathcal{P}(y) = H(-y)$. The abscissa is $\log_{10}(\frac{1}{\alpha})$ and the ordinate is $\log_{10}(\epsilon_{\mathrm{g}} - \epsilon_{\mathrm{min}})$. Dashed and solid curves denote the RS and the RSB solutions, respectively.

$z' = z'_0$ and $r = r_0$ in integrations in eqs. (39)–(42), we have the following lowest order equations,

$$\alpha^{-1} \simeq -Q \int \frac{\mathrm{d}u}{\sqrt{2\pi}} \frac{\mathcal{L}(\chi u)}{\hat{H}(\chi u)} u E(\chi u) \equiv -Q f_1(\chi, r_0, z'_0), \tag{57}$$

$$\alpha^{-1} \simeq -\frac{k}{\pi} Q \int \mathrm{d}u \mathrm{e}^{-\frac{u^2}{2}} \frac{\mathcal{L}(\chi u)}{\hat{H}(\chi u)} \equiv -\frac{k}{\pi} Q f_2(\chi, r_0, z'_0), \tag{58}$$

$$\alpha^{-1} \simeq r_0 \xi \int \frac{\mathrm{d}u}{\sqrt{2\pi}} E(\chi u) g(\chi u) \equiv r_0 \xi f_3(\chi, r_0, z'_0), \tag{59}$$

$$\alpha^{-1}\{\chi^2(1 - r_0^2) - 2\ln r_0\} \simeq -\xi \int \frac{\mathrm{d}u}{\sqrt{2\pi}} E(\chi u) \psi(\chi u) \equiv -\xi f_4(\chi, r_0, z'_0), \tag{60}$$

$$E(\chi u) \simeq 1 - k + 2kH(u),$$

$$w(\chi u) \simeq \frac{2k}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}u^2}.$$

As shown in Appendix D, $\chi \to \chi_0$=constant as $\alpha \to \infty$ follows from these equations. Then, χ_0, r_0 and z'_0 are determined by the following equations.

$$\frac{k}{\pi} f_2(\chi_0, r_0, z'_0) \simeq f_1(\chi_0, r_0, z'_0), \tag{61}$$

$$r_0 \chi_0 f_3(\chi_0, r_0, z'_0) \simeq -f_1(\chi_0, r_0, z'_0), \tag{62}$$

$$r_0 f_3(\chi_0, r_0, z'_0) \simeq \frac{f_4(\chi_0, r_0, z'_0)}{2\ln r_0 - \chi_0^2(1 - r_0^2)}. \tag{63}$$

Thus, Q and ξ are expressed by

$$Q \simeq -\frac{\pi\alpha^{-1}}{k f_2(\chi_0, r_0, z'_0)}, \tag{64}$$

$$\xi \simeq \chi_0 Q. \tag{65}$$

Therefore, $\xi \propto \alpha^{-1}$.

As shown in §3, the generalization error is expressed in terms of ζ. Now, we estimate ϵ_{g}.

Case 1.

Since $\zeta = \sqrt{\frac{Q^2+\xi^2}{1-\xi^2}}$ and we assume $\chi \gg 1$, $\zeta \sim \xi$ follows. Then,

$$\epsilon_{\mathrm{g}} - \epsilon_{\mathrm{min}} \simeq \frac{2s}{\sqrt{2\pi}(1+\delta)} \zeta^{1+\delta} \simeq \epsilon_0 \alpha^{-\frac{1+\delta}{1+2\delta}} (\ln\alpha)^{\frac{1+\delta}{2(1+2\delta)}}, \tag{66}$$

$$\epsilon_0 \equiv \frac{2s}{\sqrt{2\pi}(1+\delta)} \xi_0^{1+\delta}. \tag{67}$$

Substituting $\delta = 1$ into this relation, we obtain

$$(\epsilon_{\mathrm{g}} - \epsilon_{\mathrm{min}}) \propto \alpha^{-\frac{2}{3}} (\ln\alpha)^{\frac{1}{3}}.$$

This agrees with the previous results by Kabashima-Shinomoto and Kim-Pollard up to the logarithmic correction $(\ln\alpha)^{\frac{1}{3}}$. (See Fig. 3.)

Case 2.

Since $Q \sim \xi$, $\zeta \propto \xi$. Then,

$$\epsilon_{\mathrm{g}} - \epsilon_{\mathrm{min}} \simeq \frac{k}{\pi}\zeta \simeq \frac{k}{\pi} \frac{\sqrt{1+\chi_0^2}}{\chi_0} \xi \simeq -\frac{\sqrt{1+\chi_0^2}}{f_2} \alpha^{-1}. \tag{68}$$

This is the same α dependence as that in the RS solution although the prefactors differ. (See Fig. 4.)

§5. Summary and Discussions

We studied learning from examples by a perceptron when a return to an example is given by a stochastic relation which is represented by the conditional probability distribution $\mathcal{P}(u)$, where u is proportional to the inner product between the optimal synaptic weight vector and the example vector. Calculations were performed by using the replica method. We studied the asymptotic behavior of the learning curve as $\alpha \to \infty$ for the family of functions P_δ, which behaves as $P_\delta \simeq a\ \mathrm{sgn}(u)|u|^\delta$ around $u = 0$. To consider the minimum-error algorithm, the limit $\beta \to \infty$ is taken. Expressing the generalization error as $\epsilon_{\mathrm{g}} = \epsilon_{\mathrm{min}} + \epsilon_0\alpha^{-\gamma}$, we calculated the exponent γ. First, we studied the RS solution and obtained $\gamma = \frac{1+\delta}{1+3\delta}$. However, this solution turned out to become unstable at the finite value of α, $\alpha = \alpha_{\mathrm{AT}} = \alpha_{\mathrm{c}}$. Thus, next, we took into account the breaking of the replica symmetry. Assuming the one-step replica symmetry breaking, we obtained $\gamma = \frac{1+\delta}{1+2\delta}$.

In the course of the derivation, we assumed that $M \to$

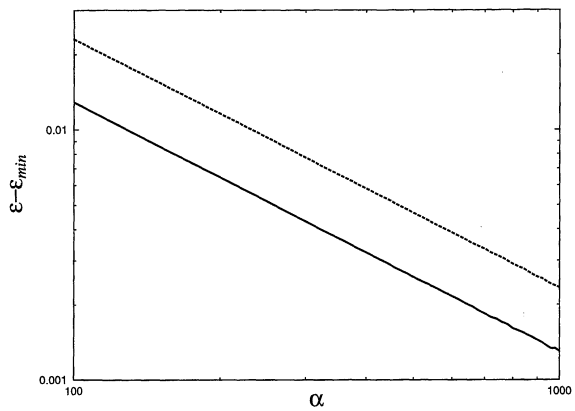

Fig. 4. log-log plot of learning curves for $\mathcal{P}(y) = \frac{1}{2}(1 + k\ \mathrm{sgn}(y))$ with $k = 0.7$. The abscissa is $\log_{10}(\frac{1}{\alpha})$ and the ordinate is $\log_{10}(\epsilon_{\mathrm{g}} - \epsilon_{\min})$. Dashed and solid curves denote the RS and the RSB solutions, respectively.

1 as $\alpha \to \infty$, and that the probability is monotonic and has the symmetry, $P(-u) = -P(u)$. This symmetry is irrelevant to the result as we discuss later.

Now, let us compare our results with previous ones.

First, we discuss the case of $\delta = 1$. This corresponds to the case of input noise and to the case of $\mathcal{P}'(0) \neq 0$. The problem is essentially the same as Györgyi-Tishby's and the result of $\gamma = \frac{1}{2}$ for the RS solution is the same as their's, the relation (3). However, this solution is unstable. On the other hand, the result of $\gamma = \frac{2}{3}$ for the one-step RSB solution agrees up to a logarithmic correction to those by Kabashima-Shinomoto and Kim-Pollard which were obtained without using the replica method in the one-dimensional model, the relation (4). Thus, we consider that the exponent in the infinite dimensional model is not $\frac{1}{2}$ but $\frac{2}{3}$, that is, the exponent in the infinite dimensional model is the same as that in the one-dimensional model.

Next, let us discuss the case of $\delta = 0$, where $P(y)$ is discontinuous at $y = 0$. This corresponds to the case of output noise. $\gamma = 1$ was obtained for both the RS and the RSB solutions. Since the one-step RSB ansatz does not change the value of γ, we consider that $\gamma = 1$ is correct for this case.

Here, let us consider the role of the temperature on the fast convergence of $\gamma = 1$ obtained in the case of $\delta = 0$. The same result is also obtained by Opper and Haussler in the following problem. They studied the learning of the stochastic relation within the RS ansatz introducing temperature. The model they treated is just the case of $P(y) = k\ \mathrm{sgn}(y)$ with the inverse temperature $\beta = \ln \frac{1+k}{1-k}$ in the present formulation. This corresponds to the presence of output noise reversing the sign of output with the probability $\frac{1-k}{2}$. The RS solution is stable for any α in this case. They insist that in stochastic models the fast convergence of $\gamma = 1$ is essential to learning with non-zero temperature. However, our result indicates that the fast convergence can be obtained in the stochastic model even if the temperature is zero.

In this paper, we treated the cases $\delta > 0$ and $\delta = 0$ separately. To see the reason, let us compare the case $\delta = 0$ with the case $\delta \to 0$.

In the RS solution, if we take the limit $\delta \to 0$ in the formula (34) for the case $\delta > 0$, the exponent γ becomes 1, but the prefactor differs from that in the formula (35) for the case $\delta = 0$. Thus, the $\delta \to 0$ limit does not agree with the result for $\delta = 0$. This is because z becomes finite in this limit and $L(z)$ can not be expanded in z.

On the other hand, in the RSB solution, the generalization error in the $\delta \to 0$ limit behaves as

$$\epsilon_{\mathrm{g}} - \epsilon_{\min} \propto \delta^{1/2}\alpha^{-1}(\ln\alpha)^{1/2},$$

since $\xi_0 \propto \delta^{1/2}$ for small δ. See eq. (66). This result is different from the result (68) in the case $\delta = 0$,

$$\epsilon_{\mathrm{g}} - \epsilon_{\min} \simeq -\frac{\sqrt{1+\chi_0^2}}{f_2}\alpha^{-1}.$$

Thus, we cannot obtain the result for the case of $\delta = 0$ by taking the $\delta \to 0$ limit in the result for $\delta > 0$. The reason is similar to the one in the RS case. If $\delta \simeq 0$ is substituted into the relations (48)–(51) for the RSB solution, it turns out that the behaviors of variables become singular at $\delta = 0$. See eq. (56). These behaviors reflect the fact that in the $\delta = 0$ case z' and r tend to non-zero constants. Since for the case of $\delta > 0$ we treat z' and r as small quantities and expand the free energy in terms of z' and r, it is necessary to treat the cases $\delta > 0$ and $\delta = 0$ differently.

Another difference between the cases $\delta > 0$ and $\delta = 0$ is the existence of the log correction for the one-step RSB solution in the case of $\delta > 0$.

When $\delta > 0$, the exponent $\gamma = (1+\delta)/(1+2\delta)$ for the RSB solution is larger than the exponent $\gamma = (1+\delta)/(1+3\delta)$ for the RS solution which is unstable, and $\epsilon_{\mathrm{g}} - \epsilon_{\min}$ for the RSB solution is much less than that for the RS solution. Since we do not investigate the stability of the one-step RSB solution, we do not know whether the one-step RSB solution is stable for any α or not. However, as discussed above, for $\delta = 1$ the same exponent $\gamma = 2/3$ as ours is also obtained by non-replica methods in one-dimensional models.[12,13] Thus, we conjecture that at least the exponent $\gamma = (1+\delta)/(1+2\delta)$ is correct irrespective of the existence of the log correction for the case of $\delta > 0$.

On the other hand, when $\delta = 0$, for both the RS and the one-step RSB solutions, the exponents are the same, $\gamma = 1$, and there is no log correction. Therefore, we conjecture that there is no log correction and $\epsilon_{\mathrm{g}} - \epsilon_{\min} \propto \alpha^{-1}$ is the exact behavior for the case of $\delta = 0$, i.e. for the case where $P(y)$ is discontinuous at $y = 0$.

Our results on the exponent γ for the RS and the one-step RSB solutions indicate that the asymptotic behavior of generalization error ϵ_{g} is determined by the exponent δ which characterizes the local behavior of $P_\delta(u)$ around $u = 0$, and that the smaller the exponent δ is, the smaller the generalization error ϵ_{g} is. We give an intuitive explanation for this.

In our model, the probability $\mathcal{P}$ that the optimal weight $\boldsymbol{J}^0$ returns $R = 1$ for an example $\boldsymbol{\xi}$ depends on the example through the form $u_0 = (\boldsymbol{\xi} \cdot \boldsymbol{J}^0)/\sqrt{N}$. Due to this dependence, in the minimum-error algorithm, the example $\boldsymbol{\xi}$ with a small value of u_0 contributes to the selection of weight vectors $\boldsymbol{J}$s which have the minimum number of errors. Thus, the behavior of $\mathcal{P}(u_0)$ around $u_0 = 0$ has a crucial influence on $\Delta M \equiv 1 - M$. Indeed, if $\mathcal{P}(u_0)$ varies rapidly around $u_0 = 0$, the minimum number of errors is small and there are only a small number of weight vectors with the minimum number of errors, and then ΔM is small. That is, as δ decreases, so does ΔM.

In regard to the relation between $\epsilon_{\rm g}$ and ΔM, we proved that $\epsilon_{\rm g} - \epsilon_{\min}$ is proportional to $(\Delta M)^{(1+\delta)/2}$ for P_δ in §2. This is intuitively derived as follows. When α is large, the angle θ between $\boldsymbol{J}$ and $\boldsymbol{J}^0$ is small, and $\theta \propto \sqrt{\Delta M}$. For the deterministic case, $\epsilon_{\rm g}$ is proportional to the probability that an example $\boldsymbol{\xi}$ satisfies the condition $u_0 u < 0$, where $u = (\boldsymbol{\xi} \cdot \boldsymbol{J})/\sqrt{N}$. In the below, we consider only the case $u_0 > 0$ and $u < 0$ for simplicity, since the contribution from the case $u_0 < 0$ and $u > 0$ is similarly estimated and has the same form as that for the case $u_0 > 0$ and $u < 0$. The probability of $u_0 > 0$ and $u < 0$ is proportional to the angle θ. Thus, $\epsilon_{\rm g} \propto \sqrt{\Delta M}$. For the stochastic cases, the $\epsilon_{\rm g}$ for the deterministic case should be multiplied by the probability that the optimal weight $\boldsymbol{J}^0$ returns $R = 1$. However, this includes the probability of the false prediction for $\boldsymbol{J} = \boldsymbol{J}^0$, which contributes to $\epsilon_{\min}$ and is proportional to $\Delta M(1 - \mathcal{P}(u_0))$. We have to subtract this term to obtain $\epsilon_{\rm g} - \epsilon_{\min}$. Therefore $\epsilon_{\rm g} - \epsilon_{\min} \propto \sqrt{\Delta M}\mathcal{P}(u_0)$. Since the typical value of u_0 under the conditions $u_0 > 0$ and $u < 0$ is proportional to $\sqrt{\Delta M}$, we obtained the above mentioned result $(\epsilon_{\rm g} - \epsilon_{\min}) \propto (\Delta M)^{(1+\delta)/2}$ for P_δ.

The symmetry we assumed for $P(y)$ is irrelevant to the results obtained in this paper. It is obvious from the derivation of the asymptotic behavior of $\epsilon_{\rm g}$ that neither the global symmetry of the probability, $P(-y) = -P(y)$, nor its local symmetry around $y = 0$, $P(y) \simeq a\ \mathrm{sgn}(y)|y|^\delta$, are necessary for the result of $\epsilon_{\rm g} - \epsilon_{\min} \propto \alpha^{-\gamma}$. In fact, if the exponents of $P(y)$ in the small regions below and above $y = 0$ are different from each other, γ is determined by the smaller one of these two exponents.

Finally, we show that the exponent $\gamma = \frac{1+\delta}{1+2\delta}$ obtained here for the one-step RSB solution can also be derived in the following one-dimensional problem applying the method by Kabashima-Shinomoto.[12] This problem corresponds to the one-dimensional version of the problem treated in this paper.

Let us consider the stochastic relation in one dimension. For a given sample ξ which is uniformly drawn from the interval [0, 1], the return is $+1$ with the probability $\mathcal{P}(\xi)$ and -1 with $1 - \mathcal{P}(\xi)$. We assume that $\mathcal{P}(\xi)$ is an increasing function of ξ and has a point $\theta_o \in (0, 1)$ which satisfies $\mathcal{P}(\theta_o) = 1/2$. Under these assumptions, the prediction $y = \mathrm{sgn}(\xi - \theta_o)$ minimizes the generalization error. A learner's task is to estimate the optimal threshold θ_o from a given set of examples. Let θ_p be the estimate of θ_o obtained through the minimum-error algorithm when the training set

$$(\xi_\mu, y_\mu), \mu = 1, \cdots, p$$

is given. We consider how θ_p fluctuates around θ_o when p is large.

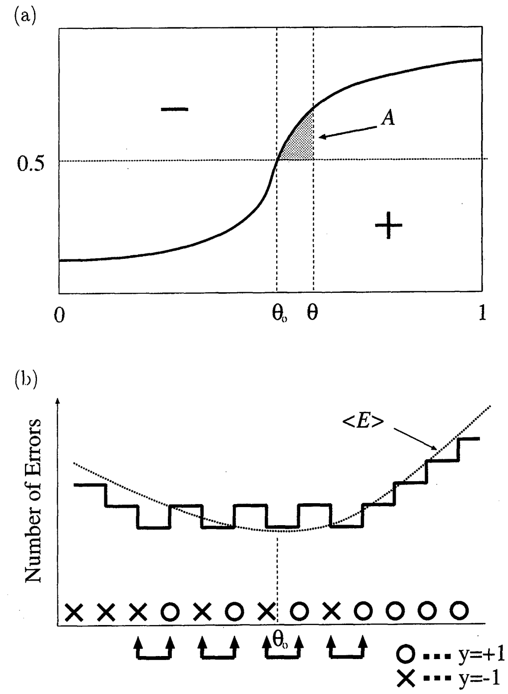

Fig. 5. Illustration of the one-dimensional model. (a) Probability $\mathcal{P}(\xi)$. (b) The solid line is the number of errors $E(\theta)$ for a realization of a set of examples. When the return to an example ξ_i is $y_i = +1$ or $y_i = +1$, a circle or a cross is marked at the position of $\xi = \xi_i$, respectively. A threshold θ_p is selected from separate intervals indicated by arrows. The dotted line is the average number of errors which has the minimum at $\theta = \theta_o$.

For a threshold θ, the number of errors $E(\theta)$ is the number of examples which satisfy the condition that $y_i = +1$ and $\xi_i < \theta$ or $y_i = -1$ and $\xi_i > \theta$ $(1 \le \mathrm{i} \le p)$. The average of the difference $\Delta E(\theta) = E(\theta) - E(\theta_o)$ is $p(\epsilon_{\rm g}(\theta) - \epsilon_{\min})$, where $\epsilon_{\rm g}(\theta)$ and $\epsilon_{\min}$ are the generalization error of the threshold θ and its minimum value obtained at $\theta = \theta_o$. The difference $\epsilon_{\rm g}(\theta) - \epsilon_{\min}$ is proportional to the area A of the shaded region in Fig. 5(a). If $\mathcal{P}(\theta) \simeq \frac{1}{2} + \frac{a}{2}\ \mathrm{sgn}(\theta - \theta_o)|\theta - \theta_o|^\delta$ around $\theta = \theta_o$, this area A is given as $A \sim |\theta - \theta_o|^{1+\delta}$. Therefore, the average of $\Delta E(\theta)$ is of the order of

$$p|\theta - \theta_o|^{1+\delta}. \tag{69}$$

Next, let us estimate the fluctuation of $\Delta E(\theta)$. Suppose that θ moves from the left end $\theta = 0$ to the right end $\theta = 1$. Each time θ comes across an example ξ_i, $E(\theta)$ increases or decreases by 1. Near $\theta = \theta_o$, this motion is like that of a random walk because the probability of $y = +1$ or that of $y = -1$ is nearly equal to $1/2$ for θ around θ_o. See Fig. 5(b). Thus, the fluctuation of the difference $\Delta E(\theta)$ is proportional to

$$\sqrt{p \cdot |\theta - \theta_o|}, \tag{70}$$

since there are about $p \cdot |\theta - \theta_o|$ examples between θ and θ_o.

The balance between quantities (69) and (70) determines the fluctuation of the estimate θ_p. This gives the scaling $|\theta_p - \theta_o| \sim p^{-\frac{1}{1+2\delta}}$, which implies that the learning curve scales as $\epsilon_{\rm g} - \epsilon_{\rm min} \sim p^{-\frac{1+\delta}{1+2\delta}}$. This result is the same as ours with $N = 1$ up to a logarithmic correction and supports our result.

Acknowledgement

We are grateful to P. Davis for valuable discussions. One of the authors (YK) was partially supported by the Japanese Grant-in-Aid for Science Research Fund from the Ministry of Education, Science and Culture No. 06260102 and No. 06740325.

Appendix A: Derivation of the Free Energy

In this Appendix, we derive the general form of the free energy by the replica method. Introducing n replicas and taking the sample average over R, we obtain

$$\begin{aligned}\langle V^n\rangle_R &= \int \left[\prod_{j=1}^{N}\prod_{\alpha=1}^{n} {\rm d}J_j^\alpha\right] [\prod_\alpha \delta(\sum_j (J_j^\alpha)^2 - N)] \prod_{\mu,\alpha} \langle \Theta(R^\mu u_\alpha^\mu)\rangle_R \Big/ \int \left[\prod_{j=1}^{N}\prod_{\alpha=1}^{n} {\rm d}J_j^\alpha\right] [\prod_\alpha \delta(\sum_j (J_j^\alpha)^2 - N)],\\ &= \int \left[\prod_{j=1}^{N}\prod_{\alpha=1}^{n} {\rm d}J_j^\alpha\right] [\prod_\alpha \delta(\sum_j (J_j^\alpha)^2 - N)] \prod_{\mu,\alpha} \int_0^\infty {\rm d}\lambda^{\mu,\alpha} \int_{-\infty}^\infty \frac{{\rm d}x^{\mu,\alpha}}{2\pi} < \exp[-{\rm i}x^{\mu,\alpha}(R^\mu u_\alpha^\mu - \lambda^{\mu,\alpha})]\rangle_R / S_N^n,\end{aligned} \tag{A·1}$$

$$S_N \equiv \int \mathcal{D}(\boldsymbol{J}) = \int \left[\prod_{j=1}^{N} {\rm d}J_j\right] \delta(\sum_j (J_j^2 - N)),\quad u_\alpha^\mu \equiv (\boldsymbol{\xi}^\mu \cdot \boldsymbol{J}^\alpha)/\sqrt{N}.$$

The average over R^μ, $\langle \exp(-{\rm i}x^{\mu,\alpha} R^\mu u_\alpha^\mu)\rangle_{R^\mu}$, is

$$\langle \exp[-{\rm i}x^{\mu,\alpha}R^\mu u_\alpha^\mu]\rangle_{R^\mu} = \frac{1}{2}[\exp({\rm i}x^{\mu,\alpha}u_\alpha^\mu) + \exp(-{\rm i}x^{\mu,\alpha}u_\alpha^\mu) - P(u_0^\mu)\{\exp({\rm i}x^{\mu,\alpha}u_\alpha^\mu) - \exp(-{\rm i}x^{\mu,\alpha}u_\alpha^\mu)\}].$$

Next, using the relation,

$$\delta\left(\sum_{j=1}^{N}(\xi_j^\mu)^2 - N\right) = \int_{-{\rm i}\infty}^{{\rm i}\infty} \frac{{\rm d}\tilde{K}_\mu}{2\pi {\rm i}} \exp\left[-\tilde{K}_\mu\left(\sum_{j=1}^{N}(\xi_j^\mu\right)^2 - N)\right],$$

we take the sample averages over $\boldsymbol{\xi}^\mu$,

$$\langle \exp(\pm {\rm i}\sum_\alpha x^{\mu,\alpha}u_\alpha^\mu)\rangle_{\xi^\mu} = \int_{-{\rm i}\infty}^{{\rm i}\infty} \frac{{\rm d}\tilde{K}_\mu}{2\pi{\rm i}} \left(\frac{\pi}{\tilde{K}_\mu}\right)^{N/2} \exp\left[\tilde{K}_\mu N - \sum_{j=1}^{N}\left(\sum_{\alpha=1}^{n} x^{\mu,\alpha}J_j^\alpha\right)^2 / 4N\tilde{K}_\mu\right] / S_N, \tag{A·2}$$

$$\begin{aligned}\langle \exp(\pm{\rm i}\sum_\alpha x^{\mu,\alpha}u_\alpha^\mu)P(u_0^\mu)\rangle_{\xi^\mu} &= \langle \exp(\pm{\rm i}\sum_\alpha x^{\mu,\alpha}u_\alpha^\mu)\rangle_{\xi^\mu}\sqrt{\frac{\tilde{K}_\mu}{\pi}}\int {\rm d}\xi P(\xi)\\ &\times \exp[-\tilde{K}_\mu\{\xi^2 \mp {\rm i}\xi\sum_\alpha x^{\mu,\alpha}M^\alpha/\tilde{K}_\mu - (\sum_\alpha x^{\mu,\alpha}M^\alpha/2\tilde{K}_\mu)^2\}],\end{aligned} \tag{A·3}$$

where $M^\alpha = \frac{1}{N}\sum_{j=1}^{N} J_j^\alpha J_j^0$ is the overlap between the weight vector of a learner and the optimal weight vector. In the above expressions $\boldsymbol{J}^0$ is included only in M^α. Then, using the following relation,

$$\delta\left(\sum_{j=1}^{N}(J_j^0)^2 - N\right) = \int_{-{\rm i}\infty}^{{\rm i}\infty}\frac{{\rm d}\tilde{E}_0}{4\pi{\rm i}}\exp\left[-\frac{1}{2}\tilde{E}_0\left(\sum_{j=1}^{N}(J_j^0)^2 - N\right)\right],$$

$$1 = \prod_\alpha \int {\rm d}M^\alpha \int_{-\infty}^{\infty}\frac{N{\rm d}\bar{M}^\alpha}{2\pi}\exp\left[{\rm i}N\bar{M}^\alpha\left(M^\alpha - \frac{1}{N}\sum_{j=1}^{N}J_j^\alpha J_j^0\right)\right],$$

we only have to calculate the average $\langle \exp(-{\rm i}\sum_\alpha \bar{M}^\alpha \sum_j J_j^\alpha J_j^0)\rangle_{J^0}$,

$$\begin{aligned}&\langle \exp(-{\rm i}\sum_\alpha \bar{M}^\alpha\sum_j J_j^\alpha J_j^0)\rangle_{J^0} =\\ &\int \frac{{\rm d}\tilde{E}_0}{4\pi{\rm i}}[\prod_j {\rm d}J_j^0]\exp\left[-\frac{1}{2}\tilde{E}_0(\sum_j (J_j^0)^2 - N) - {\rm i}\sum_\alpha \bar{M}^\alpha\sum_j J_j^\alpha J_j^0\right]/S_N,\\ &= {\rm e}^{\frac{N}{2}\tilde{E}_0}\int\frac{{\rm d}\tilde{E}_0}{4\pi{\rm i}}\prod_j\left[\sqrt{\frac{2\pi}{\tilde{E}_0}}\int {\rm D}t_0\exp({\rm i}\sum_\alpha \bar{M}^\alpha J_j^\alpha t_0/\sqrt{\tilde{E}_0})\right]/S_N.\end{aligned} \tag{A·4}$$

Finally, introducing the relations,

$$\delta(\sum_j (J_j^\alpha)^2 - N) = \int_{-i\infty}^{i\infty} \frac{d\tilde{E}_\alpha}{4\pi i} \exp\left[-\frac{1}{2}\tilde{E}_\alpha(\sum_j (J_j^\alpha)^2 - N)\right],$$

$$1 = \int \prod_{\alpha<\beta} dq^{\alpha\beta} \int_{-i\infty}^{i\infty} \frac{N dF^{\alpha\beta}}{2\pi i} \exp\left[-NF^{\alpha\beta}\left(q^{\alpha\beta} - \frac{1}{N}\sum_{j=1}^{N} J_j^\alpha J_j^\beta\right)\right],$$

we obtain the expression for $\langle V^n \rangle_{R,\xi,J^0}$,

$$\langle V^n \rangle_{R,\xi,J^0} = \int d\left(\frac{d\tilde{\boldsymbol{E}}}{4\pi i}\right) \int d\left(\frac{N\hat{F}}{2\pi i}\right) \int d\left(\frac{N\bar{\boldsymbol{M}}}{2\pi}\right) \int d\hat{q} \int d\boldsymbol{M} e^{NG}, \tag{A·5}$$

$$G = \frac{1}{2}\sum_\alpha \tilde{E}_\alpha - \sum_{\alpha<\beta} F^{\alpha\beta} q^{\alpha\beta} + i\sum_\alpha \bar{M}^\alpha M^\alpha + \sum_\mu \frac{1}{N} G_1(\hat{q}, \boldsymbol{M}) + G_2(\tilde{\boldsymbol{E}}, \bar{\boldsymbol{M}}, \hat{F}), \tag{A·6}$$

$$e^{G_1} = \left[\prod_\alpha \left(\int_0^\infty d\lambda^\alpha \int \frac{dx^\alpha}{2\pi}\right)\right] \exp\left[-\frac{1}{2}\{2\sum_{\alpha<\beta} q^{\alpha\beta} x^\alpha x^\beta + \sum_\alpha (x^\alpha)^2\} + i\sum_\alpha x^\alpha \lambda^\alpha\right]$$
$$\times \left[1 - \frac{1}{2}\{\hat{P}(\sum_\alpha x^\alpha M^\alpha) - \hat{P}(-\sum_\alpha x^\alpha M^\alpha)\}\right], \tag{A·7}$$

$$e^{G_2} = \int Dt_0 \int d\boldsymbol{J} \exp\left[-\frac{1}{2}\sum_\alpha \tilde{E}_\alpha (J^\alpha)^2 + \sum_{\alpha<\beta} F^{\alpha\beta} J^\alpha J^\beta + i\sum_\alpha \bar{M}^\alpha J^\alpha t_0\right], \tag{A·8}$$

$$\hat{P}(y) = \frac{1}{\sqrt{2\pi}} \int d\xi P(\xi) \exp\left(-\frac{\xi^2}{2} + i\xi y + \frac{y^2}{2}\right). \tag{A·9}$$

$q^{\alpha\beta} = \frac{1}{N}\sum_{j=1}^N J_j^\alpha J_j^\beta$ is the overlap between weight vectors of learners. In the final expressions we set $\tilde{K}_\mu$ and $\tilde{E}_0$ to the optimal values of $\tilde{K}_\mu = 1/2$ and $\tilde{E}_0 = 1$ respectively, and use abbreviations such as $\boldsymbol{J} = \{J^\alpha\}$, $d\boldsymbol{J} = \prod_\alpha dJ^\alpha$, $\hat{q} = \{q^{\alpha\beta}\}$, $d\hat{q} = \prod_{\alpha<\beta} dq^{\alpha\beta}$, etc. As for the partition function Z, the expression of $\langle Z^n \rangle_{R,\xi,J^0}$ is obtained by the following replacement in the expression for $\langle V^n \rangle_{R,\xi,J^0}$,

$$\int_0^\infty d\lambda^\alpha \longrightarrow e^{-\beta} \int_{-\infty}^0 d\lambda^\alpha + \int_0^\infty d\lambda^\alpha,$$

where $\beta = 1/T$. The general form of the free energy per synapse is given by

$$F = -\frac{\langle \ln Z \rangle_{\xi,J^0,R}}{N\beta}. \tag{A·10}$$

In the RS solution, we assume all quantities are independent of replica indices. In particular, putting $q^{\alpha\beta} = q$ and $M^\alpha = M$ and expressing other quantities by q and M, we obtain

$$\frac{1}{n} G_{\rm RS}(q, M, \beta) = \frac{1-M^2}{2(1-q)} + \frac{1}{2}\ln 2\pi(1-q)$$
$$+\alpha \int Dy 2\mathcal{P}(y) \int Du \ln \tilde{H}(Y), \tag{A·11}$$

$$\tilde{H}(Y) = e^{-\beta} + (1 - e^{-\beta}) H(Y),$$

$$H(Y) = \int_Y^\infty Du, \quad Du = \frac{du}{\sqrt{2\pi}} \exp\left[-\frac{u^2}{2}\right],$$

$$Y = \frac{\sqrt{q - M^2}u - My}{\sqrt{1-q}}.$$

Since we adopt the minimum-error algorithm, we take the limit $\beta \to \infty$, and to obtain non-trivial results we also take the limit $q \to 1$ fixing $x \equiv \sqrt{2\beta(1-q)}$. Let us estimate the integral $\int Du \ln \tilde{H}(Y)$ in these limits. Defining $z = \frac{x}{\sqrt{1-M^2}}$ and $\zeta = \frac{\sqrt{1-M^2}}{M}$, Y is expressed as $Y = \sqrt{2\beta}(u - y/\zeta)/z$. In the region $u < y/\zeta$, $Y \ll -1$ and $\tilde{H}(Y) \sim 1$, and so the integral in this region is neglected. In the region $u > y/\zeta$, $Y \gg 1$, and this region is further divided into two regions, in each of which $\tilde{H}(Y)$ is estimated as follows,

$$\tilde{H}(Y) \sim H(Y) \sim \frac{e^{-Y^2/2}}{\sqrt{2\pi}Y} \quad \text{for } y/\zeta < u < u_{\rm c},$$

$$\tilde{H}(Y) \sim e^{-\beta} \quad \text{for } u_{\rm c} < u.$$

The value $u_{\rm c}$ which separates these two regions is determined by the condition $e^{-\beta} \sim H(Y)$, which gives $u_{\rm c} = y/\zeta + z$. Thus, we obtain

$$\int Du \ln \tilde{H}(Y) = \int_{y/\zeta+z}^\infty Du(-\beta) + \int_{y/\zeta}^{y/\zeta+z} Du(-Y^2/2),$$
$$= -\beta H(y/\zeta + z)$$
$$- \frac{\beta}{z^2} \int_{y/\zeta}^{y/\zeta+z} Du(u - y/\zeta)^2.$$

Then, the minimum error fraction $f = -\frac{1}{\alpha\beta}\frac{G}{n}$ is given by

$$f(\alpha,\zeta,z) = -\frac{1}{\alpha z^2} - \frac{1}{2\alpha\beta}\ln 2\pi(1-q) + f_1, \qquad \text{(A·12)}$$

$$f_1 = \int \mathrm{D}y 2\mathcal{P}(y)\left[H(y/\zeta+z) + \frac{1}{z^2}\int_{y/\zeta}^{y/\zeta+z} \mathrm{D}u(u-y/\zeta)^2\right]. \qquad \text{(A·13)}$$

Appendix B: Proof of $\alpha_{\mathrm{AT}} = \alpha_{\mathrm{c}}$

Here, we show that $\alpha_{\mathrm{AT}} = \alpha_{\mathrm{c}}$.

Since we exclude the deterministic case, f is given by

$$f = -\frac{1}{\alpha z^2} + f_1.$$

Using eq. (21) this equation is transformed into

$$f = \int \mathrm{D}y 2\mathcal{P}(y) H(y/\zeta+z). \qquad \text{(B·1)}$$

By definition, the critical capacity α_{c} satisfies the condition $f(\alpha=\alpha_{\mathrm{c}}) = 0$; that is, at this value the minimum error fraction is zero. From eq. (23), $0 < \zeta < \infty$ for $\alpha < \infty$ follows, and then from the above eq. (B·1) $f > 0$ for $z < \infty$. Therefore, at $\alpha = \alpha_{\mathrm{c}}$, z should be infinity. Let ζ_0 be the value of ζ at $\alpha = \alpha_{\mathrm{c}}$. ζ_0 and α_{c} are determined by

$$0 = \int \mathrm{D}y 2\mathcal{P}(y) y \int_{y/\zeta_0}^{\infty} \mathrm{D}u(u-y/\zeta_0), \qquad \text{(B·2)}$$

$$1 = \alpha_{\mathrm{c}} \int \mathrm{D}y 2\mathcal{P}(y) \int_{y/\zeta_0}^{\infty} \mathrm{D}u(u-y/\zeta_0)^2. \qquad \text{(B·3)}$$

From these equations, we obtain

$$1 = \alpha_{\mathrm{c}} \int \mathrm{D}y 2\mathcal{P}(y) H(y/\zeta_0). \qquad \text{(B·4)}$$

On the other hand, the AT-stability for the RS solution is judged by the following quantities for any temperature.[16)]

$$\Lambda_1 = \alpha\lambda_1\lambda_1' - 1, = \alpha\frac{1-3q}{1-q}\int \mathrm{D}t \int \mathrm{D}y 2\mathcal{P}(y)\{\overline{x^2}(t) - 3\overline{x}(t)^2\}\{\overline{x^2}(t) - \overline{x}(t)^2\} - 1, \qquad \text{(B·5)}$$

$$\Lambda_3 = \alpha\lambda_3\lambda_3' - 1, = \alpha \int \mathrm{D}t \int \mathrm{D}y 2\mathcal{P}(y)\{\overline{x^2}(t) - \overline{x}(t)^2\}^2 - 1, \qquad \text{(B·6)}$$

$$\overline{x}(t) = -\mathrm{i}\tilde{\varphi}(Y), \quad \overline{x^2}(t) = Y\tilde{\varphi}(Y), \quad Y = \frac{\sqrt{q-M^2}u - My}{\sqrt{1-q}},$$

$$\tilde{H}(Y) = \mathrm{e}^{-\beta} + (1-\mathrm{e}^{-\beta})H(Y), \quad \tilde{\varphi}(Y) = \frac{\tilde{H}'(Y)}{\tilde{H}(Y)},$$

where $\lambda_1, \lambda_2 = \lambda_3$ are the eigenvalues of the matrix $\{\frac{1}{n}\frac{\partial^2 G}{\partial q_{\alpha\beta}\partial q_{\gamma\delta}}\}$ and $\lambda_1', \lambda_2' = \lambda_3'$ are those of $\{\frac{1}{n}\frac{\partial^2 G}{\partial J_{\alpha\beta}\partial J_{\gamma\delta}}\}$. The RS solution is stable when both Λ_1 and Λ_3 are negative. In our case of $\beta \to \infty$, α_{AT} is determined by $\Lambda_3 = 0$. By a similar calculation to ref. 17, Λ_3 is given by

$$\Lambda_3 = \alpha \int \mathrm{D}y \int \mathrm{D}t\, 2\mathcal{P}(y)\delta^2(t - y/\zeta - z)z^2 + \alpha \int \mathrm{D}y 2\mathcal{P}(y) \int_{y/\zeta}^{y/\zeta+z} \mathrm{D}t - 1.$$

Substituting $\alpha = \alpha_{\mathrm{c}}, z = \infty$ and $\zeta = \zeta_0$ into this equation, we obtain

$$\Lambda_3 = \alpha_{\mathrm{c}} \int \mathrm{D}y 2\mathcal{P}(y) \int_{y/\zeta_0}^{\infty} \mathrm{D}t - 1 = 0. \qquad \text{(B·7)}$$

Thus, $\alpha_{\mathrm{AT}} = \alpha_{\mathrm{c}}$.

Appendix C: Asymptotic Form of the Minimum Error Fraction

In this Appendix, we give the asymptotic form of the the minimum error fraction f in the one-step RSB ansatz for large α in the family of continuous functions $P_\delta(y)$. In eq. (38) the minimum error fraction is expressed as follows.

$$f(Q,r,\xi,z') = -\frac{1}{z'^2}\left[\alpha^{-1}\left(\frac{\xi^2}{Q^2} - \frac{2\ln r}{1-r^2}\right) + \frac{2K}{1-r^2}\right], \qquad \text{(C·1)}$$

$$K = Q \int \tilde{\mathrm{D}}u \ln \hat{H}(u) E(u),$$

$$E(u) = 1 - \int \mathrm{D}y P(\Lambda),$$

$$\Lambda = \xi y + \sqrt{1-\xi^2}Qu.$$

To obtain the asymptotic form, it is necessary to expand $\hat{H}$ and $\ln\hat{H}$ in terms of r and z'. To include possible lower order terms, we have to retain up to the fourth order terms in r and z'.

$$\hat{H}(u) = 1 + \frac{rz'^3}{3}h(u) + z'^2\left(-\frac{1-r^2}{2} + \frac{z'^2}{8}\right)H(-u), \qquad \text{(C·2)}$$

$$\ln\hat{H}(u) = \frac{rz'^3}{3}h(u) - \frac{1-r^2}{2}z'^2 H(-u) + \frac{z'^4}{8}H(-u)H(u), \qquad \text{(C·3)}$$

$$h(u) \equiv \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-u^2/2}.$$

Assuming $P(y) \simeq a\ \mathrm{sgn}(y)|y|^\delta$ around $y=0$ and substituting these expressions into eq. (C·1), we obtain the asymptotic form of the minimum error fraction as

$$f \simeq \epsilon_{\min} - \frac{\alpha^{-1}}{z'^2}\left(\frac{\xi^2}{Q^2} - 2\ln r\right) - \frac{2rz'Q}{3\sqrt{2\pi}} - \frac{Qz'^2}{4\sqrt{2\pi}} + \frac{2s\left(\dfrac{Q^2+\xi^2}{1-\xi^2}\right)^{(1+\delta)/2}}{\sqrt{2\pi}(1+\delta)}. \qquad \text{(C·4)}$$

Appendix D: The Asymptotic Behavior for χ in the Case 2

We prove that when $z' \to z'_0 (z'_0 > 0)$, $r \to r_0 (0 < r_0 < 1)$, $Q \to 0$ and $\xi \to 0$ as $\alpha \to \infty$ then $\chi \to \chi_0 (0 < \chi_0 < \infty)$.

Proof. If we assume $\chi \to 0$, the lowest order equations of eqs. (57) and (58) are

$$\alpha^{-1} = -\frac{Q^3}{\xi^2}\frac{1}{\sqrt{2\pi}\hat{H}(0)}A(r_0, z'_0), \tag{D·1}$$

$$\alpha^{-1} = -\frac{k}{\pi}Q\sqrt{2\pi}\frac{\mathcal{L}(0)}{\hat{H}(0)}, \tag{D·2}$$

$$\mathcal{L}(0) = \frac{1}{\sqrt{2\pi}}(\mathrm{e}^{-\frac{1}{2}z_0'^2} - 1), \quad \hat{H}(0) > 0,$$

$$A(r_0, z'_0) \equiv \int_0^\infty \mathrm{d}u u[(1-k)\mathcal{L}(u) - (1+k)\mathcal{L}(-u)] < \infty.$$

From these equations, we obtain

$$\chi^2 = \frac{\pi}{k}\frac{1}{\sqrt{2\pi}\mathcal{L}(0)}A(r_0, z'_0). \tag{D·3}$$

As long as $A(r_0, z'_0) \neq 0$, this is a contradiction. Next, we assume $\chi \to \infty$. Transforming $u \to \frac{u}{\chi}$, we obtain the lowest order equations,

$$\alpha^{-1}\chi^2 = -QB(r_0, z'_0), \tag{D·4}$$

$$\alpha^{-1}\chi = \frac{kQ}{2\pi}\ z_0'^2, \tag{D·5}$$

$$B(r_0, z'_0) \equiv \int_0^\infty \frac{\mathrm{d}u}{\sqrt{2\pi}} u \frac{\mathcal{L}(u)}{\hat{H}(u)} < \infty.$$

From these,

$$\frac{k}{2\pi}\chi\ z_0'^2 \simeq -B(r_0, z'_0). \tag{D·6}$$

Since $B(r_0, z'_0) < \infty$ and $z'_0 > 0$, this is a contradiction.

1) T. H. Watkin, A. Rau and M. Biehl: Rev. Mod. Phys. **65** (1993) 499.
2) H. S. Seung, H. Sompolinsky and N. Tishby: Phys. Rev. A **45** (1992) 6056.
3) I. Barkai, D. Hansel and H. Sompolinsky: Phys. Rev. A **45** (1992) 4146.
4) D. Hansel, G. Mato and C. Meunier: Europhys. Lett. A **20** (1992) 471.
5) G. Györgyi and N. Tishby: *Neural Networks and Spin Glasses*, ed. W. K. Theumann and R. Köberle (World Scientific, Singapore, 1990) p. 3.
6) M. Opper and D. Haussler: Phys. Rev. Lett. **20** (1991) 2677.
7) A. Blumer, A. Ehrenfeuncht, D. Haussler and M. K. Warmuth: *Proc. 18th ACM Symp. on Theory of Computation, Berkeley, CA., 1986.*
8) E.'B. Baum and D. Haussler: Neural Comp. **1** (1989) 151.
9) S. Amari: Neural Networks **6** (1993) 161.
10) D. Haussler, N. Litlestone and M. K. Warmuth: *Proc. the 1988 Workshop on Computational Learning Theory, Morgan Kaufmann, San Mateo.*
11) S. Amari, N. Fujita and S. Shinomoto: Neural Comp. **4** (1992) 605.
12) Y. Kabashima and S. Shinomoto: Neural Comp. **4** (1992) 712.
13) J. Kim and D. Pollard: The Annals of Statistics **18** (1990) 191.
14) T. Uezu, Y. Kabashima, K. Nokura and N. Nakamura: presented in Stat. Phys. **19**; T. Uezu and Y. Kabashima: J. Phys. A **29** (1996) L55.
15) E. Gardner: J. Phys. A **21** (1988) 257.
16) E. Gardner and B. Derrida: J. Phys. A **21** (1988) 271.
17) M. Bouten: J. Phys. A **27** (1994) 6021.