Asymptotic behaviour in learning from stochastic examples: one-step RSB calculation of the learning curve

1996 J. Phys. A: Math. Gen. 29 L55

(http://iopscience.iop.org/0305-4470/29/3/003)

J. Phys. A: Math. Gen. **29** (1996) L55–L60. Printed in the UK

LETTER TO THE EDITOR

Asymptotic behaviour in learning from stochastic examples: one-step RSB calculation of the learning curve

Tatsuya Uezu† and Yoshiyuki Kabashima‡

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

Received 30 October 1995

Abstract. The asymptotic learning curve of a N-dimensional stochastic learning model is calculated in the statistical mechanical framework. For the $N = 1$ case, Kabashima and Shinomoto (1992) gave a power law with exponent $\frac{2}{3}$. However, this is inconsistent with the RS solution with the exponent $\frac{1}{2}$ obtained by Györgyi and Tishby (1990). We show that the one-step RSB solution is consistent with Kabashima and Shinomoto's result up to a logarithmic correction.

In recent years, the problem of learning from examples has been an attractive topic in statistical mechanics. By using the replica method, learning curves of generalization error, which is the probability of a false prediction on a novel example, were calculated for various types of learning machines (in detail, see Watkin *et al* 1993). These studies revealed a rich behaviour 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 (Seung *et al* 1992, Barkai *et al* 1992, Hansel *et al* 1992). However, the most surprising result is the universality of the asymptote of learning curves when P/N is large (Györgyi and Tishby 1990, Opper and Haussler 1991). For the cases that the machine parameters are continuous and the target rules are deterministic and realizable by the students, the generalization error ε obeys the universal scaling relation

$$\varepsilon \sim N/P \tag{1}$$

independent of the architecture of machines. It should be remarked that the relation (1) is also derived by other methods than the replica trick, such as the annealed approximation (Levin *et al* 1989), the uniform convergence technique in computational learning theory (Blumer *et al* 1986, Baum and Haussler 1989) and the asymptotic analysis in statistics (Amari 1993).

However, it was pointed out by several authors that the scaling relation (1) does not yet hold when the target rule is unrealizable or seems stochastic to the learner (Haussler *et al* 1988, Györgyi and Tishby 1990, Seung *et al* 1992, Amari *et al* 1992, Kabashima and Shinomoto 1992). Among these studies, the one-dimensional binary choice problem studied by Kabashima and Shinomoto (1992) is striking because of its simplicity. The target rule considered in their model is a stochastic relation between real number input $x \in [0, 1]$ and

† E-mail address: uezu@cc.nara-wu.ac.jp
‡ E-mail address: kaba@cc.nara-wu.ac.jp

0305-4470/96/030055+06$19.50

binary output $y \in \{-1, +1\}$, $p(y|x)$. The function $p(+1|x) = 1 - p(-1|x)$ is assumed to be increasing and differentiable with respect to x, and assumed to have a point $\theta_o \in [0, 1]$ satisfying the condition $p(+1|\theta_o) = p(-1|\theta_o) = \frac{1}{2}$. Under these assumptions, the strategy returning $y = \mathrm{sign}(x - \theta_o)$ for an input x minimizes the generalization error. The learner's task is, therefore, to estimate the boundary θ_o from a given set of examples. The minimum-error algorithm which minimizes the training error, i.e. the number of false predictions on training examples, is a natural learning strategy. By using an analogy between the fluctuation of the training error and the motion of a random walk, Kabashima and Shinomoto found that the minimum-error algorithm gives a non-trivial power law

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

where $\varepsilon_{\min}$ is the minimum value of the generalization error obtained by the optimal boundary θ_o. A similar result was also discovered in the field of statistics (Kim and Pollard 1990).

The learning model studied by Györgyi and Tishby (1990) corresponds to a higher dimensional version of the above problem. In their model, the target rule is a perceptron with the weight $\boldsymbol{w}_0$ whose inputs are corrupted by noise. For an N-dimensional vector $\boldsymbol{x}$, the rule returns $y = \mathrm{sign}[\boldsymbol{w}_0 \cdot (\boldsymbol{x} + \boldsymbol{\eta})]$ where $\boldsymbol{\eta}$ is the Gaussian noise with zero mean. The learner estimates the weight $\boldsymbol{w}_0$ from a given set of input–output pairs in order to acquire a good generalization ability. It is an interesting question whether the power law (2) also holds in such an N-dimensional systems. However, one cannot directly use the same analogy as Kabashima and Shinomoto did for a higher dimensional model because their analysis depends strongly on the one-dimensional nature of the model. Györgyi and Tishby calculated the learning curve of this problem in the framework of statistical mechanics. The result obtained under the RS ansatz, however, was

$$\varepsilon - \varepsilon_{\min} \sim (N/P)^{1/2} \tag{3}$$

which, with $N = 1$, is different from equation (2). This RS solution was found to be thermodynamically unstable. Nevertheless, they conjectured that it is a good approximation because it is identical with that of the worst case analysis by Haussler *et al* (1988). The discrepancy between equations (2) and (3) is still unresolved.

The purpose of this paper is to resolve this discrepancy. In the following, we consider a stochastic relation between N-dimensional input vector $\boldsymbol{x}$ and binary output $y \in \{-1, +1\}$ as a generalized version of the problems treated by Györgyi–Tishby and Kabashima–Shinomoto. In order to perform a more precise analysis, we calculate the asymptotic behaviour of the learning curve under the one-step RSB ansatz in the statistical mechanical framework.

The main result of this paper is the following. With the one-step RSB calculation, it is found that the learning curve of the minimum-error algorithm scales as

$$\varepsilon - \varepsilon_{\min} \sim (N/P)^{2/3} \tag{4}$$

up to a logarithmic correction, when the stochastic target relation has no singularity over the input space. Thus, it is conjectured that the power law with the exponent $\frac{2}{3}$ is universal irrespective of the dimensionality of the model N.

Hereafter, we assume that an arbitrary N-dimensional vector $\boldsymbol{a}$ is normalized as $|\boldsymbol{a}| = \sqrt{N}$. We consider a stochastic target relation between N-dimensional input vector $\boldsymbol{x}$ and binary output $y \in \{-1, +1\}$ which is represented by a conditional probability $p(y|\boldsymbol{x}) = p(y \times \boldsymbol{w}_0 \cdot \boldsymbol{x}/\sqrt{N})$, where $\boldsymbol{w}_0$ is a fixed, unknown N-dimensional weight vector. We assume that the function $p(u)$ is increasing and differentiable with respect to u. Under

this assumption, the prediction $y = \mathrm{sign}[\boldsymbol{w}_0 \cdot \boldsymbol{x}/\sqrt{N}]$ minimizes the generalization error. Hence, the learner's task is to estimate the weight $\boldsymbol{w}_0$ from a given set of examples.

When inputs of a perceptron are corrupted by noise, the conditional probability that the outputs $y = \pm 1$ are generated becomes a smooth function of the inner product of the weight and the input. Therefore, the present model includes Györgyi and Tishby's learning model. Another typical noise is the output noise studied by Opper and Haussler (1991). In their model, it is assumed that the sign of each output is reversed with a probability $0 < \lambda < \frac{1}{2}$. Such a learning model can also be represented by the present model with a singular function $p(u) = \lambda + (1 - 2\lambda)\Theta(u)$. However, we concentrate on the former type of stochastic relation in this paper because another consideration is required for the latter type of problems (Uezu *et al* 1995).

By using the minimum-error algorithm, the learner estimates the weight $\boldsymbol{w}_0$ from a given set of P examples $\xi^P = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_P, y_P)\}$ which are independently and uniformly drawn from N-dimensional sphere of radius $\sqrt{N}$ centred at the origin and the conditional distribution $p(y|\boldsymbol{x})$. For a given realization of example ξ^P, the minimum-error algorithm minimizes the training error, i.e. the number of false predictions

$$E(\boldsymbol{w}|\xi^P) = \sum_{\mu=1}^{P} \Theta\left(-y_\mu \frac{\boldsymbol{w} \cdot \boldsymbol{x}}{\sqrt{N}}\right). \tag{5}$$

The performance of the learning is evaluated by the generalization error ε, which is the probability of false prediction on a novel example. Due to the assumption that the distribution of inputs is uniform on the N-dimensional sphere, ε becomes a function of the overlap between the optimal weight $\boldsymbol{w}_0$ and the estimator $\boldsymbol{w}$, $R = \boldsymbol{w}_0 \cdot \boldsymbol{w}/N$. One can show this is given by

$$\varepsilon = 2\int_{-\infty}^{+\infty} \mathrm{D}t \int_{-\infty}^{+\infty} \mathrm{D}z \; p(\sqrt{1-R}z + \sqrt{R}t) H\left(\sqrt{\frac{R}{1-R}}t\right) \tag{6}$$

where $\mathrm{D}x = \mathrm{d}x \exp[-x^2/2]/\sqrt{2\pi}$ and $H(x) = \int_x^\infty \mathrm{D}z$. In particular, when $\Delta R = 1 - R$ is small, we obtain the relation

$$\varepsilon - \varepsilon_{\min} \sim \frac{2p'(0)}{\sqrt{2\pi}} \Delta R \tag{7}$$

where $\varepsilon_{\min} = 2\int_{-\infty}^{+\infty} \mathrm{D}t \; p(t)[1 - \Theta(t)]$ is the minimum value of the generalization error which is attained by $R = 1$.

From the energy defined by the equation (5), the partition function with the inverse temperature β is given by

$$\begin{aligned} Z &= \int \mathrm{d}\boldsymbol{w} \; \delta(|\boldsymbol{w}|^2 - N) \exp[-\beta E(\boldsymbol{w}|\xi^P)] \\ &= \int \mathrm{d}\boldsymbol{w} \; \delta(|\boldsymbol{w}|^2 - N) \prod_{\mu=1}^{P} \left[\mathrm{e}^{-\beta} + (1 - \mathrm{e}^{-\beta})\Theta\left(y_\mu \frac{\boldsymbol{w} \cdot \boldsymbol{x}_\mu}{\sqrt{N}}\right)\right]. \end{aligned} \tag{8}$$

The averaged free energy can be calculated through the formula

$$\langle\langle f \rangle\rangle_{\xi^P} = -\frac{\langle\langle \ln Z \rangle\rangle_{\xi^P}}{\beta N} = -\frac{1}{\beta N} \lim_{n \to 0} \frac{\langle\langle Z^n \rangle\rangle_{\xi^P} - 1}{n} \tag{9}$$

where $\langle\langle \cdots \rangle\rangle_{\xi^P}$ means the average over the quenched variables ξ^P. This becomes a function of the order parameters $q_{ab} = \boldsymbol{w}_a \cdot \boldsymbol{w}_b/N$, $R_a = \boldsymbol{w}_0 \cdot \boldsymbol{w}_a/N$. The learning curve of the minimum-error algorithm is obtained in the limit $\beta \to \infty$. In the asymptotic region $\alpha = P/N \gg 1$, we have to take the replica-symmetry breaking (RSB) into account in the

evaluation of the equation (9) in the limit $\beta \to \infty$ (Györgyi and Tishby 1990, Bouten 1994). In the following, we investigate the solution within the one-step RSB ansatz. The one-step RSB solution is defined by four order parameters q_0, q_1, m, and R (or $\tilde{R} = R^2/q_0$). The three parameters q_0, q_1, m specify the replica overlap function $q(x)$ as $q(x) = q_0$ for $0 < x < m$ and $q(x) = q_1$ for $m < x < 1$, while single parameter R specifies the overlap between the optimal weight and a replica. In this framework, one can show that

$$\frac{\langle\langle \ln Z \rangle\rangle_{\xi^P}}{\beta N} = \mathrm{ext}_{\{\tilde{R},q_0,q_1,m\}} \left\{ \frac{2\alpha}{\beta m} \int_{-\infty}^{+\infty} \mathrm{D}t \ \Omega(\tilde{R} : t) \ln \left[\int_{-\infty}^{+\infty} \mathrm{D}s \{\Xi_\beta(q_1, q_0 : s, t)\}^m \right] \right.$$
$$+ \frac{m-1}{2\beta m} \ln(1-q_1) + \frac{1}{2\beta m} \ln[(1-q_1) + m(q_1 - q_0)]$$
$$\left. + \frac{q_0(1-\tilde{R})}{2\beta[(1-q_1) + m(q_1 - q_0)]} \right\} \tag{10}$$

where $\tilde{R} = R^2/q_0$ and

$$\Omega(\tilde{R} : t) = \int_{-\infty}^{+\infty} \mathrm{D}z \ p(\sqrt{1-\tilde{R}}z + \sqrt{\tilde{R}}t) \tag{11}$$

$$\Xi_\beta(q_1, q_0 : s, t) = \int_{-\infty}^{+\infty} \mathrm{D}z [\mathrm{e}^{-\beta} + (1 - \mathrm{e}^{-\beta}) \Theta(\sqrt{1-q_1}z + \sqrt{q_1 - q_0}s + \sqrt{q_0}t)]$$
$$= \mathrm{e}^{-\beta} + (1 - \mathrm{e}^{-\beta}) H\left(- \frac{\sqrt{q_1 - q_0}s + \sqrt{q_0}t}{\sqrt{1-q_1}} \right). \tag{12}$$

In the limit $\beta \to \infty$, a non-trivial result is obtained only when $q_1 \to 1$ and $m \to 0$ keeping both $w = \beta m$ and $c = m/(1-q_1)$ finite. With this ansatz, the equation (10) becomes

$$\lim_{\beta\to\infty} \frac{\langle\langle \ln Z \rangle\rangle_{\xi^P}}{\beta N} = \mathrm{ext}_{\{\tilde{R},q_0,c,w\}} \left\{ \frac{2\alpha}{w} \int_{-\infty}^{+\infty} \mathrm{D}t \ \Omega(\tilde{R} : t) \ln \tilde{\Xi}(q_0, c, w : t) \right.$$
$$\left. + \frac{1}{2w} \ln[1 + c(1-q_0)] + \frac{1}{2w} \frac{cq_0(1-\tilde{R})}{[1 + c(1-q_0)]} \right\} \tag{13}$$

where

$$\tilde{\Xi}(q_0, c, w : t) = \int_{-\infty}^{+\infty} \mathrm{D}z \ \tilde{\Theta}(c, w : \sqrt{1-q_0}z + \sqrt{q_0}t)$$
$$= H\left(-\sqrt{\frac{q_0}{1-q_0}}t \right) + \mathrm{e}^{-w} H\left(\sqrt{\frac{q_0}{1-q_0}}t + \sqrt{\frac{2w}{c(1-q_0)}} \right)$$
$$+ \frac{\mathrm{e}^{-\frac{cq_0t^2}{2[1+c(1-q_0)]}}}{\sqrt{1 + c(1-q_0)}} \left[H\left(\frac{1}{\sqrt{1 + c(1-q_0)}} \sqrt{\frac{q_0}{1-q_0}}t \right) \right.$$
$$\left. - H\left(\frac{1}{\sqrt{1 + c(1-q_0)}} \left\{ \sqrt{\frac{q_0}{1-q_0}}t + (1 + c(1-q_0)) \sqrt{\frac{2w}{c(1-q_0)}} \right\} \right) \right] \tag{14}$$

and

$$\tilde{\Theta}(c, w : u) = \Theta(u) + \mathrm{e}^{-w} \Theta\left(-u - \sqrt{\frac{2w}{c}} \right) + \mathrm{e}^{-cu^2/2} \left[\Theta(-u) - \Theta\left(-u - \sqrt{\frac{2w}{c}} \right) \right]. \tag{15}$$

The only solution obtained numerically from the saddle point (SP) equations behaves as $\tilde{R} \to 1$, $q_0 \to 1$, $c \to \infty$, and $w \to 0$ in the limit $\alpha \to \infty$. Further, the product $c(1-q_0)$ goes to infinity. In order to investigate how these convergences scale with α, we expand

the right-hand side of the equation (13) with respect to the small variables $\Delta\tilde{R} = 1 - \tilde{R}$, $\Delta q = 1 - q_0$, c^{-1}, and w around the above limits and obtain

$$-\alpha\left[\varepsilon_{\min} + \frac{p'(0)}{\sqrt{2\pi}}(\Delta\tilde{R} + \Delta q) - \frac{w\Delta q^{1/2}}{2\sqrt{2}\pi} - \frac{2w^{1/2}c^{-1/2}}{3\sqrt{\pi}} + \cdots\right] + \frac{1}{2w}\ln[c\Delta q] + \frac{\Delta\tilde{R}}{2w\Delta q}\,. \tag{16}$$

This gives the following SP equations:

$$\alpha \sim w^{-1}\Delta q^{-1} \tag{17}$$

$$\alpha - \alpha w\Delta q^{-1/2} \sim w^{-1}\Delta q^{-1} - w^{-1}\Delta q^{-2}\Delta\tilde{R} \tag{18}$$

$$\alpha w^{1/2}c^{-3/2} \sim w^{-1}c^{-1} \tag{19}$$

$$\alpha\Delta q^{1/2} + \alpha w^{-1/2}c^{-1/2} \sim w^{-2}\ln(c\Delta q) + w^{-2}\Delta q^{-1}\Delta\tilde{R} \tag{20}$$

which imply the following scalings:

$$\Delta\tilde{R} \sim (\ln\alpha)^{1/3}\alpha^{-2/3} \tag{21}$$

$$\Delta q \sim (\ln\alpha)^{-2/3}\alpha^{-2/3} \tag{22}$$

$$c \sim (\ln\alpha)^2\alpha \tag{23}$$

$$w \sim (\ln\alpha)^{2/3}\alpha^{-1/3}\,. \tag{24}$$

From these relations and the equation (7), we obtain the learning curve

$$\varepsilon - \varepsilon_{\min} \sim (\ln\alpha)^{1/3}\alpha^{-2/3} \sim [\ln(P/N)]^{1/3}(N/P)^{2/3} \tag{25}$$

which is consistent with the equation (2) up to a logarithmic correction (see figure 1).

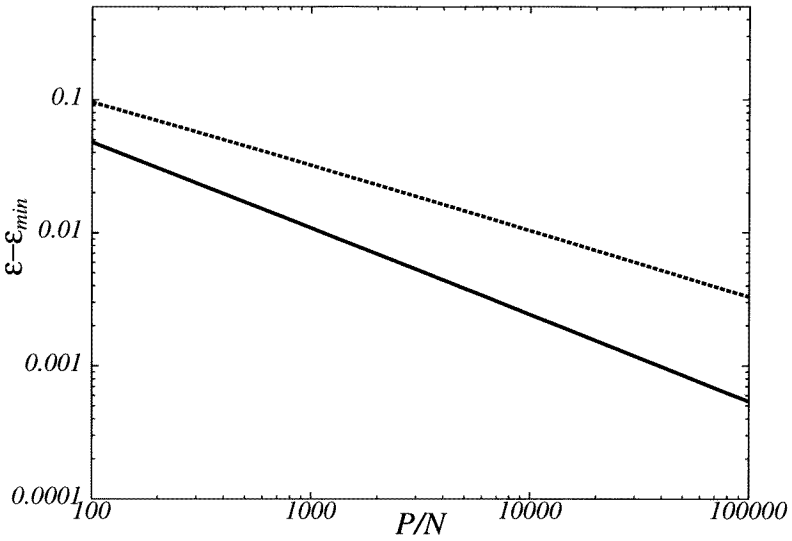

Figure 1. The learning curves for the function $p(u) = H(-u)$. The full and the broken curves represent the one-step RSB solution and the RS solution, respectively.

In summary, we calculated the asymptotic form of the learning curve of a N-dimensional stochastic learning model in the statistical mechanical framework in order to resolve the discrepancy between Kabashima and Shinomoto's result obtained in a one-dimensional system and Györgyi and Tishby's result obtained in a thermodynamic system. The

asymptotic form was obtained with the one-step RSB ansatz. It is consistent with Kabashima and Shinomoto's result for $N = 1$ up to a logarithmic correction. The discrepancy with Györgyi and Tishby's result shows that the RS calculation is insufficient†.

Finally, it is interesting to consider a singular function $p(u) = \frac{1}{2} + \mathrm{O}(\mathrm{sign}(u)|u|^{\delta})$ where $\delta \geqslant 0$ because this includes Györgyi and Tishby's input noise model ($\delta = 1$) and Opper and Haussler's output noise model ($\delta = 0$). For $\delta > 0$, it can be shown that one-step RSB calculations give the log-modified power law $\varepsilon - \varepsilon_{\min} \sim [\ln(P/N)]^{\frac{1+\delta}{2(1+2\delta)}} (N/P)^{\frac{1+\delta}{1+2\delta}}$, although RS calculations give a pure power law with the exponent $(1+\delta)/(1+3\delta)$. However, for the case $\delta = 0$, both of the RS and the one-step RSB solutions obey pure power laws $\varepsilon - \varepsilon_{\min} \sim N/P$ although the two solutions have different coefficients (Uezu *et al* 1995). Nevertheless, all the one-step RSB solutions are consistent with the power laws obtainable by Kabashima and Shinomoto's analogy in the corresponding one-dimensional problems up to logarithmic corrections. The details of these generalized cases will be reported elsewhere.

The authors especially thank N Nakamura, K Nokura and P Davis for helpful discussions and advice.

References

Amari S 1993 *Neural Networks* **6** 161
Amari S, Fujita N and Shinomoto S 1992 *Neural Comput.* **4** 605
Blumer A, Ehrenfeuncht A, Haussler D and Warmuth M K 1986 *Proc. 18th ACM Symp. on Theory of Computation (Berkeley)*
Barkai I, Hansel D and Sompolinsky H 1992 *Phys. Rev.* A **45** 4146
Baum E B and Haussler D 1989 *Neural Comput.* **1** 151
Bouten M 1994 *J. Phys. A: Math. Gen.* A **27** 6021
Györgyi G and Tishby N 1990 *Neural Networks and Spin Glasses* (Singapore: World Scientific)
Hansel D, Mato G and Meunier C 1992 *Euro. Phys. Lett.* **20** 471
Haussler D, Littlestone N and Warmuth M K 1988 *Proc. 1988 Workshop on Computational Learning Theory* (San Mateo, CA: Morgan Kaufmann)
Kim and Pollard 1990 *Ann. Stat.* **18** 191
Levin E N, Tishby N and Solla S A 1989 *Proc. 2nd Workshop on Computational Learning Theory* (San Mateo, CA: Morgan Kaufmann)
Opper M and Haussler D 1991 *Proc. 4th ACM Workshop on Computational Learning Theory* (San Mateo, CA: Morgan Kaufmann)
Seung H S 1995 *Neural Networks: The Statistical Mechanics Perspective* (Singapore: World Scientific)
Seung H S, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 6065
Uezu T, Kabashima Y, Nokura K and Nakamura N 1995 in preparation; presented in *STATPHYS 19*
Watkin T H, Rau A and Biehl M 1993 *Rev. Mod. Phys.* **65** 499

† Recently, Seung (1995) showed that a refined annealed approximation analysis gives the $\frac{2}{3}$ power law for Györgyi and Tishby's learning model. This result supports our one-step RSB calculation.