Communicated by Haim Sompolinsky

Four Types of Learning Curves

Shun-ichi Amari
Naotake Fujita
Department of Mathematical Engineering and Information Physics,
University of Tokyo, Tokyo 113, Japan

Shigeru Shinomoto
Department of Physics, Kyoto University, Kyoto 606, Japan

If machines are learning to make decisions given a number of examples, the generalization error $\varepsilon(t)$ is defined as the average probability that an incorrect decision is made for a new example by a machine when trained with t examples. The generalization error decreases as t increases, and the curve $\varepsilon(t)$ is called a learning curve. The present paper uses the Bayesian approach to show that given the annealed approximation, learning curves can be classified into four asymptotic types. If the machine is deterministic with noiseless teacher signals, then (1) $\varepsilon \sim at^{-1}$ when the correct machine parameter is unique, and (2) $\varepsilon \sim at^{-2}$ when the set of the correct parameters has a finite measure. If the teacher signals are noisy, then (3) $\varepsilon \sim at^{-1/2}$ for a deterministic machine, and (4) $\varepsilon \sim c + at^{-1}$ for a stochastic machine.

1 Introduction

A number of approaches have been proposed for machine learning. A classical example is the perceptron algorithm proposed by Rosenblatt (1961) for which a convergence theorem was given. A general theory of parametric learning was proposed by Amari (1967), Rumelhart *et al.* (1986), White (1989), and others, based on the stochastic gradient descent algorithm. See for example, Amari (1990) for a review of mathematical theory of neurocomputing.

A new framework of PAC learning was proposed by Valiant (1984), in which both the computational complexity and stochastic evaluation of performance are taken into account. The theory was successfully applied to neural networks by Baum and Haussler (1989), where the VC dimension of a dichotomy class plays an important role. However, the framework is too restrictive, and Haussler *et al.* (1988) studied the general convergence rate of a learning curve by removing the algorithmic complexity constraint, while Baum (1990) has attempted to remove the worst case constraint on the probability distribution.

A different approach is taken by Levin *et al.* (1990) in which the statistical mechanical approach is coupled with the Bayesian approach. See also Schwartz *et al.* (1990). A generalization error is defined by the probability that a machine that has been trained with t examples misclassifies a novel example. The statistical average of the generalization error over randomly generated examples is formulated using the Bayes formula. This theory can also be viewed as a straightforward application of the predictive minimum description length method proposed by Rissanen (1986). However, it is in general difficult to calculate the generalization error, so the "annealed approximation" is suggested (Levin *et al.* 1990).

The same problem has been treated by physicists (e.g., Hansel and Sompolinsky 1990; Sompolinsky *et al.* 1990; Györgyi and Tishby 1990; Seung *et al.* 1992). They use the techniques of statistical mechanics such as the thermodynamic limit, replica method, and annealed approximation to evaluate the average generalization error $\varepsilon(t)$. They have succeeded in obtaining an asymptotic form of $\varepsilon(t)$ and its phase transition for some specific models to which their methods are applicable.

In the present paper, we also discuss the average generalization error under the annealed approximation in the Bayesian framework. It is not necessary, however, to use a statistical–mechanical framework or to assume a Gibbs-type probability distribution. Our theory is statistical and is applicable to more general models beyond the limitation of the applicability of physical methods such as the thermodynamical limit and the replica method. We obtain four types of asymptotic behaviors in the way $\varepsilon(t)$ decreases with t when t is large. The results are in agreement with those obtained by other methods for specific models. The asymptotic behavior does not depend on a specific structure of the target function, or on a specific architecture of the machine; they are universal in this sense. The asymptotic behavior of a learning curve depends only on whether the teacher signals are noisy or not, whether the machine is deterministic or stochastic, and whether there is a unique correct machine.

The main concern of the present paper is the deterministic case. An exact analysis of stochastic cases will be given in a forthcoming paper (Amari and Murata 1992) without using the annealed approximation.

2 Main Results

The problem is stated as follows. Let us consider a dichotomy of an n-dimensional Euclidean space R^n,

$$R^n = D_+ \cup D_-, D_+ \cap D_- = \phi$$

where $\mathbf{x} \in D_+$ is called a positive example and $\mathbf{x} \in D_-$ a negative example. A target signal y accompanies each $\mathbf{x}$, where $y = 1$ for a positive example and $y = -1$ for a negative example. Given t randomly chosen examples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t$ independently drawn from a probability distribution $p(\mathbf{x})$ together with corresponding target signals $y_1, \ldots, y_t$, a learning machine is required to estimate the underlying dichotomy. The machine is evaluated by its generalization error $\varepsilon(t)$, that is, the probability that the next example $\mathbf{x}_{t+1}$ produced by the same probability distribution is misclassified by the machine. We evaluate the average generalization error under the so-called annealed approximation and give universal theorems on the convergence rate of $\varepsilon(t)$ as t tends to infinity.

A machine considered here is specified by a set of continuous parameters $\mathbf{w} = (w_1, \ldots, w_m) \in R^m$ and it calculates a function $f(\mathbf{x}, \mathbf{w})$. When the output of a machine is uniquely determined by the signum of $f(\mathbf{x}, \mathbf{w})$, the machine is said to be deterministic. In the deterministic case, the function $f(\mathbf{x}, \mathbf{w})$ specifies a dichotomy by

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

and

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

If the output is not deterministic but is given by a probability that is specified as a function of $f(\mathbf{x}, \mathbf{w})$, then it is said to be stochastic. A deterministic or stochastic neural network with modifiable synaptic weights gives a typical example of such a machine. For example, a layered feedforward neural network calculates a dichotomy function $f(\mathbf{x}, \mathbf{w})$, where $\mathbf{w}$ is a vector summarizing all the modifiable synaptic connection weights.

The main subject of the present paper is asymptotic learning behavior of deterministic machines. The smoothness or differentiability of dichotomy functions $f(\mathbf{x}, \mathbf{w})$ is not required in the deterministic case, so that it is applicable to multilayer deterministic thereshold-element networks as well as to analog-element neural networks. We also discuss stochastic cases to compare the difference in their asymptotic behaviors. Since we use the regular statistical estimation technique, the smoothness of $f(\mathbf{x}, \mathbf{w})$ is required in the stochastic case to guarantee the existence of the Fisher information matrix. This type of network is a generalization of the fully smooth network introduced by Sompolinsky *et al.* (1990) to the case where the network functions are smooth except for a threshold operation at the output.

Suppose that there exists parameter $\mathbf{w}_0$ such that the true machine calculates $f(\mathbf{x}, \mathbf{w}_0)$ and generates the teacher signal y based on it. In some deterministic cases, there exists a set of parameters all of which give the correct classification behavior. A typical case is that there is a neutral zone between the set of positive examples and the set of negative examples where no input signals are generated. We treat the cases both that a unique correct classifier exists and that a set of correct classifiers exists.

The teacher signal y is said to be noiseless if y is given by the sign of $f(\mathbf{x}, \mathbf{w}_0)$ and noisy if y is stochastically produced depending on the value $f(\mathbf{x}, \mathbf{w}_0)$, irrespective of the machine itself being deterministic or stochastic.

The following are the main results on the asymptotic behaviors of learning curves under the Bayesian framework and the annealed approximation.

Case 1. The average generalization error behaves asymptotically as

$$\varepsilon(t) \sim \frac{m}{t}$$

when a machine is deterministic, the teacher signal is noiseless, and the machine giving correct classification is uniquely specified by the m-dimensional parameter $\mathbf{w}_0$.

Case 2. The average generalization error behaves asymptotically as

$$\varepsilon(t) \sim \frac{c}{t^2}$$

when a machine is deterministic, the teacher signal is noiseless, and the set of correct classifiers has finite measure in the parameter space.

Case 3. The average generalization error behaves asymptotically as

$$\varepsilon(t) \sim \frac{c}{\sqrt{t}}$$

when a machine is deterministic with a unique correct machine, but the teacher signal is noisy.

Case 4. The average generalization error behaves asymptotically as

$$\varepsilon(t) \sim c_0 + \frac{c_1}{t}$$

when a machine is stochastic.

3 The Average Generalization Error

We review here the Bayesian framework of learning along the line of Levin *et al.* (1990). However, it is not necessary to use the statistical-mechanical framework or to assume a Gibbs-type distribution.

Let $p(y \mid \mathbf{x}, \mathbf{w})$ be the probability that a machine specified by $\mathbf{w}$ generates output y when $\mathbf{x}$ is input. It is given by a monotone function $k(f)$ of f in the stochastic case,

$$p(y = 1 \mid \mathbf{x}, \mathbf{w}) = k[f(\mathbf{x}, \mathbf{w})], \qquad 0 \leq k(f) \leq 1, \qquad k(0) = 1/2$$

In the deterministic case,

$$p(y \mid \mathbf{x}, \mathbf{w}) = \theta[yf(\mathbf{x}, \mathbf{w})]$$

where $\theta(z) = 1$ when $z > 0$ and 0 otherwise, that is, $p(y \mid \mathbf{x}, \mathbf{w})$ is equal to 1 when $yf(\mathbf{x}, \mathbf{w}) > 0$ and is otherwise 0. Let $q(\mathbf{w})$ be a prior distribution of parameter **w**. Then, the joint probability density that the parameter **w** is chosen and t examples of input–output pairs

$$\xi^{(t)} = [(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_t, y_t)]$$

are generated by the machine is

$$P(\mathbf{w}, \xi^{(t)}) = q(\mathbf{w}) \prod_{i=1}^{t} p(y_i \mid \mathbf{x}_i, \mathbf{w}) p(\mathbf{x}_i)$$

By using the Bayes formula, the posterior probability density of **w** is given by

$$Q(\mathbf{w} \mid \xi^{(t)}) = \frac{P(\mathbf{w}, \xi^{(t)})}{Z(\xi^{(t)}) \prod p(\mathbf{x}_i)}$$

where

$$Z(\xi^{(t)}) = \int q(\mathbf{w}) \prod_{i=1}^{t} p(y_i \mid \mathbf{x}_i, \mathbf{w})\, d\mathbf{w}$$

is the probability measure of **w**s generating $(y_1, \ldots, y_t)$ when inputs $(\mathbf{x}_1, \ldots, \mathbf{x}_t)$ are chosen.

In the deterministic case, the probability

$$Z_t = Z(\xi^{(t)}) = \int q(\mathbf{w}) \prod_{i=1}^{t} \theta[y_i f(\mathbf{x}_i, \mathbf{w})]\, d\mathbf{w}$$

is the measure of such **w** that are compatible with t examples $\xi^{(t)}$, that is, those **w** satisfying $y_i f(x_i, \mathbf{w}) > 0$, for all $i = 1, \ldots, t$. Therefore, the smaller this is, the easier it is to estimate the **w** that resolves the dichotomy. In the stochastic case, the probability $Z(\xi^{(t)})$ can also be used as a measure of identifiability of the true **w**. The quantity Z_t is related to the partition function of the Gibbs distribution in the special but important case studied by Levin *et al.* (1990) or the physicist approach with the thermodynamical limit of the dimensionality of **w** tending to infinity (Seung *et al.* 1992). The quantity defined here is more general, although we use the same notation Z_t.

The generalization error ε_t* based on t examples $\xi^{(t)}$ is defined, in the deterministic case, as the probability that a machine that classifies t examples $\xi^{(t)}$ correctly fails to correctly classify a new example $\mathbf{x}_{t+1}$. This is given by

$$\varepsilon_t* = \text{Prob}\{y_{t+1}\, f(\mathbf{x}_{t+1}, \mathbf{w}) < 0 \mid y_i f(\mathbf{x}_i, \mathbf{w}) > 0, \quad i = 1, \ldots, t\} = 1 - \frac{Z_{t+1}}{Z_t}$$

because

$$\begin{aligned} &\text{Prob}\{y_{t+1} f(\mathbf{x}_{t+1}, \mathbf{w}) > 0 \mid y_i f(\mathbf{x}_i, \mathbf{w}) > 0, \quad i = 1, \ldots, t\} \\ &= \frac{\text{Prob}\{y_i f(\mathbf{x}_i, \mathbf{w}) > 0, \quad i = 1, \ldots, t, t+1\}}{\text{Prob}\{y_i f(\mathbf{x}_i, \mathbf{w}) > 0, \quad i = 1, \ldots, t\}} \\ &= \frac{Z_{t+1}}{Z_t} \end{aligned}$$

This quantity can also be considered as the generalization error in the stochastic case, because

$$\frac{Z_{t+1}}{Z_t} = \text{Prob}\{y_{t+1} \mid \xi^{(t)}, \mathbf{x}_{t+1}\}$$

is the probability that the machine will correctly output y_{t+1} given $\mathbf{x}_{t+1}$ under the condition that t examples $\xi^{(t)}$ have been observed.

The generalization error ε_t* is a random variable depending on the randomly generated examples $\xi^{(t)}$. The average generalization error $\varepsilon(t)$ is the average of ε_t* over all the possible examples $\xi^{(t)}$ and a new pair $(y_{t+1}, \mathbf{x}_{t+1})$

$$\varepsilon(t) = \langle \varepsilon_t* \rangle = 1 - \langle Z_{t+1}/Z_t \rangle$$

$\langle \ \rangle$ denoting the expectation with respect to $\xi(t+1) = [\xi(t), (y_{t+1}, \mathbf{x}_{t+1})]$. This quantity is closely related to the stochastic complexity ε_t^c introduced by Rissanen (1986),

$$\varepsilon_t^c = -\langle \ln(1 - \varepsilon_t*) \rangle = \langle \ln Z_t \rangle - \langle \ln Z_{t+1} \rangle$$

The actual evaluation of the quantity such as $\langle Z_{t+1}/Z_t \rangle$ and $\langle \ln Z_t \rangle$ is generally a very hard problem and has been obtained only for a few model systems (see for example, Hansel and Sompolinsky 1990; Sompolinsky *et al.* 1990; Györgyi and Tishby 1990). We will show an exact example later. To obtain a rough estimate of $\varepsilon(t)$ or ε_t^c, we introduce approximations,

$$\langle Z_{t+1}/Z_t \rangle \sim \langle Z_{t+1} \rangle / \langle Z_t \rangle \qquad \text{and} \qquad \langle \ln Z_t \rangle \sim \ln \langle Z_t \rangle$$

called the "annealed average" (Levin *et al.* 1990), see also Schwartz *et al.* 1990). The approximations are valid if Z_t does not depend sensitively on the most probable $(\mathbf{x}_1, \dots, \mathbf{x}_t)$. For this reason, we may call it the random phase approximation. The validity of the approximation is still open and we will return to this point in the final section (see also Seung *et al.* 1992). It is easy to show under the approximation that the average generalization error $\varepsilon(t)$ and the stochastic complexity ε_t^c are closely related in the asymptotic limit $t \to \infty$ in that

$$\varepsilon(t) \sim \varepsilon_t^c$$

provided $\varepsilon(t) \to 0$. [It is proved in Amari and Murata (1992) that the annealed approximation for ε_t^c gives a correct result in the stochastic case. See also Amari (1992) for the deterministic case.] Thus the remaining work is based on the evaluation of the average phase volume $\langle Z_t \rangle$.

4 Case 1: A Unique Correct Deterministic Machine with a Noiseless Teacher

The expectation $\langle Z_t \rangle$ is calculated for a deterministic machine as follows. Let $s(\mathbf{w})$ be the probability that a machine specified by $\mathbf{w}$ classifies a randomly chosen $\mathbf{x}$ correctly, that is, as the true classifier specified by $\mathbf{w}_0$ does, and hence,

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

Since y_i is the signum of $f(\mathbf{x}_i, \mathbf{w}_0)$,

$$\begin{aligned}\langle Z_t\rangle &= \text{Prob}\{y_i f(\mathbf{x}_i,\mathbf{w}) > 0, \quad i = 1,\ldots,t\} \\ &= \int q(\mathbf{w})\text{Prob}\{y_i f(\mathbf{x}_i,\mathbf{w}) > 0, \quad i = 1,\ldots,t \mid \mathbf{w}\}\, d\mathbf{w} \\ &= \int q(\mathbf{w})\{s(\mathbf{w})\}^t d\mathbf{w}\end{aligned}$$

where the last equality follows because $y_i = \text{sgn} f(\mathbf{x}_i,\mathbf{w}_0)$ and because $f(\mathbf{x}_i,\mathbf{w}) \cdot f(\mathbf{x}_i,\mathbf{w}_0) > 0$, for $i = 1,\ldots,t$, are conditionally independent when $\mathbf{w}$ is fixed.

When $\mathbf{w}$ is slightly deviated from the true $\mathbf{w}_0$ in a unit direction $\mathbf{e}$, $|\mathbf{e}| = 1$,

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

the regions $D_+(\mathbf{w})$ and $D_-(\mathbf{w})$ are slightly deviated from the true $D_+(\mathbf{w}_0)$ and $D_-(\mathbf{w}_0)$. The classifier with $\mathbf{w}$ misclassifies those examples that belong to ΔD, which is the difference between $D_+(\mathbf{w})$ and $D_-(\mathbf{w}_0)$. Therefore, we have

$$s(\mathbf{w}) = 1 - \int_{\Delta D} p(\mathbf{x})\, d\mathbf{x}$$

We assume that the directional derivative

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

exists and is strictly positive for any direction $\mathbf{e}$. This holds when the probability of $\mathbf{x}$ belonging to ΔD caused by a small change $\Delta\mathbf{w}$ in $\mathbf{w}$ is in proportion to $|\Delta\mathbf{w}|$, irrespectively of the differentiability of $f(\mathbf{x},\mathbf{w})$. Note that $s(\mathbf{w})$ is not usually differentiable at $\mathbf{w} = \mathbf{w}_0$ in the deterministic case.

We use a method similar to the saddle point approximation to calculate $\langle Z_t\rangle$, namely,

$$\begin{aligned}\langle Z_t\rangle &= \int q(\mathbf{w})\{s(\mathbf{w})\}^t d\mathbf{w} \\ &= \int \exp\{t[\log s(\mathbf{w}) + \frac{1}{t}\log q(\mathbf{w})]\}\, d\mathbf{w}\end{aligned}$$

and by expanding

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

and neglecting smaller order terms when $q(\mathbf{w})$ is regular, then for large t,

$$\langle Z_t\rangle = \int \exp\{-ta(\mathbf{e})r\}\, d\mathbf{w}$$

Since the volume element $d\mathbf{w}$ can be written

$$d\mathbf{w} = r^{m-1}\, dr\, d\Omega$$

where $d\Omega$ is an angular volume element, then

$$\begin{aligned}\langle Z_t \rangle &= \int \exp\{-ta(\mathbf{e})r\} r^{m-1}\, dr\, d\Omega \\ &= \frac{C}{t^m}\end{aligned}$$

where

$$C = (m-1)! \int \frac{1}{\{a(\mathbf{e})\}^m} d\Omega$$

is a constant. From this, we have

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

proving the following theorem.

Theorem 1. *Given the annealed approximation, a noiseless teacher and that* $\mathbf{w}_0$ *is unique, the average generalization error of a deterministic machine decreases according to the universal formula*

$$\varepsilon(t) = \frac{m}{t}$$

where m is the dimension of $\mathbf{w}$.

Remark: We have assumed as regularity conditions in deriving the above result the existence of nonzero directional derivative $a(\mathbf{e})$ and a regular prior distribution $q(\mathbf{w})$. These conditions hold in usual situations, however, it is possible to extend our result to more general cases.

When the set $\mathbf{w}$ of correct classifiers forms a k-dimensional submanifold, we have,

$$\langle Z_t \rangle \propto t^{-(m-k)}$$

so that

$$\varepsilon(t) \sim \frac{m-k}{t}$$

In the case where the probability distribution $p(\mathbf{x})$ is extremely densely concentrated on or sparsely distributed in the neighborhood of the boundary of D_+ and D_-, we have the following expansion

$$s(\mathbf{w}) \sim 1 - a(\mathbf{e})r^{\alpha}, \qquad \alpha > 0$$

The result in this case is

$$\varepsilon(t) \sim \frac{m}{\alpha t}$$

so that the $1/t$ law still holds in agreement with results obtained by other methods for many models (Haussler *et al.* 1988; Sompolinsky *et al.* 1990).

5 Case 2: Deterministic Case with a Noiseless Teacher, Where a Finite Measure of Correct Classifiers Exists

In this case, $s(\mathbf{w}) = 1$ for $\mathbf{w} \in S_0$, where S_0 is the set of correct classifiers. We assume as a regularity condition that S_0 is a connected region having a piecewise smooth boundary. Moreover, we assume that if

$$\mathbf{w} = \mathbf{w}_\omega + r\mathbf{e}_\omega$$

where $\mathbf{w}_\omega$ is the value of $\mathbf{w}$ at position ω on ∂S_0 and $\mathbf{e}_\omega$ is the unit normal vector at ω, then $s(\mathbf{w})$ can be expanded as

$$s(\mathbf{w}) = \begin{cases} 1, & \mathbf{w} \in S_0 \\ 1 - a(\omega)r + O(r^2), & \mathbf{w} = \mathbf{w}_\omega + r\mathbf{e}_\omega \end{cases}$$

The calculation of $\langle Z_t \rangle$ proceeds in this case as

$$\begin{aligned} \langle Z_t \rangle &= \int q(\mathbf{w}) s(\mathbf{w})^t \, d\mathbf{w} \\ &= \int_{S_0} q(\mathbf{w}) \, d\mathbf{w} + \int\int q(\mathbf{w}) \exp\{-ta(\omega)r\} \, dr \, d\omega \\ &= P_0 + \frac{C'}{t} \end{aligned}$$

where P_0 is the measure of S_0 and

$$C' = \int_{\partial S_0} q(\mathbf{w}) \frac{1}{a(\omega)} \, d\omega$$

From this it follows that

$$\varepsilon(t) = 1 - \left(P_0 + \frac{C'}{t+1}\right) / \left(P_0 + \frac{C'}{t}\right) = \frac{B}{t^2}$$

where

$$B = \frac{C'}{P_0}$$

Hence the following theorem.

Theorem 2. *If S_0 has a finite measure $P_0 > 0$, the convergence rate of $\varepsilon(t)$ for a deterministic machine is as*

$$\varepsilon(t) \sim \frac{B}{t^2}$$

where B is a constant depending on P_0 and the function $f(\mathbf{x}, \mathbf{w})$.

Note that when S_0 tends to a point $\mathbf{w}_0$, P_0 tends to 0. This implies that B tends to infinity, and the asymptotic behavior changes to that of Theorem 1 where phase transition takes place.

Remark. The above result is obtained from the annealed approximation of $\langle Z_{t+1}/Z_t \rangle$. The above error probability $\varepsilon(t)$ is, roughly speaking, based on the learning scheme where at each time one chooses a machine randomly that correctly classifies the t examples $\xi(t)$. However, the behavior is exponential,

$$\varepsilon(t) \sim \exp\{-ct\}$$

if the learning scheme is to choose a machine randomly such that it correctly classifies $\xi^{(t)}$ and keep it if it correctly classifies the $(t+1)$st example, but if it does not then choose another machine randomly that does correctly classify the $(t+1)$ examples $\xi^{(t+1)}$. This is known as the perfect generalization (Seung *et al.* 1992).

6 Case 3: A Deterministic Machine with a Noisy Teacher

This section treats the case of where the true classifier is unique and is a deterministic machine with parameter $\mathbf{w}_0$ but teacher signals include stochastic error. The following is a typical example: The correct answer is 1 when $f(\mathbf{x}, \mathbf{w}_0) > 0$ and -1 when $f(\mathbf{x}, \mathbf{w}_0) < 0$, but the teacher signal y is 1 with probability $k[f(\mathbf{x}, \mathbf{w}_0)]$ and is -1 with probability $1 - k(f)$. A typical function k is given by

$$k(u) = \frac{1}{1 + \exp\{-\beta u\}}$$

where $1/\beta$ is the so-called "temperature."

In this case, we cannot usually find any $\mathbf{w}$ consistent with t examples $\xi^{(t)}$ when t is large. We use instead a statistical estimator $\hat{\mathbf{w}}_t$ from t examples.

From the statistical point of view, the problem is to estimate an unknown parameter vector $\mathbf{w}$ from t independent observations $(\mathbf{x}_i, y_i), i = 1, \ldots, t$ drawn from the probability distribution specified by $\mathbf{w}$,

$$r(\mathbf{x}, y; \mathbf{w}) = p(\mathbf{x})p(y \mid \mathbf{x}, \mathbf{w}) = p(\mathbf{x}) \left[\frac{1-y}{2} + yk(f(\mathbf{x}, \mathbf{w}))\right]$$

The Fisher information matrix G is defined by

$$G(\mathbf{w}) = E\left[\frac{\partial \log r(\mathbf{x}, y; \mathbf{w})}{\partial \mathbf{w}} \frac{\partial \log r(\mathbf{x}, y; \mathbf{w})^T}{\partial \mathbf{w}}\right]$$

where E is the expectation with respect to the distribution $r(\mathbf{x}, y; \mathbf{w})$, $\partial/\partial\mathbf{w}$ denotes the gradient column vector and the superscript T denotes the transposition. When the Fisher information exists, the estimation problem is regular. We assume that the problem is regular, which requires the differentiability of $f(\mathbf{x}, \mathbf{w})$. Let $\hat{\mathbf{w}}_t$ be the maximum likelihood estimator from t examples. It is well known that the covariance matrix of the maximum likelihood estimator $\hat{\mathbf{w}}_t$ is asymptotically given by

$$E[(\hat{\mathbf{w}}_t - \mathbf{w}_0)(\hat{\mathbf{w}}_t - \mathbf{w}_0)^T] = \frac{1}{t}G^{-1}$$

where G is the Fisher information matrix. The Fisher information matrix is explicitly given by

$$G = \beta^2 \int k(1-k)\frac{\partial f}{\partial \mathbf{w}}(\frac{\partial f}{\partial \mathbf{w}})^T p(\mathbf{x})\,d\mathbf{x}$$

where $k = k(f)$ and $f = f(\mathbf{x}, \mathbf{w})$ (see Amari 1991; Amari and Murata 1992).

The expectation of the generalization error is then given by

$$\varepsilon(t) = 1 - \langle s(\mathbf{w}_t)\rangle = \frac{D}{\sqrt{t}}$$

where

$$D = \int a(\mathbf{e})(\mathbf{e}G^{-1}\mathbf{e}^T)\,d\Omega$$

Theorem 3. *If their teacher signals include errors, then the average generalization error $\varepsilon(t)$ is asymptotically given by*

$$\varepsilon(t) \sim \frac{D}{\sqrt{t}}$$

This convergence rate coincides with one obtained by Hansel and Sompolinsky (1988). Here, the error probability is evaluated for a deterministic machine. When the temperature β^{-1} tends to 0, the teacher becomes noiseless. It should be noted that the Fisher information G tends to infinity in proportion to β^2 and hence, D tends to 0 in this limit. The asymptotic behavior then changes to that of Theorem 1, phase transition taking place.

7 Case 4: Stochastic Machine

In the case of a stochastic machine, the teacher signals are also stochastic. The error probability $\varepsilon(t)$ never tends to 0 in this case, but instead converges to some $\varepsilon_0 > 0$.

We have

$$\begin{aligned}\langle Z_t\rangle &= \int q(\mathbf{w})p(\xi^{(t)} \mid \mathbf{w})p(\xi^{(t)} \mid \mathbf{w}_0)\frac{1}{\prod p(\mathbf{x}_i)}\,d\mathbf{w}\,d\xi^{(t)} \\ &= \int \exp\{t \log s(\mathbf{w})\}\,d\mathbf{w}\end{aligned}$$

where

$$s(\mathbf{w}) = \int p(y \mid \mathbf{x}, \mathbf{w})p(y \mid \mathbf{x}, \mathbf{w}_0)p(\mathbf{x})\,d\mathbf{x}\,dy$$

Since $s(\mathbf{w})$ is smooth in this case, we have the following expansion at its maximum $\mathbf{w}_0'$,

$$s(\mathbf{w}) \sim c - (\mathbf{w} - \mathbf{w}_0')K(\mathbf{w} - \mathbf{w}_0')^T$$

with a constant c and a positive definite matrix K. Hence,

$$\langle Z_t\rangle \sim c_t t^{-m/2}$$

so that

$$1 - \frac{\langle Z_{t+1} \rangle}{\langle Z_t \rangle} = \varepsilon_0 + \frac{a}{t}$$

in agreement with Sompolinsky *et al.* (1990) and others.

Theorem 4. *For a stochastic machine, the generalization error behaves as*

$$\varepsilon(t) \sim \varepsilon_0 + \frac{a}{t}$$

8 Discussions

We have thus obtained four typical asymptotic laws of the generalization error $\varepsilon(t)$ under the annealed approximation. However, the validity of the annealed approximation is questionable, as is discussed in Seung *et al.* (1992). Györgyi and Tishby (1990) give a different result for a simple perceptron model based on the replica method, whose validity is not guaranteed. In order to see the validity of the approximation, we calculate the exact $\varepsilon(t)$ for the following simple example: Consider predicting a half space of R^2, where signals $\mathbf{x} = (x_1, x_2)$ are normally distributed with mean 0 and the identity covariance matrix, w is a scalar having a uniform prior $q(w)$, and

$$f(\mathbf{x}, w) = x_1 \cos w + x_2 \sin w$$

In this special case, the probability density function of Z_t is given by

$$p(Z_t) = 4t^2 Z_t \exp\{-2tZ_t\}$$

By calculating Z_{t+1} and averaging it over $\mathbf{x}_{t+1}$, the random variable e_t^* is denoted as

$$e_t^* = \frac{1}{2}(u^2 + v^2)/(u + v)$$

where u and v are independent random variables subject to the same density function

$$p(u) = t \exp\{-tu\}$$

From this, we have the asymptotically exact result

$$\varepsilon(t) = \frac{2}{3t}$$

while the annealed approximation gives $\varepsilon(t) \sim 1/t$. On the other hand, we have

$$\langle \log Z_t \rangle = c - \log t$$

so that

$$\varepsilon_t^c = \frac{1}{t}$$

where the annealed approximation holds.

This shows that the approximation gives the same order of t^{-1} but a different factor. It is interesting to see how the difference depends on the number m of parameters in $\mathbf{w}$.

Looking from the point of view of statistical inference, the deterministic case and stochastic case are quite different. The estimator $\hat{\mathbf{w}}_t$ from t example is usually subject to a normal distribution with a covariance matrix of order $1/t$ in the stochastic case. However, in the deterministic case, $\hat{\mathbf{w}}_t$ is usually not subject to a normal distribution. The squared error usually shows a stronger convergence. This is because the manifold of probability distributions has a Riemannian structure in the stochastic case (Amari 1985), while it has a Finslerian structure in the deterministic case (Amari 1987).

This suggests a difference of the validity of the annealed approximation in the two cases. We will discuss this point in more detail in a forthcoming paper (Amari and Murata 1992; Amari 1992).

Acknowledgments

The authors would like to thank Dr. Kevin Judd for valuable comments on the manuscript. This work was supported in part by Grant-in-Aid for Scientific Research on Priority Areas on "Higher-Order Brain Functions," the Ministry of Education, Science and Culture of Japan.

References

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

Amari, S. 1985. *Differential-Geometrical Methods in Statistics.* Springer Lecture Notes in Statistics, 28, Springer, New York.

Amari, S. 1987. Dual connections on the Hilbert bundles of statistical models. In *Geometrization of Statistical Theory*, C. T. J. Dodson, ed., pp. 123–152. ULDM, Lancaster, UK.

Amari, S. 1990. Mathematical foundations of neurocomputing. *Proc. IEEE* **78**, 1443–1463.

Amari, S. 1991. Dualistic geometry of the manifold of higher-order neurons. *Neural Networks* **4**, 443–451.

Amari, S. 1992. A universal theorem on learning curves. To appear.

Amari, S., and Murata, N. 1992. Predictive entropies and learning curves. To appear.

Baum, E. B. 1990. The perceptron algorithm is fast for nonmalicious distributions. *Neural Comp.* **2**, 248–260.

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

Györgyi, G., and Tishby, N. 1990. Statistical theory of learning a rule. In *Neural Networks and Spin Glasses*, W. K. Theumann and R. Koberle, eds., pp. 3–36, World Scientific, Singapore.

Haussler, D., Littlestone, N., and Warmuth, K. 1988. Predicting 0, 1 functions on randomly drawn points. *Proc. COLT'88*, pp. 280–295. Morgan Kaufmann, San Mateo, CA.

Hansel, D., and Sompolinsky, H. 1990. Learning from examples in a single-layer neural network. *Europhys. Lett.* **11**, 687–692.

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

Rissanen, J. 1986. Stochastic complexity and modeling. *Ann. Statist.* **14**, 1080–1100.

Rosenblatt, F. 1961. *Principles of Neurodynamics*. Washington, D.C.: Spartan.

Rumelhart, D., Hinton, G. E., and Williams, R. J. 1986. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 1: *Foundations*. MIT Press, Cambridge, MA.

Schwartz, D. B., Samalam, V. K., Solla, S. A., and Denker, J. S. 1990. Exhaustive learning. *Neural Comp.* **2**, 374–385.

Seung, H. S., Sompolinsky, H., and Tishby, N. 1992. Statistical mechanics of learning from examples. To appear.

Sompolinsky, H., Seung, S., and Tishby, N. 1990. Learning from examples in large neural networks. *Phy. Rev. Lett.* **64**, 1683–1686.

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

White, H. 1989. Learning in artificial neural networks: A statistical perspective. *Neural Comp.* **1**, 425–464.

Received 24 May 1991; accepted 15 November 1991.