Communicated by Haim Sompolinsky

Learning Curves for Error Minimum and Maximum Likelihood Algorithms

Y. Kabashima
S. Shinomoto
Department of Physics, Kyoto University,
Kyoto 606, Japan

For the problem of dividing the space originally partitioned by a blurred boundary, every learning algorithm can make the probability of incorrect prediction of an individual example ε decrease with the number of training examples t. We address here the question of how the asymptotic form of $\varepsilon(t)$ as well as its limit of convergence reflect the choice of learning algorithms. The error minimum algorithm is found to exhibit rather slow convergence of $\varepsilon(t)$ to its lower bound ε_0, $\varepsilon(t) - \varepsilon_0 \sim O(t^{-2/3})$. Even for the purpose of minimizing prediction error, the maximum likelihood algorithm can be utilized as an alternative. If the true probability distribution happens to be contained in the family of hypothetical functions, then the boundary estimated from the hypothetical distribution function eventually converges to the best choice. Convergence of the prediction error is then $\varepsilon(t) - \varepsilon_0 \sim O(t^{-1})$. If the true distribution is not available from the algorithm, however, the boundary generally does not converge to the best choice, but instead $\varepsilon(t) - \varepsilon_1 \sim \pm O(t^{-1/2})$, where $\varepsilon_1 > \varepsilon_0 > 0$.

1 Introduction

The original purpose of machine learning is to adjust the machine parameters so as to reproduce the input–output relationship implied by the examples. Learning situations can be classified into two cases depending upon whether or not the machine is in principle able to completely reproduce the individual examples. In the case that the machine is able to reproduce examples, the remaining interest is the estimate of generalization error: the probability of the incorrect prediction ε of a novel example provided that the machine has succeeded to reproduce t examples. The problem has currently been resolved by two means: computational theoretical, and statistical mechanical. First, the idea of PAC learning by Valiant (1984) was applied by Baum and Haussler (1989) to the worst case estimate of the generalization error of the neural network models. Second, a statistical mechanical theory for typical case estimate of the generalization error is formulated under the Bayes formula by Levin *et* *al.* (1990). Amari *et al.* (1992) classified the asymptotic scaling forms of the learning curves $\varepsilon(t)$ into four types. The statistical theory is not restricted to the case that a machine can reproduce the raw examples. Actually, two among the four types of the scaling forms are concerned with the case that the examples shown by the supervisor are more or less noisy.

We take up here the convergence of prediction error for dividing the space originally partitioned by a blurred boundary. The purpose of the learning would not be unique in this case; one may seek the probability distribution of the classification, or one may seek the best boundary so as to minimize the prediction error for individual examples. The maximum likelihood algorithm and the error minimum algorithm are the corresponding standard strategies for these motivations. The two strategies are identical if the family of hypothetical distribution functions for the maximum likelihood algorithm are stepwise and symmetrical [see Rissanen (1989); Levin *et al.* (1990)]. In the case that the hypothetical distribution functions are smooth, however, the two strategies are generally different from each other.

We found that the convergence of the error minimum algorithm is rather slow. In this algorithm, $\varepsilon(t)$ converges to the lower bound ε_0 with the asymptotic form $\varepsilon(t)-\varepsilon_0 \sim O(t^{-2/3})$. We will explain the source of the fractional exponent 2/3 theoretically. Even for the purpose of minimizing prediction error, we can use the maximum likelihood algorithm as an alternative. In this case, the boundary can be defined as a hypersurface on which the hypothetical probabilities for alternative classes balance with each other. If the true probability distribution is available from the algorithm, the prediction error converges rapidly as $\varepsilon(t)-\varepsilon_0 \sim O(t^{-1})$. In the case that the true distribution is not available from the algorithm, the boundary generally does not converge to the best choice, but $\varepsilon(t) - \varepsilon_1 \sim \pm O(t^{-1/2})$, where $\varepsilon_1 > \varepsilon_0 > 0$.

2 Numerical Simulation

We first show the result of numerical simulation of the following simple partition problem. Every example consists of the real input $x \in [0,1]$ and the binary output $s = \pm 1$. Real number x is drawn independently from the uniform distribution over the interval, $p(x) = 1$. The probability of getting $s = \pm 1$ depends on x as $p(s = +1 \mid x) = 0.1 + 0.7x$ and $p(s = -1 \mid x) = 1-p(s = +1 \mid x)$ (Fig. 1a). We examined the following three strategies for the partition of the interval: (1) the error minimum algorithm, (2) the maximum likelihood algorithm with the family of probability functions $q_w(s = +1 \mid x) = w + 0.7x$, and (3) the maximum likelihood with $q_w(s = +1 \mid x) = w + 0.4x$.

The error minimum algorithm seeks the partition that minimizes the total number of the left points with $s = +1$ and the right points with $s = -1$. As the number of examples increases the partition point x_o is

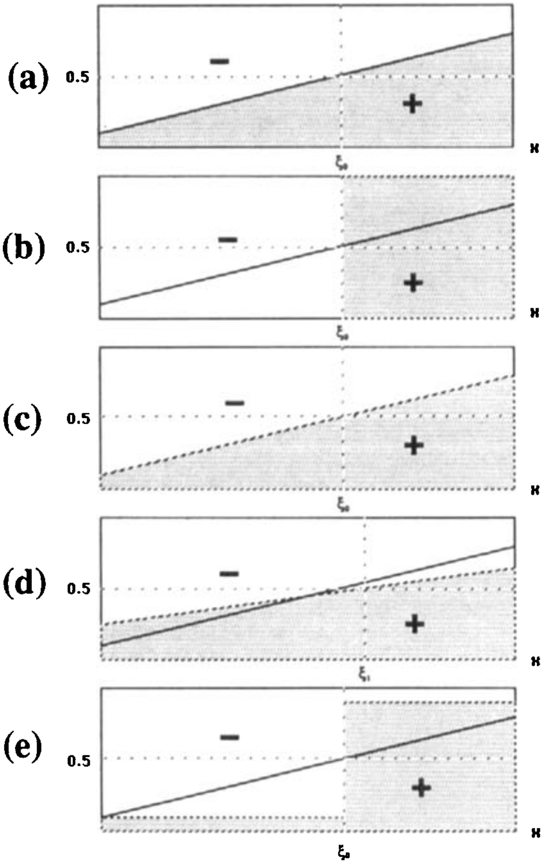

Figure 1: (a) The original probability distribution $p(s = +1 \mid x) = 0.1 + 0.7x$. (b) The best partition $\xi_0 = 4/7$ for the error minimum. (c–e) The best hypothetical distribution functions for the maximum likelihood with $p_w(s = +1 \mid x) = w + 0.7x, w + 0.4x$, and $\delta + (1 - 2\delta)\theta(x - w)$.

expected to approach the point $\xi_0 = 4/7$ at which the probabilities for the alternative classes balance: $p(s = +1 \mid \xi_0) = p(s = -1 \mid \xi_0)$ (Fig. 1b). For the given partition at x_o, the probability of incorrect prediction is $\varepsilon = \varepsilon_0 + a(x_o - \xi_0)^2$, where $\varepsilon_0 = 9/28$ and $a = 0.7$. In this algorithm, the possible position of the optimal partition x_o is given by the interval of adjacent points (x_i, x_j) and the error measure has to be averaged over the interval.

The maximum likelihood algorithm seeks the optimal parameter value w_o, which maximizes the likelihood function,

$$l_w = \sum_{s,x} \log q_w(s \mid x) \tag{2.1}$$

The original probability distribution is available from the algorithm (2), and the optimal parameter w_o is expected to approach $\omega_0 = 0.1$, which minimizes the Kullback divergence. As a result, the optimal partition x_o estimated by $q_{w_o}(s = +1 \mid x_o) = q_{w_o}(s = -1 \mid x_o)$ eventually approaches to the best choice, $\xi_0 = 4/7$ (Fig. 1c). On the other hand, algorithm (3) does not contain the true distribution function, and the optimal parameter is expected to approach to a value ω_1, which minimizes the Kullback divergence. In this case, the optimal partition x_o approaches to a point ξ_1 remote from $\xi_0 = 4/7$ (Fig. 1d). Note again that the maximum likelihood is identical to the error minimum if the family of hypothetical functions is stepwise and symmetrical (Fig. 1e), although this is rather exceptional as a maximum likelihood algorithm.

In the numerical simulation, the three algorithms are carried out to obtain the optimal partition x_o and the prediction error ε for the set of t examples drawn from the distribution, $p(s \mid x)p(x)$. The average of the prediction error ε taken over 1000 sets of examples is plotted in Figure 2. The plots of $\varepsilon(t) - \varepsilon_0$ for (1) and (2) exhibit the scaling $\varepsilon(t) - \varepsilon_0 \sim O(t^{-\alpha})$, with the exponents $\alpha = 0.670 \pm 0.004$ and $\alpha = 1.012 \pm 0.008$, respectively. The prediction error according to the algorithm (3) does not converge to the lower bound ε_0 but to $\varepsilon_1 (> \varepsilon_0)$. The mean square deviation of $\varepsilon(t)$ from ε_1 in case (3) is depicted in Figure 3, where we can see $\varepsilon(t) - \varepsilon_1 \sim \pm O(t^{-\alpha})$ with the exponent $\alpha = 0.520 \pm 0.004$. These results are examined in the next section.

3 Theoretical Interpretation

In order to elucidate the nontrivial exponent obtained from the error minimum (1), we wish to consider here the simpler situation that the examples are arranged at regular intervals of $1/t$, assuming the same form for $p(s \mid x)$. Let each example be denoted by the sequence j from 1 to t. The probability of s_j taking the value $s = \pm 1$ is given by $p(s \mid x = j/t)$. The total number of errors for the partition between i and $i + 1$ is

$$E_i = \sum_{j=1}^{i} (1 + s_j)/2 + \sum_{j=i+1}^{t} (1 - s_j)/2 \tag{3.1}$$

The expectation value of the number of errors is estimated as

$$\langle E_i \rangle \sim \langle E_m \rangle + (a/t)(i - m)^2 \tag{3.2}$$

where m is the best partition which minimizes the difference of the alternative probabilities, $|p(s = +1 \mid x = m/t) - p(s = -1 \mid x = m/t)|$. On

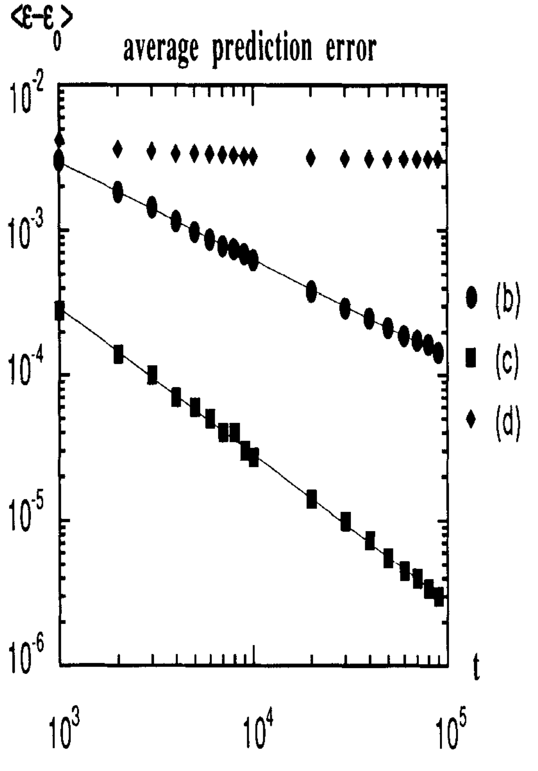

Figure 2: Average of $\varepsilon(t) - \varepsilon_0$. (b), (c), and (d) correspond to cases (1), (2), and (3), respectively. The lines for (1) and (2) were drawn from the least square fit: $\varepsilon(t) - \varepsilon_0 \propto t^{-\alpha}$ with the exponents $\alpha = 0.670 \pm 0.004$ and 1.012 ± 0.008, respectively.

the other hand, the mean square deviation of the difference $E_i - E_m$ is approximated as

$$\Delta E^2 = \langle (E_i - E_m)^2 \rangle - (\langle E_i - E_m \rangle)^2 \sim |i - m| \tag{3.3}$$

This is the result of a "random walk" of E_i (see Fig. 4). Thus the optimal partition i that minimizes the number of errors E_i can fluctuate around m. The order of the deviation is estimated by the balance $|\Delta E| \sim \langle E_i - E_m \rangle$, which implies $|i - m| \sim O(t^{2/3})$, or

$$|x_o - \xi_0| \sim O(t^{-1/3}) \tag{3.4}$$

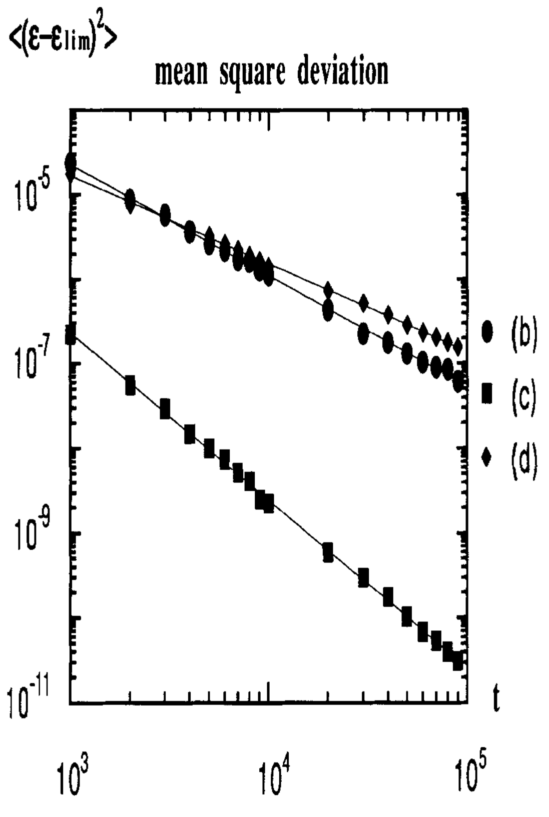

Figure 3: Average of $[\varepsilon(t)-\varepsilon_0]^2$ for (1) and (2), and $[\varepsilon(t)-\varepsilon_1]^2$ for (3). They exhibit the scaling $t^{-\alpha}$ with the exponents $\alpha = 1.313 \pm 0.016, 1.976 \pm 0.014$, and 1.040 ± 0.007, respectively.

The prediction error is thus estimated as

$$\varepsilon(t) = \langle E_i\rangle/t = \langle E_m\rangle/t + (a/t^2)(i-m)^2 = \varepsilon_0 + O(t^{-2/3}) \tag{3.5}$$

The numerical result of (1) is consistent with this fractional exponent 2/3.

The remaining two asymptotic scaling forms for the maximum likelihood algorithms (2) and (3) can be explained by the conventional theory. The variance of the maximum likelihood estimator w_o is known to obey the asymptotic scaling,

$$\langle[w_o(t)-\omega]^2\rangle \propto t^{-1} \tag{3.6}$$

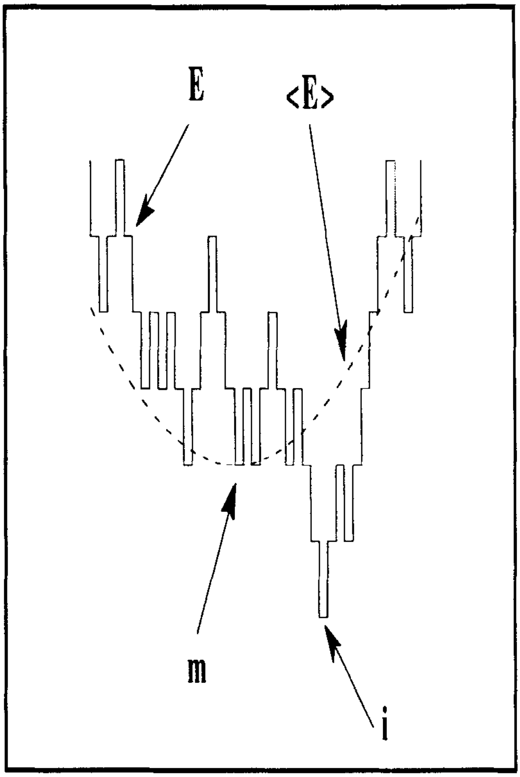

Figure 4: Schematic representation of the fluctuation of E_i around $\langle E_i \rangle$.

where $\omega = \lim_{t\to\infty}\langle w_o(t)\rangle$. Thus the deviation of w_o from ω is of the order of $t^{-1/2}$. Deviation of the position of the boundary x_o from $\xi = \lim_{t\to\infty}\langle x_o(t)\rangle$ is proportional to the one of w_o from ω. In the case that the limit ξ is identical to the best choice, $\xi = \xi_0$, the prediction error is estimated as

$$\varepsilon(t) - \varepsilon_0 \propto (x_o - \xi_0)^2 = O(t^{-1}) \tag{3.7}$$

On the other hand, if the limit $\xi = \xi_1$ is remote from ξ_0, the prediction error is

$$\varepsilon(t) \sim \varepsilon_1 + 2c(\xi_1 - \xi_0)(x_o - \xi_1) = \varepsilon_1 \pm O(t^{-1/2}) \tag{3.8}$$

where $\varepsilon_1 = \varepsilon_0 + c(\xi_1 - \xi_0)^2 > \varepsilon_0$. The numerical results of (2) and (3) are, respectively, consistent with the scaling forms of equations 3.7 and 3.8.

It is not so difficult to show that these three types of learning curves are not sensitive to the choice of the problems as well as the dimensionality of the space x. The point will be discussed elsewhere. In this paper we did not take into account the computational complexity of these algorithms as well as the estimate of the error measure ε itself such as discussed by Haussler (1991). The problems are left to future studies.

Acknowledgments

We thank Shun-ichi Amari, Hideki Asoh, Kenji Yamanishi, and Michael Crair for helpful discussion. The present work is partly supported by a Grant-in-Aid for Scientific Research by the Ministry of Education, Japan, No. 0325105.

References

Amari, S., Fujita, N., and Shinomoto, S. 1992. Four types of learning curves. *Neural Comp.* **4**, 605–618.

Baum, E. B., and Haussler, D. 1989. What size net gives valid generalization? *Neural Comp.* **1**, 151–160.

Haussler, D. 1991. Decision theoretic generalization of the PAC model for neural net and other learning applications. UCSC-CRL-91-02.

Levin, E., Tishby, N., and Solla, S. A. 1990. A statistical approach to learning and generalization in layered neural network. *Proc. IEEE* **78**, 1568–1574.

Rissanen, J. 1989. *Stochastic Complexity in Statistical Inquiry.* World Scientific, Singapore.

Valiant, L. G. 1984. A theory of learnable. *Commun. ACM* **27**(11), 1134–1142.

Received 16 September 1991; accepted 5 February 1992.