Statistical Theory of Learning Curves

Shun-ichi AMARI
Faculty of Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo, Tokyo 113 Japan
RIKEN Frontier Research Program
on Brain Information Processing
2-1 Hirosawa, Wako 351-01, Japan
E-mail: amari@sat.t.u-tokyo.ac.jp

Noboru MURATA
Faculty of Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo, Tokyo 113 Japan
E-mail: mura@sat.t.u-tokyo.ac.jp

Kazushi IKEDA
Faculty of Engineering, Kanazawa University
2-40-20 Kodatsuno, Kanazawa 920 Japan
E-mail: kazushi@ec.t.kanazawa-u.ac.jp

Abstract

Behaviors of a learning machine depends on its complexity and the number of training examples. A learning curve shows how fast a neural network or a general learning machine can improve its behavior as the number of training examples increases. This is also related to the complexity of a learning machine. The characteristic of the learning curve is studied from the statistical-mechanical, information theoretic and statistical points of view. The present paper summarizes universal as well as specific properties of learning curves of both deterministic and stochastic pattern classifiers from the statistical point of view.

1 Introduction

Learning is an important subject of research studied by various methods of approach such as algorithm theory, stochastic gradient method, statistical mechanics, information theory, statistics, etc. Statistical mechanics and information theory have proved its wide applicability in the field of machine learning. The present paper intends to elucidate the characteristics of learning machines from the statistical point of view.

Various learning algorithms have so far been proposed in the field of neural networks and pattern recognition, and their characteristics have been studied extensively. For example, Amari [1967] (see also Amari [1993a]) gave the stochastic-descent learning algorithm for a general continuous machine and its on-line learning dynamics was studied. On the other hand, more basic concepts of learnability and complexity of learning have been studied in the field of machine learning (Valiant, [1984]). A basic capability of learning is shown by the learning curve which indicates how fast the behavior of a learning machine is improved as training examples increase. Learning

curves have so far been studied from various points of view (see, e.g., Cover [1964], Haussler, Kearns and Shapire [1991], Levin, Tishby and Solla [1990], Hansel and Sompolinski [1990], Seung, Tishby and Sompolinski [1992], Opper and Haussler [1991]; see also a number of papers in this volume). It is one of the remarkable subjects, as is seen from the many related papers in this volume. We review results obtained by a new statistical approach completely different from the statistical-mechanical one (Amari, Fujita and Shinomoto [1992], Amari and Murata [1993], Amari [1993b], and Murata, Yoshizawa and Amari [1994], Ikeda, Amari and Yoshizawa [1995], Mühler et al. [this volume]). The present paper is an improved version of the paper (Amari, Murata and Ikeda [1994]).

The present paper considers learning dichotomy machines which classify given input signals $\mathbf{x}$ into two classes C_+ and C_-. The behavior of machine is specified by a p-dimensional vector parameter $\mathbf{w}$ which is modified through learning. A deterministic machine calculates the value of smooth a function $f(\mathbf{x}, \mathbf{w})$ and classifies $\mathbf{x}$ into C_+ or C_- according to the signum of f. A stochastic machine classifies $\mathbf{x}$ stochastically into C_+ or C_- according to a probability specified by $f(\mathbf{x}, \mathbf{w})$. A deterministic machine and a stochastic machine have different characteristics.

Given t input-output pairs $D_t = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_t, y_t)\}$ of examples, they determine a set $A(D_t)$ in the parameter space of a deterministic machine such that any machine specified $\mathbf{w} \in A(D_t)$ outputs the correct y_i when $\mathbf{x}_i$ is input $(i = 1, \cdots, t)$. In other words, $A(D_t)$ is the set of parameters $\mathbf{w}$ that can explain the input-output data D_t correctly. The $A(D_t)$ is called the consistent or admissible region. The behavior of a deterministic machine is explained how fast the area $Z_t = |A(D_t)|$ converges to 0 as t increases, because when Z_t is smaller it is easier to estimate the true $\mathbf{w}$. We give some mathematical analysis on the distribution of the shape and the area of $A(D_t)$. This includes some new results on stochastic geometry.

In the case of a stochastic machine, the consistent $A(D_t)$ soon becomes null as t increases, because the input-output data are noisy. This is also the case when the target machine is unrealizable by any function $f(\mathbf{x}, \mathbf{w})$. We need to discuss a stochastic version of the consistent region from the Bayesian standpoint. We can define a stochastic version of Z_t. In order to obtain the probability distribution of Z_t, its stochastic expansion is given.

Let P_t be the probability that a machine trained by t examples D_t correctly classifies a new example. Here all the examples are assumed to be generated independently from a fixed but unknown probability distribution $p(\mathbf{x})$. The $P_t(D_t)$ depends on the past examples D_t. We denote by $\langle \quad \rangle$ the average over D_t.

The expected error probability is the expectation of $e_t = 1 - P_t$,

$$\langle e_t \rangle = 1 - \langle P_t \rangle.$$

The function $\langle e_t \rangle$ is called the learning curve for classification error, when it is regarded as a function of t. On the other hand, when the loss of a machine is measured by the entropy $\langle -\log P_t \rangle$ of the correct classification, we have the learning curve

$$\langle e_t^* \rangle = \langle -\log P_t \rangle$$

for the entropic loss. The inequality

$$\langle e_t \rangle \leq \langle e_t^* \rangle$$

holds in general.

We give a rigorous asymptotic evaluation of the entropic learning curves for both deterministic and stochastic machines. Universal properties hold on the learning curves of both deterministic and stochastic machines. When we consider a non-faithful model, that is when the true teacher machine is not realizable by our model, the result is a little different. The learning curve defines the complexity of a machine, and this in turn leads us to an extension of Akaike's information criterion (AIC) to be used for model selection from training data.

2 Bayesian framework for learning

We study a dichotomy problem where input signals $\mathbf{x} \in \mathbf{R}^m$ are to be classified into two categories C_+ and C_-. There are two cases, the deterministic case and the stochastic case, depending on the behaviors of the networks or machines. When a signal $\mathbf{x}$ is input, a machine or a neural network emits a binary output y taking values on 1 and -1, where $y = 1$ implies that $\mathbf{x}$ belongs to C_+ and $y = -1$ to C_-.

Let us consider a family of machines parameterized by a vector $\mathbf{w} \in \mathbf{R}^p$. There are two types of machines, deterministic and stochastic machines. A deterministic machine calculates a smooth function $f(\mathbf{x}, \mathbf{w})$ and its signum is the output,

$$y = \mathrm{sgn} f(\mathbf{x}, \mathbf{w}).$$

A stochastic machine emits $y = 1$ or -1 stochastically subject to the conditional probability distribution $p(y|\mathbf{x}, \mathbf{w})$ depending on $\mathbf{w}$. Here, we assume that $p(y|\mathbf{x}, \mathbf{w})$ is given typically by

$$p(y|\mathbf{x}, \mathbf{w}) = \frac{1}{1 + \exp\{-f(\mathbf{x}, \mathbf{w})\}}.$$

The true machine is said to be realizable, when it is included in the parameterized family of machines, that is, when there exists a parameter $\mathbf{w}_0$ called the true parameter such that the output y is given by

$$g(\mathbf{x}) = \mathrm{sgn}\ f(\mathbf{x}, \mathbf{w}_0)$$

in the deterministic case, and that the probability of $y = \pm 1$ is given by the conditional probability

$$q(y|\mathbf{x}) = p(y|\mathbf{x}, \mathbf{w}_0)$$

determined by $\mathbf{w}_0$ in the stochastic case. Otherwise, the true system is said to be unrealizable or the family of machines is said to be unfaithful.

Let

$$D_t = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_t, y_t)\} \tag{2.1}$$

be the set of t observed examples, where $\mathbf{x}_i$ $(i = 1, 2, \cdots)$ are generated independently subject to a prescribed probability distribution having a density $p(\mathbf{x}) > 0$ and y_i are the outputs from the true system given $\mathbf{x}_i$. The set D_t is called the training set for a learning machine. The D_t can be regarded as i.i.d. (independently and identically distributed) observations from the parametric probability distribution

$$p(\mathbf{x}, y|\mathbf{w}) = p(\mathbf{x})p(y|\mathbf{x}, \mathbf{w}), \tag{2.2}$$

where

$$p(y|\mathbf{x},\mathbf{w}) = \delta\{y - \mathrm{sgn} f(\mathbf{x},\mathbf{w})\}$$

in the deterministic case.

In order to obtain the predictive distribution of a machine trained by the observed data D_t, we take the Bayesian standpoint and assume that the prior distribution of the true $\mathbf{w}$ is given by $q_{pr}(\mathbf{w})$. After observing t examples D_t, we have the posterior distribution $Q(\mathbf{w}|D_t)$ of $\mathbf{w}$ from the Baysian standpoint. To obtain the posterior distribution, we put

$$Q(\mathbf{w}, D_t) = q_{pr}(\mathbf{w}) \prod_{i=1}^{t} p(y_i|\mathbf{x}_i,\mathbf{w}) p(\mathbf{x}_i), \tag{2.3}$$

which is the joint probability density that a machine $\mathbf{w}$ is chosen from the prior distribution $q_{pr}(\mathbf{w})$, $\mathbf{x}$'s are selected as inputs and corresponding y's are produced from that machine. In the deterministic case,

$$p(y_i|\mathbf{x}_i,\mathbf{w}) = 1$$

when $y_i f(\mathbf{x}_i,\mathbf{w}_i) > 0$ and is otherwise equal to 0. The average probability, over all the machines, of examples D_t being generated is given by

$$\mathrm{Prob}\{D_t\} = \int Q(\mathbf{w}, D_t) d\mathbf{w} = Z(D_t) \prod_{i=1}^{t} p(\mathbf{x}_i),$$

where

$$\begin{aligned} Z(D_t) &= \int q_{pr}(\mathbf{w}) \prod_{i=1}^{t} p(y_i|\mathbf{x}_i,\mathbf{w}) d\mathbf{w} \\ &= \int q_{pr}(\mathbf{w}) p(D_t|\mathbf{x}_1,\cdots,\mathbf{x}_t,\mathbf{w}) d\mathbf{w} \end{aligned} \tag{2. 4}$$

is the average of the conditional distribution $p(D_t|\mathbf{x}_1,\cdots,\mathbf{x}_t,\mathbf{w})$ over all machines.

In the deterministic case, we define the set $A(D_t)$ in the parameter space by

$$A(D_t) = \{\mathbf{w}| \quad y_i f(\mathbf{x}_i,\mathbf{w}) > 0, \quad i = 1,\cdots,t\}.$$

This is the set of $\mathbf{w}$'s which can produce $y_1,\cdots,y_t$ from the input examples $\mathbf{x}_1,\cdots,\mathbf{x}_t$. This is called the admissible set or the consistent set. The true $\mathbf{w}_0$ is included in this set. We have $p(D_t|\mathbf{x}_1,\cdots,\mathbf{x}_t,\mathbf{w}) = 1$ or 0 according to whether D_t is consistent with $\mathbf{w}$ or not. Hence it is the indicator function of the set $A(Z_t)$ of $\mathbf{w}$. Consequently, Z_t is the measure of those $\mathbf{w}$ which are compatible with D_t,

$$Z_t = |A(D_t)| = \int_{\mathbf{w}\in A(D_t)} q_{pr}(\mathbf{w}) d\mathbf{w}.$$

The posterior distribution is simply given by

$$Q(\mathbf{w}|D_t) = \frac{Q(\mathbf{w}, D_t)}{Z_t \prod_{i=1}^{t} p(\mathbf{x}_i)} \tag{2.5}$$

in both cases. Given t examples D_t, we define the predictive distribution of the next output y_{t+1} when the next input $\mathbf{x}_{t+1}$ is given. Given D_t, we first choose a candidate machine $\mathbf{w}_t^*$ randomly subject to the posterior distribution $Q(\mathbf{w}|D_t)$. This is one learning algorithm, which is called the Gibbs algorithm. The predictive distribution is the conditional probability of y_{t+1} under the condition that D_t has been observed and the next input is $\mathbf{x}_{t+1}$. It is given by the average of the distribution of the output y_{t+1}, when $\mathbf{x}_{t+1}$ is input, over all the candidate machines subject to the posterior distribution $Q(\mathbf{w}|D_t)$. It is easy to prove

Lemma 1. The predictive distribution is given by

$$P(y_{t+1}|\mathbf{x}_{t+1}, D_t) = \frac{Z_{t+1}}{Z_t}, \tag{2.6}$$

both in the deterministic and stochastic cases.

3 Universal learning curve under entropic loss — deterministic case

It is not easy to evaluate the generalization error

$$\langle e_t \rangle = 1 - \left\langle \frac{Z_{t+1}}{Z_t} \right\rangle$$

in the deterministic case in general. This is because of the correlations of Z_t and Z_{t+1}. The replica method gives a good answer in a specific model under the thermodynamical limit which elucidates interesting phase transition phenomena.

Here, we calcualte the entropic loss or the information gain instead of the error probability. The entropic loss e_t^* is given by

$$\begin{aligned} e_t^* &= -\log P(y_{t+1}|\mathbf{x}_{t+1}, D_t) \\ &= \log Z_t - \log Z_{t+1}. \end{aligned}$$

This is a random variable depending on D_t. Its expectation $\langle e_t^* \rangle$ over D_{t+1} gives the learning curve of the entropic loss. This is called the entropic generalization error because the average behavior of a machine is evaluated by the new example $\mathbf{x}_{t+1}$ which is not included in D_t.

Let p be the dimension number of $\mathbf{w}$. We assume that no redundant parameter is included in $\mathbf{w}$. More precisely, when $\mathbf{w}$ is changed to $\mathbf{w} + \Delta\mathbf{w}$, the dichotomy region $D_+(\mathbf{w})$ is changed to $D_+ + \Delta D_+$, and we assume that

$$|\Delta D_t| = O(|\Delta \mathbf{w}|),$$

when $\Delta\mathbf{w}$ is infinitesimally small. Here the dichotomy region is defined by

$$D_+(\mathbf{w}) = \{\mathbf{x}|\ f(\mathbf{x}, \mathbf{w}) > 0\}.$$

We also assume that $q_{pr}(\mathbf{w})$ is non-singular, that is, $q_{pr}(\mathbf{w}) > 0$ and continuous.

It is rather surprising that the following universal theorem holds concerning the entropic learning curve: The curve depends only on the dimension of the parameter $\mathbf{w}$ and does not depend on the specific architecture of the machine.

Theorem 1 (**Universal Theorem for Deterministic Machine**). The generalization entropic loss is asymptotically given by

$$\langle e_t^* \rangle = \frac{p}{t} \tag{3.1}$$

irrespectively of the architecture of the model networks.

This gives an upper bound to the prediction error $\langle e_t \rangle$. It is also shown that the entropic loss $\langle e_t^* \rangle$ gives the amount of information which the tth example carries (Haussler, Kearns and Shapire [1991]).

We give a sketch of the proof of the theorem. See Amari [1993b] for the details of the proof.

Proof. We use a function $s(\mathbf{w})$ defined in the deterministic case (Amari, Fujita and Shinomoto [1992]). It is the probability that a machine with $\mathbf{w}$ gives the same output as the true machine,

$$s(\mathbf{w}) = \text{Prob}\{f(\mathbf{x}, \mathbf{w}) f(\mathbf{x}, \mathbf{w}_0) > 0\}. \tag{3.2}$$

The average $\langle Z_t \rangle$ is then written as

$$\langle Z_t \rangle = \int q_{pr}(\mathbf{w}) \{s(\mathbf{w})\}^t d\mathbf{w}. \tag{3.3}$$

This is because

$$\begin{aligned} \langle Z_t \rangle &= \left\langle \int q_{pr}(\mathbf{w}) 1[f(\mathbf{x}_i, \mathbf{w}_0) f(\mathbf{x}_i; \mathbf{w}) > 0, i = 1, \cdots, t] d\mathbf{w} \right\rangle \\ &= \int q_{pr}(\mathbf{w}) \prod_{i=1}^{t} \langle \text{Prob}\{f(\mathbf{x}_i, \mathbf{w}_0) f(\mathbf{x}_i, \mathbf{w}) > 0\} \rangle d\mathbf{w} \\ &= \int q_{pr}(\mathbf{w}) \{s(\mathbf{w})\}^t d\mathbf{w}, \end{aligned}$$

where $\langle \quad \rangle$ denotes the expectation with respect to $\mathbf{x}_1, \cdots, \mathbf{x}_t$ and $1[\cdot]$ is the indicator function. We see that $s(\mathbf{w})$ takes its maximum value 1 at $\mathbf{w}_0$ and that $s(\mathbf{w})$ has directional derivative at $\mathbf{w}_0$, so that

$$s(\mathbf{w}) = 1 - a(\mathbf{e})|\mathbf{w} - \mathbf{w}_0| + O(|\mathbf{w} - \mathbf{w}_0|^2),$$

where

$$\mathbf{e} = \frac{\mathbf{w} - \mathbf{w}_0}{|\mathbf{w} - \mathbf{w}_0|}.$$

This gives the following asymptotic evaluation of the average volume by expanding (3.30 and by using the Laplace method of integration,

$$\langle Z_t \rangle = \frac{c}{t},$$

showing how $\langle Z_t \rangle$ scales. We now generalize this function to $s(\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_k)$, which is the probability measure of those $\mathbf{x}$ for which the outputs of k machines with parameters $\mathbf{w}_1$, $\cdots$, $\mathbf{w}_k$ are all equal to the output of the true machine specified by $\mathbf{w}_0$. It is shown that the k-th moment of Z_t is given by

$$\langle (Z_t)^k \rangle = \int q(\mathbf{w}_1) \cdots q(\mathbf{w}_k) \{ s(\mathbf{w}_1, \cdots, \mathbf{w}_k) \}^t d\mathbf{w}_1 \cdots d\mathbf{w}_k, \tag{3.4}$$

which can be interpreted as the replica calculation. Obviously, s is maximized at $\mathbf{w}_1 = \mathbf{w}_2 = \cdots \mathbf{w}_k = \mathbf{w}_0$,

$$s(\mathbf{w}_0, \mathbf{w}_0, \cdots, \mathbf{w}_0) = 1.$$

This $s(\mathbf{w}_1, \cdots \mathbf{w}_k)$ is a very complicated function of $\mathbf{w}_1, \cdots, \mathbf{w}_k$. However, we see that it has directonal derivatives at $\mathbf{w}_1 = \cdots = \mathbf{w}_k = \mathbf{w}_0$,

$$s(\mathbf{w}_1, \cdots \mathbf{w}_k) = 1 - a(\mathbf{E}) |\mathbf{W} - \mathbf{W}_0|, \qquad a(\mathbf{E}) > 0,$$

where $\mathbf{W} = (\mathbf{w}_1, \cdots, \mathbf{w}_k)$, and $\mathbf{W}_0 = (\mathbf{w}_0 \cdots, \mathbf{w}_0)$ and

$$\mathbf{E} = \frac{\mathbf{W} - \mathbf{W}_0}{|\mathbf{W} - \mathbf{W}_0|}.$$

By complicated calculations and using the Laplace integration technique, we prove that, when t is large,

$$\langle (Z_t)^k \rangle \sim \frac{c_k}{t^{pk}}, \tag{3.5}$$

where c_k is a coefficient (see Amari [1993]). In other words, for a sequence of random variables

$$Y_t = t^p Z_t, \qquad t = 1, 2, \cdots$$

its k-th moment ($k = 1, 2, \cdots$) converges, under a certain regularity condition, to c_k. Hence, Y_t converges to a random variable Y in distribution. This implies that

$$\langle \log Z_t \rangle \sim \langle \log(t^{-p} Y) \rangle = \langle \log Y \rangle - p \log t. \tag{3.6}$$

The theorem is easily proved from this.

Now, we evaluate the generalization error $\langle e_t \rangle$ by this method. We have asymptotically

$$\langle e_t \rangle = 1 - \left\langle \frac{Z_{t+1}}{Z_t} \right\rangle = 1 - \left(1 - \frac{p}{t} \right) \left\langle \frac{Y_{t+1}}{Y_t} \right\rangle.$$

Here, Y_t converges to Y in distribution. If we can assume that

$$\left\langle \frac{Y_{t+1}}{Y_t} \right\rangle = 1 - \frac{b}{t},$$

we have

$$\langle e_t \rangle = \frac{p-b}{t},$$

where b may depend on the architecture of the machine and input distribution $p(\mathbf{x})$. Here, the curve is not universal in this case but the power law of t^{-1} holds.

When the teacher machine is not realizable or stochastic (that is, noisy), we see that $s(\mathbf{w})$, and also $\mathbf{s}(\mathbf{w}_1, \cdots, \mathbf{w}_k)$, are differentiable at $\mathbf{w} = \mathbf{w}_0$ or at $\mathbf{w}_1 = \cdots = \mathbf{w}_k = \mathbf{w}_0$. In this case, we have the following scaling law

$$Z_t = (\sqrt{t})^p Y_t, \qquad Y_t \to Y.$$

Hence, the entropic error $\langle e_t^* \rangle$ becomes

$$\langle e_t^* \rangle = \frac{p}{2t},$$

as will be shown in the next section by the stochastic expansion method rigorously.

If we can assume

$$\left\langle \frac{Y_{t+1}}{Y_t} \right\rangle = 1 - \frac{b}{t},$$

the generalization error is asymptotically given by

$$\langle e_t \rangle = \frac{p}{\sqrt{t}},$$

which is universal. This shows the power law of $1/\sqrt{t}$. It is an interesting conjecture to prove or disprove that the learning curve is universal under the unrealizable teacher.

4 Universal learning curve under entropic loss — stochastic case

In the stochastic case, we can use the ordinary technique of asymptotic inference in statistics. Let $\hat{\mathbf{w}}_t$ be the maximum likelihood estimator from the observed data D_t. Its distribution is given by the following lemma, where the Fisher information matrix $G = (g_{ij})$ is defined by

$$g_{ij} = E\left[\frac{\partial}{\partial w_i} \log p(y|\mathbf{x}, \mathbf{w}) \frac{\partial}{\partial w_j} \log p(y|\mathbf{x}, \mathbf{w}) \right], \tag{4.1}$$

E denoting the expectation with respect to the random variables $(y, \mathbf{x})$ and w_i is the ith component of $\mathbf{w}$. The Baysian (maximum posterior probability) estimator has asymptotically the same distribution.

Lemma 2. The maximum likelihood estimator $\hat{\mathbf{w}}_t$, and the posterior estimator, based on D_t are asymptotically normally distributed,

$$\hat{\mathbf{w}}_t \sim N(\mathbf{w}_0, \frac{1}{t} G^{-1}).$$

This fact is well known in statistics. The following important lemma is calculated by stochastic expansion (see Amari and Murata [1993]).

Lemma 3.

$$\log Z_t \sim -tH_0 - \frac{p}{2}\log t - \frac{1}{2}\log |G| + \frac{1}{2}\chi_p^2, \tag{4.2}$$

where $|G|$ is the determinant of the Fisher information matrix at $\mathbf{w}_0$ and χ_p^2 is a random variable subject to the χ^2-distribution of degree p.

Now we calculate $\langle \log Z_t \rangle$, and $\langle e_t^* \rangle$ is obtained therefrom. It should be noted that even the true machine with parameter $\mathbf{w}_0$ is not free of loss, because its behavior is stochastic. Its loss is given by the conditional entropy,

$$\begin{aligned} H_0 &= H(Y|X;\mathbf{w}_0) \\ &= E_X[-\sum_y p(y|\mathbf{x},\mathbf{w}_0)\log p(y|\mathbf{x},\mathbf{w}_0)] \end{aligned} \tag{4. 3}$$

where E_X is the expectation with respect to input signals $\mathbf{x}$. In general, $\langle e_t^* \rangle$ is larger than H_0, and it converges to H_0 as the number t of examples increases. The present paper studies how fast it converges to H_0.

Theorem 2 (**Universal Theorem for Stochastic Machine**). The asymptotic learning curve is given for any regular stochastic machine by

$$\langle e_t^* \rangle = H_0 + \frac{p}{2t}, \tag{4.4}$$

where p is the dimension number of parameters $\mathbf{w}$.

We can also evaluate the estimator $\hat{\mathbf{w}}_t$ based on the training data D_t. This is called the training entropic loss because we use the average over D_t instead of the average over new examples. That is, it evaluates the behavior of the trained system by using again the training data,

$$e_{\text{train}}^*(t) = -\frac{1}{t}\sum_{i=1}^{t}\log \text{Prob}\{y_i = \hat{y}_i|D_t\}, \tag{4.5}$$

where $\hat{y}_i$ is the output of the system whose parameter $\hat{\mathbf{w}}_t$ is estimated from D_t. Its expectation with respect to D_t gives the expected training loss. Since the machine is trained by using D_t, it overfits the data set D_t, that is, its behavior is particularly good for the training data but might not be so good for a new example. Therefore, we need to evaluate the discrepancy between the expected training loss and the expected generalization loss. We can again prove (Amari and Murata [1993]), in the case of the entropic loss,

$$\langle e_t^* \rangle_{\text{train}} \approx H_0 - \frac{p}{2t}. \tag{4.6}$$

It is interesting that the generalization error (5.4) and training error (5.6) have a symmetric form, both approaching H_0 from the opposite sides. This property is again universal.

The two results combine into

$$\langle e_t^* \rangle \approx \langle e_t^* \rangle_{\text{train}}(t) + \frac{p}{t}, \tag{4.7}$$

showing the relation between the entropic generalization error and the entropic training error. This is the same as the information criterion AIC (see, e.g., Sakamoto, Ishiguro and Kitagawa [1986]) if each term is multiplied by $2t$. These are again universal in the sense that they do not depend on specific architectures of networks but depend on only the number of modifiable parameters.

The result is further generalized to give a relation between the generalization loss and training loss under a general loss $l(\mathbf{x}, y; \mathbf{w})$ which is the loss of processing $\mathbf{x}$ by the machine with $\mathbf{w}$ (Murata, Yoshizawa and Amari [1994]). We define

$$\begin{aligned} G &= E[\frac{\partial}{\partial w_i} l(\mathbf{x}, y; \mathbf{w}) \frac{\partial}{\partial w_j} l(\mathbf{x}, y; \mathbf{w})], \\ Q &= E[\frac{\partial^2}{\partial w_i \partial w_j} l(\mathbf{x}, y; \mathbf{w})]. \end{aligned}$$

We have in this case, the following general result (Murata, Yoshizawa and Amari [1994].)

Theorem 3. (**Generalization Error Theorem**). The generalization loss and training loss satisfy asymptotically

$$L_{\text{gen}}(t) \approx L_{\text{train}}(t) + \frac{p^*}{t}, \tag{4.8}$$

where p^* is a quantity showing the complexity of the network model defined by

$$p^* = \text{tr}(GQ^{-1}), \tag{4.9}$$

and tr is the trace of a matrix.

When the log likelihood loss is used but the model is not faithful, this result reduces to Takeuchi information criterion (Takeuchi [1976]).

Moreover, when the true machine is realizable, p^* is equal to the dimension number of $\mathbf{w}$, that is the number of the modifiable parameters. This result coincides completely with the information criterion (AIC).

This result can be applied to model selection. Given two network models, we train the networks based on a given D_t. We then evaluate the behaviors of the models in terms of the training loss. However, even if one model has a smaller training loss, this does not imply that its generalization error is better. In order to compare the estimated generalization errors, we can use the following *Network Information Criterion*

$$\text{NIC} = L_{\text{train}}(t) + \frac{p^*}{t}. \tag{4.10}$$

See Murata, Yoshizawa and Amari [1994]

5 Stochastic geometry of admissible region in the deterministic case

It is interesting to know how the shape and the volume of the admissible region $A(D_t)$ changes with t. Here, we summarize the results given in Ikeda and Amari, this volume. Consider a simplest case of the simple perceptron, where

$$f(\mathbf{x}, \mathbf{w}) = \mathbf{w} \cdot \mathbf{x}.$$

We assume without loss of generality that the parameter space is the $(n-1)$-dimensional unit sphere,

$$W = \{\mathbf{w} | \ |\mathbf{w}| = 1\}$$

so that $p = n - 1$. We also assume that $\mathbf{x}$ is uniformly distributed on the unit sphere $|\mathbf{x}| = 1$.

Even for a general machine, some properties may be shared with the simple perceptron case, although the perceptron case is not yet solved. The admissible region $A(D_t)$ is given by

$$A(D_t) = \{\mathbf{w} | \ y_i f(\mathbf{x}_i, \mathbf{w}) > 0, \quad i = 1, \cdots, t\}.$$

In the perceptron case, each example $(\mathbf{x}_i, y_i)$ divides the parameter space (the unit sphere) into two parts by the separating hyperplane

$$\mathbf{w} \cdot \mathbf{x}_i = 0.$$

Define

$$A_i = \{\mathbf{w} | \quad y_i \mathbf{w} \cdot \mathbf{x}_i > 0\}.$$

Then, $A(D_t)$ is the intersection of A_i,

$$A(D_t) = \bigcap_{i=1}^{t} A_i,$$

which forms a polyhedron on the unit sphere. However, not all the hyperplane $\mathbf{w} \cdot \mathbf{x}_i = 0$ is the boundary of $A(D_t)$ because A_i sometimes includes $A(D_t)$ properly. That is, some examples $(\mathbf{x}_i, y_i)$ are covered by other sharper examples. We say that example $(\mathbf{x}_i, y_i)$ is effective when the hyperplane $\mathbf{w} \cdot \mathbf{x}_i = 0$ is a boundary of $A(D_t)$. It is otherwise ineffective. Let r be the number of effective examples in D_t. Then, $A(D_t)$ is a polyhedron surrounded by r spherical faces.

It is interesting to know the characteristic features of the polyhedron $A(D_t)$. This gives an interesting problem of stochastic geometry on the sphere. In this connection, we consider the average F_t of the number of faces of the convex hull generated by t randomly chosen points on the p-halfsphere. This F_t is the number of vertices of the polyhedron $A(D_t)$. We then have the following theorem.

Theorem 4.

$$F_t = \frac{\pi^p}{pI_{p-1}},$$

where

$$I_j = \int_0^{\pi/2} \mathrm{sgn}^j \phi \; d\phi.$$

With the help of Euler's equation, we have

$$\begin{aligned} \langle r \rangle &= 2, && \text{when} \quad p = 1 \\ \langle r \rangle &= \pi^2/2, && \text{when} \quad p = 2 \\ \langle r \rangle &= 2 + \frac{2}{3}\pi^2, && \text{when} \quad p = 3. \end{aligned}$$

It is rather surprising that $\langle r \rangle$ remains finite even when the number t of points increases infinitely.

The expectation of the area of A_t is calculated as

$$\langle Z_t \rangle = \frac{(p-1)!}{2I_{p-1}} \frac{\pi^p}{t^p}.$$

On the other hand, in order to know the shape of $A(D_t)$, we can calculate the expected value of the radius d_t of $A(D_t)$, that is, the distance from $\mathbf{w}_0$ to the boundary of $A(D_t)$. Let $d_t(\Theta)$ be the distance from $\mathbf{w}_0$ to the boundary of $A(D_t)$ in the direction of Θ, where Θ is the angular direction. Then, we have the following theorem.

Theorem 5. The radius d_t is asymptotically subject to the exponential distribution with parameter $\lambda = \dfrac{t}{\pi}$,

$$p(d_t) = \frac{t}{\pi} \exp\left\{ -\frac{t}{\pi} d_t \right\}.$$

The error probability is written as

$$e_t = \frac{p}{(p+1)\pi} \frac{\int d_t(\Theta)^{p+1} d\Theta}{\int d_t(\Theta)^p d\Theta}.$$

6 Conclusions

The learning curve of a learning machine has some universal properties of convergence. We showed universal aspects of the learning curve both in the deterministic and stochastic cases in terms of the entropic loss. From the statistical point of view, the deterministic case is quite different from the stochastic case, since the Fisher information is infinite in the former case so that it defines a non-regular statistical model. On the other hand, the stochastic case is regular so that we can apply the standard asymptotic expansion of the estimator. We can in this case obtain a new information criterion for the model selection which is a generalization of the AIC.

We need further studies on the learning curve in the deterministic case along the lines we proposed here.

References

Amari, S. [1967] : Theory of Adaptive Pattern Classifiers, *IEEE Trans.*, **EC-16**, 299–307.

Amari, S. [1993a] : Backpropagation and Stochastic Gradient Descent Method, *Neurocomputing*, **5**, 185–196.

Amari, S. [1993b] : A Universal Theorem on Learning Curves. *Neural Networks*, **6**, pp.161–166.

Amari, S., Fujita, N. and Shinomoto, S. [1992] : Four Types of Learning Curves. *Neural Computation*, **4**, pp.604–618.

Amari, S. and Murata, N. [1993] : Statistical Theory of Learning Curves under Entropic Loss Criterion. *Neural Computation*, **5**, pp.140–153.

Amari, S., Murata, N. and Ikeda, K. [1993] : Universal Properties of Learning Curves, in *Cognitive Processing for Vision and Voice*, ed. Ishiguro, T., SIAM, 77–87.

Cover, T. M. [1964] : Geometrical and Statistical Properties of Linear Threshold Devices. *Technical Report* No.6107-1, Stanford Electronics Laboratories, Stanford University.

Hansel, D. and Sompolinsky, H. [1990] : Learning from Examples in a single-;ayer neural network. *Europhys. Lett.*, **11**, pp.687–692.

Haussler, D., Kearns, M. and Schapire, R. [1991] : Bounds on the Sample Complexity of Bayesian Learning Using Information Theory and the VC Dimension. *Proc. Fourth Ann. Workshop on Comp. Learning Theory*, Morgan Kaufmann, pp.61–74.

Ikeda, K., Amari, S. and Yoshizawa, S. [1995] : Prediction Error and Consistent Parameter Area in Neural Learning. to be published.

Levin, E., Tishby, N. and Solla, S. A. [1990] : A Statistical Approach to Learning and Generalization in Layered Neural Networks. *Proceedings of the IEEE*, vol. **78**, No.10, pp.1568–1574.

Murata, N., Yoshizawa, S. and Amari, S. [1994] : Network Information Criterion – Determining the Number of Hidden Units for an Artificial Neural Network Model. *IEEE Trans. on Neural Networks*, **NN5**, 865–872.

Opper, M. and Haussler, D. [1991] : Calculation of the Learning Curve of Bayes Optimal Classification Algorithm for Learning Perceptron with Noise. *Proc. Fourth Ann Workshop on Comp. Learning Theory*, Morgan-Kaufmann, pp.75–87.

Sakamoto, Y., Ishiguro, M. and Kitagawa, G [1986] : *Akaike Information Criterion Statistics.* Reidel-Kluwer Academic.

Seung, S., Tishby, N. and Sompolinsky, H. [1992] : Learning from Examples in Large Neural Networks. *Physical Review*, **A45**, pp.6058–6091.

Takeuchi, K. [1976] : Distribution of Information Statistics and Validity Criterion of Models (in Japanese), *Mathematical Sciences*, No.153, 12–18.

Valiant, L. G. [1984] : A Theory of the Learnable, *Communications of ACM*, **27**, 1134–1142.