TECHNICAL REPORTS

Statistical Theory of Learning Curves under Entropic Loss Criterion

Shun-ichi Amari and Noboru Murata

METR 91 - 12 November 1991

MATHEMATICAL ENGINEERING SECTION

DEPARTMENT OF MATHEMATICAL ENGINEERING AND INSTRUMENTATION PHYSICS

FACULTY OF ENGINEERING, UNIVERSITY OF TOKYO

BUNKYO-KU, TOKYO, JAPAN

Statistical Theory of Learning Curves under Entropic Loss Criterion

Shun-ichi Amari and Noboru Murata

Department of Mathematical Engineering and Information Physics
University of Tokyo, Bunkyo-ku, Tokyo 113 Japan

Abstract

A learning curve shows how fast a learning machine improves its behavior as the number of training examples increases. The present paper studies universal asymptotic behaviors of learning curves for stochastic dichotomy machines from the statistical point of view. The behavior of a trained machine is evaluated by its entropic loss, that is, the negative of the logarithm of its predictive distribution. It is important to study a difference in learing curves for the generalization error and training error. The generalization entropic error is the average loss when a machine trained with t examples classifies a new examples. The training entropic error is the average loss when the trained machine classifies the previous t examples that have been used in the training process. It is proved that the generalization error converges to H_0, the entropy of the conditional distribution of the true machine, as $H_0 + m/(2t)$, while the training error converges to $H_0 - m/(2t)$, where t is the number of examples and m is the number of modifiable parameters. This is a universal law because it holds for any regular machines irrespective of their structure. It shows an important relation between the training error and generalization error. The learning curves are also given when the underlying statistical model is not faithful.

1. Introduction

It is an important subject of research of neural networks and machine learning to study general characteristics of learning curves, which represent how fast the behavior of a learning machine is improved by learning from examples. This is an interdisciplinary problem related to neural networks, machine learning, algorithms, statistical inference, etc.

There has been a lot of research on learning algorithms in neural networks based on the stochastic descent method (see, e.g., Rosenblatt [1961], Widrow [1966], Amari [1967], Rumelhart, Hinton & Williams [1986], White [1989]). Even in an old paper by Amari [1967], the asymptotic dynamic behavior of learning curves was discussed, and the trade-off between the learning speed and the accuracy was studied (see Heskes and Kappen [1991] for recent developments).

A new framework of research was opened by Valiant [1984] in which the learning performance was evaluated stochastically under computational complexity constraints on algorithms. This approach was successfully applied to neural networks (Baum and Haussler [1989]). Haussler, Littlestone, and Warmuth [1988] studied the convergence rate of general learning curves by relaxing algorithmic constraints. Yamanishi [1990, 1991] among others extended the framework to noisy or stochastic machines. Levin, Tishby and Solla [1990] presented a Bayesian statistical-physical approach to study learning curves, where behaviors of generalization errors, predictive-entropic errors, and stochastic complexity of Rissanen [1986] were discussed. There are also a number of statistical-mechanical research on this problem (see, for example, Hansel and Sompolinsky [1990], Györgi and Tishby [1990], Seung, Sompolinsky and Tishby [1991], Opper and Haussler [1991]). The statistical-mechanical method can give

some deep theory for specific models, in which the replica method is typically used in the "thermodynamical limit" situation, implying that both the number m of the parameters of a machine and the number t of training examples tend to infinity with its ratio $\alpha = t/m$ being fixed.

Amari, Fujita and Shinomoto [1992] studied several types of universal properties of learning curves for dichotomy machines from a different point of view. However, the so-called annealed approximation is used so that the results are not necessary exact, and the discussions are mostly concentrated on deterministic machines but not on stochastic machines. The present paper uses the statistical approach to elucidate the universal property of learning curves for generalization error and training error. We consider a stochastic machine parametrized by a vector parameter $\mathbf{w}$, which when an input $\mathbf{x}$ is given, emits a binary output y with probability $p(y|\mathbf{x}, \mathbf{w})$. Given t examples $\xi_t = \{(y_1, \mathbf{x}_1), \cdots, (y_t, \mathbf{x}_t)\}$, where $\mathbf{x}_i$ is randomly generated from a fixed probability distribution $p(\mathbf{x})$ and y_i is the corresponding output generated by the true machine with parameter $\mathbf{w}_0$, the maximum likelihood estimator $\hat{\mathbf{w}}_t$ is calculated as a candidate machine whose behavior is given by the predictive distribution $p(y|\mathbf{x}, \mathbf{w})$ of the output y given $\mathbf{x}$. There are two different methods of evaluating the behavior of a machine : One is the average rate of error that the candidate machine predicts an output different from that of the true machine. The other loss is measured by the average predictive entropy evaluated by $-\log p(y|\mathbf{x}, \hat{\mathbf{w}}_t)$ for an input-output pair $(\mathbf{x}, y)$, which is zero if the prediction is 100% sure. We use this entropic loss to evaluate the learning behavior of machine (see also Yamanishi [1991]).

The generalization error is the average entropic loss or the average predictive entropy of a trained machine for a new example $(y_{t+1}, \mathbf{x}_{t+1})$. It is proved that the average predictive entropy for the generalization error converges to the entropy H_0 of the true machine asymptotically as

where $\langle\rangle$ denotes the expectation and m is the number of parameters in $\mathbf{w}$.

$$\langle e(t)\rangle_{\text{gen}} = H_0 + \frac{m}{2t},$$

This is in agreement with Yamanishi's result (Yamanishi [1991]). On the other hand, the training error is the average entropic loss of the estimated machine for the training examples $(y_i, \mathbf{x}_i)$, $i=1, \cdots, t$, which are used to estimate $\hat{\mathbf{w}}_t$. It is proved that the training error converges to

$$\langle e(t)\rangle_{\text{train}} = H_0 - \frac{m}{2t}.$$

These results coincide with those obtained in Seung, Sompolinsky and Tishby [1991] under the thermodynamical limit. Our statistical theory is universal in the sense that the results are valid for any regular stochastic machines, irrespective of the architecture of the machines or their sizes.

It is possible to obtain similar results when the true distribution is not included in a statistical model, where the $1/t$ convergence is the same but its coefficient is different. The precise coefficient is given in this case. Similar learning curves are obtainable from the Bayesian point of view, where we use the Bayesian predictive distribution or a randomly chosen one subject to the posterior distribution (Gibbs learning algorithm in Opper and Haussler [1991]).

The results are the same for the Bayesian predictive distribution. The Boltzmann learning algorithm gives

$$\langle e(t)\rangle_{\text{gen}} = H_0 + \frac{m}{t}$$

for the generalization error, and

$$\langle e(t)\rangle_{\text{train}} = H_0$$

for the training error.

2. Statistical theory of stochastic machines

Let us consider a machine which receives an n-dimensional input signal $\mathbf{x} \in \mathbf{R}^n$ and emits an binary output $y=1$ or -1. A machine is stochastic when y is not a function of $\mathbf{x}$ but y takes on 1 and -1 subject to a probability $p(y|\mathbf{x})$ specified by $\mathbf{x}$.

Let us consider a parametric family of machines where a machine is specified by an m-dimensional parameter $\mathbf{w} \in \mathbf{R}^m$ such that the probability of output y, given an input $\mathbf{x}$, is specified by $p(y|\mathbf{x}, \mathbf{w})$.

A typical form of $p(y|\mathbf{x}, \mathbf{w})$ is as follows: A machine first calculates a smooth function $f(\mathbf{x}, \mathbf{w})$ and then specifies the probabilities by

$$p(y=1|\mathbf{x}, \mathbf{w}) = k\{f(\mathbf{x}, \mathbf{w})\}, \tag{2.1}$$

$$p(y=-1|\mathbf{x}, \mathbf{w}) = 1 - k\{f(\mathbf{x}, \mathbf{w})\},$$

where

$$k(f) = \frac{1}{1+e^{-\beta f}}. \tag{2.2}$$

When $f(\mathbf{x}, \mathbf{w})>0$, it is more likely that the output of the machine is $y=1$, and when $f(\mathbf{x}, \mathbf{w})<0$, it is more likely that the output is $y=-1$. The parameter $1/\beta$ is the so-called "temperature" parameter . When $\beta=\infty$, the machine is deterministic, emitting $y=1$ when $f(\mathbf{x}, \mathbf{w})>0$ and $y=-1$ when $f(\mathbf{x}, \mathbf{w})<0$.

Let $\mathbf{w}_0$ be the true machine which generates examples. More specifically, let $p(\mathbf{x})$ be a non-singular probability distribution of input signals $\mathbf{x}$, and let $\mathbf{x}_1$, $\mathbf{x}_2$, $\cdots$, $\mathbf{x}_t$ be t randomly and independently chosen input signals subject to $p(\mathbf{x})$. The true machine generates answers y_1, $\cdots$, y_t by using the probability distribution $p(y_i|\mathbf{x}_i, \mathbf{w}_0)$, $i=1, \cdots, t$.

Let ξ_t be t pairs of examples thus generated,

$$\xi_t = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_t, y_t)\}, \tag{2.3}$$

from which we guess the true machine.

Let $\hat{\mathbf{w}}_t$ be the maximum likelihood estimator from t observed data ξ_t. Since the probability of obtaining ξ_t from a machine is

$$p(\xi_t|\mathbf{w}) = \prod_{i=1}^{t} p(\mathbf{x}_i)p(y_i|\mathbf{x}_i, \mathbf{w}),$$

by taking the logarithm, the $\hat{\mathbf{w}}_t$ maximizes

$$\log p(\xi_t|\mathbf{w}) = \sum_{i=1}^{t} l(y_i|\mathbf{x}_i, \mathbf{w}),$$

where

$$l(y|\mathbf{x}, \mathbf{w}) = \log p(y|\mathbf{x}, \mathbf{w}). \tag{2.4}$$

Hence, it satisfies

$$\sum \nabla l(y_i|\mathbf{x}_i, \hat{\mathbf{w}}_t) = 0, \tag{2.5}$$

where ∇ is the gradient with respect to $\mathbf{w}$,

$$\nabla l = \frac{\partial}{\partial \mathbf{w}} l = \left(\frac{\partial l}{\partial w_i} \right).$$

3. Generalization error and training error in terms of predictive distribution

Given t examples ξ_t, we estimate the true parameter $\hat{\mathbf{w}}_t$. The behavior of the estimated machine is given by the conditional probability $p(y|\mathbf{x}, \hat{\mathbf{w}}_t)$. Given the next example $\mathbf{x}_{t+1}$ randomly chosen subject to $p(\mathbf{x})$, the next output y_{t+1} is predicted with the probability $p(y_{t+1}|\mathbf{x}_{t+1}, \hat{\mathbf{w}}_t)$. The best prediction in the sense of the minimum expected error is that the predicted output y^*_{t+1} is 1 when

$$p(1|\mathbf{x}_{t+1}, \hat{\mathbf{w}}_t) > p(-1|\mathbf{x}_{t+1}, \hat{\mathbf{w}}_t),$$

and is -1 otherwise. The prediction error is given by $u_t = 0.5|y_{t+1} - y^*_{t+1}|$. This is a random variable depending on the t training examples ξ_t and $\mathbf{x}_{t+1}$.

Its expectation $\langle u_t \rangle_{\text{gen}}$ with respect to ξ_t and $\mathbf{x}_{t+1}$ is called the generalization error, because it denotes the average error when the machine trained by t examples predicts the output of a new example.

On the other hand, the training error is evaluated by the average of u_i ($i = 1, \cdots, t$) which are the errors when the machine with $\hat{\mathbf{w}}_t$ predicts the past outputs y_i for the past training inputs $\mathbf{x}_i$, retrospectively, by using the distribution $p(y_i|\mathbf{x}_i, \hat{\mathbf{w}}_t)$,

$$\langle u_t \rangle_{\text{train}} = \frac{1}{t} \langle \sum u_i \rangle .$$

These error never converge to 0 when a machine is stochastic, because even when $\hat{\mathbf{w}}_t$ converges to the true parameter $\mathbf{w}_0$ the machine cannot be free from stochastic errors.

The prediction error can be measured by the logarithm of the predictive probability for the new input output pair $(y_{t+1}, \mathbf{x}_{t+1})$,

$$e(t) = -\log p(y_{t+1}|\mathbf{x}_{t+1}, \hat{\mathbf{w}}_t). \tag{3.1}$$

This is called the entropic loss, log loss or stochastic complexity (Rissanen [1986], Yamanishi [1991]). The generalization entropic error is its expectation over the randomly generated training examples ξ_t, $\mathbf{x}_{t+1}$ and y_{t+1},

$$\langle e(t) \rangle_{\text{gen}} = \langle -\log p(y_{t+1}|\mathbf{x}_{t+1}, \hat{\mathbf{w}}_t) \rangle. \tag{3.2}$$

Since the expectation of $-\log p(y|\mathbf{x})$ is the conditional entropy,

$$H(Y|X) = E[-\log p(y|\mathbf{x})] = -\int \sum_y p(y|\mathbf{x}) \log p(y|\mathbf{x}) d\mathbf{x},$$

the generalization entropic loss is the expectation of the conditional entropy

$H(Y|X;\ \hat{\mathbf{w}}_t)$ over the estimator $\hat{\mathbf{w}}_t$. The entropic error of the true machine, specified by $\mathbf{w}_0$, is given by the conditional entropy,

$$H_0 = H(Y|X\ ;\ \mathbf{w}_0) = E[\ -\log p(y|\mathbf{x},\ \mathbf{w}_0)]. \tag{3.3}$$

Similarly, the training entropic error is the average of the entropic loss over the past examples $(y_i,\ \mathbf{x}_i)$ that are used for obtaining $\hat{\mathbf{w}}_t$,

$$\langle e(t)\rangle_{\text{train}} = -\ \frac{1}{t}\sum_{i=1}^{t}\langle \log\ \ p(y_i|\mathbf{x}_i,\ \hat{\mathbf{w}}_t)\rangle. \tag{3.4}$$

Obviously, the training error is smaller than the generalization error. It is interesting to know the difference between the two errors. The present paper studies the universal behaviors of the training and generalization entropic errors from the statistical point of view.

Universal Convergence Theorem for Training and Generalization Errors.

The asymptotic learning curve is given by

$$\langle e(t)\rangle_{\text{train}} = H_0 - \ \frac{m}{2t}, \tag{3.5}$$

for the entropic training error, and by

$$\langle e(t)\rangle_{\text{gen}} = H_0 + \ \frac{m}{2t} \tag{3.6}$$

for the entropic generalization error, where m is the number of parameters in $\mathbf{w}$.

The result of $1/t$ convergence is in good agreement with the results obtained by the statistical-mechanical approach (e.g., Seung, Sompolinsky and Tishby [1991].) It is possible to compare our result with Yamanishi [1991], where the cumulative log loss

is used. Here $\hat{\mathbf{w}}_i$ is the maximum likelihood estimator based on the i observations ξ_i. From (3. 6), we easily have

$$\langle e(t)\rangle_{\mathrm{cum}} = \frac{1}{t}\sum_{i=1}^{t}\langle -\log p(y_i|\mathbf{x}_i, \hat{\mathbf{w}}_i)\rangle$$

$$\langle e(t)\rangle_{\mathrm{cum}} = H_0 + \frac{m\log t}{2t},$$

in agreement with Yamanishi [1991], because of

$$\sum_{i=1}^{t}\frac{1}{i} = \log t + o(\log t).$$

The proof uses the following fundamental lemma in statistics.

Lemma. The maximum likelihood estimator $\hat{\mathbf{w}}_t$ based on t observations ξ_t is asymptotically normally distributed with mean $\mathbf{w}_0$ and covariance matrix $(tG)^{-1}$,

$$\hat{\mathbf{w}}_t \sim N(\mathbf{w}_0, \frac{1}{t}G^{-1}), \tag{3.7}$$

where $\mathbf{w}_0$ is the true parameter and $G=(g_{ij})$ is the Fisher information matrix defined by

$$g_{ij} = E[\frac{\partial}{\partial w_i}\log p(y|\mathbf{x}, \mathbf{w})\frac{\partial}{\partial w_j}\log p(y|\mathbf{x}, \mathbf{w})] \tag{3.8}$$

where E denotes the expectation with respect to the distribution $p(\mathbf{x})p(y|\mathbf{x}, \mathbf{w})$.

When the probability distribution is of the form (2. 1), the Fisher information matrix can be calculated to be

$$g_{ij} = \beta^2\int \sum_y k(1-k)\frac{\partial f}{\partial w_i}\frac{\partial f}{\partial w_j}p(\mathbf{x})d\mathbf{x}, \tag{3.9}$$

(see Amari [1991]). This shows that G diverges to ∞ as the temperature tends to 0, the estimator $\hat{\mathbf{w}}_t$ becoming more and more accurate.

Proof of the theorem. In order to calculate

$$\langle e(t)\rangle_{\text{gen}} = -E[\log p(y|\mathbf{x}, \hat{\mathbf{w}}_t)],$$

we expand

$$l(y|\mathbf{x}, \hat{\mathbf{w}}_t) = \log p(y|\mathbf{x}, \hat{\mathbf{w}}_t)$$

at $\mathbf{w}_0$,

$$l(y|\mathbf{x}, \hat{\mathbf{w}}_t) = l(y|\mathbf{x}, \mathbf{w}_0) + \nabla l(y|\mathbf{x}, \mathbf{w}_0)(\hat{\mathbf{w}}_t - \mathbf{w}_0)$$

$$+\frac{1}{2}(\hat{\mathbf{w}}_t - \mathbf{w}_0)^T \nabla\nabla l(y|\mathbf{x}, \mathbf{w}_0)(\hat{\mathbf{w}}_t - \mathbf{w}_0) + \cdots$$

where ∇l is the gradient with respect to $\mathbf{w}$, $\nabla\nabla l = (\partial^2 l / \partial w_i \partial w_j)$ is the Hessian matrix and the superscript T denotes the transposition of a column vector. By taking the expectation with respect to the new input-output pair $(y, \mathbf{x})$, we have

$$E[l(y|\mathbf{x}, \mathbf{w}_0)] = -H_0, \tag{3.10}$$

$$E[\nabla l(y|\mathbf{x}, \mathbf{w}_0)] = 0, \tag{3.11}$$

$$E[\nabla\nabla l(y|\mathbf{x}, \mathbf{w}_0)] = -G, \tag{3.12}$$

because of the identity

$$-E[\nabla\nabla l(y|\mathbf{x}, \mathbf{w}_0)] = E[(\nabla l)(\nabla l)^T].$$

Taking the expectation with respect to $\hat{\mathbf{w}}_t$, we have

$$E[\hat{\mathbf{w}}_t - \mathbf{w}_0] = O(1/t^2),$$

$$E[(\hat{\mathbf{w}}_t - \mathbf{w}_0)(\hat{\mathbf{w}}_t - \mathbf{w}_0)^T] = \frac{1}{t}G^{-1} + O(1/t^2),$$

or

$$E[(\hat{\mathbf{w}}_t - \mathbf{w}_0)^T G(\hat{\mathbf{w}}_t - \mathbf{w}_0)] = \frac{m}{t} + O(1/t^2).$$

Therefore, we have (3. 6).

We next evaluate the training error. To this end, expanding $l(y_i|\mathbf{x}_i, \hat{\mathbf{w}}_t)$, we have

$$l(y_i|\mathbf{x}_i, \hat{\mathbf{w}}_t) = l(y_i|\mathbf{x}_i, \mathbf{w}_0) + \nabla l(y_i|\mathbf{x}_i, \mathbf{w}_0)(\hat{\mathbf{w}}_t - \mathbf{w}_0)$$

$$+\frac{1}{2}(\hat{\mathbf{w}}_t - \mathbf{w}_0)^T(\nabla\nabla l)(\hat{\mathbf{w}}_t - \mathbf{w}_0) + \cdots. \tag{3.13}$$

We then expand

$$\nabla l(y_i|\mathbf{x}_i, \mathbf{w}_0) = \nabla l(y_i|\mathbf{x}_i, \hat{\mathbf{w}}_t) - (\hat{\mathbf{w}}_t - \mathbf{w}_0)^T\nabla\nabla l(y_i|\mathbf{x}_i, \mathbf{w}_0) + \cdots,$$

and substituting this in (3.13), and then summing over i, we have

$$\sum_{i=1}^{t} l(y_i|\mathbf{x}_i, \hat{\mathbf{w}}_t) = \sum l(y_i|\mathbf{x}_i, \mathbf{w}_0) - \frac{1}{2}\sum(\hat{\mathbf{w}}_t - \mathbf{w}_0)^T\nabla\nabla l(y_i|\mathbf{x}_i, \mathbf{w}_0)(\hat{\mathbf{w}}_t - \mathbf{w}_0)$$

because $\hat{\mathbf{w}}_t$ satisfies

$$\sum_{i=1}^{t} \nabla l(y_i|\mathbf{x}_i, \hat{\mathbf{w}}_t) = 0.$$

Since the $\mathbf{x}_i$'s are independently generated, then by the law of large numbers, we have

$$\frac{1}{t}\sum_{i=1}^{t} l(y_i|\mathbf{x}_i, \mathbf{w}_0) \sim -H_0.$$

On the other hand,

$$\frac{1}{t}\sum_{i=1}^{t} \nabla\nabla l(y_i|\mathbf{x}_i, \mathbf{w}_0) \sim E[\nabla\nabla l(y|\mathbf{x}, \mathbf{w}_0)] = -G.$$

Since $(\hat{\mathbf{w}}_t - \mathbf{w}_0)/\sqrt{t}$ is normally distributed with mean 0 and covariance matrix G^{-1},

$$(\hat{\mathbf{w}}_t - \mathbf{w}_0)^T G(\hat{\mathbf{w}}_t - \mathbf{w}_0)$$

can be expressed as a sum of squares of m independent normal random variables with mean 0 and variance 1, implying that it is subject to the χ^2-distribution of degree m. Therefore, we have

$$-\frac{1}{t}\sum_{i=1}^{t}\log p(y_i|\mathbf{x}_i, \hat{\mathbf{w}}_t) = H_0 - \frac{1}{2t}\chi^2_m, \tag{3.14}$$

where χ^2_m is a random variable subject to the χ^2-distribution of degree m. Since its expectation is m,

$$\langle e(t)\rangle_{\text{train}} = H_0 - \frac{m}{2t}.$$

4. Learning curves for unfaithful model

It has so far been assumed that there exists $\mathbf{w}_0$ such that the true dstribution $p(y|\mathbf{x})$ is written as

$$p(y|\mathbf{x}) = p(y|\mathbf{x}, \mathbf{w}_0). \tag{4.1}$$

This implies that the model $M = \{p(y|\mathbf{x}, \mathbf{w})\}$ of the distribution parameterized by $\mathbf{w}$ is faithful. When the true distribution is not in M, that is, there exists no $\mathbf{w}_0$ satisfying (4. 1), the model M is said to be unfaithful.

We can obtain learning curves in the case of unfaithful models, in a quite similar manner as in the faithful case. Let $p(y|\mathbf{x}, \mathbf{w}^*_0)$ be the best approximation of the true distribution $p(y|\mathbf{x})$ in the sense that $\mathbf{w}^*_0$ minimizes the Kullback-Leibler divergence

$$D[p(y|\mathbf{x}), p(y|\mathbf{x}, \mathbf{w})] = E[\log\frac{p(y|\mathbf{x})}{p(y|\mathbf{x}, \mathbf{w})}],$$

where the expectation E is taken with respect to the true distribution $p(\mathbf{x})p(y|\mathbf{x})$. We define the following quantities,

$$H^*_0 = E[-\log p(y|\mathbf{x}, \mathbf{w}^*_0)], \tag{4.2}$$

$$G^* = E[\{\nabla l(y|\mathbf{x}, \mathbf{w}^*_0)\}\{\nabla l(y|\mathbf{x}, \mathbf{w}^*_0)\}^T], \tag{4.3}$$

$$K^* = -E[\nabla\nabla l(y|\mathbf{x}, \mathbf{w}^*_0)]. \tag{4.4}$$

In the faithful case, $\mathbf{w}^*_0 = \mathbf{w}_0$, $H^*_0 = H_0$, and $G^* = K^* = G$ is the Fisher information matrix. However,

$$G^* = K^*$$

does not in general hold in the unfaithful case.

Universal Convergence Theorem for Learning Curves (Unfaithful Case).

The asymptotic learning curve is given by

$$\langle e(t) \rangle_{\text{train}} = H^*_0 - \frac{m^*}{2t}, \tag{4.5}$$

for the entropic training error, and by

$$\langle e(t) \rangle_{\text{gen}} = H^*_0 + \frac{m^*}{2t} \tag{4.6}$$

for the entropic generalization error, where

$$m^* = \text{tr}(K^{*-1} G^*)$$

is the trace of $K^{*-1} G^*$.

The proof uses the follwoing lemma.

Lemma. The maximum likelihood estimator $\hat{\mathbf{w}}_t$ under an unfaithful model is asymptotically normally distributed with mean $\mathbf{w}^*_0$ and covariance matrix $t^{-1}H^{*-1}GH^{*-1}$,

$$\hat{\mathbf{w}}_t \sim N(\mathbf{w}^*_0, \frac{1}{t} K^{*-1} G^* K^{*-1}). \tag{4.7}$$

The proof of the theorem is almost parallel to the faithful case, if we replace $\mathbf{w}_0$ by $\mathbf{w}^*_0$ and taking account that $K^* \neq G^*$.

5. Bayesian approach

The Bayesian approach uses a prior distribution $q(\mathbf{w})$, and then calculates the posterior probability distribution $Q(\mathbf{w}|\xi_t)$ based on t observations (training examples). The predictive distribution based on ξ_t is defined by

$$p(y|\mathbf{x};\xi_t)=\int p(y|\mathbf{x},\mathbf{w})Q(\mathbf{w}|\xi_t)d\mathbf{w}. \tag{5.1}$$

One idea is to use this predictive distribution for predicting the output. Another idea is to choose one candidate parameter $\mathbf{w}^*_t$ from the posterior distribution $Q(\mathbf{w}|\xi_t)$ and to use $p(y|\mathbf{x};\mathbf{w}^*_t)$ for predicting the output. This is called the Gibbs algorithm (Opper and Haussler [1991]).

The entropic generalization loss is evaluated by the expectation of $-\log p(y|\mathbf{x};\xi_t)$ for a new example $(y,\mathbf{x})$ or of $-\log p(y|\mathbf{x};\mathbf{w}^*_t)$, while the entropic training loss is given by

$$-\frac{1}{t}\sum_{i=1}^{t}\log p(y_i|\mathbf{x}_i,\xi_t) \quad \text{or} \quad -\frac{1}{t}\sum_{i=1}^{t}\log p(y_i|\mathbf{x}_i,\mathbf{w}^*_t).$$

We first study the case of using the predictive distribution $p(y|\mathbf{x};\xi_t)$. By putting

$$Z_t(\xi_t)=\int q(\mathbf{w})\prod_{i=1}^{t}p(y_i|\mathbf{x}_i,\mathbf{w})d\mathbf{w}, \tag{5.2}$$

the predictive distribution is written as

$$p(y_{t+1}|\mathbf{x}_{t+1},\xi_t)=Z_{t+1}/Z_t \tag{5.3}$$

(Amari, Fujita and Shinomoto [1992], see also the statistical-mechanical approach, for example, Levin, Tishby and Solla [1990], Seung, Tishby and Sompolinsky [1991], Opper and Haussler [1991]). Therefore,

$$\langle e(t)\rangle_{\text{gen}}=\langle\log Z_t\rangle-\langle\log Z_{t+1}\rangle. \tag{5.4}$$

By using the maximum likelihood estimator, we have

$$p(\mathbf{w}|\xi_t)\sim q(\mathbf{w})t^{m/2}|G|^{1/2}\exp\{-\frac{t}{2}(\mathbf{w}-\hat{\mathbf{w}}_t)G(\mathbf{w}-\hat{\mathbf{w}}_t)\}, \tag{5.5}$$

and

$$Z_t \sim t^{-m/2}|G|^{1/2}\prod_{i=1}^{t} p(y_i|\mathbf{x}_i, \hat{\mathbf{w}}_t) \tag{5.6}$$

or

$$\log Z_t \sim -H_0 - \frac{m}{2}\log t - \frac{1}{2}\log|G| + \frac{1}{2t}\chi_2^m. \tag{5.7}$$

From this we have

Theorem 3. The learniang curves for the Bayesian predictive distribution are the same as those for the maximum likelihood estimation.

We can perform similar calculations in the case of the Gibbs algorithm.

Theorem 4. The learning curves for the Gibbs algorithm is

$$\langle e(t)\rangle_{\text{train}} = H_0 \tag{5.8}$$

for the training error and

$$\langle e(t)\rangle_{\text{gen}} = H_0 + \frac{m}{t} \tag{5.9}$$

for the generalization error.

Conclusions

We have presented a statistical theory of learning curves. The characteristics of learning curves for stochastic machines can easily be analyzed by the ordinary asymptotic method of statistics. We have shown the universal $1/t$ convergence rule under the faithful and unfaithful statistical models. The difference between the training error and the generalization error is also given in detail. These results are in terms of the entropic loss, which fits very well with

the maximum likelihood estimator. The present theory is closely related with the AIC approach (Akaike [1974]) and the MDL approach (Rissanen [1986]).

Our statistical method cannot be applied to deterministic machines, because the statistical model is non-regular in this case, where the Fisher information diverges to infinity. However, we can prove

$$\langle e(t) \rangle_{\text{gen}} = \frac{m}{t}$$

for the entropic loss without using the annealed approximation. But this does not hold for the expected error u_t.

References

Akaike, H. [1974] : A new look at the statistical model identification, *IEEE Trans.* AC-19, pp. 716-723

Amari, S. [1967]: Theory of adaptive pattern classifiers, *IEEE Trans.*, EC-16, No. 3, pp.299-307.

Amari, S. [1991] : Dualistic geometry of the manifold of higher-order neurons, *Neural Networks*, pp. 443-445

Amari, S., Fujita, N. and Shinomoto, S. [1992] : Four types of learning curves, *Neural Computation*, to appear

Baum, E. B. and Haussler, D. [1989] : What size net gives valid generalization, *Neural Computation*, vol. 1, pp.151-160.

Györgyi, G. and Tishby, N. [1990] : Statistical theory of learning a rule, In *Neural Networks and Spin Glasses*, Thuemann, K. and Koeberle, R. eds., pp. 3-36, World Scientific

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

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

Heskes, T. M . and Kappen, B. [1991] : Learning proceses in neural networks, *Physical Review, A.*

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.

Opper, M. and Haussler, D. [1991] : Caluculation of the learning curve of Bayes optimal classification algorithm for learning a perceptron with noise, *Proc. fourth Ann Workshop on Comp. Learning Theory*, Morgan-Kaufmann

Rissanen, J. [1986] : Stochastic complexity and modeling. *Ann. Statist.*, vol. 14, pp.1080-1100.

Rosenblatt, F. [1961] : *Principles of Neurodynamics.* Spartan.

Rumelhart, D., Hinton, G. E. and Williams, R. J. [1986] : Learning internal representations by error propagation. *Parallel Distributed Processing : Explorations in the microstructure of congnition,* vol. 1 : *Foundations.* MIT Press.

Seung, S., Sompolinsky, H. and Tishby, N [1991] : Learning from examples in large neural networks. to be published.

Valiant, L. G. [1984] ; A theory of the learnable. *Comm. ACM.* vol. 27, No. 11, pp.1134-1142.

White, H. [1989] : Learning in artificial neural networks : A statistical perspective. *Nerual Computation*,vol. 1, pp.425-464.

Widrow, B. [1966] : *A statistical Theory of Adaptation*. Pergamon Press.

Yamanishi, K. [1990] : A learning criterion for stochastic rules, *Proc. Third Ann. Workshop on Comp. Learning Theory*, pp. 67-81, Morgan-Kaufman

Yamanishi, K. [1991] : A loss bound model for on-line stochastic prediction strategies. *Proc. Fourth Ann. Workshop on Comp. Learning Theory*, Morgan-Kaufmann