Network Information Criterion—Determining the Number of Hidden Units for an Artificial Neural Network Model

Noboru Murata, Shuji Yoshizawa, *Member, IEEE,* and Shun-ichi Amari, *Member, IEEE*

Abstract— **The problem of model selection, or determination of the number of hidden units, can be approached statistically, by generalizing Akaike's information criterion (AIC) to be applicable to unfaithful (i.e., unrealizable) models with general loss criteria including regularization terms. The relation between the training error and the generalization error is studied in terms of the number of the training examples and the complexity of a network which reduces to the number of parameters in the ordinary statistical theory of the AIC. This relation leads to a new Network Information Criterion (NIC) which is useful for selecting the optimal network model based on a given training set.**

I. Introduction

In engineering fields, one of the most important application of artificial neural networks is modeling a system with an unknown input–output relation. Usually, we do not have accurate information of the system and we can utilize only observations form the system. Usually in such a case, given a fixed architecture of networks, parameters are modified by the stochastic gradient descent method which eventually minimizes a certain loss function. Learning is carried out based on a training set which consists of a number of examples observed from the actual system [1]–[3]. For instance, the backpropagation method is used for learning of multilayered perceptrons with sigmoidal functions [4].

An important but difficult problem is to determine the optimal number of parameters. In other words, we wish to determine the number of hidden units needed to mimic the system by using only input–output examples. The difficulty is because an increase in the number of the parameters lessens the output errors for the training examples, but increases the errors for novel examples. Such a phenomena is often called "over-fitting." For instance, let us think about multilayered perceptrons. Suppose that we have two networks, one which has ten hidden units and another which has one hundred hidden units. After enough training with a thousand examples, the large network, which has a hundred hidden units, may produce better outputs for the training examples than the small one, but it may emit worse outputs for inexperienced inputs. Generating accurate outputs for known inputs competes against predicting accurate outputs for unknown inputs. Such a disparity between known and unknown inputs increases as the number of parameters to be estimated is getting large. In order to solve this problem, we need to consider the relation among the complexity of a model, the performance for the training data and the number of examples, for example using Akaike's Information Criterion (AIC) [5] and the Minimum Description Length (MDL) [6]. There has been some research intending to apply these principles (e.g., [7]–[10]). The present paper is a detailed version of a short note by Murata *et al.* [11], giving a most general solution to this problem.

The present paper treats a family of stochastic neural networks of the feed forward type, which means that a network does not have any recurrent connections. Systems and networks are regarded as stochastic machines. A system, which is mimicked, produces an output y for an input x, based on a conditional probability $q(\mathbf{y}|\mathbf{x})$, and a network, which mimics the system, produces an output y for an input x, based on a conditional probability $p(\mathbf{y}|\mathbf{x},\theta)$, where θ is the parameter vector which specifies the network, that is, weights and thresholds. The problem is to find the optimal model in a family of networks and find the optimal parameter to approximate the system's conditional distribution from which a set of training examples is produced.

Within this framework, this problem is regarded as a statistical problem. We take the AIC approach, but generalize it as two points. The first is that the true distribution $q(\mathbf{y}|\mathbf{x})$ is not necessarily included in any of the model $\{p(\mathbf{y}|\mathbf{x},\theta)\}$. In such a case, the true distribution is said to be an unrealizable rule and the model is said to be unfaithful. The second is that we introduce a notion of general loss functions including regularization terms. The regularization term was introduced to the loss by Moody [8], and it gives the smoothness condition and fits well with the purpose of neural information processing. These loss functions include the negative of the likelihood as a special case which leads to the maximum likelihood estimator.

In Section II, we define a general loss function which measures differences between the system and the model. Once the training set is fixed, we can only use the empirical distribution of the training set instead of the true unknown distribution. We then formulate a leaning procedure based on a repeated resampling plan from a fixed training set of examples sampled from the true distribution.

Section III is devoted to the evaluation of the network parameters after learning. Two kinds of evaluations are

Manuscript received June 29, 1992; revised December 2, 1993. This work was supported in part by grant-in-aid 03251102 and 03234109 from the Ministry of Education of Japan.

The authors are with the Department of Mathematical Engineering and Information Physics, Faculty of Engineering, University of Tokyo, Bunkyo-ku, Tokyo 113, Japan.

IEEE Log Number 9400108.

1045–9227/94$04.00 © 1994 IEEE

Authorized licensed use limited to: University of Minnesota. Downloaded on June 21,2010 at 15:16:55 UTC from IEEE Xplore. Restrictions apply.

necessary: one is in how well the learned parameter approximates the quasi-optimal parameter which is optimal to the empirical distribution of the training set, and the other is in how well the quasi-optimal parameter approximates the true optimal parameter for the unknown distribution. These evaluations elucidate the relation between the training error and the generalization error in terms of the complexity of a network and the number of training examples (see also [12], [13]).

In Section IV, based on this relation, we propose the Network Information Criterion (NIC), which reduces to the AIC in an ordinary statistical setting. The criterion leads to the effective number $m*$ of parameters which is the same as one introduced by Moody [8] in the case of additive noise. We finally give an important remark on how the criterion is applicable only for comparing models within a hierarchical set. This constraint originates from the fact that the stochastic fluctuation in the training error cannot be evaluated solely by its ensemble average. However, in the case of a hierarchical set, this fluctuation term is common to all members, and hence the terms cancel out. This restriction, on evaluating the generalization error in terms of the training error, is still not well recognized.

II. Discrepancy Function and Learning Rule

Let us consider a stochastic system which receives an input vector $\mathbf{x} \in \mathbf{R}^k$ and emits an output vector $\mathbf{y} \in \mathbf{R}^l$. An input vector $\mathbf{x}$ is generated subject to a probability $q(\mathbf{x})$ and an output vector $\mathbf{y}$ is emitted subject to a conditional probability $q(\mathbf{y}|\mathbf{x})$ specified by $\mathbf{x}$. In the following discussion, we identify the target system with the conditional probability $q(\mathbf{y}|\mathbf{x})$, or a joint probability distribution $q(\mathbf{x}, \mathbf{y})$ which is a product of the given input distribution $q(\mathbf{x})$ and the conditional distribution $q(\mathbf{y}|\mathbf{x})$.

In the same way, a network is considered to have a conditional distribution $p(\mathbf{y}|\mathbf{x}, \theta)$, where $\theta \in \mathbf{R}^m$ is an m-dimensional parameter vector that specifies the network, such as a set of weights and thresholds. Thus, a group of artificial neural networks that have the same architecture can be regarded as a parametric family of conditional distributions. For example, we can think that inside of a conventional multilayered perceptron the output probability is calculated and the average is emitted as the output of the network. Hereafter, an individual network $p(\mathbf{y}|\mathbf{x}, \theta)$ shall be called a "machine," and this family of conditional distributions $\{p(\mathbf{y}|\mathbf{x}, \theta); \theta \in \mathbf{R}^m\}$ shall be called a "model."

When the target system $q(\mathbf{y}|\mathbf{x})$ belongs to the model $\{p(\mathbf{y}|\mathbf{x}, \theta); \theta \in \mathbf{R}^m\}$, that is, when there exists a $\hat{\theta}$ which satisfies

$$q(\mathbf{y}|\mathbf{x}) = p(\mathbf{y}|\mathbf{x}, \hat{\theta})$$

the system is said to be realizable and the model is said to be faithful. In the present paper, we do not assume the realizability of the system or the faithfulness of the model.

The following is a typical form of $p(\mathbf{y}|\mathbf{x}, \theta)$ for a multilayered perceptron. The network calculates a function $f(\mathbf{x}, \theta)$, where components of the parameter vector θ correspond to weights and thresholds, and then a noise term $\xi(\mathbf{x})$ is added to produce the output

$$\mathbf{y} = f(\mathbf{x}, \theta) + \xi(\mathbf{x}).$$

The noise is said to be additive when its distribution is independent of $\mathbf{x}$. In this additive noise case, the conditional distribution is given by

$$p(\mathbf{y}|\mathbf{x}, \theta) = \psi(\mathbf{y} - f(\mathbf{x}, \theta))$$

where $\psi(\xi)$ is the probability density function of the noise ξ. In the general case, the noise distribution $\psi(\xi|\mathbf{x})$ depends on $\mathbf{x}$, so that

$$p(\mathbf{y}|\mathbf{x}, \theta) = \psi(\mathbf{y} - f(\mathbf{x}, \theta)|\mathbf{x}).$$

When the network is noiseless, it is deterministic and the function $\psi(\xi|\mathbf{x})$ reduces to the delta function.

In order to evaluate the performance of the network, let us define a discrepancy function $D(q, p(\theta))$ which measures the difference between the target conditional distribution $q(\mathbf{y}|\mathbf{x})$ and the conditional distribution $p(\mathbf{y}|\mathbf{x}, \theta)$ of the machine. To this end, we first introduce a loss function $s(\mathbf{x}, \mathbf{y}, \theta)$ which measures a loss when an input $\mathbf{x}$ is processed by a machine specified by parameter θ, where $\mathbf{y}$ is the true output. In the case of a multilayered perceptron we usually take the mean square error as the loss

$$\begin{aligned} s(\mathbf{x}, \mathbf{y}, \theta) &\equiv \frac{1}{2} \int \|\mathbf{y} - \mathbf{y}'\|^2 p(\mathbf{y}'|\mathbf{x}, \theta) d\mathbf{y}' \\ &= \frac{1}{2} \int \|\mathbf{y} - f(\mathbf{x}, \theta) - \xi\|^2 \psi(\xi|\mathbf{x}) d\xi \\ &= \frac{1}{2} \{\|\mathbf{y} - f(\mathbf{x}, \theta)\|^2 + \text{ variance of } (\xi(\mathbf{x}))\}. \end{aligned}$$

In the deterministic case, the squared loss reduces to

$$s(\mathbf{x}, \mathbf{y}, \theta) \equiv \frac{1}{2} \|\mathbf{y} - f(\mathbf{x}, \theta)\|^2.$$

Another candidate is the log likelihood ratio or the log loss,

$$s(\mathbf{x}, \mathbf{y}, \theta) \equiv \log \frac{q(\mathbf{y}|\mathbf{x})}{p(\mathbf{y}|\mathbf{x}, \theta)}$$

or

$$s(\mathbf{x}, \mathbf{y}, \theta) \equiv -\log p(\mathbf{y}|\mathbf{x}, \theta).$$

We can treat many other types of loss functions [2], [3]. It is possible to add a regularization term $r(\theta)$ to the loss. The regularization term was introduced by Moody [8]. It gives a penalty to complex machines and makes the parameter vector stay in an appropriate region. And so we have $d(\mathbf{x}, \mathbf{y}, \theta) = s(\mathbf{x}, \mathbf{y}, \theta) + r(\theta)$ as a new loss function.

Definition 1: A discrepancy function or the expected loss $D(q, p(\theta))$ between two distributions, a target q and a machine $p(\theta)$, is defined by the expectation of a loss plus a regularization term

$$\begin{aligned} D(q, p(\theta)) &\equiv \int d(\mathbf{x}, \mathbf{y}, \theta) q(\mathbf{x}) q(\mathbf{y}|\mathbf{x}) d\mathbf{x} d\mathbf{y} \\ &= \int (s(\mathbf{x}, \mathbf{y}, \theta) + r(\theta)) q(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}. \end{aligned} \quad (1)$$

Authorized licensed use limited to: University of Minnesota. Downloaded on June 21,2010 at 15:16:55 UTC from IEEE Xplore. Restrictions apply.

In the simplest case, the discrepancy function is

$$D(q,p(\theta)) = \frac{1}{2}\int \|\mathbf{y}-\mathbf{y}'\|^2 p(\mathbf{y}'|\mathbf{x},\theta) q(\mathbf{y}|\mathbf{x})q(\mathbf{x})d\mathbf{y}'d\mathbf{y}d\mathbf{x} = \frac{1}{2}\int \|\mathbf{y}-f(\mathbf{x},\theta)-\xi\|^2\psi(\xi) q(\mathbf{x},\mathbf{y})d\xi d\mathbf{x}d\mathbf{y}$$

where the mean square error is taken as the loss, the noise is additive and no regularization term is added. However, our theory holds in the general case.

The optimal parameter, the one which gives the optimal machine with respect to the discrepancy function, depends on the true distribution $q(\mathbf{x},\mathbf{y})$ of the target system. However, we do not know this distribution, and instead we can use only a training set consisting of t examples generated by the target distribution. In other words, we can use only the empirical distribution constructed from the training set of t examples

$$q^*(\mathbf{x},\mathbf{y}) \equiv \frac{1}{t}\sum_{i=1}^{t}\delta(\mathbf{x}-\mathbf{x}_i,\mathbf{y}-\mathbf{y}_i) \tag{2}$$

where $\{(\mathbf{x}_i,\mathbf{y}_i); i = 1,\cdots,t\}$ is a training set. It is well known that, if t is large enough, then the empirical distribution $q^*(\mathbf{x},\mathbf{y})$ approximates the true distribution $q(\mathbf{x},\mathbf{y})$ in the weak sense, and hence it is reasonable to evaluate the network model by using $q^*(\mathbf{x},\mathbf{y})$ instead of $q(\mathbf{x},\mathbf{y})$. However, for a finite number t of examples, it is necessary to take the difference between $q(\mathbf{x},\mathbf{y})$ and $q^*(\mathbf{x},\mathbf{y})$ into account.

The difference is briefly shown as follows. Let θ_{opt} be the optimal parameter in the sense of minimizing the discrepancy function $D(q,p(\theta))$, that is

$$D(q,p(\theta_{\text{opt}})) = \min_{\theta} D(q,p(\theta)).$$

Similarly, let θ^* be the quasi-optimal parameter which minimizes the discrepancy between the empirical distribution and the machine

$$D(q^*,p(\theta^*)) = \min_{\theta} D(q^*,p(\theta)).$$

Usually θ_{opt} and θ^* are different, and it is quite important to evaluate the difference.

Before evaluating this difference, we describe the stochastic descent learning procedure for searching for the quasi-optimal parameter vector as follows.

Definition 2: In an each learning step, a new example is randomly chosen from the training set. This is called the resampling plan. The parameter vector θ_n at time n is modified according to the following rule,

$$\theta_{n+1} = \theta_n - \varepsilon_n \nabla d(\mathbf{x}_n,\mathbf{y}_n,\theta_n) \tag{3}$$

where ∇ denotes the gradient with respect to the parameter vector θ and ε_n is a positive value called a learning coefficient. Here $(\mathbf{x}_n,\mathbf{y}_n)$ is an example at time n independently chosen from the training set.

This learning rule, which is called the stochastic gradient descent method, was studied by many researchers (e.g., [2]–[4]) for multilayered perceptrons and more general models. Generally the learning coefficient ε_n may be changed depending on time n, but in the following we fix ε_n at a positive value ε ($\varepsilon_n = \varepsilon, \varepsilon > 0$). The asymptotic accuracy of this learning is discussed in the next section.

In the case of the deterministic multilayered perceptron, this method leads to the backpropagation method. For a training set of t observed examples, the discrepancy function is given by

$$D(q^*,p(\theta)) \equiv \int \frac{1}{2}\|\mathbf{y}-f(\mathbf{x},\theta)\|^2 \cdot \frac{1}{t}\sum_{i=1}^{t}\delta(\mathbf{x}-\mathbf{x}_i,\mathbf{y}-\mathbf{y}_i)d\mathbf{y}d\mathbf{x} = \frac{1}{t}\sum_{i=1}^{t}\frac{1}{2}\|\mathbf{y}_i-f(\mathbf{x}_i,\theta)\|^2 = \frac{1}{t}E(\theta)$$

where $E(\theta)$ is sometimes called the energy function. The modification rule of the parameter θ is given by

$$\theta_{n+1} = \theta_n - \varepsilon(\mathbf{y}-f(\mathbf{x},\theta_n))^T\nabla f(\mathbf{x},\theta_n)$$

and it is shown that θ_n approaches θ^* as $n \to \infty$ and $\varepsilon \to 0$.

III. Asymptotic Accuracy of Learning

In this section we consider the asymptotic properties of the estimated parameter θ_n after n modifications by using t examples repeatedly. We also study the relation among $\theta_n, \theta_{\text{opt}}$ and θ^*.

The parameter θ_n obtained at time n is a random variable which depends on the resampling plan. In other words, θ_n differs according to the order of examples which are resampled during the learning procedure. Hence θ_n is a random variable, even when the training set is fixed. The training set itself, or its empirical distribution $q^*(\mathbf{x},\mathbf{y})$, is also a random variable, because it depends on observed t examples subject to the true distribution $q(\mathbf{x},\mathbf{y})$. Let $\pi_n(\theta_n)$ be the probability distribution of θ_n. The distribution $\pi_n(\theta_n)$ converges to some probability distribution $\tilde{\pi}(\theta)$, that is

$$\tilde{\pi}(\theta) = \lim_{n\to\infty}\pi_n(\theta). \tag{4}$$

Therefore, when n is large enough, the random variable θ_n is subject to the distribution $\tilde{\pi}(\theta_n)$. In the following, E_P and V_P denote the expectation and the variance of a probability distribution $P(X)$, respectively. We now evaluate the behavior of the learned parameter θ_n when the learning time n is large. The following lemma shows how θ_n deviates from the quasi-optimal parameter θ^*. We fix the learning coefficient ε at a small positive constant and we assume that an initial value of the parameter θ before learning is taken in a neighborhood of the optimal parameter θ_{opt} by certain means.

Lemma 1: Let $\tilde{\theta}$ be the parameter after sufficient learning. Then $\tilde{\theta}$ is subject to $\tilde{\pi}$ and the expectation and the variance of $\tilde{\theta}$ are given by

$$E_{\tilde{\pi}}[\tilde{\theta}] = \theta^* + O(\varepsilon), \text{ and } V_{\tilde{\pi}}[\tilde{\theta}] = \varepsilon\Xi_{Q^*}^{-1}G^* + O(\varepsilon^2) \tag{5}$$

where

$$G^* \equiv V_{q^*}[\nabla d(\mathbf{x},\mathbf{y},\theta^*)],$$
$$Q^* \equiv E_{q^*}[\nabla\nabla d(\mathbf{x},\mathbf{y},\theta^*)]$$

Authorized licensed use limited to: University of Minnesota. Downloaded on June 21,2010 at 15:16:55 UTC from IEEE Xplore. Restrictions apply.

and Ξ_{Q^*} is a linear operator defined as

$$\Xi_{Q^*} M \equiv Q^* M + (Q^* M)^T, \quad M \in \mathbf{R}^{m \times m}.$$

The proof is given by Amari [2]. It should be noted that Q^* can be written as

$$Q^* = \nabla\nabla D(q^*, p(\theta^*)).$$

Moreover it can be shown that the distribution of $\tilde{\theta}$ approaches a normal distribution as $\varepsilon \to 0$ [14].

Lemma 2: The distribution $\tilde{\pi}(\tilde{\theta})$ approaches the normal distribution

$$N(\theta^*, \varepsilon \Xi_{Q^*}^{-1} G^*) \tag{6}$$

as $t \to \infty$ and $\varepsilon \to 0$.

A brief proof of this lemma is given in Appendix-A.

Roughly speaking, this lemma shows that the estimated parameter $\tilde{\theta}$ $(= \theta_n)$ has a Gaussian distribution whose expectation is the quasi-optimal parameter θ^* and variance is proportional to ε.

Since θ^* is different from θ_{opt}, we next consider the fluctuation of θ^* around θ_{opt} caused by substituting $q^*(\mathbf{x}, \mathbf{y})$ for $q(\mathbf{x}, \mathbf{y})$ in the training procedure. The empirical distribution $q^*(\mathbf{x}, \mathbf{y})$ is composed of t examples which are randomly and independently generated subject to $q(\mathbf{x}, \mathbf{y})$, hence the quasi-optimal parameter θ^* is a random variable which depends on the choice of the training set. Let $\pi^*(\theta^*)$ be the probability distribution of θ^*. It can easily be shown that the distribution $\pi^*(\theta^*)$ has the following properties.

Lemma 3: When t is sufficiently large, then

$$E_{\pi^*}[\theta^*] = \theta_{\text{opt}}, \text{ and } V_{\pi^*}[\theta^*] = \frac{1}{t} Q^{-1} G Q^{-1} \tag{7}$$

where

$$G \equiv V_q[\nabla d(\mathbf{x}, \mathbf{y}, \theta_{\text{opt}})], \text{ and } Q \equiv E_q[\nabla\nabla d(\mathbf{x}, \mathbf{y}, \theta_{\text{opt}})].$$

This lemma is known in statistics and a sketch of the proof is given in Appendix-B.

When the model is faithful, there are some special cases in which simple relations hold between G and Q, and between G^* and Q^*. One example is the case of a single output multilayered perceptron with unbiased additive noise. Under the mean-square error loss, if the model is faithful, then we can easily deduce the relations

$$G = \sigma^2 Q, \text{ and } G^* = \sigma^2 Q^* + O\left(\frac{1}{\sqrt{t}}\right)$$

where σ^2 is the variance of the additive noise. Another example is the case when the negative of the log likelihood is taken as the loss,

$$d(\mathbf{x}, \mathbf{y}, \theta) = -\log p(\mathbf{y}|\mathbf{x}, \theta).$$

If the model is faithful, the matrices G and Q coincide and the matrices G^* and Q^* approximately coincide, namely

$$G = Q, \text{ and } G^* = Q^* + O\left(\frac{1}{\sqrt{t}}\right).$$

IV. Network Information Criterion

Since we have studied the asymptotic behavior of the estimator $\tilde{\theta}$ obtained by resampling learning, we can apply these results to the problem of model selection for learning from given examples. Let us consider two parametric models, one denoted by $\{p_1(\mathbf{y}|\mathbf{x}, \theta_1); \theta_1 \in \mathbf{R}^{m_1}\}$, and the other by $\{p_2(\mathbf{y}|\mathbf{x}, \theta_2); \theta_2 \in \mathbf{R}^{m_2}\}, (m_1 < m_2)$. We assume that one is a submodel of the other:

$$\{p_1(\mathbf{y}|\mathbf{x}, \theta_1)\} \subset \{p_2(\mathbf{y}|\mathbf{x}, \theta_2)\}.$$

This implies that, by projecting θ_2 into $\mathbf{R}^{m_1}$ using some appropriate relation, we obtain the first submodel. In the case of multilayered perceptrons, for example, the numbers of units in the input layer and output layer are the same in two models, but the numbers of units in the hidden layers of the second model are larger than those of the first model. By putting the connection weights and thresholds of the extra units equal to 0, we obtain a smaller submodel.

When the parameters $\tilde{\theta}_1$ and $\tilde{\theta}_2$ of the two models are estimated by using the common training set, the problem is to decide which model is better. We have already defined the discrepancy function $D(q, p(\theta))$ and the learning rule for minimizing this discrepancy. Therefore, one idea is to choose the model which has a smaller discrepancy value $D(q, p_i(\tilde{\theta}_i))$. However we do not know the true system $q(\mathbf{x}, \mathbf{y})$ so that we cannot calculate $D(q, p_i(\tilde{\theta}_i))$ directly. We know only the empirical distribution $q^*(\mathbf{x}, \mathbf{y})$, and so we estimate $D(q, p_i(\tilde{\theta}_i))$ by using $D(q^*, p_i(\tilde{\theta}_i))$. Hence our purpose can be reduced to evaluating the difference between $D(q^*, p_i(\tilde{\theta}_i))$ and $D(q, p_i(\tilde{\theta}_i))$.

From lemma 1 and lemma 3 we can derive the following relation.

Theorem 1: The average discrepancy between the system $q(\mathbf{x}, \mathbf{y}) = q(\mathbf{y}|\mathbf{x})\, q(\mathbf{x})$ and the machine $p(\mathbf{y}|\mathbf{x}, \tilde{\theta})$ learned from t examples is given by

$$\langle D(q, p(\tilde{\theta})) \rangle = \langle D(q^*, p(\tilde{\theta})) \rangle + \frac{1}{t} \operatorname{tr}\left(G Q^{-1}\right) + O(t^{-\frac{3}{2}}) \tag{8}$$

where $\langle \cdot \rangle$ denotes the expectation subject to both $\tilde{\pi}(\tilde{\theta})$ and $\pi^*(\theta^*)$.

Proof is given in Appendix-C. This relation gives the following Network Information Criterion for selecting the optimal architecture of neural networks.

A. Network Information Criterion: Let $M_i = \{p_i(\mathbf{y}|\mathbf{x}, \theta_i); \theta_i \in \mathbf{R}^{m_i}\}$ be a hierarchical series of models:

$$M_1 \subset M_2 \subset M_3 \subset \ldots$$

where M_i is a submodel of M_j $(i < j)$. Let $\tilde{\theta}_i$ be the parameter of model M_i obtained by learning based on a common training set of t examples. We call

$$\text{NIC}(p_i) = D(q^*, p_i(\tilde{\theta}_i)) + \frac{1}{t} \operatorname{tr}\left(G_i^*(\tilde{\theta}_i) Q_i^*(\tilde{\theta}_i)^{-1}\right) \tag{9}$$

the Network Information Criterion, where

$$G_i^*(\tilde{\theta}_i) \equiv V_{p^*}[\nabla d(\mathbf{x}, y, \tilde{\theta}_i)],$$

and

$$Q_i^*(\tilde{\theta}_i) \equiv E_{p^*}[\nabla\nabla d(\mathbf{x}, \mathbf{y}, \tilde{\theta}_i)].$$

Authorized licensed use limited to: University of Minnesota. Downloaded on June 21,2010 at 15:16:55 UTC from IEEE Xplore. Restrictions apply.

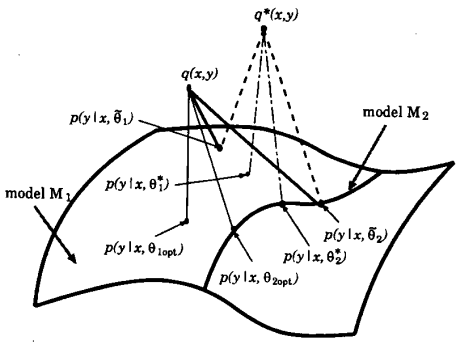

Fig. 1. The geometrical relationship between the system and network models. This figure shows the simple image of the relation among the true system, the empirical system, the model M_1 and the model M_2. Distributions $q(\mathbf{x}, \mathbf{y})$ and $q^*(\mathbf{x}, \mathbf{y})$ denote the system and empirical distributions. In this case, the smaller model is enough to approximate the true system and the parameter $\{\theta_{i\text{opt}},\ i = 1, 2\}$ represents the optimal parameter for the true distribution and $\{\theta_i^*,\ i = 1, 2\}$ represents the quasi-optimal parameters for the models $M_i = \{p(\mathbf{y}|\mathbf{x}, \theta_i, i = 1, 2\}$ under the empirical distribution and $\{\tilde{\theta}_i, i = 1, 2\}$ are obtained by learning. The thick solid lines show the discrepancy between the true distribution and those of the learned networks.

The model which minimizes NIC is optimal in the minimum averaging loss sense, that is, minimizing the expected discrepancy $\langle D(q, p_i(\tilde{\theta}_i))\rangle$.

Fig. 1 shows the geometrical relationship between the system and the models.

V. Discussion

It is important to compare the NIC with other results related to this problem. When the model is faithful and the loss is given by the negative log loss, the problem is exactly the same as selecting a statistical model for estimating the joint distribution $q(\mathbf{x}, \mathbf{y})$. In this case, the matrices G and Q coincide with the Fisher information matrix so that

$$\mathrm{tr}\left(G_i^* Q_i^{*-1}\right) = m_i + O\left(\frac{1}{\sqrt{t}}\right)$$

is the number of parameters in the model M_i. The NIC is exactly the same as the AIC except for the coefficient. Hence the NIC is a natural generalization of the AIC.

Moody [8] proposed a generalization of the AIC by introducing the effective dimension m^* of a model with additive noises. It is given by

$$m^* = \frac{\sigma^2}{t}\mathrm{tr}(TQT).$$

It is easy to show that $G = \sigma^2 Q$ and that

$$\frac{1}{t}\mathrm{tr}(TT) = Q^{-1} + O\left(\frac{1}{t}\right).$$

Therefore, the NIC reduces to Moody's in the case of additive noise. See also [9].

The NIC, as well as the AIC, is effective for model selection among a sequence of hierarchical models where one is included in another as a lower-dimensional submodel. This remark is usually not stated explicitly when discussing the AIC, or generalizations thereof (e.g., [8]). This restriction originates from the following fact.

In deriving the NIC, we have evaluated the difference between $D(q, p(\tilde{\theta}))$ and $D(q^*, p(\tilde{\theta}))$ in the sense of the ensemble average of training sets. However, when we apply the criterion in selecting a model, we need to evaluate $D(q, p(\tilde{\theta}))$ and $D(q^*, p(\tilde{\theta}))$ by using only one training set. Roughly speaking, $D(q, p(\tilde{\theta}))$ can be decomposed into following (see appendix C):

$$D(q, p(\tilde{\theta})) \simeq D(q^*, p(\tilde{\theta})) + \frac{1}{\sqrt{t}}U + \frac{1}{t}\ \mathrm{tr}\ GQ^{-1} + O_p\left(\frac{1}{t\sqrt{t}}\right)$$

where

$$U = \sqrt{t}\{D(q, p(\theta_{\text{opt}})) - D(q^*, p(\theta_{\text{opt}}))\}$$

is a random variable of order 1 with zero mean, so that $U/\sqrt{t}$ dominates the effective dimension term of order $1/t$. However, we can prove that the U is common to all the models within a hierarchical structure. Therefore, it is not effective to apply this type of criteria to non hierarchical models. This fact was pointed out by Takeuchi [15] and is known to specialists of the AIC but is still not well known by those who apply the AIC.

Recently, Hagiwara *et al.* [16] have cast doubt on validity of applying the AIC method to multilayered perceptrons. They pointed out that there are some critical values of parameters where multilayered perceptrons are reduced to smaller models. For example, when two hidden units have the same weight values from all units of the lower layer, the multilayered perceptron perform the same as the perceptron of fewer hidden units. A similar situation can occur when some weight values between hidden units and output units vanish. In these cases, matrices G and Q degenerate, and the effective dimension term can not be calculated because there does not exist Q^{-1}. But we expect that the effective dimension can be defined as a limiting value when parameters approach critical values, i.e.,

$$m^* \equiv \lim_{\theta \to \text{critical values}} GQ^{-1}$$

and that it is still valid.

VI. Conclusion

We investigated the problem of determining the optimal number of the parameters in neural networks from a statistical point of view. We have generalized Akaike's Information Criterion to be applicable to non-faithful models under a general loss function including a regularization term. The proposed NIC criterion measures the relative merits of two models which have the same structure but different number of parameters. In other words, when applied to neural networks, the criterion determines whether or not more neurons should be added to a network.

Authorized licensed use limited to: University of Minnesota. Downloaded on June 21,2010 at 15:16:55 UTC from IEEE Xplore. Restrictions apply.

APPENDIX

A. Proof of Lemma 2

Let $\phi_n(z)$ be the characteristic function of $\pi_n(\theta - \theta^*)$. According to Amari's lemma [2], we can obtain the equation

$$\varphi_{n+1}(z) = \varphi_n(z) - \varepsilon z^T Q^* \frac{\partial}{\partial z}\varphi_n(z) - \frac{\varepsilon^2}{2} z^T G^* z \varphi_n(z) + O(\varepsilon^3). \tag{10}$$

Letting $\varphi(z)$ be the characteristic function at $n \to \infty$:

$$\varphi(z) \equiv \lim_{n\to\infty} \varphi_n(z) \tag{11}$$

we have

$$z^T Q^* \frac{\partial}{\partial z}\varphi(z) = -\frac{\varepsilon}{2} z^T G^* z \varphi(z) + O(\varepsilon^2). \tag{12}$$

Suppose we wish to express $\varphi(z)$ in the following form:

$$\varphi(z) = \exp\{h_0(z) + \varepsilon h_1(z) + O(\varepsilon^2)\}. \tag{13}$$

Then we find that $h_0(z)$ and $h_1(z)$ must satisfy the following:

$$z^T Q \frac{\partial}{\partial z} h_0(z) = 0, \tag{14}$$

$$z^T Q \frac{\partial}{\partial z} h_1(z) = -\frac{\varepsilon}{2} z^T G^* z. \tag{15}$$

Solving these equations, we obtain

$$h_0(z) = 0, \tag{16}$$

$$h_1(z) = -\frac{\varepsilon}{2} z^T \Xi_{Q^*}^{-1} G^* z. \tag{17}$$

These indicate that the distribution $\pi_n(\theta - \theta^*)$ approaches a normal distribution $N(0, \varepsilon \Xi_{Q^*}^{-1} G^*)$ as $n \to \infty$ and $\varepsilon \to 0$. □

B. Proof of Lemma 3

Let us consider the expansion of $\nabla D(q^*, p(\theta^*))$ at θ_{opt}

$$\begin{aligned}\nabla D(q^*, p(\theta^*)) &= E_{q^*}[\nabla d(\mathbf{x}, \mathbf{y}, \theta^*)] \\ &= \frac{1}{t}\sum_{i=1}^{t} \nabla d(\mathbf{x}_i, \mathbf{y}_i, \theta^*) \\ &\simeq \frac{1}{t}\sum_{i=1}^{t} \nabla d(\mathbf{x}_i, \mathbf{y}_i, \theta_{\text{opt}}) \\ &\quad + \frac{1}{t}\sum_{i=1}^{t} \nabla\nabla d(\mathbf{x}_i, \mathbf{y}_i, \theta_{\text{opt}}) \cdot (\theta^* - \theta_{\text{opt}}).\end{aligned}$$

If the parameter θ^* minimizes the discrepancy function $D(q^*, p(\theta^*))$ and satisfies

$$\nabla D(q^*, p(\theta^*)) = 0,$$

then

$$\begin{aligned}\frac{1}{t}\sum_{i=1}^{t} \nabla\nabla d(\mathbf{x}_i, \mathbf{y}_i, \theta_{\text{opt}}) \cdot \sqrt{t}(\theta^* - \theta_{\text{opt}}) \\ \simeq -\frac{1}{\sqrt{t}}\sum_{i=1}^{t} \nabla d(\mathbf{x}_i, \mathbf{y}_i, \theta_{\text{opt}}).\end{aligned}$$

By the law of large numbers

$$\begin{aligned}\frac{1}{t}\sum_{i=1}^{t} \nabla\nabla d(\mathbf{x}_i, \mathbf{y}_i, \theta_{\text{opt}}) &\simeq E_q[\nabla\nabla d(\theta_{\text{opt}})] \\ &= Q(\theta_{\text{opt}}).\end{aligned}$$

From the central limit theorem, we know that $\frac{1}{\sqrt{t}} \sum_{i=1}^{t} \nabla d(\mathbf{x}, \mathbf{y}, \theta_{\text{opt}})$ has a normal distribution

$$\begin{aligned}\frac{1}{\sqrt{t}}\sum_{i=1}^{t} \nabla d(\mathbf{x}, \mathbf{y}, \theta_{\text{opt}}) \\ \sim N(\sqrt{t} E_q[\nabla d(\mathbf{x}, \mathbf{y}, \theta_{\text{opt}})], V_q[\nabla d(\mathbf{x}, \mathbf{y}, \theta_{\text{opt}})]) \\ = N(0, G)\end{aligned}$$

where the equation follows because

$$E_q[\nabla d(\mathbf{x}, \mathbf{y}, \theta_{\text{opt}})] = \nabla D(q, p(\theta_{\text{opt}})) = 0,$$

and

$$V_q[\nabla d(\mathbf{x}, \mathbf{y}, \theta_{\text{opt}})] = G.$$

Therefore

$$(\theta^* - \theta_{\text{opt}}) \sim N\left(0, \frac{1}{t} Q^{-1} G Q^{-1}\right).$$

□

C. Proof of Theorem 1

The expansion of $D(q, p(\tilde{\theta}))$ at θ_{opt} is given by

$$\begin{aligned}D(q, p(\tilde{\theta})) &\simeq D(q, p(\theta_{\text{opt}})) \\ &\quad + \nabla D(q, p(\theta_{\text{opt}}))(\tilde{\theta} - \theta_{\text{opt}}) \\ &\quad + \frac{1}{2}(\tilde{\theta} - \theta_{\text{opt}})^T \nabla \\ &\quad \nabla D(q, p(\theta_{\text{opt}}))(\tilde{\theta} - \theta_{\text{opt}}) + \ldots\end{aligned}$$

where T denotes the transpose of a vector. Since

$$\nabla D(q, p(\theta_{\text{opt}})) = 0$$

Authorized licensed use limited to: University of Minnesota. Downloaded on June 21,2010 at 15:16:55 UTC from IEEE Xplore. Restrictions apply.

holds, the second term vanishes. By neglecting the higher order terms, the first term can be transformed as follows

$$\begin{aligned}
D(q,p(\theta_{\rm opt})) &= D(q,p(\theta_{\rm opt})) - D(q^*,p(\theta_{\rm opt})) \\
&\quad + D(q^*,p(\theta_{\rm opt})) - D(q^*,p(\theta^*)) \\
&\quad + D(q^*,p(\theta^*)) - D(q^*,p(\tilde{\theta})) \\
&\quad + D(q^*,p(\tilde{\theta})) \\
&\simeq \{D(q,p(\theta_{\rm opt})) - D(q^*,p(\theta_{\rm opt}))\} \\
&\quad + D(q^*,p(\tilde{\theta})) + \frac{1}{2}(\theta_{\rm opt}-\theta^*)^T\nabla \\
&\quad \nabla D(q^*,p(\theta^*))(\theta_{\rm opt}-\theta^*) \\
&\quad - \frac{1}{2}(\tilde{\theta}-\theta^*)^T\nabla\nabla D(q^*,p(\theta^*))(\tilde{\theta}-\theta^*) \\
&= \frac{1}{\sqrt{t}}U + D(q^*,p(\tilde{\theta})) \\
&\quad + \frac{1}{2}\,\mathrm{tr}[Q^*\{(\theta_{\rm opt}-\theta^*)(\theta_{\rm opt}-\theta^*)^T \\
&\quad - (\tilde{\theta}-\theta^*)(\tilde{\theta}-\theta^*)^T\}] \\
&\simeq \frac{1}{\sqrt{t}}U + D(q^*,p(\tilde{\theta})) \\
&\quad + \frac{1}{2}\,\mathrm{tr}[Q\{(\theta_{\rm opt}-\theta^*)(\theta_{\rm opt}-\theta^*)^T \\
&\quad - (\tilde{\theta}-\theta^*)(\tilde{\theta}-\theta^*)^T\}]
\end{aligned}$$

where

$$U = \sqrt{t}\{D(q,p(\theta_{\rm opt})) - D(q^*,p(\theta_{\rm opt}))\}$$

and

$$\begin{aligned}
E_{\pi^*}[U] &= 0, \\
V_{\pi^*}[U] &= O(1).
\end{aligned}$$

The third term can be rewritten as

$$\begin{aligned}
&\frac{1}{2}(\tilde{\theta}-\theta_{\rm opt})^T\nabla\nabla D(q,p(\theta_{\rm opt}))(\tilde{\theta}-\theta_{\rm opt}) \\
&\quad = \frac{1}{2}\,\mathrm{tr}[Q\{(\tilde{\theta}-\theta^*)-(\theta_{\rm opt}-\theta^*)\} \\
&\qquad \{(\tilde{\theta}-\theta^*)-(\theta_{\rm opt}-\theta^*)\}^T] \\
&\quad = \frac{1}{2}\,\mathrm{tr}[Q\{(\tilde{\theta}-\theta^*)(\tilde{\theta}-\theta^*)^T \\
&\qquad + (\theta_{\rm opt}-\theta^*)(\theta_{\rm opt}-\theta^*)^T \\
&\qquad + (\tilde{\theta}-\theta^*)(\theta_{\rm opt}-\theta^*)^T \\
&\qquad + (\theta_{\rm opt}-\theta^*)(\tilde{\theta}-\theta^*)^T\}].
\end{aligned}$$

Averaging each term subject to $\tilde{\pi}(\tilde{\theta})$ and $\pi^*(\theta^*)$ and using

$$\begin{aligned}
\langle(\tilde{\theta}-\theta^*)(\tilde{\theta}-\theta^*)^T\rangle &= \frac{\varepsilon}{2}Q^{*-1}G^*, \\
\langle(\theta_{\rm opt}-\theta^*)(\theta_{\rm opt}-\theta^*)^T\rangle &= \frac{1}{t}Q^{-1}GQ^{-1}, \\
\langle(\tilde{\theta}-\theta^*)(\theta_{\rm opt}-\theta^*)^T\rangle &= 0
\end{aligned}$$

we get the required relation. □

The value U reflects the basic structure of the model. For example, the values for a multilayered perceptron and a radial basis function network differ, but the value for two multilayered perceptrons is equivalent if one model includes the other. If the smaller model $\{p_1(\mathbf{y}|\mathbf{x},\theta_1)\}$ approximates the true system sufficiently, the U_1 and the U_2 for models $\{p_1(\mathbf{y}|\mathbf{x},\theta_1)\}$ and $\{p_2(\mathbf{y}|\mathbf{x},\theta_2)\}$ respectively are approximately equivalent. That is

$$U_1 \simeq U_2.$$

In the special case where the condition

$$p_1(\mathbf{y}|\mathbf{x},\theta_{1\rm opt}) = p_2(\mathbf{y}|\mathbf{x},\theta_{2\rm opt})$$

holds, U_1 and U_2 are exactly equivalent

$$U_1 = U_2.$$

On the other hand, when the smaller model cannot approximate the true system sufficiently, the discrepancy $D(q,p_1(\tilde{\theta}_1))$ is much larger than the discrepancy $D(q,p_2(\tilde{\theta}_2))$, the difference between the discrepancy functions is dominant, and so we do not have to consider U_i in this case. Thus U can be ignored in evaluating the performance of models which have the same structure but differ in the number of parameters.

REFERENCES

[1] B. Widrow, *A Statistical Theory of Adaptation*. Pergamon Press, 1963.

[2] S. Amari, "Theory of adaptive pattern classifiers," *IEEE Trans. Elect. Comp.*, vol. 16, pp. 299–307, 1967.

[3] H. White, "Learning in artificial neural networks: A statistical perspective," *Neural Computation*, vol. 1, pp. 425–464, 1989.

[4] D. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, D . Rumelhart, J. L. McClelland, and the PDP Research Group, ed. Cambridge, MA: The MIT Press, 1986, ch. 8, pp. 318–362.

[5] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Appl. Comp.*, vol. AC-19, pp. 716–723, 1974.

[6] J. Rissanen, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, pp. 1080–1100, 1986.

[7] D. B. Forgel, "An information criterion for optimal neural network selection," *IEEE Trans. Neural Net.*, vol. 2, pp. 490–497, 1991.

[8] J. E. Moody, "The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems," in *Advances in Neural Information Processing Systems 4*, J. E. Moody, S. J. Hanson, and R. P. Lippmann, ed. San Mateo, CA: Morgan Kaufmann, 1992.

[9] G. Wahba, "Three topics in ill-posed problems," in *Inverse and Ill-Posed Problems*, M. Engl and G. Groetsch, Ed. London, U.K.: Academic Press, 1987, pp. 48–48.

[10] Y. Wada and M. Kawato, "Estimation of generalization capability by combination of new information criterion and cross validation," *Trans. IEICE*, vol. J74-D-II, no. 7, pp. 955–965, July 1 991, in Japanese.

[11] N. Murata, S. Yoshizawa, and S. Amari, "A criterion for determining the number of parameters in an artificial neural network model," in *Artificial Neural Networks*, T. Kohonen, *et al.*, ed. Amsterdam: The Netherlands: Elsevier, 1991, pp. 9–14.

[12] S. Amari and N. Murata, "Statistical theory of learning curves under entropic loss criterion," *Neural Computation*, vol. 5, no. 1, pp. 140–153, 1993.

[13] N. Murata, S. Yoshizawa, and S. Amari, "Learning curves, model selection and complexity of neural networks," in *Advances in Neural Information Processing Systems 5*, S. José Hanson, J. D. Cowan, and C. Lee Giles, ed. San Mateo, CA: Morgan Kaufmann, 1993, pp. 607–614.

[14] N. Murata, "A statistical study on asymptotic theory of learning," Ph. D. dissertation, Dept. of Math. Eng. and Info. Physics, Univ. of Tokyo, Japan, Mar. 1992, in Japanese.

[15] K. Takeuchi, "Distribution of information statistic and validity criterion of models," *Math. Sci.*, no. 153, pp. 12–18, 1976, in Japanese.

[16] K. Hagiwara, N. Toda, and S. Usui, "Nonuniqueness of connecting weights and aic in multilayered neural networks," *Trans. IEICE*, vol. J76-D-II, no. 9, pp. 2058–2065, Sept. 1993, in Japanese.

Authorized licensed use limited to: University of Minnesota. Downloaded on June 21,2010 at 15:16:55 UTC from IEEE Xplore. Restrictions apply.

Noboru Murata received the B. Eng., M. Eng., and Dr. Eng. degrees in mathematical engineering and instrumentation physics from the University of Tokyo, Japan in 1987, 1989, and 1992, respectively.

Dr. Murata has been a Research Associate at the University of Tokyo since 1992. His research interests include the theoretical aspects of neural networks, focusing on the dynamics and statistical properties of learning.

Shuji Yoshizawa (S'62–M'62) received the B. Eng., M. Eng., and Dr. Eng. degrees in mathematical engineering and instrumentation physics in 1962, 1964, and 1971, respectively, from the University of Tokyo, Japan.

Dr. Yoshizawa was engaged with Central Laboratory of Hitachi Ltd. from 1964 to 1967. In 1972, he moved to the Department of Mathematical Engineering and Instrumentation Physics, the University of Tokyo. He was Associate Professor with the same department from 1974 to 1992. He has been Professor with the Department of Machano-Informatics, the University of Tokyo since 1992. His research interest is in nonlinear dynamics of distributed active media, neural networks and bio-cybernetics.

Dr. Yoshizawa is the Vice-President of the Japan Neural Network Society.

Shun-ichi Amari (M'88– SM'82–F'94) graduated from the University of Tokyo in 1958 majoring in mathematical engineering, and received the Dr. Eng. degree from the University of Tokyo in 1963.

He was a Professor at Hyushu University and is now a Professor at the University of Tokyo. He has been engaged in research in many areas of mathematical engineering and applied mathematics, such as topological network theory, differential geometry of continuum mechanics, pattern recognition, and information sciences. In particular, he has devoted himself to mathematical foundations of neural network theory. Another main subject of his research is the information geometry that he himself proposed.

Dr. Amari is a founding governing board member of International Neural Network Society and a member of the editorial or advisory editorial boards of more than 15 international journals. He was given the IEEE Neural Networks Pioneer Award in 1992 .

Authorized licensed use limited to: University of Minnesota. Downloaded on June 21,2010 at 15:16:55 UTC from IEEE Xplore. Restrictions apply.