Neural Networks, Vol. 6, pp. 161–166, 1993
Printed in the USA. All rights reserved.

0893-6080/93 $6.00 + .00
Copyright © 1993 Pergamon Press Ltd.

ORIGINAL CONTRIBUTION

A Universal Theorem on Learning Curves

SHUN-ICHI AMARI

University of Tokyo

(*Received* 9 *March* 1992; *revised and accepted* 10 *July* 1992)

Abstract—*A learning curve shows how fast a learning machine improves its behavior as the number of training examples increases. This paper proves a universal asymptotic behavior of learning curves for general noiseless dichotomy machines, or neural networks. It is proved that irrespective of the architecture of a machine, the average predictive entropy or the information gain* $\langle e^*(t) \rangle$ *converges to* 0 *as* $\langle e^*(t) \rangle \sim d/t$ *as the number* t *of training examples increases, where* d *is the number of modifiable parameters of a machine.*

Keywords—Learning curve, Generalization error, Entropic error, Information gain, Universal theorem.

1. INTRODUCTION

Learning is an important subject of research, not only from the biological point of view but from the information-theoretic point of view. Recently, researchers have had an interest in properties of learning curves. Learning curves show how fast the behavior of a machine improves as the number of training examples increases. There are several approaches to this problem. An old example is (Amari, 1967) where the trade-off between the speed and accuracy of learning is studied in the general stochastic descent dynamics. (See also Heskes & Kappen, 1991, for more accurate treatment.) Valiant (1984) proposed a new paradigm to study learning behaviors and a number of research efforts have appeared along this line (for example, Baum & Haussler, 1989; Haussler, Littlestone, & Warmuth, 1988). The statistical-mechanical approach is another origin of research. It provides in-depth results for specific models when such tools as the thermodynamical limit, the replica symmetry, and the annealed approximation can be used. (See, e.g., Levin, Tishby, & Solla, 1990; Seung, Sompolinsky, & Tishby, 1992; Opper & Haussler, 1991, etc.) There are statistical and information-theoretic methods of approach too (Rissanen, 1986; Yamanishi, (1991); Haussler, Kearns, & Schapire, (1991), and others.)

Amari, Fujita, and Shinomoto (1992) used a Bayesian statistical approach to obtain four types of universal results on the asymptotic behaviors of learning under the annealed approximation. Amari and Murata (1992) studied the learning of stochastic machines and obtained universal results on learning curves for the generalization entropic error and the training entropic error (see also Murata, Yoshizawa, & Amari, 1992).

All of these approaches suggest that the average error decreases universally in the order of $1/t$. However, its coefficient remained unknown except for very special cases. This paper solves this problem in the logarithmic error case and proves universal property of learning curves for deterministic dichotomy machines. We study the behavior of the average entropic error or loss $\langle e^*(t) \rangle$, which is the average of the logarithm of the probability of correct classification for the next new signal after a machine has been trained by t training examples. This quantity is called the information gain by Haussler, Kearns, and Schapire (1991) because it represents the average amount of information carried by the t-th example. We prove the universal property that irrespective of the machine architecture, the average entropic loss decreases asymptotically as $\langle e^*(t) \rangle = d/t$, where d is the number of parameters.

2. MAIN RESULT

Consider a dichotomy of an n-dimensional Euclidean space $\mathbf{R}^n$ given by

$$\mathbf{R}^n = D_+ \cup D_-, \quad D_+ \cap D_- = \phi. \tag{2.1}$$

The present research was partially supported by Grant-in-aid in Special Priority Area (#03251105) of the Ministry of Education, Science and Culture of Japan.

The author would like to thank Dr. Kevin Judd for his comments on the manuscript.

Requests for reprints should be sent to Shun-ichi Amari, Department of Mathematical Engineering and Information Physics, Faculty of Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113, Japan.

A signal $\mathbf{x} \in D_+$ is called a positive example and has a target signal $y = 1$ associated with it. A signal $\mathbf{x} \in D_-$ is a negative example with target $y = -1$. Examples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t$ are randomly and independently chosen subject to a nonsingular probability density $p(\mathbf{x})$, where the nonsingularity implies that $p(\mathbf{x})$ is positive and bounded. Each example has associated with it a corresponding target signal $y_1, \ldots, y_t$.

Let us consider a set of machines (or neural networks) which are parameterized by d-dimensional real parameter $\mathbf{w} = (w_1, \ldots, w_d) \in \mathbf{R}^d$. A machine calculates a function $f(\mathbf{x}, \mathbf{w})$ and emits $y = 1$ when $f(\mathbf{x}, \mathbf{w}) > 0$ and $y = -1$ when $f(\mathbf{x}, \mathbf{w}) \leq 0$. Therefore, it defines a dichotomy

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

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

Let $\mathbf{w}_0$ be the parameter of the true dichotomy by which the target signals y_i, $i = 1, \ldots, t$, are generated for each $\mathbf{x}_i$.

Given t examples $\Xi_t = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_t, \mathbf{y}_t)\}$, let $A(\Xi_t)$ be the set of $\mathbf{w}$'s that classify the given examples Ξ_t correctly,

$$A(\Xi_t) = \{\mathbf{w} | \, y_i f(\mathbf{x}_i, \mathbf{w}) \geq 0, i = 1, \ldots, t\}.$$

This set will be called the admissible set of machines. At any stage of learning, a candidate machine $\mathbf{w}_t^*$ should be chosen from $A(\Xi_t)$. To this end, we assume a nonsingular prior distribution $q(\mathbf{w})$ of $\mathbf{w}$, and select $\mathbf{w}_t^*$ randomly from $A(\Xi_t)$ subject to the prior $q(\mathbf{w})$. This is called the Gibbs algorithm.

The performance of learning for a machine trained on Ξ_t can be evaluated by the probability of classifying the new example $\mathbf{x}_{t+1}$ correctly. A machine sometimes classifies $\mathbf{x}_{t+1}$ correctly but may not depending on which $\mathbf{x}_{t+1}$ is chosen as the next example and which $\mathbf{w}_t^*$ is chosen from $A(\Xi_t)$. More precisely, it classifies $\mathbf{x}_{t+1}$ correctly when $\mathbf{w}_t^*$ happens to be included in $A(\Xi_{t+1})$, where $\Xi_{t+1} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{t+1}, y_{t+1})\}$ and

$$A(\Xi_{t+1}) \subseteq A(\Xi_t).$$

Let

$$p(y_{t+1} \,|\mathbf{x}_{t+1}, \Xi_t)$$

be the probability that a randomly chosen machine from $A(\Xi_t)$ outputs the correct target signal y_{t+1}, that is, the probability that the machine is included in $A(\Xi_{t+1})$. The error probability is given by

$$e(t) = 1 - p(y_{t+1} \,|\mathbf{x}_{t+1}, \Xi_t),$$

which is called the generalization error because it is the probability of error in classifying a new signal $\mathbf{x}_{t+1}$.

The entropic error or entropic loss $e^*(t)$ is given by the logarithm of the probability of correct classification, that is,

$$e^*(t) = -\log p(y_{t+1} \,|\mathbf{x}_{t+1}, \Xi_t).$$

This is also a generalization error measure because it tends to 0 when the probability of correct classification tends to 1.

Moreover,

$$e^*(t) = -\log\{1 - e(t)\},$$

so that when $e(t)$ is small, they are almost identical.

The entropic error $e^*(t)$, as well as $e(t)$, is a random variable depending on the t examples Ξ_t and the next example $\mathbf{x}_{t+1}$. Its average over all the training examples Ξ_t and all $\mathbf{x}_{t+1}$ is denoted by $\langle e^*(t)\rangle$ and is called the average entropic error. When $\langle e^*(t)\rangle$ is regarded as a function of t, it is called the entropic learning curve. The main result of this paper is the following universal theorem which holds for any regular deterministic learning machine.

2.1. Universal Theorem

Under certain regularity conditions stated later, the entropic learning curve is asymptotically given by

$$\langle e^*(t)\rangle \sim \frac{d}{t},$$

irrespective of the structure of the machine, where d is the number of parameters.

The regularity conditions are imposed to ensure that each component of $\mathbf{w}$ makes some contribution to the machine performance, that is, the machine does not have superfluous parameters.

3. FUNDAMENTAL CONCEPTS AND A SKETCH OF PROOF

Let $Z_t = Z(\Xi_t)$ be the measure of the admissible set $A(\Xi_t)$, that is,

$$Z(\Xi_t) = \text{Prob}\{\mathbf{w} | \mathbf{w} \in A(\Xi_t)\}$$

$$= \int_{A(\Xi_t)} q(\mathbf{w})\, d\mathbf{w}.$$

This is the probability that a randomly chosen $\mathbf{w}$ satisfies the conditions $y_i f(\mathbf{x}_i, \mathbf{w}) > 0$, $i = 1, \ldots, t$, where Ξ_t is fixed,

$$Z(\Xi_t) = \text{Prob}\{\mathbf{w} : y_i f(\mathbf{x}_i, \mathbf{w}) \geq 0, i = 1, \ldots, t | \Xi_t\},$$

where $\mathbf{w}$ is the random variable. $Z_t = Z(\Xi_t)$ is a random variable depending on the examples Ξ_t. The smaller Z_t is, the easier it is to guess the true parameter $\mathbf{w}_0$, which is included in $A(\Xi_t)$. The probability of a randomly chosen machine from $A(\Xi_t)$ classifying the next example $\mathbf{x}_{t+1}$ correctly is given by

$$p(y_{t+1} \,|\mathbf{x}_{t+1}, \Xi_t) = \text{Prob}\{y_{t+1} f(\mathbf{x}_{t+1}, \mathbf{w}) \geq 0 |$$

$$\times \, y_i f(\mathbf{x}_i, \mathbf{w}) \geq 0, i = 1, \ldots, t, \Xi_t, \mathbf{x}_{t+1}\}$$

$$= \frac{\text{Prob}\{y_i f(\mathbf{x}_i, \mathbf{w}) \geq 0,\ i = 1, \ldots, t, t+1 | \Xi_{t+1}\}}{\text{Prob}\{y_i f(\mathbf{x}_i, \mathbf{w}) \geq i = 1, \ldots, t | \Xi_t\}}$$

$$= \frac{Z(\Xi_{t+1})}{Z(\Xi_t)}.$$

This depends on Ξ_t and $\Xi_{t+1} = \{\Xi_t, (\mathbf{x}_{t+1}, y_{t+1})\}$. We have

$$\langle e^*(t) \rangle = \langle \log Z_t \rangle - \langle \log Z_{t+1} \rangle.$$

This quantity can also be regarded as the expected information gain in the sense of C. E. Shannon carried by the $(t+1)$st example (Haussler, Kearns, & Schapire, 1991).

It is, in general, difficult to calculate $\langle \log Z_t \rangle$ except in special cases where the replica method is applicable. We are searching for the universal result in general cases. Since Z_t converges to 0 in distribution as $t \to \infty$, we first search for the scaling law such that

$$Y_t = a(t) Z_t,$$

where $a(t) \to \infty$ and Y_t converges to a random variable Y in distribution as t tends to ∞. We then have

$$\langle \log Z_t \rangle = \langle \log Y_t \rangle - \log a(t).$$

In this case, we obtain asymptotically

$$\langle \log Z_t \rangle - \langle \log Z_{t+1} \rangle = (\langle \log Y_t \rangle - \log \langle Y_{t+1} \rangle) + \log \frac{a(t+1)}{a(t)}.$$

If the first term converges to 0 faster than the second term, we have

$$\langle \log Z_t \rangle - \langle \log Z_{t+1} \rangle \approx \log \frac{a(t+1)}{a(t)}$$

without calculating $\langle \log Z_t \rangle$ explicitly.

In order to obtain the scaling law of Z_t, that is, $a(t)$, explicitly, we calculate the moments of Z_t, that is, $\langle (Z_t)^k \rangle$, $k = 1, 2, \ldots$. Here we use the replica-like method. However, we do not use the replica symmetry self-averaging assumption, which unfortunately does not hold in our general case. Nor do we use the replica trick of

$$\langle \log Z_t \rangle = \lim_{k \to 0} \frac{1}{k} \{\langle (Z_t)^k \rangle - 1\}.$$

By the calculation of moments, we have the universal scaling law,

$$a(t) = t^d,$$

from which the theorem is proved. The proof is given in the next section.

4. PROOF OF THE THEOREM

In order to evaluate $\langle \log Z_t \rangle$, we search for the probability distribution of Z_t. Amari, Fujita and Shinomoto (1992) introduced a function $s(\mathbf{w})$ for calculating $\langle Z_t \rangle$. This $s(\mathbf{w})$ is the probability that the machine specified by $\mathbf{w}$ behaves in the same manner as the true machine specified by the parameter $\mathbf{w}_0$, that is,

$$s(\mathbf{w}) = \text{Prob}\{f(\mathbf{x}, \mathbf{w}) \cdot f(\mathbf{x}, \mathbf{w}_0) \geq 0\}$$

for a randomly chosen input $\mathbf{x}$. Obviously,

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

and it decreases as $\mathbf{w}$ deviates from $\mathbf{w}_0$. We show properties of $s(\mathbf{w})$.

If $\mathbf{w}$ is different from $\mathbf{w}_0$, then its dichotomy regions $D_+(\mathbf{w})$ and $D_-(\mathbf{w})$ are different from those by $\mathbf{w}_0$. We denote by ΔD the difference of the dichotomy regions by $\mathbf{w}_0$ and $\mathbf{w}$. It is given by the symmetric difference of $D_+(\mathbf{w}_0)$ and $D_+(\mathbf{w})$, or of $D_-(\mathbf{w}_0)$ and $D_-(\mathbf{w})$, that is,

$$\Delta D = D_+(\mathbf{w}_0) \ominus D_+(\mathbf{w})$$

$$= D_+(\mathbf{w}_0) \cup D_+(\mathbf{w}) - D_+(\mathbf{w}_0) \cap D_+(\mathbf{w}).$$

We then have

$$s(\mathbf{w}) = 1 - |\Delta D|,$$

where

$$|\Delta D| = \int_{\Delta D} p(\mathbf{x})\, d\mathbf{x}$$

is the probability measure of ΔD.

The function $s(\mathbf{w})$ is usually not differentiable at $\mathbf{w}_0$ but it has directional derivatives. To show this, we put

$$\mathbf{w} = \mathbf{w}_0 + r\mathbf{e},$$

implying that $\mathbf{w}$ deviates from $\mathbf{w}_0$ in the direction $\mathbf{e}$, where $|\mathbf{e}| = 1$, by an amount r. If the size $|\Delta D|$ of the deviation of the dichotomy region is proportional to r when r is infinitesimally small, then we say that the distribution $p(\mathbf{x})$ is regular and the parameter $\mathbf{w}$ is effective. Specifically, we can assume as a regularity condition that

$$a(\mathbf{e}) = \lim_{r \to 0} \frac{1}{r} \int_{\Delta D} p(\mathbf{x})\, d\mathbf{x}$$

exists and is positive. This holds for many typical machines, including multilayer neural networks. We then have the expansion

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

where $-a(\mathbf{e})$ is the directional derivative of $s(\mathbf{w})$ in the direction $\mathbf{e}$.

Generalizing this function, we introduce a series of functions $s_1(\mathbf{w}_1)$, $s_2(\mathbf{w}_1, \mathbf{w}_2)$, $\ldots$, $s_k(\mathbf{w}_1, \ldots, \mathbf{w}_k)$, where $s_1(\mathbf{w}_1) = s(\mathbf{w}_1)$. In order to define $s_k(\mathbf{w}_1, \ldots, \mathbf{w}_k)$, we introduce k machines, each specified by $\mathbf{w}_1, \ldots, \mathbf{w}_k$. They may be replicas of the original machine $\mathbf{w}_0$, but with different parameters. This is the same technique as the replica method of calculating the moments of Z_t. The $s_k(\mathbf{w}_1, \ldots, \mathbf{w}_k)$ is defined by the

probability that all of the k machines, each specified by $\mathbf{w}_1, \ldots, \mathbf{w}_k$, classify a randomly chosen signal $\mathbf{x}$ correctly, that is,

$$s_k(\mathbf{w}_1, \ldots, \mathbf{w}_k) = \text{Prob}\{f(\mathbf{x}, \mathbf{w}_1)f(\mathbf{x}, \mathbf{w}_0) \geq 0, \ldots, f(\mathbf{x}, \mathbf{w}_k)f(\mathbf{x}, \mathbf{w}_0) \geq 0\}.$$

Obviously, s_k takes its maximum value at $\mathbf{w}_1 = \ldots = \mathbf{w}_k = \mathbf{w}_0$,

$$s_k(\mathbf{w}_0, \ldots, \mathbf{w}_0) = 1.$$

Now let us put

$$\Delta D_i = D_+(\mathbf{w}_0) \ominus D_+(\mathbf{w}_i).$$

A machine specified by $\mathbf{w}_i$ misclassifies those signals $\mathbf{x}$ belonging to ΔD_i. Therefore, we have

$$s_k(\mathbf{w}_1, \ldots, \mathbf{w}_k) = 1 - \left|\bigcup_{i=1}^{k} \Delta D_i\right|,$$

where $|D|$ is the probability measure of a set D,

$$|D| = \int_D p(\mathbf{x})\, d\mathbf{x}.$$

It should be noted that $s_k(\mathbf{w}_1, \ldots, \mathbf{w}_k)$ is a nonanalytic function and is not differentiable at $\mathbf{w}_1 = \mathbf{w}_2 = \ldots = \mathbf{w}_k = \mathbf{w}_0$.

We put

$$\mathbf{w}_i = \mathbf{w}_0 + r_i\mathbf{e}_i.$$

However, we cannot have such an expansion as

$$s_k(\mathbf{w}_1, \ldots, \mathbf{w}_k) = 1 - \sum_{i=1}^{k} a(\mathbf{e}_i)r_i + O(|r_i|^2)$$

even when all r_i are small.

We calculate the k-th moments of Z_t, searching for the scaling factor $a(t)$ such that $\langle\{a(t)\}^k((Z_t)^k\rangle$ never converge to 0 nor diverge to infinity for all $k = 1, 2, \ldots$.

LEMMA 1. *The k-th moment of Z_t is given by*

$$\langle(Z_t)^k\rangle = \int q(\mathbf{w}_1)\ldots q(\mathbf{w}_k)\{s_k(\mathbf{w}_1, \ldots, \mathbf{w}_k)\}^t\, d\mathbf{w}_1 \ldots d\mathbf{w}_k.$$

Proof. We prove the case of $k = 1$ first. We have

$$\begin{aligned}\langle Z_t\rangle &= \left\langle \int_{A_t} q(\mathbf{w})\, d\mathbf{w}\right\rangle \\ &= \int q(\mathbf{w})1\{y_if(\mathbf{x}_i, \mathbf{w}) \geq 0, i = 1, \ldots, t\} \times \prod_{i=1}^{t} p(\mathbf{x}_i)\, d\mathbf{x}_i\, d\mathbf{w} \\ &= \int q(\mathbf{w})\text{Prob}\{y_if(\mathbf{x}_i, \mathbf{w}) \geq 0, i = 1, \ldots, t|\mathbf{w}\}\, d\mathbf{w} \\ &= \int q(\mathbf{w})\{s(\mathbf{w})\}^t\, d\mathbf{w},\end{aligned}$$

where $1\{\cdot\}$ is the indicator function of the event in the bracket. We used the fact that

$$y_if(\mathbf{x}_i, \mathbf{w}) \geq 0$$

is equivalent to

$$f(\mathbf{x}_i, \mathbf{w}_0)f(\mathbf{x}_i, \mathbf{w}) \geq 0$$

and that the events $y_if(\mathbf{x}_i, \mathbf{w}) \geq 0$ are conditionally independent under the condition that $\mathbf{w}$ is fixed so that

$$\text{Prob}\{y_if(\mathbf{x}_i, \mathbf{w}) \geq 0, i = 1, \ldots, t|\mathbf{w}\} = \prod_{i=1}^{t} \text{Prob}\{y_if(\mathbf{x}_i, \mathbf{w}) \geq 0|\mathbf{w}\} = \{s(\mathbf{w})\}^t.$$

We next prove the case of $k = 2$.

$$\begin{aligned}\langle Z_t^2\rangle &= \left\langle \int_{A_t} q(\mathbf{w}_1)\, d\mathbf{w}_1 \int_{A_t} q(\mathbf{w}_2)\, d\mathbf{w}_2\right\rangle \\ &= \int q(\mathbf{w}_1)q(\mathbf{w}_2)1\{y_if(\mathbf{x}_i, \mathbf{w}_1) \geq 0, y_if(\mathbf{x}_i, \mathbf{w}_2) \geq 0, t = 1, \ldots, t\} \times \prod_{i=1}^{t} p(\mathbf{x}_i)\, d\mathbf{x}_i\, d\mathbf{w}_1\, d\mathbf{w}_2 \\ &= \int q(\mathbf{w}_1)q(\mathbf{w}_2)\text{Prob}\{y_if(\mathbf{x}_i, \mathbf{w}_1) \geq 0, y_if(\mathbf{x}_i, \mathbf{w}_2) \geq 0, i = 1, \ldots, t|\mathbf{w}_1, \mathbf{w}_2\}\, d\mathbf{w}_1\, d\mathbf{w}_2 \\ &= \int q(\mathbf{w}_1)q(\mathbf{w}_2)\{s_2(\mathbf{w}_1, \mathbf{w}_2)\}^t\, d\mathbf{w}_1\, d\mathbf{w}_2.\end{aligned}$$

The general case can be proved in a similar way so that for the k-th moment of Z_t we have

$$\begin{aligned}\langle(Z_t)^k\rangle &= \int \prod_{i=1}^{k} q(\mathbf{w}_i)\{s_k(\mathbf{w}_1, \ldots, \mathbf{w}_k)\}^t \prod_{i=1}^{k} d\mathbf{w}_i \\ &= \int \exp\{t \log s_k(\mathbf{w}_1, \ldots, \mathbf{w}_k) + \sum \log q(\mathbf{w}_i)\} \prod_{i=1}^{k} d\mathbf{w}_i.\end{aligned}$$

Now we evaluate the integral asymptotically. Taking into account that t is large and $q(\mathbf{w})$ is nonsingular, we have the asymptotic evaluation

$$\langle(Z_t)^k\rangle \sim \int \exp\{t \log s_k(\mathbf{w}_1, \ldots, \mathbf{w}_k)\} \prod d\mathbf{w}_i.$$

Although s_k is not differentiable, the integrand rapidly decreases to 0 as $\mathbf{w}_i$ deviates from $\mathbf{w}_0$. This is the same situation as we evaluate the integrand by the saddle point approximation. However, since s_k is not expandable at $\mathbf{w}_1 = \ldots = \mathbf{w}_k = \mathbf{w}_0$, we cannot simply apply the saddle point method. The replica symmetry technique is also not applicable because the interaction cannot be decomposed into a sum of the typical interaction of two $\mathbf{w}$'s, say, $\mathbf{w}_i$ and $\mathbf{w}_j$.

To evaluate the integral, we take the polar coordinates for each $\mathbf{w}_i$

$$\mathbf{w}_i = \mathbf{w}_0 + r_i\mathbf{e}_i,$$

so that

$$s_k = s_k(r_1, \ldots, r_k; \mathbf{e}_1, \ldots, \mathbf{e}_k)$$

and the volume element is

$$d\mathbf{w}_i = r_i^{d-1}\, dr_i d\mathbf{e}_i$$

where $d\mathbf{e}_i$ is the infinitesimal angular volume. We next transform $(r_1, \ldots, r_k)$ into the polar coordinates $(R, \mathbf{E})$, where

$$R = (r_1^2 + r_2^2 + \ldots + r_k^2)^{1/2}$$

and $\mathbf{E}$ represents the $(k-1)$-dimensional angular direction where $|\mathbf{E}| = 1$. We then have

$$s_k = s_k(R; \mathbf{E}, \mathbf{e}_1, \ldots, \mathbf{e}_k),$$

$$r_i = Rk_i(\mathbf{E}),$$

$$dr_1 \ldots dr_k = R^{k-1} dR d\mathbf{E}.$$

As $\mathbf{E}$ and $\mathbf{e}_i$ change, the shape of ΔD changes in a complex manner. However, since $|\Delta D_i|$ increases linearly in r_i when r_i is small, $|\Delta D|$ increases linearly in proportion to R when R is small and $\mathbf{E}$ and $\mathbf{e}_i$ are fixed. Hence, we have the following expansion,

$$s_k = 1 - c(\Omega)R + O(R^2),$$

where $\Omega = (\mathbf{E}, \mathbf{e}_1, \ldots, \mathbf{e}_k)$.

From

$$\log s_k \approx -c(\Omega)R,$$

provided t is large, we have asymptotically

$$\begin{aligned}\langle (Z_t)^k \rangle &= \int (R^{d-1})^k R^{k-1} \exp\{-tc(\Omega)R\} \Pi k_i(\mathbf{E})\, d\Omega \\ &= \frac{(dk-1)!}{t^{kd}} \int c(\Omega)^{-dk} \Pi k_i(\mathbf{E})\, d\Omega \\ &= \frac{\alpha_k}{t^{kd}},\end{aligned}$$

where α_k is a constant.

Since the k-th moment of Z_t converges to 0 as $\alpha_k t^{-kd}$, the required scaling is

$$a(t) = t^d$$

and $t^d Z_t$ converges to a random variable Y, which is of order 1.

Define a new random variable by rescaling Z_t as

$$Y_t = t^d Z_t.$$

The above result shows that Y_t converges to a random variable Y in distribution. Since Y_t is positive, we put

$$\log Y_t = \log Y + \varepsilon_t,$$

in distribution, where ε_t is of $o_p(1)$. This means

$$\langle \varepsilon_t \rangle = cb(t),$$

where $b(t) \to 0$ as $t \to \infty$. We assume that $b(t)$ monotonically converges to 0, such as $b(t) = 1/t$, $1/\log t$, etc. We then have

$$b(t+1) - b(t) = O\left(\frac{b(t)}{t}\right).$$

Now it is easy to prove the theorem. From

$$\langle \log Z_t \rangle \sim \langle \log(t^{-d} Y_t) \rangle = \langle \log Y \rangle - d \log t + \langle \varepsilon_t \rangle,$$

we have the asymptotic relation

$$\langle e^*(t) \rangle = \langle \log Z_t \rangle - \langle \log Z_{t+1} \rangle \sim \frac{d}{t} + o\left(\frac{1}{t}\right),$$

proving the main theorem.

5. CONCLUDING REMARKS

A universal asymptotic law of learning curves $\langle e^*(t) \rangle \sim d/t$ has been proved, which holds for any nonsingular learning machine, for example, a neural network. It is not necessary to rely on the annealed approximation, the replica method, the thermodynamical limit, or some other approximation. However, we have not yet obtained the rigorous asymptotic result of the generalization error $\langle e(t) \rangle$. Because of the concavity of the log function, the inequality

$$\langle e(t) \rangle \leq \langle e^*(t) \rangle$$

holds so that d/t is an upper bound (see Haussler, Kearns, & Schapire, 1991). It should be noted that even though $\langle e(t) \rangle$ may be small, this does not imply that $e(t)$ itself is small. It may happen, for example, that $e(t)$ is 0 with a large probability but when $e(t)$ is not 0, it is not small. Hence, we cannot use the approximation

$$-\log(1-e) \approx e.$$

For the simple perceptron model with 2-dimensional signals $\mathbf{x}$, we have $f(\mathbf{x}, \mathbf{w}) = \mathbf{w} \cdot \mathbf{x}$, $|\mathbf{w}| = 1$. We can prove in this special case the rigorous asymptotic result

$$\langle e(t) \rangle = \frac{2}{3t}.$$

On the other hand, the replica symmetry technique is applicable in the thermodynamical limit of $d \to \infty$, and a strong result,

$$\langle e(t) \rangle = 0.62 \frac{d}{t},$$

was proved by the statistical-mechanical method (Györgyi & Tishby, 1990; Opper & Haussler, 1991). It was shown by Haussler, Littlestone, and Warmuth (1989) that $\langle e(t) \rangle$ decreases in the order of t^{-1}. Various upper and lower bounds of order t^{-1} are given in Haussler, Kearns, and Schapire (1991). Therefore, it is conjectured that a universal theorem independent of the machine's architecture also holds for $\langle e(t) \rangle$, but such a result remains to be proved.

REFERENCES

Amari, S. (1967). Theory of adaptive pattern classifiers, *IEEE Transactions*, **EC-16**(3), 299–307.

Amari, S., Fujita, N., & Shinomoto, S. (1992). Four types of learning curves. *Neural Computation,* **4,** 605–618.

Amari, S., & Murata, N. (1992). Statistical theory of learning curves under entropic loss criterion. *Neural Computation* (in press).

Baum, E. B., & Haussler, D. (1989). What size net gives valid generalization. *Neural Computation,* **1,** 151–160.

Györgyi, G., & Tishby, N. (1990). Statistical theory of learning a rule. In K. Thuemann & R. Koeberle (Eds.), *Neural networks and spin glasses* (pp. 31–36). Singapore: World Scientific Publishers.

Haussler, D., Kearns, M., & Schapire, R. (1991). Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. *Proceedings of the Fourth Annual Workshop on Computer Learning Theory,* 61–74.

Haussler, D., Littlestone, N., & Warmuth, K. (1988). Predicting {0, 1} functions on randomly drawn points. *Proceedings of COLT* '88. 280–295.

Heskes, T. M., & Kappen, B. (1991). Learning processes in neural networks. *Physical Review A* **440,** 2718–2726.

Levin, E., Tishby, N., & Solla, S. A. (1990). A statistical approach to learning and generalization in layered neural networks. *Proceedings of the IEEE,* **78**(10), 1568–1574.

Murata, N., Yoshizawa, S., & Amari, S. (1992). Network information criterion—Determining the number of hidden units for an artificial neural network model. *METR,* **92-05,** University of Tokyo.

Opper, M., & Haussler, D. (1991). Calculation of the learning curve of Bayes optimal classification algorithm for learning perceptron with noise. *Proceedings of the Fourth Annual Workshop on Computer Learning Theory,* 75–87.

Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics,* **14,** 1080–1100.

Seung, S., Sompolinsky, H., & Tishby, N. (1992). Learning from examples in large neural networks, *Physical Review,* **A45,** 6058–6091.

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

Yamanishi, K. (1991). A loss bound model for on-line stochastic prediction strategies. *Proceedings of the Fourth Annual Workshop on Computer Learning Theory,* 290–302.