Communicated by Gerald Tesauro

Statistical Theory of Learning Curves under Entropic Loss Criterion

Shun-ichi Amari
Noboru Murata
*Department of Mathematical Engineering and Information Physics,
University of Tokyo, Bunkyo-ku, Tokyo 113, Japan*

The present paper elucidates a universal property of learning curves, which shows how the generalization error, training error, and the complexity of the underlying stochastic machine are related and how the behavior of a stochastic machine is improved as the number of training examples increases. The error is measured by the entropic loss. 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 as $H_0 - m^*/(2t)$, where t is the number of examples and m^* shows the complexity of the network. When the model is faithful, implying that the true machine is in the model, m^* is reduced to m, the number of modifiable parameters. This is a universal law because it holds for any regular machine irrespective of its structure under the maximum likelihood estimator. Similar relations are obtained for the Bayes and Gibbs learning algorithms. These learning curves show the relation among the accuracy of learning, the complexity of a model, and the number of training examples.

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. It is also important to evaluate the performance of a trained machine in terms of that for the old training examples. This is given by the relation between the generalization error and the training error, in terms of the complexity of the network. This is an interdisciplinary problem related to neural networks, machine learning, algorithms, statistical inference, etc.

There are a number of approaches to learning machines. One is the stochastic descent learning algorithm (see, e.g., Widrow 1966; Amari 1967; Rumelhart, Hinton, and Williams 1986; White 1989). Even in an old paper by Amari (1967) where the stochastic descent method was proposed for general layered neural networks, 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].

Another approach is a computational one (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 *et al.* (1988) studied the convergence rate of general learning curves by relaxing algorithmic constraints. See also Haussler *et al.* (1991) for recent developments. Here, the VC dimension plays a major role. Yamanishi (1990, 1991) extended the framework to noisy or stochastic machines.

The third approach is statistical-mechanical. Levin *et al.* (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 papers using a statistical–mechanical approach to this problem [see, for example, Hansel and Sompolinsky (1990); Györgyi and Tishby (1990); Seung *et al.* (1991); Opper and Haussler (1991)]. The statistical–mechanical approach can give some deep theory for specific simple models such as the simple perceptron, in which the replica method is typically used in the "thermodynamic limit" situation.

The present paper uses the fourth approach of statistical inference to elucidate the asymptotic learning behavior of a general stochastic learning dichotomy machine. The predictive entropic loss is used for evaluating the machine performance, where the maximum likelihood estimator of the Bayes and the Gibbs algorithms is used to choose a candidate machine based on training examples. The statistical approach is based on the asymptotic expansion of estimators [see, e.g., Amari (1985) for the higher order asymptotic expansion].

Before comparing the results of the present paper with others, we state the problems treated here and the main results. We consider a stochastic machine or stochastic multilayer neural network parameterized by a vector parameter $\mathbf{w}$, which, when an input $\mathbf{x}$ is given, emits a binary ouput y with probability $p(y \mid \mathbf{x}, \mathbf{w})$. Suppose we are given t examples $\xi_t = \{(y_1, \mathbf{x}_1), \ldots, (y_t, \mathbf{x}_t)\}$, where $\mathbf{x}_i$ is randomly generated from a fixed but unknown probability distribution $p(\mathbf{x})$ and y_i is a corresponding output generated by the true machine that has parameter $\mathbf{w}_0$. The maximum likelihood estimator $\hat{\mathbf{w}}_t$ is calculated as a candidate machine in the beginning. This machine predicts an output y for given $\mathbf{x}$ by the predictive distribution $p(y \mid \mathbf{x}, \hat{\mathbf{w}}_t)$. There are two different methods of evaluating the behavior of a machine. One is the average error rate at which the candidate machine predicts an output different from that of the true machine. The other is the average predictive entropy evaluated by the expectation of $-\log p(y \mid \mathbf{x}, \hat{\mathbf{w}}_t)$ for an input–output pair $(\mathbf{x}, y)$, which is zero if the prediction is 100% correct. We use this entropic loss to evaluate the learning behavior of a machine (see also Yamanishi 1991).

The generalization error is the average entropic loss, or 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 $\langle e(t)\rangle_{\text{gen}}$ converges to the entropy H_0 of the true machine asymptotically as in the following Theorems, where $\langle\ \rangle$ denotes the expectation and m is the number of parameters in $\mathbf{w}$. This is in agreement with Yamanishi's result (Yamanishi 1991). On the other hand, the training error $\langle e(t)\rangle_{\text{train}}$ is the average entropic loss of the candidate machine for the training examples $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, t$, which are used to estimate $\hat{\mathbf{w}}_t$. It is proved that the training error also converges as in the Theorems.

Theorem. *Universal Convergence Theorem.*

$$\begin{aligned}\langle e(t)\rangle_{\text{gen}} &= H_0 + \frac{m}{2t}\\ \langle e(t)\rangle_{\text{train}} &= H_0 - \frac{m}{2t}\end{aligned}$$

Since H_0 is unknown, we can obtain $\langle e(t)\rangle_{\text{gen}}$ by the relation

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

This is in good agreement with the AIC approach (Akaike 1974).

Instead of using the maximum likelihood estimator $\hat{\mathbf{w}}_t$, we can use the Bayes approach. When the behavior of a trained machine is evaluated by the Bayes posterior distribution (the Bayes algorithm), the learning curves are exactly the same as the previous Theorem. When we choose a candidate machine from the posterior distribution [the Gibbs learning algorithm (Opper and Haussler 1991)], we obtain the following result.

Theorem. *Bayesian Convergence Theorem.*

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

for the generalization error, and

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

for the training error.

The above results hold under the assumption that there exists $\mathbf{w}_0$ by which the true machine is specified. However, in many cases there is no $\mathbf{w}_0$ that specifies the true machine. The model is said to be unfaithful in this case. Let $\mathbf{w}_0^*$ be the best approximation to the true machine in the sense of Kullback divergence and let H_0^* be its entropy. By using the maximum likelihood estimator, we prove the following theorem, where

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

to be defined later plays a role of the effective dimensions.

Theorem. *Convergence Theorem for Unfaithful Model.*

$$\begin{aligned} \langle e(t)\rangle_{\text{gen}} &= H_0^* + \frac{m^*}{2t} \\ \langle e(t)\rangle_{\text{train}} &= H_0^* - \frac{m^*}{2t} \end{aligned}$$

Now we compare our methods and results with others. The $1/t$ convergence law was first proved by Haussler *et al.* (1988). However, its coefficients were not exactly known. Their exact values are still unknown even for the simple perceptron in the case of the error rate loss (Haussler *et al.* 1991). By using the entropic loss, the Theorem proves the universal coefficient of the convergence rate. This is universal in the sense that the theorem holds irrespective of the machine architecture. This implies that the VC dimension seems to be irrelevant for *stochastic* machines.

The statistical–mechanical approach is useful for determining the coefficient for the $1/t$ convergence. However, it uses the replica method which is unjustified. Moreover, it is applicable only to simple models like the simple perceptron and only in the case of the thermodynamic limit, implying that both t and m tends to infinity with a fixed ratio $\alpha = t/m$. Our method does not use the statistical–mechanical assumptions such as the replica method, annealed approximation, and the thermodynamic limit. Instead, we use the standard technique of asymptotic statistical inference, which is valid under the regularity conditions such as the existence of the moments of random variables and the existence of the Fisher information. The statistical technique is not applicable to deterministic machines, because they violate the regularity conditions. Therefore, the present paper complements the result by Amari *et al.* (1992) where learning curves are obtained for deterministic machines under the annealed approximation. Amari (1992) succeeded in obtaining a similar result for deterministic machines without the annealed approximation.

The present results are closely related to the model selection by AIC (Akaike 1974) and its generalization to general nonlinear neural networks (Murata *et al.* 1991; Moody 1992). The first Theorem can be regarded as a detailed version of the original AIC, while the third Theorem corresponds to its generalization. Moody (1992) proposes a similar generalization of AIC under a more general loss criterion in an unfaithful model. This approach is more general in the sense that it includes a regularization term, but is less general than Murata *et al.* (1991) in the sense that the latter treats a more general nonlinear model including non-additive noises. It should be pointed out that these papers give essentially the same effective number m^* of parameters, although they are different in their expressions.

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 a 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 $p(y \mid \mathbf{x})$ 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 $p(y \mid \mathbf{x}, \mathbf{w})$.

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

$$\begin{aligned} p(y = 1 \mid \mathbf{x}, \mathbf{w}) &= k[f(\mathbf{x}, \mathbf{w})] \\ p(y = -1 \mid \mathbf{x}, \mathbf{w}) &= 1 - k\{f(\mathbf{x}, \mathbf{w})\} \end{aligned} \tag{2.1}$$

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 us consider the case where the true machine that generates examples is specified by $\mathbf{w}_0$. More specifically, let $p(\mathbf{x})$ be a nonsingular probability distribution of input signals $\mathbf{x}$, and let $\mathbf{x}_1, \dots, \mathbf{x}_t$ be t randomly and independently chosen input signals subject to $p(\mathbf{x})$. The true machine generates answers $y_1, \dots, y_t$ using the probability distribution $p(y_i \mid \mathbf{x}_i, \mathbf{w}_0)$, $i = 1, \dots, t$.

Let ξ_t be t pairs of examples thus generated,

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

from which we guess the true machine.

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

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

by taking the logarithm,

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

should be maximized by the maximum likelihood estimator $\hat{\mathbf{w}}_t$, where

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

3 Generalization Error and Training Error in Terms of the Predictive Distribution

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

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

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 with t examples predicts the output of a new example.

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

$$\langle u_t \rangle_{\text{train}} = \frac{1}{t} \left\langle \sum_{i=1}^{t} u_i \right\rangle$$

This error never converges 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 also be measured by the logarithm of the predictive probability for the new input–output pair $(y_{t+1}, \mathbf{x}_{t+1})$,

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

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 and new input–output pair $(\mathbf{x}_{t+1}, y_{t+1})$,

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

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

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

the generalization entropic loss is the expectation of the conditional entropy $H(Y \mid X; \hat{\mathbf{w}}_t)$ 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,

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

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

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

Obviously, the training error is smaller than the generalization error. It is interesting to know the difference between the two errors. The following theorem gives the universal behaviors of the training and generalization entropic errors in a faithful model, that is, when there is a $\mathbf{w}_0$ specifying the true machine.

Theorem 1. *Universal Convergence Theorem for Training and Generalization Errors. The asymptotic learning curve for the entropic training error is given by*

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

and for the entropic generalization error by

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

where m is the number of parameters in $\mathbf{w}$.

The result of $1/t$ convergence is in good agreement with the results obtained for another model by the statistical-mechanical approach (e.g., Seung *et al.* 1991). It is possible to compare our result with Yamanishi (1991), where the cumulative log loss,

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

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

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

in agreement with Yamanishi (1991), because of

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

The proof of Theorem 1 uses the standard techniques of asymptotic statistics and is given in the Appendix.

4 Learning Curves for Unfaithful Model

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

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

This implies that the model $M = \{p(y \mid \mathbf{x}, \mathbf{w})\}$ of the distribution parameterized by $\mathbf{w}$ is faithful. When the true distribution is not in M, that is, there exisits 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 $p(y \mid \mathbf{x}, \mathbf{w}_0^*)$ be the best approximation of the true distribution $p(y \mid \mathbf{x})$ in the sense that $\mathbf{w}_0^*$ minimizes the Kullback–Leibler divergence

$$D[p(y \mid \mathbf{x}) : p(y \mid \mathbf{x}, \mathbf{w})] = E\left[\log \frac{p(y \mid \mathbf{x})}{p(y \mid \mathbf{x}, \mathbf{w})}\right]$$

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

$$\begin{aligned} H_0^* &= E[-\log p(y \mid \mathbf{x}, \mathbf{w}_0^*)] && (4.2)\\ G^* &= E[\{\nabla l(y \mid \mathbf{x}, \mathbf{w}_0^*)\}\{\nabla l(y \mid \mathbf{x}, \mathbf{w}_0^*)\}^T] && (4.3)\\ K^* &= -E[\nabla\nabla l(y \mid \mathbf{x}, \mathbf{w}_0^*)] && (4.4) \end{aligned}$$

where ∇ is the gradient operator, ∇ implying the column vector

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

the suffix T denotes the transposition of a vector, and $\nabla\nabla l$ is the Hessian matrix. In the faithful case, $\mathbf{w}_0^* = \mathbf{w}_0$, $H_0^* = H_0$, and $G^* = K^* = G$ is the Fisher information matrix. However, in general,

$$G^* \neq K^*$$

in the unfaithful case.

Theorem 2. *Convergence Theorem for Learning Curves : Unfaithful Case. The asymptotic learning curve for the entropic training error is given by*

$$\langle e(t)\rangle_{\text{train}} = H_0^* - \frac{m^*}{2t} \tag{4.5}$$

and for the entropic generalization error by

$$\langle e(t)\rangle_{\text{train}} = H_0^* + \frac{m^*}{2t} \tag{4.6}$$

where

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

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

See the Appendix for the proof. It is easy to see that $m^* = m$ in the faithful case, because of $K^* = G^*$. The above relations can be used for selecting an adequate model (see Murata *et al.* 1991; Moody 1992).

5 Bayesian Approach

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

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

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 $Q(\mathbf{w} \mid \xi_t)$ and to use $p(y \mid \mathbf{x}; \mathbf{w}_t^*)$ for predicting the output. The former one is called the Bayes algorithm and the latter is called the Gibbs algorithm (Opper and Haussler 1991).

The entropic generalization loss is evaluated by the expectation of $-\log p(y \mid \mathbf{x}; \xi_t)$ for a new example $(y, \mathbf{x})$ in the Bayes algorithm case and the expectation of $-\log p(y \mid \mathbf{x}; \mathbf{w}_t^*)$ in the Gibbs algorithm case. The entropic training loss is given, correspondingly, by

$$-\frac{1}{t}\sum_{i=1}^{t} \log p(y_i \mid \mathbf{x}_i, \xi_t) \qquad \text{and} \qquad -\frac{1}{t}\sum_{i=1}^{t} \log p(y_i \mid \mathbf{x}_i, \mathbf{w}_t^*)$$

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

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

the predictive distribution is written as

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

[Amari *et al.* (1992); see also the statistical-mechanical approach, for example, Levin *et al.* (1990); Seung *et al.* (1991); Opper and Haussler 1991)]. Therefore,

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

We can evaluate these quantities by statistical techniques (see the Appendix).

Theorem 3. *The learning 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 for the training error*

$$\langle e(t) \rangle_{\text{train}} = H_0 \tag{5.5}$$

and for the generalization error

$$\langle e(t) \rangle_{\text{gen}} = H_0 + \frac{m^*}{t} \tag{5.6}$$

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 a universal $1/t$ convergence rule for 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; Murata *et al.* 1991; Moody 1992) and the MDL approach (Rissanen 1986).

Our statistical method cannot be applied to deterministic machines, because the statistical model is nonregular 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 (Amari 1992). But this does not hold for the expected error u_t.

Appendix: Mathematical Proofs

In order to prove Theorem 1, we use 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 $(tG)^{-1}$,

$$\hat{\mathbf{w}}_t \sim N\left(\mathbf{w}_0, \frac{1}{t}G^{-1}\right) \tag{A1}$$

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

$$g_{ij} = E\left[\frac{\partial}{\partial w_i}\log p(y \mid \mathbf{x}, \mathbf{w}_0)\frac{\partial}{\partial w_j}\log p(y \mid \mathbf{x}, \mathbf{w}_0)\right] \tag{A2}$$

where E denotes the expectation with respect to the distribution $p(\mathbf{x})p(y \mid \mathbf{x}, \mathbf{w}_0)$.

When the probability distribution is of the form (2.1), the Fisher information matrix can be calculated to be

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

(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 Theorem 1. In order to calculate

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

we expand

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

at $\mathbf{w}_0$, giving

$$\begin{aligned} l(y \mid \mathbf{x}, \hat{\mathbf{w}}_t) &= l(y \mid \mathbf{x}, \mathbf{w}_0) + \nabla l(y \mid \mathbf{x}, \mathbf{w}_0)(\hat{\mathbf{w}}_t - \mathbf{w}_0) \\ &\quad + \frac{1}{2}(\hat{\mathbf{w}}_t - \mathbf{w}_0)^T \nabla\nabla l(y \mid \mathbf{x}, \mathbf{w}_0)(\hat{\mathbf{w}}_t - \mathbf{w}_0) + \cdots \end{aligned} \tag{A3}$$

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

$$E[l(y \mid \mathbf{x}, \mathbf{w}_0)] = -H_0, \tag{A4}$$

$$E[\nabla l(y \mid \mathbf{x}, \mathbf{w}_0)] = 0, \tag{A5}$$

$$E[\nabla\nabla l(y \mid \mathbf{x}, \mathbf{w}_0)] = -G \tag{A6}$$

because of the identity

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

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

$$\begin{aligned} E[\hat{\mathbf{w}}_t - \mathbf{w}_0] &= O(1/t) \\ E[(\hat{\mathbf{w}}_t - \mathbf{w}_0)(\hat{\mathbf{w}}_t - \mathbf{w}_0)^T] &= \frac{1}{t} G^{-1} + O(1/t^2) \end{aligned} \tag{A7}$$

and hence

$$E[(\hat{\mathbf{w}}_t - \mathbf{w}_0)^T G (\hat{\mathbf{w}}_t - \mathbf{w}_0)] = \frac{m}{t} + O(1/t^2) \tag{A8}$$

Therefore, using (A1) $\sim$ (A7), we obtain (3.6).

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

$$\begin{aligned} l(y_i \mid \mathbf{x}_i, \hat{\mathbf{w}}_t) &= l(y_i \mid \mathbf{x}_i, \mathbf{w}_0) + \nabla l(y_i \mid \mathbf{x}_i, \mathbf{w}_0)(\hat{\mathbf{w}}_t - \mathbf{w}_0) \\ &\quad + \frac{1}{2}(\hat{\mathbf{w}}_t - \mathbf{w}_0)^T [\nabla\nabla l(y_i \mid \mathbf{x}_i, \mathbf{w}_0)](\hat{\mathbf{w}}_t - \mathbf{w}_0) + \cdots \end{aligned} \tag{A9}$$

We then expand

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

and substituting this in (A9), and then summing over i, we have

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

because the maximum likelihood estimator $\hat{\mathbf{w}}_t$ satisfies

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

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

$$\begin{aligned}\frac{1}{t}\sum_{i=1}^{t} l(y_i \mid \mathbf{x}_i, \mathbf{w}_0) &\sim -H_0 \\ \frac{1}{t}\sum_{i=1}^{t} \nabla\nabla l(y_i \mid \mathbf{x}_i, \mathbf{w}_0) &\sim E[\nabla\nabla l(y \mid \mathbf{x}, \mathbf{w}_0)] = -G\end{aligned}$$

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

$$-\frac{1}{t}\sum_{i=1}^{t} \log p(y_i \mid \mathbf{x}_i, \hat{\mathbf{w}}_t) = H_0 - \frac{1}{2t}\chi_m^2$$

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

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

This proves Theorem 1.

In order to prove Theorem 2, we use the following 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 $t^{-1}K^{*-1}GK^{*-1}$,

$$\hat{\mathbf{w}}_t \sim N\left(\mathbf{w}_0^*, \frac{1}{t}K^{*-1}GK^{*-1}\right)$$

We do not give the proof of the lemma, because it is too technical. Refer to Murata *et al.* (1991). 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 $K^* \neq G^*$.

The Bayesian case can be proved by using the relations

$$\begin{aligned} p(\mathbf{w} \mid \xi_t) &\sim q(\mathbf{w}) t^{m/2} |G|^{1/2} \exp\{-\frac{t}{2}(\mathbf{w} - \hat{\mathbf{w}}_t)^T G (\mathbf{w} - \hat{\mathbf{w}}_t)\} \\ Z_t &\sim t^{m/2} |G|^{1/2} \prod_{i=1}^{t} p(y_i \mid \mathbf{x}_i, \hat{\mathbf{w}}_t) \\ \log Z_t &\sim -H_0 - \frac{m}{2} \log t - \frac{1}{2} \log |G| + \frac{1}{2t} \chi_m^2 \end{aligned}$$

However, the proof is much more complicated and we omit it. One can complete it by using the asymptotic statistical techniques.

Acknowledgments

The authors would like to thank Dr. K. Judd for comments on the manuscript. The present research is supported by the Japanese Ministry of Education, Science and Culture under Grant-in-Aid on Special Priority Area of Higher Order Brain Functioning.

References

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

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. 1991. Dualistic geometry of the manifold of higher-order neurons. *Neural Networks* **4**, 443–445.

Amari, S. 1992. Universal property of learning curves. *METR92-03*, Univ. of Tokyo.

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

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*, K. Thuemann and R. Koeberle, eds., pp. 3–36. World Scientific, Singapore.

Haussler, D., Kearns, M., and Shapire, R. 1991. Bounds on the sample complexity and the VC dimension. *Proc. 4th Ann. Workshop on Computational Learning Theory*, pp. 61–73. Morgan Kaufmann, San Mateo, CA.

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.

Heskes, T. M., and Kappen, B. 1991. Learning processes in neural networks. *Phys. Rev. A* **440**, 2718–2726.

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

Moody, J. E. 1992. The effective number of parameters: An analysis of generalization and regularization in nonlinear systems. In *Advances in Neural Information Processing Systems*, J. E. Moody, S. J. Hanson, and R. P. Lippmann, eds. Morgan Kaufmann, San Mateo, CA.

Murata, N., Yoshizawa, S., and Amari, S. 1991. A criterion for determining the number of parameters in an artificial neural networks model. In *Artificial Neural Networks*, T. Kohonen, K. Mäkisara, O. Simula, and J. Kangas, eds. Elsevier Science Publishers B. V., North-Holland.

Opper, M., and Haussler, D. 1991. Calculation of the learning curve of Bayes optimal classfication algorithm for learning a perceptron with noise. *Proc. 4th Ann. Workshop on Computational Learning Theory.* Morgan Kaufmann, San Mateo, CA.

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

Rosenblatt, F. 1961. *Principles of Neurodynamics.* Spartan, New York.

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*, Vol. 1. *Foundations.* MIT Press, Cambridge, MA.

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.* **27**(11), 1134–1142.

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

Widrow, B. 1966. *A Statistical Theory of Adaptation.* Pergamon Press, Oxford.

Yamanishi, K. 1990. A learning criterion for stochastic rules. *Proc. 3rd Ann. Workshop on Computational Learning Theory*, pp. 67–81. Morgan-Kaufmann, San Mateo, CA.

Yamanishi, K. 1991. A loss bound model for on-line stochastic prediction strategies. *Proc. 4th Ann. Workshop on Computational Learning Theory.* Morgan-Kaufmann, San Mateo, CA.

Received 14 November 1991; accepted 18 August 1992.

This article has been cited by:

2. Yu Nishiyama, Sumio Watanabe. 2007. Stochastic complexity of complete bipartite graph-type Boltzmann machines in mean field approximation. *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)* **90**:9, 1-9. [CrossRef]
3. Shun-ichi Amari , Hyeyoung Park , Tomoko Ozeki . 2006. Singularities Affect Dynamics of Learning in NeuromanifoldsSingularities Affect Dynamics of Learning in Neuromanifolds. *Neural Computation* **18**:5, 1007-1065. [Abstract] [PDF] [PDF Plus]
4. Koichiro Nishiue, Sumio Watanabe. 2005. Effects of priors in model selection problem of learning machines with singularities. *Electronics and Communications in Japan (Part II: Electronics)* **88**:2, 47-58. [CrossRef]
5. Kazushi Ikeda. 2004. An Asymptotic Statistical Theory of Polynomial Kernel MethodsAn Asymptotic Statistical Theory of Polynomial Kernel Methods. *Neural Computation* **16**:8, 1705-1719. [Abstract] [PDF] [PDF Plus]
6. Koji Tsuda, Shotaro Akaho, Motoaki Kawanabe, Klaus-Robert Müller. 2004. Asymptotic Properties of the Fisher KernelAsymptotic Properties of the Fisher Kernel. *Neural Computation* **16**:1, 115-137. [Abstract] [PDF] [PDF Plus]
7. Toshiaki Aida. 2001. Reparametrization-covariant theory for on-line learning of probability distributions. *Physical Review E* **64**:5. . [CrossRef]
8. Sumio Watanabe . 2001. Algebraic Analysis for Nonidentifiable Learning MachinesAlgebraic Analysis for Nonidentifiable Learning Machines. *Neural Computation* **13**:4, 899-933. [Abstract] [PDF] [PDF Plus]
9. Didier Herschkowitz, Manfred Opper. 2001. Retarded Learning: Rigorous Results from Statistical Mechanics. *Physical Review Letters* **86**:10, 2174-2177. [CrossRef]
10. Wenxin Jiang, M.A. Tanner. 2000. On the asymptotic normality of hierarchical mixtures-of-experts for generalized linear models. *IEEE Transactions on Information Theory* **46**:3, 1005-1013. [CrossRef]
11. Toshiaki Aida. 1999. Field Theoretical Analysis of On-Line Learning of Probability Distributions. *Physical Review Letters* **83**:17, 3554-3557. [CrossRef]
12. Silvia Scarpetta, Magnus Rattray, David Saad. 1999. *Journal of Physics A: Mathematical and General* **32**:22, 4047-4059. [CrossRef]
13. S. Guarnieri, F. Piazza, A. Uncini. 1999. Multilayer feedforward networks with adaptive spline activation function. *IEEE Transactions on Neural Networks* **10**:3, 672-683. [CrossRef]
14. Terrence L. Fine , Sayandev Mukherjee . 1999. Parameter Convergence and Learning Curves for Neural NetworksParameter Convergence and Learning Curves for Neural Networks. *Neural Computation* **11**:3, 747-769. [Abstract] [PDF] [PDF Plus]

15. Magnus Rattray, David Saad. 1999. Analysis of natural gradient descent for multilayer neural networks. *Physical Review E* **59**:4, 4523-4532. [CrossRef]
16. Didier Herschkowitz, Jean-Pierre Nadal. 1999. Unsupervised and supervised learning: Mutual information between parameters and observations. *Physical Review E* **59**:3, 3344-3360. [CrossRef]
17. A. Uncini, L. Vecci, P. Campolucci, F. Piazza. 1999. Complex-valued neural networks with adaptive spline activation function for digital-radio-links nonlinear equalization. *IEEE Transactions on Signal Processing* **47**:2, 505-514. [CrossRef]
18. Magnus Rattray, David Saad, Shun-ichi Amari. 1998. Natural Gradient Descent for On-Line Learning. *Physical Review Letters* **81**:24, 5461-5464. [CrossRef]
19. Jianfeng Feng. 1998. *Journal of Physics A: Mathematical and General* **31**:17, 4037-4048. [CrossRef]
20. A.J. Zeevi, R. Meir, V. Maiorov. 1998. Error bounds for functional approximation and estimation using mixtures of experts. *IEEE Transactions on Information Theory* **44**:3, 1010-1025. [CrossRef]
21. Shun-ichi Amari . 1998. Natural Gradient Works Efficiently in LearningNatural Gradient Works Efficiently in Learning. *Neural Computation* **10**:2, 251-276. [Abstract] [PDF] [PDF Plus]
22. S. Raudys. 1997. On dimensionality, sample size, and classification error of nonparametric linear classification algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**:6, 667-671. [CrossRef]
23. A. Atiya, Chuanyi Ji. 1997. How initial conditions affect generalization performance in large networks. *IEEE Transactions on Neural Networks* **8**:2, 448-451. [CrossRef]
24. Sepp Hochreiter, Jürgen Schmidhuber. 1997. Flat MinimaFlat Minima. *Neural Computation* **9**:1, 1-42. [Abstract] [PDF] [PDF Plus]
25. S. Amari, N. Murata, K.-R. Muller, M. Finke, H.H. Yang. 1997. Asymptotic statistical theory of overtraining and cross-validation. *IEEE Transactions on Neural Networks* **8**:5, 985-996. [CrossRef]
26. Manfred Opper. 1996. On-line versus Off-line Learning from Random Examples: General Results. *Physical Review Letters* **77**:22, 4671-4674. [CrossRef]
27. K.-R. Müller, M. Finke, N. Murata, K. Schulten, S. Amari. 1996. A Numerical Study on Learning Curves in Stochastic Multilayer Feedforward NetworksA Numerical Study on Learning Curves in Stochastic Multilayer Feedforward Networks. *Neural Computation* **8**:5, 1085-1106. [Abstract] [PDF] [PDF Plus]
28. Manfred Opper, David Haussler. 1995. Bounds for Predictive Errors in the Statistical Mechanics of Supervised Learning. *Physical Review Letters* **75**:20, 3772-3775. [CrossRef]
29. Florence d'Alché-Buc, Jean-Pierre Nadal. 1995. Asymptotic performances of a constructive algorithm. *Neural Processing Letters* **2**:2, 1-4. [CrossRef]

30. M. B Gordon, D. R Grempel. 1995. Learning with a Temperature-Dependent Algorithm. *Europhysics Letters (EPL)* **29**:3, 257-262. [CrossRef]

31. Peter Sollich. 1994. Query construction, entropy, and generalization in neural-network models. *Physical Review E* **49**:5, 4637-4651. [CrossRef]