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

Noboru Murata, Shuji Yoshizawa, Shun-ichi Amari,
University of Tokyo*

June 22, 1992.

Abstract

The problem of model selection or determination of the number of hidden units is elucidated by the statistical approach, 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 or determining the number of hidden units based on a given training set.

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

1 Introduction

In engineering fields, one of the most important applications of artificial neural networks is modeling a system with an unknown input-output relation. Given a fixed architecture of networks, parameters are usually modified by the stochastic gradient descent method which eventually minimizes a loss function. Learning is carried out based on a training set which consists of a number of examples observed from the actual system (Widraw [1], Amari [2], White [3], etc.). For instance, the back-propagation method is used for multi-layered networks [4].

An important but difficult problem is to determine the number of parameters or the number of hidden units needed to model the system by using only input-output examples. This 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. For instance, in the case of a multi-layered network, when we add some units in hidden layers, the network can emit more precise outputs for the training inputs, but it may emit worse outputs for inexperienced inputs. Performance for generating accurate outputs for the training inputs competes against that for predicting appropriate outputs for unknown inputs. Such discrepancy increases with the number of parameters to be estimated. 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, such as the AIC [5] and the MDL [6]. There are some researches intending to apply these principles (e.g. Forgel [7], Moody [8], Wahba [9], Wada and Kawato [10]). The present paper is a detailed version of a short note by Murata, Yoshizawa and Amari [11], giving a most general solution to this problem.

The present paper treats a hierarchy of stochastic neural networks of feed forward type. A network is regarded as a machine, which produces an output y, when x is an input, based on a conditional probability $p(y|x,\theta)$, where θ is the parameter vector specifying the network. The problem is to find the optimal model and the optimal parameter value to approximate a given true conditional distribution from which a set of training examples is chosen.

This is regarded as a statistical problem. However, we need to generalize the AIC approach in two points: The first is that the true distribution $q(y|x)$ is not necessarily included in any of the model $\{p(y|x,\theta)\}$. The true distribution is said to be an unrealizable rule and the model is said to be unfaithful in such a case. The second is that we use a general loss function includ-

ing a regularization term, including the negative of likelihood as a special case which leads to the maximum likelihood estimator. The regularization term was introduced in the loss by Moody [8]. This gives the smoothness condition and fits well the neural information processing purpose.

In section 2, we define a general loss function by which we evaluate how the behavior given by the true input-output distribution is different from that given by a network model. We then formulate a leaning procedure based on a repeated resampling plan from a fixed training set of examples sampled from the true distribution. Once the training set is fixed, we can only use the empirical distribution of the training set instead of the true unknown distribution.

Section 3 is devoted to the evaluation of the network parameter after learning. Double evaluations are necessary, one showing how the learned parameter approximates the quasi optimal parameter which is optimal to the training data or the empirical distribution of the training set, and the other showing how the quasi optimal parameter for the empirical distribution approximates the true optimal parameter for the unknown distribution. This elucidates the relation between the training error and the generalization error in terms of the complexity of a network and the number of training examples.

Based on this relation, we propose in section 4 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 that the criterion is applicable only for comparing a hierarchical set of models in which one model is included in another as a submodel. This originates from a stochastic fluctuation in the training error, which cannot be evaluated by its ensemble average. However, this fluctuation term is common to all the members of a hierarchical model, so that they are canceled out. This is a restriction when the generalization error is evaluated in terms of the training error, but is not still well recognized.

2 A Discrepancy Function and the Learning Rule

Let us consider a stochastic system which receives an input vector $x \in \boldsymbol{R}^{n_{in}}$ and emits an output vector $y \in \boldsymbol{R}^{n_{out}}$. An input vector x is generated subject to a probability $q(x)$ and an output vector y is emitted subject to a conditional probability $q(y|x)$ specified by x. In the following discussion we

identify the system with the conditional probability $q(y|x)$, or a joint probability distribution $q(x,y)$ which is a product of the given input distribution $q(x)$ and the conditional distribution $q(y|x)$.

An artificial layered stochastic neural networks are regarded as a parametric family of conditional distributions. A network has a conditional distribution $p(y|x,\theta)$, where $\theta \in \boldsymbol{R}^m$ is an m-dimensional parameter that specifies the model network. When the true distribution $q(y|x)$ belongs to the model, that is, when there is a $\hat{\theta}$ for which

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

the distribution is said to be realizable and the model is said to be faithful. The present paper does not assume the realizability or the faithfulness of the model.

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

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

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

$$p(y|x,\theta) = a(y - f(x,\theta)),$$

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

$$p(y|x,\theta) = a(y - f(x,\theta)|x).$$

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

In order to evaluate the performance of the network, we define a discrepancy function which measures the difference $D(q,p(\theta))$ between the true conditional distribution $q(y|x)$ and the conditional distribution $p(y|x,\theta)$ of the model. To this end, we first introduce a loss function $k(x,y,\theta)$ which is a loss when an input x is processed by a network specified by parameter θ, where y is the true output. In the case of a multi-layered network we usually take the mean square error as the loss,

$$\begin{aligned} k(x,y,\theta) &= \frac{1}{2}\int \|y - y'\|^2 p(y'|x,\theta)dy' \\ &= \frac{1}{2}\int \|y - f(x,\theta) - \xi\|^2 a(\xi|x)d\xi. \end{aligned}$$

The squared loss reduces to

$$k(x,y,\theta) = \frac{1}{2}\|y - f(x,\theta)\|^2$$

in the deterministic case. Another candidate is the log likelihood ratio or the log loss,

$$k(x,y,\theta) = \log \frac{q(y|x)}{p(y|x,\theta)},$$

or

$$k(x,y,\theta) = -\log p(y|x,\theta).$$

We can treat many other types of loss functions (Amari [2], White [3]). Following Moody [8], it is possible to add a regularization term $s(\theta)$ to the loss, which gives a penalty to a complex network. In this case, we have $d(x,y,\theta) = k(x,y,\theta) + s(\theta)$ as a new loss function.

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

$$\begin{aligned} D(q,p(\theta)) &= \int d(x,y,\theta)q(x)q(y|x)dxdy \\ &= \int (k(x,y,\theta) + s(\theta))\, q(x,y)dxdy. \end{aligned} \tag{1}$$

In the simplest case, when the mean square error is taken as the loss, the noise is additive and no regularization term is added, the discrepancy function is

$$\begin{aligned} D(q,p(\theta)) &= \frac{1}{2}\int \|y - y'\|^2 p(y'|x,\theta)q(y|x)q(x)dy'dydx \\ &= \frac{1}{2}\int \|y - f(x,\theta) - \xi\|^2 a(\xi)q(x,y)d\xi dxdy. \end{aligned}$$

However, our theory holds in the general case.

The optimal parameter, that is the optimal model with respect to the discrepancy function or the expected loss, depends on the unknown true distribution $q(x,y)$. However, we do not know it and instead we can use only a training set consisting of t examples generated from the true distribution. In other words, we can use only the empirical distribution constructed from the training set of t examples,

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

where (x_i, y_i) are the observed input-output pairs. It is well known that, if t is large enough, then $q^*(x,y)$ approximates $q(x,y)$ in the weak sense, and it is reasonable to evaluate the the network model by using $q^*(x,y)$ instead of $q(x,y)$. However, for a finite number t of examples, it is necessary to take the difference between $q(x,y)$ and $q^*(x,y)$ into account.

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

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

Similarly, let θ^* be the quasi optimal parameter when the underlying distribution is the empirical distribution constructed from t examples,

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

Usually θ_o and θ^* are different, and it is important to evaluate their difference.

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

Definition 2 *In each learning step, a new example is chosen from the training set, that is, subject to the empirical distribution. This is called the resampling plan. The parameter θ_n at time n is modified according to the following rule,*

$$\theta_{n+1} = \theta_n - \varepsilon_n \nabla d(x_n, y_n, \theta_n), \tag{3}$$

where ∇ denotes the gradient with respect to the parameter θ, ε_n is a positive value called a learning coefficient. Here (x_n, y_n) is an example at time n independently chosen subject to $q^*(x,y)$.

This learning rule called the stochastic gradient descent method was studied by many researchers (e.g., Amari [2], Rumelhart et al. [4], White [3]) for multi-layered neural networks and more general models. In the following we fix the learning coefficient ε_n at a positive value ε. The asymptotic accuracy of learning is discussed in the next section.

In the case of the deterministic multi-layered network, this method leads to the back-propagation method. In the case where a training set consists of t stored examples, the discrepancy function is

$$D(q^*, p(\theta)) = \int \frac{1}{2}\|y - f(x,\theta)\|^2 \frac{1}{t}\sum_{i=1}^{t} \delta(x - x_i, y - y_i) dy dx$$

$$= \frac{1}{t}\sum_{i=1}^{t}\frac{1}{2}\|y_i - f(x_i,\theta)\|^2$$
$$= \frac{1}{t}E(\theta),$$

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

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

and θ_n approaches θ^* as $n \to \infty$ and $\varepsilon \to 0$.

3 Asymptotic Accuracy of Learning

In this section we consider the asymptotic properties of the estimated parameter θ_n at the n-th modification by using t stored examples repeatedly. We also study the relation among θ_n, θ_o and θ^*.

The parameter θ_n obtained at the n-th modification depends on the resampling plan, that is, the order in which the examples are selected from the training set during the learning procedure. Hence θ_n is a random variable, even when the training set is fixed. The training set itself, or its sample empirical distribution $q^*(x,y)$, is also a random variable, because it depends on randomly chosen t examples. Let $r_n(\theta_n)$ be the probability distribution of θ_n. The distribution $r_n(\theta_n)$ converges to some probability distribution $\tilde{r}(\theta)$ in law, that is,

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

Therefore, when n is large enough, the random variable θ_n is subject to the distribution $\tilde{r}(\theta_n)$. In the following, E_P and V_P denote, respectively, the expectation and the variance with respect to a probability distribution $P(X)$. We now evaluate the behavior of θ_n when n is large. The following lemma shows how θ_n deviates from θ^*, the optimal parameter for a given training set. We fix the learning coefficient ε at a small positive constant and an initial value of the parameter θ before learning is taken in a neighborhood of the optimal parameter θ_o. Let $\tilde{\theta}$ be the parameter after enough times of learning. Then $\tilde{\theta}$ is subject to $\tilde{r}$ and the expectation and the variance of $\tilde{\theta}$ are given by the following lemma.

Lemma 1

$$E_{\tilde{r}}[\tilde{\theta}] = \theta^* + O(\varepsilon), \tag{5}$$

and

$$V_{\tilde{r}}[\tilde{\theta}] = \frac{\varepsilon}{2} Q^{*-1} G^* + O(\varepsilon^2), \tag{6}$$

where

$$G^* = V_{q^*}[\nabla d(x, y, \theta^*)],$$

and

$$Q^* = E_{q^*}[\nabla\nabla d(x, y, \theta^*)]$$

The proof is given by Amari [2]. Moreover it is shown that the distribution of $\tilde{\theta}$ approaches a normal distribution as $\varepsilon \to 0$ [12]. It should be noted that Q^* is written as

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

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 whose variance is proportional to ε.

Since θ^* is different from θ_o, we next consider the fluctuation of θ^* around θ_o caused by a random choice of the training set, that is, by substituting $q^*(x, y)$ for $q(x, y)$ in the training procedure. The empirical distribution $q^*(x, y)$ is composed of t examples which are randomly and independently generated subject to $q(x, y)$, and the quasi optimal parameter θ^* depends on this random empirical distribution. Let $r^*(\theta)$ be the probability distribution of θ^*. The distribution $r^*(\theta)$ has the following properties.

Lemma 2 *When t is large enough, then*

$$E_{r^*}[\theta^*] = \theta_o, \tag{7}$$

and

$$V_{r^*}[\theta^*] = \frac{1}{t} Q^{-1} G Q^{-1}, \tag{8}$$

where

$$G = V_q[\nabla d(x, y, \theta_o)],$$

and

$$Q = E_q[\nabla\nabla d(x, y, \theta_o)].$$

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

In the case of a single output multi-layered network with an unbiased additive noise whose variance is σ^2, if the model is faithful, then we can easily deduce the relations

$$G = \sigma^2 Q$$

and

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

under the mean square error loss. In the special case when the negative of log likelihood is taken as the loss and the model is faithful, that is

$$d(x, y, \theta) = -\log p(y|x, \theta),$$

the matrices G and Q coincide and the matrices G^* and Q^* approximately coincide, namely

$$G = Q$$

and

$$G^* = Q^* + O(\frac{1}{\sqrt{t}}).$$

4 A Network Information Criterion

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

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

This implies that, by restricting some components of θ_2 to fixed values or within fixed relations, we obtain the first submodel. In the case of multi-layered networks, the numbers of units in the input layer and output layer are the same in two models, but the number of units in the hidden layers are larger in the second 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 data set, the problem is to decide which model is better. We have already defined the discrepancy function $D(q, p(\theta))$ and defined the learning rule as minimizing this discrepancy value. Therefore, one idea is to select the model which has a smaller discrepancy $D(q, p_i(\tilde{\theta}_i)$ value. However we do not know the true system $q(x, y)$ so that we cannot calculate $D(q, p_i(\tilde{\theta}_i))$. We can only know the empirical distribution $q^*(x, y)$, and so we estimate $D(q, p_i(\tilde{\theta}_i))$ by using $D(q^*, p_i(\tilde{\theta}_i))$. When we use $D(q^*, p_i(\tilde{\theta}_i))$ instead of $D(q, p_i(\tilde{\theta}_i))$, we need to evaluate their difference. This is because $D(q, p_i(\tilde{\theta}_i))$ is tend to be overestimated by using $D(q^*, p_i(\tilde{\theta}_i))$ when the number n_i of parameter in the model is larger. From lemma 1 and lemma 2 we can derive the following relation.

Theorem 1 *The average discrepancy between the true system $q(x, y) = q(y|x)q(x)$ and the trained model $p(y|x, \tilde{\theta})$ by using a set of t examples is given by*

$$\langle D(q, p(\tilde{\theta}))\rangle = \langle D(q^*, p(\tilde{\theta}))\rangle + \frac{1}{t}\ \mathrm{tr}\left(GQ^{-1}\right) + O(t^{-\frac{3}{2}}), \tag{9}$$

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

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

Network Information Criterion

Let $M_i = \{p_i(y|x, \theta^i)\}$ be a hierarchical series of models

$$M_1 \subset M_2 \subset M_3 \subset \cdots$$

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

$$\mathrm{NIC}(p_i) = D(q^*, p_i(\tilde{\theta}_i)) + \frac{1}{t}\ \mathrm{tr}\left(G_i^*(\tilde{\theta}_i)Q_i^*(\tilde{\theta}_i)^{-1}\right) \tag{10}$$

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

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

and

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

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

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 $q(x, y)$. In this case

$$G = Q$$

is the Fisher information matrix so that

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

is the number of parameters in the model M_i. Hence the NIC is exactly the same as the AIC divided by $2t$. 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

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

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

The NIC as well as the AIC is effective for the 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 in the AIC case nor other generalizations of the AIC (e.g., in Moody [8]). This restriction originates from the following fact.

We have evaluated the difference in the ensemble average between $D(q, p(\tilde{\theta}))$ and $D(q^*, p(\tilde{\theta}))$ in deriving the NIC. However, when we apply the criterion in selecting a model, the training set is fixed, and we need to evaluate $D(q, p(\tilde{\theta}))$ and $D(q^*, p(\tilde{\theta}))$. We have the stochastic expansion (see the appendix) of D as

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

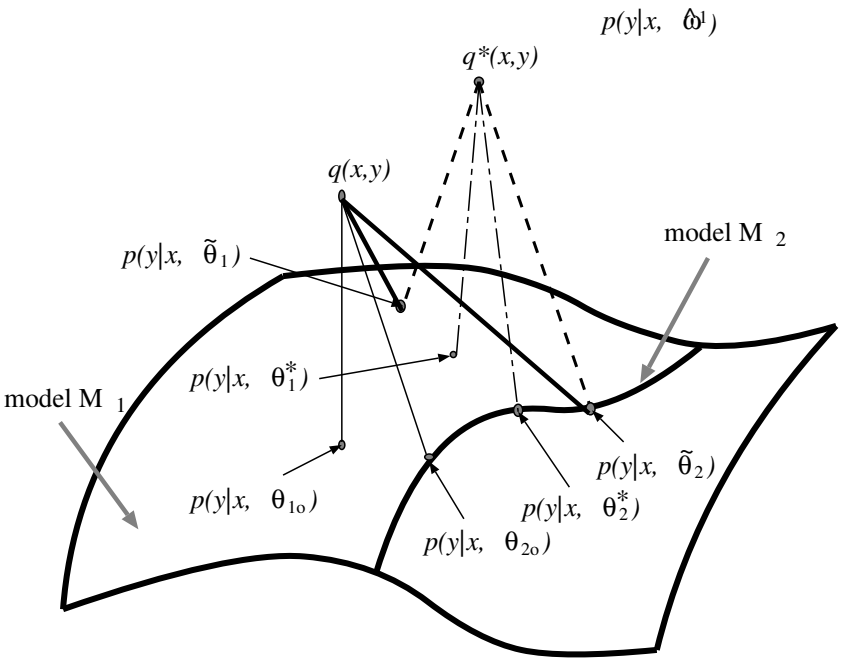

Figure 1: The geometrical relationship between the system and network models. This figure shows the simple image of the relation among the true system, the empirical system, model M_1 and model M_2. Distributions $q(x,y)$ and $q^*(x,y)$ denote the system and the empirical distribution. In this case, the smaller model is enough to approximate the true system, and the parameter $\{\theta_{io},\ i=1,2\}$ represents the optimal parameter for the true distribution and $\{\theta_i^*,\ i=1,2\}$ represent the quasi optimal parameters of models $M_i = \{p(y|x,\theta_i),\ i=1,2\}$ for 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 learned networks.

Here

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

is a random variable of order 1, so that it dominates the effective dimension term which is of order $1/t$. However, we can prove that the U is common except for the term of order $t^{-3/2}$ 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 (1976) [13] and is known to specialists of the AIC but is still not well known to those who apply the AIC.

5 Conclusion

We investigated the problem of determining the optimal number of the parameters in neural networks from the statistical point of view. We have generalized Akaike's information criterion to be applicable to non-faithful models (or the approximation of an unrealizable input-output relation) 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 a different number of parameters. In other words, the criterion determines whether more neurons should be added to a network or not.

This study is partly supported by grant-in-aid #03251102 and #03234109 from Ministry of Education of Japan.

A Proof of Lemma 2

Let us consider the expansion of $\nabla D(p', p(\theta'_o))$ at θ_o,

$$\begin{aligned}
\nabla D(q^*, p(\theta^*)) &= E_{q^*}[\nabla d(x, y, \theta^*)] \\
&= \frac{1}{t}\sum_{i=1}^{t} \nabla d(x_i, y_i, \theta^*) \\
&\simeq \frac{1}{t}\sum_{i=1}^{t} \nabla d(x_i, y_i, \theta_o) + \frac{1}{t}\sum_{i=1}^{t} \nabla\nabla d(x_i, y_i, \theta_o)\cdot(\theta^* - \theta_o).
\end{aligned}$$

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

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

then

$$\frac{1}{t}\sum_{i=1}^{t}\nabla\nabla d(x_i,y_i,\theta_o)\cdot\sqrt{t}(\theta^*-\theta_o)\simeq-\frac{1}{\sqrt{t}}\sum_{i=1}^{t}\nabla d(x_i,y_i,\theta_o).$$

Because of the law of large numbers,

$$\begin{aligned}\frac{1}{t}\sum_{i=1}^{t}\nabla\nabla d(x_i,y_i,\theta_o) &\simeq E_q[\nabla\nabla d(\theta_o)]\\ &= Q(\theta_o).\end{aligned}$$

From the central limit theorem, $\frac{1}{\sqrt{t}}\sum_{i=1}^{t}\nabla d(x,y,\theta_o)$ is normally distributed,

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

because

$$E_q[\nabla d(x,y,\theta_o)] = \nabla D(q,p(\theta_o))=0,$$

and

$$V_q[\nabla d(x,y,\theta_o)] = G.$$

Therefore

$$(\theta^*-\theta_o)\sim N(0,\frac{1}{t}Q^{-1}GQ^{-1}).$$

∎

B Proof of Theorem 1

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

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

where T denotes the transpose of a vector. The second term vanishes since

$$\nabla D(q,p(\theta_o))=0.$$

The first term is transformed as follows, by neglecting the higher order terms,

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

where

$$
U = \sqrt{t} \left\{D(q, p(\theta_o)) - D(q^*, p(\theta_o))\right\},
$$

and

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

The third term can be rewritten as

$$
\begin{aligned}
& \frac{1}{2}(\tilde{\theta} - \theta_o)^T \nabla\nabla D(q, p(\theta_o))(\tilde{\theta} - \theta_o) \\
&= \frac{1}{2} \operatorname{tr}\left[Q \{(\tilde{\theta} - \theta^*) - (\theta_o - \theta^*)\}\{(\tilde{\theta} - \theta^*) - (\theta_o - \theta^*)\}^T\right] \\
&= \frac{1}{2} \operatorname{tr}\left[Q \{(\tilde{\theta} - \theta^*)(\tilde{\theta} - \theta^*)^T + (\theta_o - \theta^*)(\theta_o - \theta^*)^T \right. \\
&\qquad \left. + (\tilde{\theta} - \theta^*)(\theta_o - \theta^*)^T + (\theta_o - \theta^*)(\tilde{\theta} - \theta^*)^T\}\right].
\end{aligned}
$$

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

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

we can get the required relation. ∎

The value U reflects the basic structure of a model. For example, the values for a multi-layered network and a competitive network differ, but the values for two multi-layered networks are equivalent if one model includes the other. If the smaller model $p_1(y|x,\theta_1)$ approximates the true system adequately, U_1 and U_2 for models $p_1(y|x,\theta_1)$ and $p_2(y|x,\theta_2)$ respectively are approximately equivalent, that is

$$U_1 \simeq U_2.$$

In special case, if following condition

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

holds, then U_1 and U_2 are exactly equivalent, that is

$$U_1 = U_2.$$

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

References

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

[2] S.-i. Amari, "Theory of adaptive pattern classifiers," *IEEE Trans. EC*, vol. 16, no. 3, 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, eds.), ch. 8, pp. 318–362, The MIT Press, 1986.

[5] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. AC*, vol. 19, no. 6, 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. NN*, vol. 2, no. 5, 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, eds.), San Mateo, CA: Morgan Kaufmann Publishers, 1992.

[9] G. Wahba, "Three topics in ill-posed problems," in *Inverse and Ill-Posed Problems* (M. Engl and G. Groetsch, eds.), pp. 48–48, Academic Press, 1987.

[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, pp. 955–965, July 1991. in Japanese.

[11] N. Murata, S. Yoshizawa, and S.-i. Amari, "A criterion for determining the number of parameters in an artificial neural network model," in *Artificial Neural Networks* (T. Kohonen *et al.*, eds.), (Holland), pp. 9–14, ICANN, Elsevier Science Publishers, July 1991.

[12] N. Murata, *A Statistical Asymptotic Study on Learning.* PhD thesis, University of Tokyo, Department of Mathematical Engineering and Information Physics, Faculty of Engineering, Mar. 1992.

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