Learning from stochastic rules under finite temperature - optimal temperature and asymptotic learning curve

J. Phys. A: Math. Gen. **30** (1997) L777–L784. Printed in the UK PII: S0305-4470(97)83645-1

LETTER TO THE EDITOR

Learning from stochastic rules under finite temperature—optimal temperature and asymptotic learning curve

Tatsuya Uezu[†]

Department of Physics, Nara Women's University, Nara 630, Japan

Received 24 April 1997, in final form 8 September 1997

Abstract. In learning under external disturbance, it is expected that some tolerance in the system will optimize the learning process. In this paper, we give one example of this in learning from stochastic rules by the Gibbs algorithm. Using the replica method, we show that for the case of output noise, there exists an optimal temperature at which the generalization error is a minimum. This temperature exists even in the limit of large training sets and is determined by the stable replica symmetric solution. On the other hand, for other types of noise no such temperature exists and the asymptotic behaviour is determined by the one-step replica symmetric breaking solution. Further, the asymptotic expressions for learning curves are derived. They are precisely the same as those for the minimum-error algorithm.

In recent years, the problem of learning from examples in neural networks has been extensively studied from the viewpoint of statistical mechanics [1, 2] since Gardner opened the way to treat the problem by means of the replica method [3]. Since there are many learning algorithms, it is important to evaluate generalization ability acquired through learning. For this purpose, the learning curves of the generalization error ϵ_{g}, which is the probability of false prediction on a novel example, have been calculated for various types of networks, and it has been clarified that there are rich behaviours of learning curves depending on the details of the networks [1, 2].

Among these learning behaviours, it seems very interesting that there may exist an optimal procedure for learning. In particular, we expect the presence of an optimal temperature in learning stochastic rules by perceptrons for the Gibbs algorithm. The reason is that in general the existence of some proper tolerance will allow the system to adapt to an external disturbance more easily, and in the Gibbs algorithm, the temperature represents the measure of tolerance in selecting suitable network parameters, i.e. synaptic weights.

On this subject, there have been several studies. Györgyi [4] and Györgyi and Tishby [5] studied the perceptron learning for the spherical synaptic weight under finite temperature by the replica method. They treated the input-noise model in which input examples are subject to external disturbance and found there is an optimal temperature at which the generalization error takes the minimum value. Since the replica symmetric (RS) solution becomes unstable for large α, they considered the optimal temperature only for the restricted region of α where the RS solution is stable. Here, α is defined as $\alpha \equiv p/N$, where N is the dimension of synaptic weights and p is the number of examples.

[†] E-mail address: uezu@cc.nara.wu-ac.jp

0305-4470/97/220777+08$19.50

The other study by Opper and Haussler is on the output-noise model in which the binary output by a teacher is reversed by a given probability [6]. They studied learning behaviour at the special temperature $T_{\rm OH}$ for which the RS solution is stable for any value of α. They show that the temperature $T_{\rm OH}$ is optimal for the Bayes algorithm. Further they state that the fast convergence of the learning curve, $\Delta\epsilon_{\rm g} \equiv (\epsilon_{\rm g} - \epsilon_{\rm min}) \propto \alpha^{-1}$, where $\epsilon_{\rm min}$ is the minimum of the generalization error, is due to learning under the finite temperature. However, it is not obvious whether there exists an optimal temperature and, even if it exists, whether it is equal to $T_{\rm OH}$ for the Gibbs algorithm.

The purpose of this paper is to investigate whether an optimal temperature exists in the problem of learning from stochastic examples by perceptron with spherical synaptic weights for the Gibbs algorithm, especially in the asymptotic region of $\alpha \to \infty$. We calculate the learning curve for general types of noise including input and output noise by replica method under the ansatz up to the one-step replica symmetry breaking (1RSB).

Let us describe the problem we treat in this paper. For details, see [7, 8]. Hereafter, the norm of any vector is set to $\sqrt{N}$. We consider a stochastic target relation between a N-dimensional input vector $\boldsymbol{x}$ and a binary output $r^{\rm o} \in \{1, -1\}$ which is represented by a conditional probability $p_r(r^{\rm o}|\boldsymbol{x})$. We assume that the probability is a function of the inner product between the input $\boldsymbol{x}$ and the optimal spherical weight $\boldsymbol{w}^{\rm o}$ as

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

We further assume that the function $P(u)$ is increasing and behaves as

$$P(u) \simeq a\, \mathrm{sgn}(u)|u|^{\delta} \qquad (2)$$

near $u = 0$. Further, $P(-u) = -P(u)$ is assumed for brevity. The generalization error $\epsilon_{\rm g}$ is expressed as

$$\epsilon_{\rm g} = \epsilon_{\rm min} + \int_0^{\infty} \mathrm{D}y\, P(y) H\left(\frac{Ry}{\sqrt{1 - R^2}}\right) \qquad \epsilon_{\rm min} = \tfrac{1}{2} - \int_0^{\infty} \mathrm{D}y\, P(y)$$

where $H(x) = \int_x^{\infty} \mathrm{D}y$, $\mathrm{D}y = \exp(-y^2/2)\, \mathrm{d}y/\sqrt{2\pi}$ and R is the overlap between the optimal weight and the weight of a learner.

When the training set $\xi_p = \{(\boldsymbol{x}_1, r_1^{\rm o}), (\boldsymbol{x}_2, r_2^{\rm o}), \ldots, (\boldsymbol{x}_p, r_p^{\rm o})\}$ is given, we define the energy of a weight $\boldsymbol{w}$ as the number of discrepancy between outputs by the teacher and by the learner, $E[w, \xi_p] = \sum_{\mu=1}^{p} \Theta(-r_\mu u_\mu)$, where $\Theta(u)$ is the Heaviside function and $u_\mu = (\boldsymbol{x}_\mu \cdot \boldsymbol{w})/\sqrt{N}$.

The learning strategy we adopt is the Gibbs algorithm, in which a synaptic weight $\boldsymbol{w}$ is selected according to the probability proportional to $\mathrm{e}^{-\beta E[w,\xi_p]}$, where β is the inverse temperature, $\beta = 1/T$. The algorithm for the limit $T \to +0$ corresponds to the minimum-error algorithm, in which only the synaptic weights whose energies are minimum are selected. Therefore, the temperature represents a measure of tolerance in selecting synaptic weights. The partition function Z is expressed as

$$Z = \int \mathrm{d}\boldsymbol{w}\, \delta(\boldsymbol{w}^2 - N)\mathrm{e}^{-\beta E[w,\xi_p]} = \int \mathrm{d}\boldsymbol{w}\, \delta(\boldsymbol{w}^2 - N)\Pi_{\mu=1}^{p}[\mathrm{e}^{-\beta} + (1 - \mathrm{e}^{-\beta})\Theta(r_\mu u_\mu)]$$

and we calculate the average free energy per synapse f by the standard replica recipe

$$-\beta N f = \langle \ln Z \rangle = \lim_{n \to 0} \frac{1}{n}(\langle Z^n \rangle - 1)$$

where $\langle\rangle$ denotes the average over quenched variables $\boldsymbol{x}_\mu, r_\mu^{\rm o}$ and $\boldsymbol{w}^{\rm o}$.

The results obtained in this paper are summarized as follows.

(1) The asymptotic behaviours of learning curves as $\alpha \to \infty$ in the Gibbs algorithm are characterized by the local property of the probability $\mathcal{P}$, the exponent δ. The expressions of the generalization error for the RS solutions are

$$\Delta\epsilon_{\mathrm{g}} \simeq \psi_0^{(RS)}(T)\alpha^{-1} \qquad \text{for } \delta = 0 \tag{3}$$

$$\Delta\epsilon_{\mathrm{g}} \simeq \{\psi_1^{(\mathrm{RS})}(T)\}^{\frac{2(1+\delta)}{1+3\delta}}\alpha^{-\frac{1+\delta}{1+3\delta}} \qquad \text{for } \delta > 0 \tag{4}$$

and the ones for the 1RSB solutions are

$$\Delta\epsilon_{\mathrm{g}} \simeq \psi_0^{(\mathrm{1RSB})}(T)\alpha^{-1} \qquad \text{for } \delta = 0 \tag{5}$$

$$\Delta\epsilon_{\mathrm{g}} \simeq \psi_\delta^{(\mathrm{1RSB})}(T)\alpha^{-\frac{1+\delta}{1+2\delta}}(\ln\alpha)^{\frac{1+\delta}{2(1+2\delta)}} \qquad \text{for } \delta > 0. \tag{6}$$

The α-dependence of these expressions are precisely the same as those obtained by the minimum-error algorithm both for the RS and 1RSB solutions [7, 8].

(2) For the output-noise model ($\delta = 0$), there exists an optimal temperature for α larger than some critical value $\bar{\alpha}_0$, and this temperature really gives the minimum generalization error and is determined by the stable RS solution. $\psi_0^{(\mathrm{RS})}(T)$ attains its minimum value at the finite value of T_0^* where the RS solution is stable. For other types of noise ($\delta > 0$), the asymptotic behaviour is determined by the 1RSB solution and the optimal temperature does not exist. That is, although $\psi_1^{(\mathrm{RS})}(T)$ behaves as

$$\psi_1^{(\mathrm{RS})}(T) \propto \sqrt{1/T} \qquad \text{for } T \ll \alpha^{-1}$$

the RS solution is asymptotically unstable for fixed T. On the other hand, $\psi_\delta^{(\mathrm{1RSB})}(T)$ is independent of temperature for a wide range of T,

$$\psi_\delta^{(\mathrm{1RSB})}(T) \propto \delta^{\frac{1+\delta}{2(1+2\delta)}} \qquad \text{for } \alpha^{-\frac{\delta}{3(1+2\delta)}} \ll T \ll \alpha^{\frac{\delta}{1+2\delta}}.$$

That is, no special temperature exists.

Now, let us go into detailed calculations. In this paper, we use the following notation. $q^{\alpha\beta} = \frac{(\boldsymbol{w}^\alpha\cdot\boldsymbol{w}^\beta)}{N}$ denotes the overlap between weights of learners and $R^\alpha = \frac{(\boldsymbol{w}^{\mathrm{o}}\cdot\boldsymbol{w}^\alpha)}{N}$ the overlap between the weight of a learner and the optimal weight. Since the asymptotic behaviours of these quantities are different in the two cases of $\delta = 0$ and $\delta > 0$, we treat these cases separately.

The case of $\delta = 0$. Assuming the RS ansatz, we put $q^{\alpha\beta} = q$ and $R^\alpha = R$. For general $\mathcal{P}$, the RS free energy $f^{(\mathrm{RS})}$ is expressed as

$$-\frac{1}{T}f^{(\mathrm{RS})} = \frac{1-R^2}{2(1-q)} + \frac{1}{2}\ln 2\pi(1-q) + \alpha\int \mathrm{D}y\, 2\mathcal{P}(y)\int \mathrm{D}u\, \ln\tilde{H}\left(\frac{\sqrt{q-R^2}u - Ry}{\sqrt{1-q}}\right)$$

where $\tilde{H}(u) = \mathrm{e}^{-\beta} + (1-\mathrm{e}^{-\beta})H(u)$. As an example, we treat $P(u) = k\,\mathrm{sgn}(u)$ for several values of k and calculate $\Delta\epsilon_{\mathrm{g}}$. $\Delta\epsilon_{\mathrm{g}}$ attains the minimum value at the finite temperature $T_0(\alpha)$ for $\alpha > \bar{\alpha}_0$. $T_0(\alpha)$ increases to a finite value T_0^* as α increases to infinity. In this case, by investigating the AT instability, we can show that there exists a region of temperature T where the RS solution is stable for arbitrary α [10]. Let $T_0^{\mathrm{AT}}(\alpha)$ be the temperature where the AT-instability takes place. We found that $T_0^{\mathrm{AT}}(\alpha) \leqslant T_0(\alpha)$, and then $T_0^{\mathrm{AT}*} \equiv \lim_{\alpha\to\infty} T_0^{\mathrm{AT}}(\alpha) \leqslant T_0^*$. The difference $T_0^* - T_0^{\mathrm{AT}*}$ tends to zero as the magnitude of noise $(1-k)$ increases. As for the asymptotic behaviour, from the saddle-point

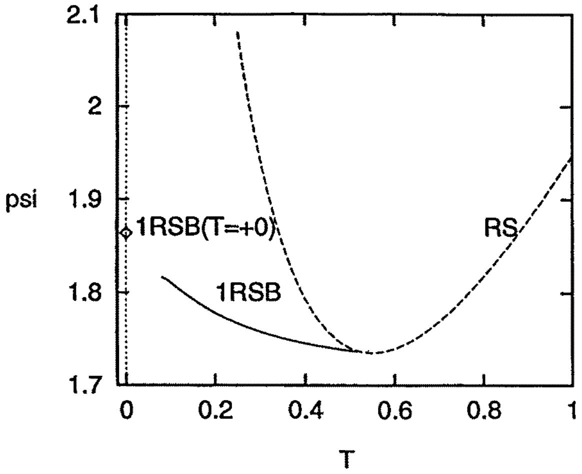

Figure 1. Temperature dependence of prefactors of asymptotic learning curves for the output-noise model, $P(u) = k$ sgn (u) with $k = 0.5$. Broken and full curves denote $\psi_0^{(\mathrm{RS})}(T)$ and $\psi_0^{(\mathrm{1RSB})}(T)$, respectively. $\psi_{\mathrm{MEA}}^{(\mathrm{1RSB})}$ for 1RSB solution in the minimum-error algorithm is also shown by a diamond symbol.

equations we can prove that $\chi \equiv \sqrt{\frac{q-R^2}{1-q}}$ tends to a constant, $\Delta q \equiv (1-q) \propto \alpha^{-2}$ and $\Delta R \equiv (1-R) \propto \alpha^{-2}$. Then the asymptotic form of generalization error becomes

$$\Delta\epsilon_{\mathrm{g}} \simeq \frac{k}{\pi}(2\Delta R)^{1/2} \simeq \psi_0^{(\mathrm{RS})}(T)\alpha^{-1}. \tag{7}$$

$\psi_0^{(\mathrm{RS})}(T)$ attains its minimum value at $T = T_0^*$. See figure 1. To see whether the minimum obtained above is really the absolute minimum, we calculate the RSB solutions outside the stable region of the RS solution and also in the minimum-error algorithm of the $T \to +0$ limit. According to Parisi, we put $q^{\alpha\beta} = q_0$ for $I(\alpha/m) \neq I(\beta/m), q^{\alpha\beta} = q_1$ for $I(\alpha/m) = I(\beta/m)$ and $\alpha \neq \beta, R^\alpha = R$, where $I(x)$ is the Gauss function [9]. For general $\mathcal{P}$, the free energy $f^{(\mathrm{1RSB})}$ for 1RSB is expressed as follows.

$$-\frac{1}{T}f^{(\mathrm{1RSB})} = \frac{1}{2} + \frac{\varphi}{2}(q_0 - R^2) + \frac{1}{2}\ln 2\pi(1-q_1) - \frac{1}{2m}\ln\varphi(1-q_1) + \frac{\alpha}{m}\int \mathrm{D}y 2\mathcal{P}(y)$$
$$\times \int \mathrm{D}z_0 \ln\left\{\int \mathrm{D}z_1\, \tilde{H}\left(\frac{\sqrt{q_0-R^2}z_0 + \sqrt{q_1-q_0}z_1 - Ry}{\sqrt{1-q_1}}\right)^m\right\} \tag{8}$$

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

From the numerical results, we find that $m, \tilde{\chi} \equiv \frac{\xi}{\tilde{\eta}} = \sqrt{\frac{q_0-R^2}{q_1-q_0}}$ and $\tilde{\varepsilon} \equiv \sqrt{\frac{1-q_1}{q_1-q_0}}$ tend to constants as $\alpha \to \infty$, where $\xi \equiv \sqrt{1-\frac{R^2}{q_0}}$ and $\tilde{\eta} \equiv \sqrt{\frac{q_1-q_0}{q_0}}$. In these limits, the saddle-point equations for $m, \tilde{\chi}$ and $\tilde{\varepsilon}$ are given by

$$m\tilde{\chi}^2 - m - \tilde{\varepsilon}^2 = -\tilde{\varepsilon}(m+\tilde{\varepsilon}^2)\frac{\tilde{A}_2}{2k\tilde{A}_3}$$
$$\left(\frac{m\tilde{\chi}}{m+\tilde{\varepsilon}^2}\right)^2 = \frac{m}{m+\tilde{\varepsilon}^2} + \ln\frac{\tilde{\varepsilon}^2}{m+\tilde{\varepsilon}^2} - \frac{m^2\tilde{\varepsilon}}{k(m+\tilde{\varepsilon}^2)}\frac{\tilde{A}_4}{\tilde{A}_3}$$

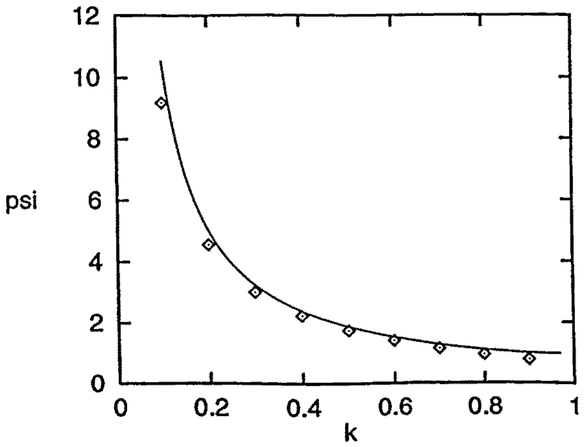

Figure 2. Noise amplitude dependence of prefactors of asymptotic learning curves for $P(u) = k$ sgn (u). Full curve and diamonds denote $\psi_{\text{MEA}}^{(1\text{RSB})}$ and $\psi_0^{(\text{RS})}(T_0^*)$, respectively.

$$\tilde{A}_3 = \frac{\tilde{\varepsilon}^2}{4k(\tilde{\chi}^2 - m - \tilde{\varepsilon}^2)}\tilde{A}_1$$

where the $\tilde{A}_i$s are expressed by integrations and functions of $m, \tilde{\chi}, \tilde{\varepsilon}$ and β. By solving these equations, we obtain the following scalings,

$$\Delta q_1 \equiv (1 - q_1) \simeq \Delta q_{1,0}\alpha^{-2} \qquad \Delta q_0 \equiv (1 - q_0) \simeq \Delta q_{0,0}\alpha^{-2} \qquad \Delta R \simeq \Delta R_0\alpha^{-2}$$
$$\Delta\epsilon_{\text{g}} \simeq \psi_0^{(1\text{RSB})}(T)\alpha^{-1}.$$

We find that $\psi_0^{(1\text{RSB})}(T)$ gradually increases as T decreases from $T_0^{\text{AT}}(\alpha)$. See figure 1. In the minimum-error algorithm, $\Delta\epsilon_{\text{g}}$ for the 1RSB solution is expressed as $\Delta\epsilon_{\text{g}} \simeq \psi_{\text{MEA}}^{(1\text{RSB})}\alpha^{-1}$ [7, 8]. As shown in figure 2, $\psi_{\text{MEA}}^{(1\text{RSB})} > \psi_0^{(\text{RS})}(T_0^*)$ holds for any value of k. Thus T_0^* really gives the absolute minimum.

The case of $\delta > 0$. This case includes the input-noise case ($\delta = 1$). As a typical example for $\delta = 1$, we treat $P(u) = k(1 - 2H(u))$ with $k = 1$. Let us consider the RS solution. As in the case of $\delta = 0$, there exists the optimal temperature $T_1(\alpha) > 0$ for $\alpha > \bar{\alpha}_1$. The numerical results suggest that the optimal temperature $T_1(\alpha)$ tends to infinity as α tends to infinity. In fact, this is the case as is shown in the following. For general δ, we can prove that in the RS solution χ tends to infinity as $\alpha \to \infty$ and the asymptotic form is expressed as

$$\Delta q \propto \alpha^{-\frac{2(1+\delta)}{1+3\delta}} \qquad \Delta R \propto \alpha^{-\frac{2}{1+3\delta}}$$
$$\Delta\epsilon_{\text{g}} \simeq \frac{2s}{\sqrt{2\pi}(1+\delta)}(2\Delta R)^{\frac{1+\delta}{2}} \simeq \{\psi_1^{(\text{RS})}(T)\}^{\frac{2(1+\delta)}{1+3\delta}}\alpha^{-\frac{1+\delta}{1+3\delta}}$$

where $s = a\int_0^\infty \text{D}x\, x^{1+\delta}$. The behaviour of $\psi_1^{(\text{RS})}(T)$ near $\beta = 1/T \sim 0$ is

$$\psi_1^{(\text{RS})}(T) \propto \sqrt{\beta} \qquad \text{for } \alpha^{-1} \ll \beta.$$

See figure 3. Thus, $\lim_{\alpha\to\infty} T_1(\alpha) = \infty$ follows. This is rather a strange result and the RS solution seems to be inadequate. Indeed, from numerical calculations we observe $T_1^{\text{AT}}(\alpha) \geqslant T_1(\alpha)$, that is, the RS solution is unstable at the optimal temperature $T_1(\alpha)$ for

$\delta = 1$. Further, we can prove that the RS solution is always asymptotically unstable for fixed T and for any δ. Therefore, we consider the 1RSB solution. From numerical results, we assume the following asymptotic behaviours, $m \ll 1, \tilde{\varepsilon} \ll 1, \tilde{\eta} \ll 1, \tilde{\chi} \gg 1$. Then, the saddle-point equations become

$$\varphi^2 q_0 m \xi^2 - \varphi + 1 \simeq -\frac{\alpha\beta}{\tilde{\eta}} \left(\frac{\beta m}{2\pi\sqrt{2}} - \frac{2s}{\sqrt{2\pi}} \xi^{\delta-1} \tilde{\eta} \right) \tag{9}$$

$$\varphi \xi^2 - 1 \simeq -\frac{\alpha\tilde{\eta}}{\sqrt{2\pi}} \left(2s\beta\tilde{\eta}\xi^{\delta-1} - a_{00}\tilde{\beta}\tilde{\varepsilon} - m\frac{\beta^2}{2\sqrt{\pi}} \right) \tag{10}$$

$$\varphi \frac{1}{2}(q_1 - q_0) \left(\varphi q_0 \xi^2 - \frac{1}{m} \right) - \frac{1}{2m^2} \ln\{\varphi(1 - q_1)\} \simeq \frac{\alpha\beta^2\tilde{\eta}}{2\pi\sqrt{2}} \tag{11}$$

$$\varphi R \xi \simeq \frac{2s}{\sqrt{2\pi}} \alpha\beta\xi^{\delta-1} \tag{12}$$

where $\tilde{\beta} = 1 - \mathrm{e}^{-\beta}$, $a_{jk} = \int \mathrm{D}x \, \frac{x^{j+1}}{\tilde{H}(x)} \{\ln \tilde{H}(x)\}^k$. From these equations, we obtain

$$\begin{aligned}
\Delta q_1 &\simeq \Delta q_{1,0} \alpha^{-\frac{2(1+\delta)}{1+2\delta}} (\ln\alpha)^{-\frac{1+3\delta}{1+2\delta}} \\
\Delta q_0 &\simeq \Delta q_{0,0} \alpha^{-\frac{2}{1+2\delta}} (\ln\alpha)^{-\frac{2\delta}{1+2\delta}} \\
\Delta R &\simeq \Delta R_0 \alpha^{-\frac{2}{1+2\delta}} (\ln\alpha)^{\frac{1}{1+2\delta}} \\
m &\simeq m_0 \alpha^{-\frac{\delta}{1+2\delta}} (\ln\alpha)^{-\frac{1+3\delta}{2(1+2\delta)}} .
\end{aligned}$$

Therefore, the generalization error becomes

$$\Delta\epsilon_{\mathrm{g}} \simeq \psi_\delta^{(1\mathrm{RSB})}(T) \alpha^{-\frac{1+\delta}{1+2\delta}} (\ln\alpha)^{\frac{1+\delta}{2(1+2\delta)}} .$$

The prefactor $\psi_\delta^{(1\mathrm{RSB})}(T)$ is temperature independent for a wide range of T. That is,

$$\psi_\delta^{(1\mathrm{RSB})}(T) \propto \delta^{\frac{1+\delta}{2(1+2\delta)}} \,\mathrm{for}\, \alpha^{-\frac{\delta}{3(1+2\delta)}} \ll T \ll \alpha^{\frac{\delta}{1+2\delta}} .$$

Thus, no optimal temperature exists for 1RSB solution.

From the above results, we note that the asymptotic behaviours of $\Delta\epsilon_{\mathrm{g}}$ (3)–(6) and other variables for the Gibbs algorithm are the same as those for the minimum-error algorithm in each case [7, 8].

Here, we make some comments on the results obtained in this paper.

Noise versus temperature. We can make a plausible argument for the existence of an optimal temperature at least for some region of α in the situation we consider. Let $\mathcal{V}$ be the set which consists of the weight vectors whose energies are the minimum value E_{min} and let α_c be the maximum value of α for which the minimum energy E_{min} is 0. For $\alpha \ll \alpha_c$, since the volume V of the set $\mathcal{V}$ is a large fraction of the total volume of the space of $\boldsymbol{w}$, the overlap R is determined by the vectors in $\mathcal{V}$ not only for $T = 0$ but also for $0 < T \ll 1$. Therefore, $\frac{\partial R}{\partial T}|_{T=0} \simeq 0$. On the other hand, for rather large values of α for which E_{min} is much smaller than the energy of $\boldsymbol{w}^{\mathrm{o}}$, $E[w^{\mathrm{o}}, \xi_p] \simeq p\epsilon_{\mathrm{min}}$, V is a very small fraction of the total volume. Since $\boldsymbol{w}^{\mathrm{o}} \notin \mathcal{V}$ and V is very small, for $T \ll 1$ among the vectors which contribute to the average R, the vectors outside $\mathcal{V}$ will bear more resemblance to $\boldsymbol{w}^{\mathrm{o}}$ than those inside $\mathcal{V}$. Thus, as T increases from 0, the participation of vectors outside the space $\mathcal{V}$ will make R increase. Thus, $\frac{\partial R}{\partial T}|_{T=0}$ becomes positive. In the case of the $T \to \infty$ limit, for any α R tends to zero since almost all students are selected equally. Therefore, at least for some region of α there seems to exist a moderate temperature T at which R

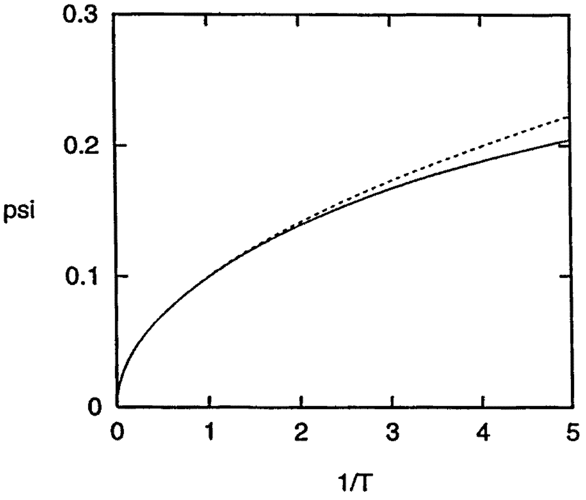

Figure 3. $\beta = 1/T$ dependence of prefactors of asymptotic learning curves for the input-noise model, $P(u) = k(1 - H(u))$ with $k = 1$. Full and broken curves denote $\psi_1^{(\mathrm{RS})}(\beta)$ and its approximation $c\sqrt{\beta}$ for $\beta \ll 1$, respectively.

takes the maximum value and then ϵ_{g} takes the minimum value. Our results in this paper are beyond this speculation and for α greater than some critical value $\bar{\alpha}_\delta$ we found the optimal temperature in the RS solution. Although this result is obtained for two classes of functions $P(u) = k\mathrm{sgn}(u)$ and $P(u) = k(1 - 2H(u))$, we consider that this is valid at least for increasing functions $P(u)$.

The model treated by Opper and Haussler corresponds to the case $P(u) = k\mathrm{sgn}(u)$ and they set the temperature to $T_{\mathrm{OH}} = (\ln \frac{1+k}{1-k})^{-1}$. This temperature is optimal for the Bayes algorithm [6]. From numerical calculation, we obtain the relation $T_0^{\mathrm{AT}*} \leqslant T_0^* < T_{\mathrm{OH}}$. Thus T_{OH} is in the stable region for the RS solution but not the optimal temperature for the Gibbs algorithm.

The reason that an optimal temperature exists only in the case of $\delta = 0$ in the asymptotic region is not obvious. However, the cause of different asymptotic behaviours of the generalization error for the cases $\delta = 0$ and $\delta > 0$ is considered as follows. As learning advances, the examples $\boldsymbol{x}$ which give crucial influence on the choice of weight vectors are those with $u^{\mathrm{o}} = (\boldsymbol{x} \cdot \boldsymbol{w}^{\mathrm{o}})/\sqrt{N} \sim 0$. The probability that such a sample answers $r^{\mathrm{o}} = +1$ is $\mathcal{P}(u^{\mathrm{o}}) = (1 + P(u^{\mathrm{o}}))/2$. Since $P(\pm 0)$ is finite for $\delta = 0$ and is equal to 0 for $\delta > 0$, it is obvious that learning advances faster in the case of $\delta = 0$ than in the case of $\delta > 0$.

Our result for $\delta = 0$ implies that the Gibbs algorithm is more efficient than the minimum-error algorithm. This is a remarkable result because we need not search for the lowest energy states which are very difficult to find in general, especially in systems which have many low energy states. Our result also implies the existence of an optimal schedule $T = T_0(\alpha)$, when we perform the Monte Carlo simulation by the Gibbs algorithm and increase the value of α. On the other hand, the result for $\delta > 0$ implies that there exists no special temperature in the Gibbs algorithm. Thus, we do not have to tune the temperature to achieve optimal performance in this case. It would be very interesting to perform these simulations and to see how learning advances. However, it is beyond the scope of this paper and left as a future problem.

I am grateful to Y Kabashima, K Nokura and P Davis for valuable discussions.

References

[1] Watkin T H, Rau A and Biehl M 1993 *Rev. Mod. Phys.* **65** 499

[2] Seung H S, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 6056

[3] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257

[4] Györgyi G 1990 *Phys. Rev.* A **41** 7097

[5] Györgyi G and Tishby N 1990 *Neural Networks and Spin Glasses* ed W K Theumann and R Köberle (Singapore: World Scientific) p 3

[6] Opper M and Haussler D 1991 *Proc. 4th Annual Workshop on Computational Learning Theory (COLT91)* (San Maeto, CA: Morgan Kaufmann) pp 75–87

[7] Uezu T and Kabashima Y 1996 *J. Phys. A: Math. Gen.* **29** L55

[8] Uezu T, Kabashima Y, Nokura K and Nakamura N 1996 *J. Phys. Soc. Japan* **65** 3797

[9] Parisi G 1980 *J. Phys. A: Math. Gen.* **13** 1101
Parisi G 1980 *J. Phys. A: Math. Gen.* **13** L115
Parisi G 1980 *J. Phys. A: Math. Gen.* **13** 1887

[10] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271