Res. Rep. Fac. Eng. Mie Univ., Vol. 20, pp. 63-81(1995)

Original Paper

On the Statistical Properties of Function Representation with Discrete Variable Basis under Squared Error Loss

Katsuyuki HAGIWARA, Naohiro TODA † and Shiro USUI †
Department of Electrical and Electronic Engineering

(Received September 18, 1995)

Abstract

One of the most important property of 3-layered neural networks is the selectability of the basis functions. In this paper, to focus on the selectability in the context of the regression model, we restricted our attention to function representations in which the basis functions are modified according to the associated discrete parameters. For such function representations, we derived lower and upper bounds for the expectations of the empirical loss and the expected loss with respect to the distribution of the set of samples by taking the squared error as a loss function, provided that the given set of samples is a Gaussian noise sequence and the basis functions satisfy the orthonormality condition. Based on these results, we showed that the statistical properties of the function representations with adaptive basis functions is different from conventional function representations with fixed basis functions.

Key words

discrete variable basis type function representation, regression model, squared error, empirical loss, expected loss

1. Introduction

In wide-ranging applications of layered neural networks, a basic and an important task of networks is to identify an unknown input-output relation of a target system with stochastic nature. When there are few a priori knowledges about the system, the relation should be identified based on a finite set of pairs of input-output samples observed according to the system. Usually, the given samples are utilized to estimate connection weights of a network with a predetermined complexity, so as to decrease the average squared error on the samples by an algorithm such as the well-known backpropagation[13]. This procedure is intended to minimize

† Department of Information and Computer Sciences, Toyohashi University of Technology

the empirical loss by taking the squared error as a loss function.. In the task, although we expect the estimated network performs with low error on unseen samples ; i.e. the network generalize to data outside the given samples, the above procedure does not necessarily give us the network with a better generalization performance. Indeed, the generalization ability depends on the complexity of a network as follows. If the network is too complex, one obtains the estimated network with the outputs which are very close to the given output samples at the corresponding input samples. The input-output relation of such network, however, may deviate from the invariant input-output relation of the target system. Conversely, when the network is too simple, the error on unseen samples also will be large because the small size network cannot represent the underlying input-output relation enough. This is often called bias/variance dilemma[7]. Under the minimization of the empirical loss, thus, it is necessary to determine the optimal complexity of the network. In statistics, the issue mentioned above is known as the problem of model selection. There have been proposed criteria for the model selection since Akaike's pioneer attempt of AIC (Akaike Information Criterion)[1][2]. Also, in neural network field, there are some attempts to apply traditional criteria such as AIC directly to neural networks[9][6] and to develop considerable criteria[10][11].

The generalization performance of a network can be naturally measured by the expected loss, which corresponds to the error on unseen samples. The basic idea of the model selection given by AIC is to estimate the expected loss based on the empirical loss. Therefore, once the relation between the expected loss and the empirical loss is derived, one can construct a model selection criterion. Generally, the estimation of the expected loss based on the empirical loss yields the estimation bias. The AIC is the unbiased estimator of the expected loss based on the empirical loss, in which the loss function is taken to be the negative log-likelihood. In the derivation of the AIC, the bias is approximately given as the number of modifiable parameters in an assumed statistical model, which can be regarded as penalizing the model complexity. In the case of taking the squared error as a loss function, Barron[4] has derived a criterion called PSE (Predicted Squared Error) for linear regression models. The penalty term of PSE is given by $2 \cdot s \cdot \sigma_*^2/N$, where s, σ_*^2 and N denote the number of coefficients in the model, the variance of additive noise and the number of samples respectively. Recently, based on the above viewpoint of model selection, Murata et. al[11] have proposed the most extended criterion called NIC (Network Information Criterion) under a general condition. Under the assumption of a certain smoothness condition on loss functions, the NIC has derived as the unbiased estimator of the expected loss, in which the bias term is given by tr GQ^{-1}/N, where G is the variance-covariance matrix of the first partial derivatives of a loss function with respect to the parameters and Q is the expectation of the second partial derivative matrix of a loss. If a loss is taken to be the negative log-likelihood and the true distribution which defines a sample generating mechanism is in an assumed model family, the NIC reduces to the AIC. However, the authors[8] have pointed out that there exists the case where AIC cannot be derived for 3-layered neural networks under the latter condition because the matrix G and Q degenerate and Q^{-1} does not exist in such case. This is due to the nonuniqueness of connection weights originated from the nonlinear parameters of a network; i.e. connection weights form input units to hidden units and thresholds at hidden units. Therefore, although the model selection problem looked to be given a solution by the NIC, there remains a vagueness in the problem.

On the other hand, in the framework of the function approximation, the capability of 3-layered neural networks to avoid the curse of dimensionality has been revealed by taking

account of the selectability of basis functions in function representations given by 3-layered neural networks[5]. The selectability is achieved by the nonlinear parameters associated with the basis functions and is important because they essentially characterize the nonlinearity of the parameters of 3-layered neural networks, which is hard to analyze directly. Actually, the hardness of the model selection problem of 3-layered neural networks is caused by the existence of the nonlinear parameters as mentioned above. Thus, the viewpoint of the selectability of the basis functions may enable us to develop the theory of neural network regression.

In this paper, to find a solution to the model selection problem of 3-layered neural networks, we focus on the selectability of basis functions in the context of regression models. Here, to simplify the problem, we deal with the function representations, in which the parameters associated with the basis functions are restricted to a finite set of allowed values ; i.e. the function representations whose basis functions are modified according to the associated discrete parameters. We first give the definition of such function representations and conventional function representations with fixed basis functions. And we define the expectations of the minimum of the empirical loss and the expected loss with respect to the distribution of samples by taking the squared error as a loss. Also we briefly summarize the statistical properties of regression models using the function representations with fixed basis functions. Next, we give the least square parameter estimation algorithm for regression models using the function representations with adaptive basis functions and derive lower and upper bounds of the above expectations for such models. Lastly, based on these results, we compare the statistical properties between the function representations with fixed basis functions and the function representations with adaptive basis functions.

2. Function Representation

We deal with a function representation (FR) whose output for an input $x \in \boldsymbol{R}^d$ is given by

$$g(x;\omega_s) = \sum_{j=1}^{s} c_j \phi(x;b_j), \tag{1}$$

where $\omega_s = (\boldsymbol{c}_s, \boldsymbol{b}_s) \in \Omega_s \subseteq \boldsymbol{R}^{s(t+1)}$ is a parameter vector of the FR and $\boldsymbol{c}_s = (c_1,\ldots,c_s)$, $c_j \in \boldsymbol{R}$, $\boldsymbol{b}_s = (b_1,\ldots,b_s)$, $b_j \in \boldsymbol{B}_j \subseteq \boldsymbol{R}^t$; $j = 1,\ldots,s$. Ω_s is a parameter space of the FR. $\phi(x;b_j)$ denotes a basis function and a set of basis functions $\Phi_s(x) = (\phi(x;b_1),\ldots,\phi(x;b_s))$ is called a basis, where s is the number of basis functions . We call $\boldsymbol{c}_s$ a coefficient vector and $\boldsymbol{b}_s$ a basis parameter vector. A family of FRs is denoted by $\mathcal{G}_s = \{g(\,\cdot\,;\omega_s)\ ;\ \omega_s \in \Omega_s\}$. The FR given by (1) is a linear combination of basis functions in a basis $\Phi_s(x)$. We call a set $\boldsymbol{B}_j$ a basis parameter space, whose elements are t dimensional vectors. If $\boldsymbol{B}_j = \boldsymbol{R}^t$ then t can be regarded as the number of basis parameters. FRs are characterized by its basis parameter space.

For a FR, if $|\boldsymbol{B}_j| = 1$; $j = 1,\ldots,s$ and the basis $\Phi_s(x)$ is linearly independent, the FR is said to be fixed basis type, where $|\ \cdot\ |$ denotes the number of elements in a set. In other words, when we set $\boldsymbol{B}_j = \left\{b'_j\right\}$; $j = 1,\ldots,s$, by which a basis $\Phi_s(x) = (\phi(x;b'_1),\ldots,\phi(x;b'_s))$ is determined, the FR is fixed basis type if the basis is linearly independent. The parameters of a fixed basis type FR are only the coefficients ; i.e. $\omega_s = \boldsymbol{c}_s$. On the other hand, if $|\boldsymbol{B}_j| > 1$ then the FR is said to be variable basis type. The parameters of a variable basis type FR are pairs of the coefficients and the basis parameters ; i.e. $\omega_s = (\boldsymbol{c}_s, \boldsymbol{b}_s)$. The basis of a variable

basis type FR can be modified according to the basis parameter vector $\boldsymbol{b}_s$. When we choose $b_j = b'_j$, $j = 1, \ldots, \overline{s}$ so as to hold linearly independency of $\phi(x; b_j)$, $j = 1, \ldots, \overline{s}$ and set $\boldsymbol{B}_j = \{b'_1, \ldots, b'_{\overline{s}}\}$; $j = 1, \ldots, s$ for an integer $\overline{s}$, the FR is said to be discrete variable basis type. That is, the basis parameter space of discrete variable basis type FR is a finite set while that of a 3-layered neural network is uncountable, which can be described as $\boldsymbol{B}_j = \boldsymbol{R}^{d+1}$ for all j by regarding the input connection weights and the thresholds as the basis parameters. Thus, discrete variable basis type FRs give a natural restriction of 3-layered neural networks.

Example Let us set $d = 1$ and $\Phi_s(x) = (x^{b_1}, \ldots, x^{b_s})$. If $\boldsymbol{B}_j = \{b'_j\}$, $b'_j = j - 1$; $j = 1, \ldots, s$ then we have a basis $\Phi_s(x) = (x^0, \ldots, x^{s-1})$ and the basis functions are linearly independent. Thus, the FR is fixed basis type. On the other hand, if we set $\boldsymbol{B}_j = \{0, \ldots, \overline{s}\}$, $s < \overline{s} < \infty$ for all j then the FR is discrete variable type. Moreover, in this example, the family of the discrete variable basis type FR includes that of fixed basis type FR. □

For a discrete variable type FR, a basis is determined by choosing a basis parameter vector from $\boldsymbol{B}_1 \times \cdots \times \boldsymbol{B}_s$. Because $\boldsymbol{B}_1 = \cdots = \boldsymbol{B}_s$ and $|\boldsymbol{B}_1| = \overline{s}$, the choice corresponds to the sampling of s elements from a set of size $\overline{s}$ with replacement. Hence, the number of ways of the choice is equal to $\overline{s}^s$. In the ways of the choice of a basis parameter vector, there exist the cases that a corresponding basis functions will be linearly dependent ; e.g. in the above example, $\boldsymbol{b}_s = (1, \ldots, 1)$, and so on. Indeed, the number of ways of the choice in which a corresponding basis functions will be linearly independent is equal to the number of ways of the sampling without replacement. Furthermore, the ordering of elements in a basis parameter vector is not important because the outputs of the FRs with the same elements of the basis parameter vector except their ordering are exactly the same. As we shall see later in the section 6., we deal with the ways of a choice of the basis parameter vector, in which the linear independency of the corresponding basis functions holds and the ordering of the elements in the basis parameter vector is not taken into consideration. The corresponding number of ways of the choice is equal to ${}_{\overline{s}}C_s$.

Throughout this paper, we assume that the number of input points N and input values $\{x_i; x_i \in \boldsymbol{R}^d, 1 \leq i \leq N\}$ are predetermined. We denote a input vector by $\boldsymbol{x}_N = (x_1, \ldots, x_N)$[1]. Under this assumption, a FR can be described by a vector formulation as follows.

$$\boldsymbol{g}_s = (g(x_1; \boldsymbol{\omega}_s), \ldots, g(x_N; \boldsymbol{\omega}_s))^T = \boldsymbol{\Phi}_s \boldsymbol{c}_s \tag{2}$$

$$\boldsymbol{\Phi}_s = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_s) \tag{3}$$

$$\boldsymbol{\phi}_j = (\phi(x_1; b_j), \phi(x_2; b_j), \ldots, \phi(x_N; b_j))^T \ ; \ j = 1, \ldots, s, \tag{4}$$

where ^{T} denotes the transpose of a matrix[2]. Since each $\boldsymbol{\phi}_j$ is N dimensional vector, $\operatorname{rank} \boldsymbol{\Phi}_s \leq N$ holds. And, it is possible to choose $b_j = b'_j$; $j = 1, \ldots, N$ so as to hold linearly independency of $\boldsymbol{\phi}_j$; $j = 1, \ldots, N$. By the choice, $N \times N$ matrix $\boldsymbol{\Phi}_N$ will be nonsingular. In this case, we reform the definition of the discrete variable basis type as follows. When we choose $b_j = b'_j$; $j = 1, \ldots, N$ so as to hold linear independency of $\boldsymbol{\phi}_j$; $j = 1, \ldots, N$ and set $\boldsymbol{B}_j = \{b'_1, \ldots, b'_N\}$; $j = 1, \ldots, s$, the FR is said to be discrete variable basis type. That is, we set $\overline{s} = N$ in the previous definition.

[1]The linear independency of basis functions depends on not only basis functions but also an input vector. Throughout this paper, the input vector is assumed to be predetermined properly.

[2]Although we previously defined a coefficient vector as a row vector, we deal with the coefficient vector as a column vector in the vector formulation of the FR. But there may be no confusion with this notation.

For basis functions, the condition

$$\sum_{i=1}^{N} \phi(x_i; b_m)\phi(x_i; b_n) = \begin{cases} 1 & (m = n) \\ 0 & (m \neq n) \end{cases} \tag{5}$$

or, similarly,

$$\boldsymbol{\phi}_m^T \boldsymbol{\phi}_n = \begin{cases} 1 & (m = n) \\ 0 & (m \neq n) \end{cases} \tag{6}$$

is called orthonormality. If $\boldsymbol{\phi}_j$; $j = 1, \ldots, s$ satisfy the orthonormality condition for a fixed type FR then the FR is said to be orthonormal fixed basis type. On the other hand, when we choose the $b_j = b'_j$; $j = 1, \ldots, N$ so as to hold orthonormality of $\boldsymbol{\phi}_j$; $j = 1, \ldots, N$ and set $\boldsymbol{B}_j = \{b'_1, \ldots, b'_N\}$; $j = 1, \ldots, s$, the FR is said to be orthonormal discrete variable basis type. Namely, when we set $\boldsymbol{\phi}_j = (\phi(x_1; b'_j), \ldots, \phi(x_N; b'_j))$; $j = 1, \ldots, N$, ${\boldsymbol{\phi}'_m}^T \boldsymbol{\phi}'_n = 1$ $(m = n)$; $= 0$ $(m \neq n)$ for any m, n $(1 \leq m, n, \leq N)$, the FR is orthonormal variable basis type.

3. Regression Model

We assume that the outputs of a target system for inputs $\boldsymbol{x}_i \in \boldsymbol{R}^d$; $i = 1, 2, \ldots$ are generated by adding noise to outputs of a function h ; i.e.

$$y_i = h(x_i) + \varepsilon_i; \ i = 1, 2, \ldots \quad , \tag{7}$$

where ε_is are independent random variables with a common normal distribution $N(0, \sigma_*^2)$. The h is called a true function, which determines an invariant input-output relation of the system. Now we assume that we observed a set of N pairs of input-output samples, which is denoted by $(\boldsymbol{x}, \boldsymbol{y})_N = \{(x_i, y_i); 1 \leq i \leq N\}$. Here, we denote an output vector by $\boldsymbol{y}_N = (y_1, \ldots, y_N)^T$ whose ordering corresponds to the ordering of $\boldsymbol{x}_N$. Under the setting,

$$\boldsymbol{y}_N = \boldsymbol{h}_N + \boldsymbol{\varepsilon}_N \tag{8}$$

and the distribution of the random vector $\boldsymbol{y}_N$ is given by

$$\boldsymbol{y}_N \sim N(\boldsymbol{h}_N, \sigma_*^2 \boldsymbol{I}_N), \tag{9}$$

where $\boldsymbol{h}_N = (h(x_1), \ldots, h(x_N))$ and $\boldsymbol{\varepsilon}_N = (\varepsilon_1, \ldots, \varepsilon_N)$.

Now we describe a generating rule of samples $(\boldsymbol{x}, \boldsymbol{y})_N$ as follows.

$$y_i = g(x_i; \boldsymbol{\omega}_s) + e_i \ ; \ i = 1, 2, \ldots, N, \tag{10}$$

where e_i ; $i = 1, \ldots, N$ are independent random variables with a common normal distribution $N(0, \sigma_s^2)$. The above representation of the sample generating rule is known as a regression model. The parameter vector of the model is denoted by $\boldsymbol{\theta}_k = (\boldsymbol{\omega}_s, \sigma_s^2)$, which consists of the parameter vector of the FR and the unknown variance of the normal distribution. We assume that $\mathcal{G}_{s_1} \subset \mathcal{G}_{s_2}$ for $s_1 < s_2$ and there exists s^* such that $h \in \mathcal{G}_{s^*}$ holds ; i.e. there exists $\boldsymbol{\omega}_s^* \in \boldsymbol{\Omega}_s$ for $s \geq s^*$, which satisfies

$$h(x_i) = g(x_i; \boldsymbol{\omega}_s^*) \ ; \ i = 1, \ldots, N. \tag{11}$$

$\boldsymbol{\omega}_s^*$ and $g(\ \cdot\ ; \boldsymbol{\omega}_s^*)$ will be refered to as the true parameter vector of the FR and the true FR respectively.

We denote a fixed basis type FR and an orthonormal fixed basis type FR by FFR and OFFR respectively. The regression model using FFR and OFFR are denoted by $\mathcal{M}^F(s)$ and $\mathcal{M}^{OF}(s)$ respectively. Since the parameters of both FRs are only coefficients, $\mathcal{M}^F(s)$ and $\mathcal{M}^{OF}(s)$ are linear regression models[3]. On the other hand, We denote a discrete variable basis type FR and an orthonormal discrete variable basis type FR by DVFR and ODVFR respectively. The regression model with DVFR and ODVFR are denoted by $\mathcal{M}^{DV}(s)$ and $\mathcal{M}^{ODV}(s)$ respectively. Since both FRs have basis parameters, $\mathcal{M}^{DV}(s)$ and $\mathcal{M}^{ODV}(s)$ are nonlinear regression models[3].

4. Loss Function

We employ the squared error as a loss function ;

$$r(x_i, y_i; \omega_s) = \{y_i - g(x_i; \omega_s)\}^2 ; \ i = 1, \ldots, N, \tag{12}$$

and define the empirical loss by

$$r_{\text{emp}}(\omega_s) = \frac{1}{N}(\boldsymbol{y}_N - \boldsymbol{g}_s)^T(\boldsymbol{y}_N - \boldsymbol{g}_s) = \frac{1}{N}\sum_{i=1}^{N} r(x_i, y_i; \omega_s). \tag{13}$$

The estimator of the parameter vector based on the set of samples $(\boldsymbol{x}, \boldsymbol{y})_N$ is denoted by $\widehat{\boldsymbol{\theta}}_k = (\widehat{\omega}_s, \widehat{\sigma}_s^2)$, which is defined as the minimizing parameter vector of the empirical loss, that is,

$$\widehat{\omega}_s = \underset{\omega_s \in \boldsymbol{\Omega}_s}{\operatorname{argmin}}\, r_{\text{emp}}(\omega_s) \tag{14}$$

$$\widehat{\sigma}_s^2 = r_{\text{emp}}(\widehat{\omega}_s). \tag{15}$$

The estimator $\widehat{\boldsymbol{\theta}}_k$ is the least square estimator. We define the expectation of the minimum of the empirical loss with respect to the distribution of the samples by

$$R_{\text{emp}}(s, N) \equiv E_{\boldsymbol{y}_N}\left[r_{\text{emp}}(\widehat{\omega}_s)\right] = E_{\boldsymbol{y}_N}\left[\widehat{\sigma}_s^2\right], \tag{16}$$

where $E_{\boldsymbol{y}_N}[\,\cdot\,]$ denotes the expectation with respect to the distribution of $\boldsymbol{y}_N$; i.e. the joint distribution of $y_1, \ldots, y_N$. On the other hand, we define the expected loss by

$$R(\omega_s, N) \equiv E_{\boldsymbol{z}_N}\left[\frac{1}{N}(\boldsymbol{z}_N - \boldsymbol{g}_s)^T(\boldsymbol{z}_N - \boldsymbol{g}_s)\right] = E_{\boldsymbol{z}_N}\left[\frac{1}{N}\sum_{i=1}^{N} r(x_i, z_i; \omega_s)\right], \tag{17}$$

where $\boldsymbol{z}_N = (z_1, \ldots, z_N)$ denotes an independent and identically distributed random vector with $\boldsymbol{y}_N$ and $E_{\boldsymbol{z}_N}[\,\cdot\,]$ denotes the expectation with respect to the distribution of $\boldsymbol{z}_N$. The expected loss at the least square estimator, $R(\widehat{\omega}_s, N)$, is also a random variable since it is a function of $\boldsymbol{y}_N$. Therefore, by taking the expectation of $R(\widehat{\omega}_s, N)$ with respect to the distribution of $\boldsymbol{y}_N$, we define the expectation of the expected loss at the least square estimator by

$$R(s, N) \equiv E_{\boldsymbol{y}_N}\left[R(\widehat{\omega}_s, N)\right]. \tag{18}$$

As mentioned in the introduction, the model selection criterion can be constructed as the unbiased estimator of $R(s, N)$ based on $r_{\text{emp}}(\widehat{\omega}_s)$. Generally, the estimate of $R(s, N)$ based on $r_{\text{emp}}(\widehat{\omega}_s)$ is biased. The expectation of the bias is given by the difference between the expectation of $R(\widehat{\omega}_s, N)$ and that of $r_{\text{emp}}(\widehat{\omega}_s)$, which are $R(s, N)$ and $R_{\text{emp}}(s, N)$ respectively. For example,

in case of linear regression models, it can be shown that

$$R(s,N) = R_{\text{emp}}(s,N) + 2\sigma_*^2 \frac{s}{N}. \tag{19}$$

By using this relation, the model selection criterion PSE[4] is given by

$$\text{PSE}(s) = r_{\text{emp}}(\widehat{\omega}_s) + 2\widehat{\sigma}^2 \frac{s}{N}, \tag{20}$$

where $\widehat{\sigma}^2$ is a proper estimate of σ_*^2. Thus, it is essential to derive $R(s,N)$ and $R_{\text{emp}}(s,N)$ for an assumed model or a FR. In the following, $R_{\text{emp}}(s,N)$ and $R(s,N)$ are simply called the expectation of empirical loss and the expectation of expected loss.

5. The Statistical Properties of $\mathcal{M}^F(s)$ and $\mathcal{M}^{OF}(s)$

In this section, as a preliminary, we briefly summarize the statistical properties of the $\mathcal{M}^F(s)$ and $\mathcal{M}^{OF}(s)$, which are linear regression models. Moreover, $\mathcal{M}^{OF}(s)$ is a special case of $\mathcal{M}^F(s)$. In the following, we assume that $h \in \mathcal{G}_s$.

By solving the normal equation :

$$(\boldsymbol{\Phi}_s^T \boldsymbol{\Phi}_s)\widehat{\boldsymbol{c}}_s = \boldsymbol{\Phi}_s^T \boldsymbol{y}_N, \tag{21}$$

the least square estimator of coefficient vector of $\mathcal{M}^F(s)$ is given by

$$\widehat{\boldsymbol{c}}_s = (\boldsymbol{\Phi}_s^T \boldsymbol{\Phi}_s)^{-1} \boldsymbol{\Phi}_s^T \boldsymbol{y}_N. \tag{22}$$

Under the assumption of normality of the noise term in (7), the distribution of the least square estimator $\widehat{\boldsymbol{c}}_s$ is given by

$$\widehat{\boldsymbol{c}}_s \sim N(\boldsymbol{c}_s^*, \sigma_*^2 (\boldsymbol{\Phi}_s^T \boldsymbol{\Phi}_s)^{-1}). \tag{23}$$

In case of $\mathcal{M}^{OF}(s)$, $\boldsymbol{\Phi}_s^T \boldsymbol{\Phi}_s = \boldsymbol{I}_s$ holds by the orthonormality of the column vectors $\boldsymbol{\phi}_j$; $j = 1, \ldots, s$ of $\boldsymbol{\Phi}_s$, where $\boldsymbol{I}_s$ denotes the $s \times s$ unit matrix. Thus, the least square estimator of the coefficient vector of $\mathcal{M}^{OF}(s)$ is given by

$$\widehat{\boldsymbol{c}}_s = \boldsymbol{\Phi}_s^T \boldsymbol{y}_N. \tag{24}$$

The elements of (24) is written by

$$\widehat{c}_j = \sum_{i=1}^{N} y_i \phi(x_i; b_j) \ ; \ j = 1, \ldots, s. \tag{25}$$

From the relation (23), the distribution of $\widehat{\boldsymbol{c}}_s$ of $\mathcal{M}^{OF}(s)$ is given by

$$\widehat{\boldsymbol{c}}_s \sim N(\boldsymbol{c}_s^*, \sigma_*^2 \boldsymbol{I}_s), \tag{26}$$

where $\boldsymbol{\omega}_s^* = \boldsymbol{c}_s^* = (c_1^*, \ldots, c_s^*)$ in (11). Hence, by the property of multivariate normal distributions, each $\widehat{c}_j$ has the normal distribution $N(c_j^*, \sigma_*^2)$ and those are independent.

From this fact, $\left(\widehat{c}_j - c_j^*\right)^2 / \sigma_*^2$; $j = 1, \ldots, s$ are independent random variables with a common χ^2 distribution with 1 degree of freedom. Thus, we obtain

$$\sum_{j=1}^{s} \left(\widehat{c}_j - c_j^*\right)^2 / \sigma_*^2 \sim \chi_s^2, \tag{27}$$

where the χ_s^2 denotes the χ^2 distribution with s degrees of freedom. We denote $R_{\rm emp}(s,N)$ and $R(s,N)$ of $\mathcal{M}^{OF}(s)$ by $R_{\rm emp}^{OF}(s,N)$ and $R^{OF}(s,N)$ respectively.

On the other hand, by using (24), the empirical loss of $\mathcal{M}^{OF}(s)$ given by (13) is easily reformulated as follows.

$$r_{\rm emp}(\widehat{\boldsymbol{c}}_s) = \frac{1}{N}\boldsymbol{y}_N^T\boldsymbol{y}_N - \frac{1}{N}\widehat{\boldsymbol{c}}_s^T\widehat{\boldsymbol{c}}_s. \tag{28}$$

Then, taking the expectation of (28) with respect to the distribution of $\boldsymbol{y}_N$, we easily obtain

$$R_{\rm emp}^{OF}(s,N) = \sigma_*^2 + \frac{1}{N}\boldsymbol{h}_N^T\boldsymbol{h}_N - \frac{1}{N}E_{\boldsymbol{y}_N}\left[\widehat{\boldsymbol{c}}_s^T\widehat{\boldsymbol{c}}_s\right]. \tag{29}$$

On the other hand, from (17), we obtain

$$R(\widehat{\boldsymbol{\omega}}_s,N) = \sigma_*^2 + \frac{1}{N}\boldsymbol{h}_N^T\boldsymbol{h}_N - \frac{2}{N}\boldsymbol{h}_N^T\boldsymbol{\Phi}_s\widehat{\boldsymbol{c}}_s + \frac{1}{N}\widehat{\boldsymbol{c}}_s^T\widehat{\boldsymbol{c}}_s. \tag{30}$$

Hence,

$$R^{OF}(s,N) = \sigma_*^2 + \frac{1}{N}\boldsymbol{h}_N^T\boldsymbol{h}_N - \frac{2}{N}\boldsymbol{h}_N^T E_{\boldsymbol{y}_N}\left[\boldsymbol{\Phi}_s\widehat{\boldsymbol{c}}_s\right] + \frac{1}{N}E_{\boldsymbol{y}_N}\left[\widehat{\boldsymbol{c}}_s^T\widehat{\boldsymbol{c}}_s\right]. \tag{31}$$

In the remainder of the paper, we assume that $\boldsymbol{h}_N = \boldsymbol{0}$; $h(x_i) = 0$; $i = 1,\ldots,N$. In this case, AIC can not be derived for 3-layered neural networks[8]. Therefore the assumed case is important in the analysis of FRs with adaptive basis while it is the simplest case. Under the assumption, $\boldsymbol{y}_N = \boldsymbol{\varepsilon}_N$, thus, the distribution of $\boldsymbol{y}_N$ is given by

$$\boldsymbol{y}_N \sim N(\boldsymbol{0}, \sigma_*^2\boldsymbol{I}_N), \tag{32}$$

which means that y_i ; $i = 1,\ldots,N$ are independent random variables with a common normal distribution $N(0,\sigma_*^2)$. We will use the term "the set of samples is a Gaussian noise sequence" to refer this assumption. Furthermore, $\boldsymbol{c}_s^* = \boldsymbol{0}$ because the $\boldsymbol{\phi}_j$; $j = 1,\ldots,s$ are linearly independent. Thus, the following relation holds by (26).

$$\widehat{\boldsymbol{c}}_s \sim N(\boldsymbol{0}, \sigma_*^2\boldsymbol{I}_N), \tag{33}$$

which means that $\widehat{c}_j$; $j = 1,\ldots,s$ are independent random variables with a common normal distribution $N(0,\sigma_*^2)$. Under the above assumption, we can easily obtain the following theorem by putting $\boldsymbol{h}_N = \boldsymbol{0}$ into (29) and (31) and by taking into account (33).

Theorem 1 *Let us define*

$$C^{OF}(s,N) \equiv E_{\boldsymbol{y}_N}\left[\frac{1}{\sigma_*^2}\sum_{j=1}^{s}\widehat{c}_j^2\right]. \tag{34}$$

Then, $C^{OF}(s,N)$ is given by

$$C^{OF}(s,N) = s \tag{35}$$

and the following equations hold.

$$R_{\rm emp}^{OF}(s,N) = \sigma_*^2 - \frac{\sigma_*^2}{N}C^{OF}(s,N) \tag{36}$$

$$R^{OF}(s,N) = \sigma_*^2 + \frac{\sigma_*^2}{N}C^{OF}(s,N). \tag{37}$$

It can be shown that this theorem holds for $\mathcal{M}^{F}(s)$, thus for $\mathcal{M}^{OF}(s)$, without assuming

that $\boldsymbol{h}_N = \mathbf{o}$ or the normality of the noise term. We however omit the detail here.

6. Parameter Estimation Algorithm of $\mathcal{M}^{DV}(s)$

As mentioned in the section 2., in case of $\mathcal{M}^{DV}(s)$, there exist the situations that the basis functions of a FR will be linearly dependent according to a choice of the basis parameter vector. The error given by such FR must be larger than the error given by the FR with the linearly independent s basis functions. Thus, in the choice of a basis parameter vector, it is sufficient to take into account the basis parameter vector by which the basis functions will be linearly independent. The parameter estimation algorithm of $\mathcal{M}^{DV}(s)$ is as follows.

The parameter estimation algorithm of $\mathcal{M}^{DV}(s)$

step 1 We determine a basis parameter vector $\boldsymbol{b}_{s,m} = (b_{1,m}, \ldots, b_{s,m})$, $b_{j,m} \in \boldsymbol{B}_j$ and obtain the least square estimate of the coefficient vector $\widehat{\boldsymbol{c}}_{s,m} = (\widehat{c}_{1,m}, \ldots, \widehat{c}_{s,m})$ at the basis parameter vector $\boldsymbol{b}_{s,m}$ by solving the normal equation (21). Remark that, for any j_1, j_2, if $j_1 \neq j_2$ then $b_{j_1,m} \neq b_{j_2,m}$ and, for any m_1, m_2, if $m_1 \neq m_2$ then there exists at most one element of $\boldsymbol{b}_{s,m_1}$ which is different form every element of $\boldsymbol{b}_{s,m_2}$. Hence, $1 \leq m \leq {}_N C_s \equiv M_s$. Set $\widehat{\boldsymbol{\theta}}_{k,m} = (\widehat{\omega}_{s,m}, \widehat{\sigma}^2_{s,m})$, $\widehat{\omega}_{s,m} = (\widehat{\boldsymbol{c}}_{s,m}, \boldsymbol{b}_{s,m})$. Repeat the above procedure for all m.

step 2 Define m^* so as to satisfy

$$m^* = \underset{1 \leq m \leq M_s}{\operatorname{argmin}} \widehat{\sigma}^2_{s,m}. \tag{38}$$

Then, the least square estimate of the parameter vector of $\mathcal{M}^{DV}(s)$ $\widehat{\boldsymbol{\theta}}_k = (\widehat{\omega}_s, \widehat{\sigma}^2_s)$ is given by putting $\widehat{\omega}_s = (\widehat{\boldsymbol{c}}_s, \widehat{\boldsymbol{b}}_s)$, $\widehat{\boldsymbol{c}}_s = \widehat{\boldsymbol{c}}_{s,m^*}$, $\widehat{\boldsymbol{b}}_s = \boldsymbol{b}_{s,m^*}$, $\widehat{\sigma}^2_s = \widehat{\sigma}^2_{s,m^*}$. □

Obviously, this algorithm gives us the global least square estimate of the parameter vector of $\mathcal{M}^{DV}(s)$ in the parameter space $\boldsymbol{\Omega}_s$. The step 1 corresponds to the estimation of a coefficient vector, which is the same procedure for $\mathcal{M}^{F}(s)$, and the step 2 corresponds to the estimation of a basis parameter ; i.e. the selection of the basis which gives the least square error for a given set of samples.

7. Parameter Estimation Algorithm of $\mathcal{M}^{ODV}(s)$

In case of applying the parameter estimation algorithm of $\mathcal{M}^{DV}(s)$ to $\mathcal{M}^{ODV}(s)$, by recalling (28), the minimum of the empirical loss at m is given by

$$r_{\text{emp}}(\widehat{\boldsymbol{c}}_{s,m}) = \frac{1}{N}\boldsymbol{y}_N^T\boldsymbol{y}_N - \frac{1}{N}\widehat{\boldsymbol{c}}_{s,m}^T\widehat{\boldsymbol{c}}_{s,m} = \frac{1}{N}\sum_{i=1}^{N} y_i^2 - \frac{1}{N}\sum_{j=1}^{s}\widehat{c}^2_{j,m}. \tag{39}$$

Because the first term does not depend on the FR, we determine m^* so as to maximize $\sum_{j=1}^{s}\widehat{c}^2_{j,m}$ in the step 2. That is, (38) in the step 2 can be rewritten as

$$m^* = \underset{1 \leq m \leq M_s}{\operatorname{argmax}} \sum_{j=1}^{s}\widehat{c}^2_{j,m}. \tag{40}$$

Moreover, by using the orthonormality of the basis, the parameter estimation algorithm of $\mathcal{M}^{ODV}(s)$ can be simplified as follows. Because $b_j \in \boldsymbol{B}_j = \{b'_1, \ldots, b'_N\}$; $j = 1, \ldots, s$, every

element of $\boldsymbol{b}_{s,m} = (b_{1,m}, \ldots, b_{s,m})$ is in $\{b'_1, \ldots, b'_N\}$. On the other hand, by setting $s = N$ and $\boldsymbol{b}_N = (b'_1, \ldots, b'_N)$ in (24), we can obtain the least square estimate of the coefficient vector $\boldsymbol{c}_N = (c_1, \ldots, c_N)$. The estimated coefficient vector is denoted by $\tilde{\boldsymbol{c}}_s = (\tilde{c}_1, \ldots, \tilde{c}_N)$. Because $\tilde{c}_j$ depends only on $\phi(x; b'_j)$ as seen in (24), the estimate of a coefficient vector $\hat{\boldsymbol{c}}_{s,m}$ at a basis parameter vector $\boldsymbol{b}_{s,m} = (b_{1,m}, \ldots, b_{s,m}) = (b'_{j_1}, \ldots, b'_{j_s})$ is given by $\hat{\boldsymbol{c}}_{s,m} = (\hat{c}_{1,m}, \ldots, \hat{c}_{s,m}) = (\tilde{c}_{j_1}, \ldots, \tilde{c}_{j_s})$. That is, for any m, each element of $\hat{\boldsymbol{c}}_{s,m}$ is sure to be in $\{\tilde{c}_1, \ldots, \tilde{c}_N\}$. Hence, we denote by $\{\tilde{c}_{l_1}, \ldots, \tilde{c}_{l_N}\}$ the estimated coefficients $\{\tilde{c}_1, \ldots, \tilde{c}_N\}$ rearranged in increasing order of $\{\tilde{c}_1^2, \ldots, \tilde{c}_N^2\}$. Then, in the step 1, there exists the m such that $\hat{\boldsymbol{c}}_{s,m} = (\hat{c}_{1,m}, \ldots, \hat{c}_{s,m}) = (\tilde{c}_{l_1}, \ldots, \tilde{c}_{l_s})$ holds and such m satisfies (40) ; i.e.

$$\max_{1 \le m \le M_s} \sum_{j=1}^{s} \hat{c}_{j,m}^2 = \sum_{j=1}^{s} \hat{c}_{j,m^*}^2 = \sum_{j=1}^{s} \tilde{c}_{l_j}^2. \tag{41}$$

We summarize the parameter estimation procedure of $\mathcal{M}^{ODV}(s)$ for $(\boldsymbol{x}, \boldsymbol{y})_N$ as follows.

The parameter estimation procedure of $\mathcal{M}^{ODV}(s)$

step 1′ By applying (24) under the setting of $s = N$ and $\boldsymbol{b}_N = (b'_1, \ldots, b'_N)$, we obtain the least square estimate of the coefficient vector $\tilde{\boldsymbol{c}}_s = (\tilde{c}_1, \ldots, \tilde{c}_N)$

step 2′ Reorder $\tilde{c}_1^2, \ldots, \tilde{c}_N^2$ in increasing order of magnitude and obtain $\tilde{c}_{l_1}^2, \ldots, \tilde{c}_{l_N}^2$, where $\tilde{c}_{l_1}^2 \ge \ldots \ge \tilde{c}_{l_N}^2$. Set

$$\hat{c}_1 = \tilde{c}_{l_1}, \quad \ldots \quad, \hat{c}_s = \tilde{c}_{l_s} \tag{42}$$

$$\hat{b}_1 = \tilde{b}_{l_1}, \quad \ldots \quad, \hat{b}_s = \tilde{b}_{l_s}, \tag{43}$$

and obtain $\hat{\boldsymbol{c}}_s = (\hat{c}_1, \ldots, \hat{c}_s)$, $\hat{\boldsymbol{b}}_s = (\hat{b}_1, \ldots, \hat{b}_s)$. Hence, set $\hat{\boldsymbol{\omega}}_s = (\hat{\boldsymbol{c}}_s, \hat{\boldsymbol{b}}_s)$ and obtain $\hat{\boldsymbol{\theta}}_k = (\hat{\boldsymbol{\omega}}_s, \hat{\sigma}_s^2)$, where $\hat{\sigma}_s^2 = r_{\mathrm{emp}}(\hat{\boldsymbol{\omega}}_s)$. □

For $\mathcal{M}^{ODV}(s)$, we define

$$C^{ODV}(s, N) \equiv E_{\boldsymbol{y}_N}\left[\frac{1}{\sigma_*^2}\sum_{j=1}^{s}\hat{c}_j^2\right] = E_{\boldsymbol{y}_N}\left[\frac{1}{\sigma_*^2}\sum_{j=1}^{s}\tilde{c}_{l_j}^2\right] \tag{44}$$

and denote $R_{\mathrm{emp}}(s, N)$ and $R(s, N)$ by $R_{\mathrm{emp}}^{ODV}(s, N)$ and $R^{ODV}(s, N)$ respectively. Then, under the assumption that the set of samples is a Gaussian noise sequence, the following holds.

Theorem 2

$$R_{\mathrm{emp}}^{ODV}(s, N) = \sigma_*^2 - \frac{\sigma_*^2}{N} C^{ODV}(s, N) \tag{45}$$

$$R^{ODV}(s, N) = \sigma_*^2 + \frac{\sigma_*^2}{N} C^{ODV}(s, N) \tag{46}$$

Proof By applying the parameter estimation algorithm of $\mathcal{M}^{DV}(s)$ to $\mathcal{M}^{ODV}(s)$, the empirical loss is given by (39) at every m. Hence,

$$\begin{aligned} r_{\mathrm{emp}}(\hat{\boldsymbol{c}}_{s,m^*}) &= \frac{1}{N}\sum_{i=1}^{N}{\varepsilon_i}^2 - \frac{1}{N}\sum_{j=1}^{s}\hat{c}_{s,m^*}^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}{\varepsilon_i}^2 - \frac{1}{N}\sum_{j=1}^{s}\tilde{c}_{l_j}^2 \end{aligned} \tag{47}$$

holds because $\boldsymbol{y}_N = \boldsymbol{\varepsilon}_N$ under the assumption. Taking the expectation of the both side of the above equation with respect to the distribution of $\boldsymbol{y}_N$, we obtain (45). On the other hand, because $\boldsymbol{z}_N$ is an independent and identically distributed random vector with $\boldsymbol{y}_N$,

$$\begin{aligned} R(\widehat{\omega}_s, N) &= E_{\boldsymbol{z}_N}\left[\frac{1}{N}\sum_{i=1}^{N}\{z_i - g(x_i;\widehat{\omega}_s)\}^2\right] \\ &= \sigma_*^2 + \frac{1}{N}\sum_{i=1}^{N} g(x_i;\widehat{\omega}_s)^2. \end{aligned} \tag{48}$$

Because the estimated output of the ODVFR is given by

$$g(x_i;\widehat{\omega}_s) = \sum_{j=1}^{s} \widetilde{c}_{l_j}\phi(x_i;b_{l_j}) \tag{49}$$

and from the orthonormality of the basis,

$$\frac{1}{N}\sum_{i=1}^{N} g(x_i;\widehat{\omega}_s)^2 = \frac{1}{N}\sum_{j=1}^{s}\widetilde{c}_{l_j}^2 \tag{50}$$

holds. Thus, (48) can be rewritten as

$$R(\widehat{\omega}_s, N) = \sigma_*^2 + \frac{1}{N}\sum_{j=1}^{s}\widetilde{c}_{l_j}^2. \tag{51}$$

By taking the expectation of the both side of (51) with respect to the distribution of $\boldsymbol{y}_N$, we obtain (46). □

By (33), we know that $\widetilde{c}_j$; $j = 1,\ldots,N$ are independent random variables with a common normal distribution $N(0,\sigma_*^2)$. Hence, $\widetilde{c}_j^2/\sigma_*^2$; $j = 1,\ldots,N$ are independent random variables with a common χ_1^2 distribution. Therefore, by taking into account the way of the choice of $\widetilde{c}_{l_1},\ldots,\widetilde{c}_{l_s}$ in the parameter estimation algorithm and the definition of $C^{ODV}(s,N)$, the calculation of $C^{ODV}(s,N)$ reduces to the calculation of the expectation of the sum of the s largest values in an sequence of the independent N random variables with the common χ_1^2 distribution.

Let us define

$$\overline{U}_s = \frac{1}{\sigma_*^2}\sum_{j=1}^{s}\widetilde{c}_{l_j}^2, \tag{52}$$

and

$$C^{ODV}(s,N) = E_{\boldsymbol{y}_N}\left[\,\overline{U}_s\,\right]. \tag{53}$$

In this paper, we give lower and upper bounds for $C^{ODV}(s,N)$, by which we obtain lower and upper bounds for $R_{\mathrm{emp}}^{ODV}(s,N)$ and $R^{ODV}(s,N)$. In the following, we omit $\boldsymbol{y}_N$ from $E_{\boldsymbol{y}_N}[\,\cdot\,]$. The following lemma is basic throughout this paper.

Lemma 1 *Let U and V be random variables with finite mean. If $U \geq V$ then $E[U] \geq E[V]$.*

We may use above lemma without proof of the finite mean property of random variables because the random variables, for which we apply the above lemma, are obviously finite if the number of samples is finite as will be seen in the following.

Lemma 2 *Let $U_1, \ldots, U_M$ be independent random variables with a common χ_2^2 distribution. Let us define*

$$\overline{U}^M = \max_{1 \le m \le M} U_m \tag{54}$$

and

$$C^*(2, M) = E\left[\overline{U}^M\right]. \tag{55}$$

Furthermore, let $\delta(M)(> 0)$ be a proper function of M such that $\delta(M) \to 0$ as $M \to \infty$. Then

$$2 \cdot \{\gamma + \log M - \delta(M)\} \le C^*(2, M) \le 2 \cdot \{\gamma + \log M + \delta(M)\} \tag{56}$$

and $\delta(M)$ can be taken arbitrarily small, where γ is the Euler constant ; i.e. $\gamma = 0.5772156 \cdots$.

Proof Since the χ_2^2 distribution is the same as the exponential distribution, the density and the distribution of U_m are respectively given by

$$f(u) = \frac{1}{2} e^{-u/2} \tag{57}$$

$$F(u) = \int_0^u f(u)du = (1 - e^{-u/2}). \tag{58}$$

Because $U_1, \ldots, U_M$ are independent and identically distributed, the density of $\overline{U}^M$ is given by

$$f_{\overline{U}^M}(u) = M \{F(u)\}^{M-1} f(u). \tag{59}$$

Thus, by applying the binomial formula and the integration by parts, we obtain

$$\begin{aligned} C^*(2, M) &= \int_0^\infty u f_{\overline{U}^M}(u)du \\ &= 2M \sum_{m=0}^{M-1} (-1)^m \begin{pmatrix} M-1 \\ m \end{pmatrix} \frac{1}{(m+1)^2}, \end{aligned} \tag{60}$$

where

$$\begin{pmatrix} M \\ m \end{pmatrix} = \frac{M!}{m!(M-m)!} \tag{61}$$

denotes the binomial coefficient. Since

$$\frac{1}{m+1} \begin{pmatrix} M-1 \\ m \end{pmatrix} = \frac{1}{M} \begin{pmatrix} M \\ m+1 \end{pmatrix}$$

and

$$\sum_{m=1}^{M} (-1)^{m+1} \begin{pmatrix} M \\ m \end{pmatrix} \frac{1}{m} = \sum_{m=1}^{M} \frac{1}{m} \tag{62}$$

hold, the following equation is obtained.

$$C^*(2, M) = 2 \sum_{m=1}^{M} \frac{1}{m}.$$

Finally, by using the relation

$$\lim_{M\to\infty}\left\{\sum_{m=1}^{M}\frac{1}{m}-\log M\right\}=\gamma, \tag{63}$$

and taking $\delta(M)$ properly, (56) is proved. □

First, we will give a lower bound for $C^{ODV}(s,N)$, which is denoted by $\underline{C}^{ODV}(s,N)$.

Lemma 3

$$C^{ODV}(2,N)\geq\underline{C}^{ODV}(2,N)=2\cdot\{\gamma+\log\underline{M}-\delta(\underline{M})\}, \tag{64}$$

where $\underline{M}=\lceil\frac{N}{2}\rceil$. *Here,* $\lceil q\rceil$ *denotes the integer which does not exceed* q.

Proof When we define

$$V_1=(\tilde{c}_1^2+\tilde{c}_2^2)/\sigma_*^2,\ V_2=(\tilde{c}_3^2+\tilde{c}_4^2)/\sigma_*^2,\ \cdots, \tag{65}$$

the number of the above random variables is at least $\underline{M}$. Since $\tilde{c}_j^2/\sigma_*^2$; $j=1,\ldots,N$ are independent random variables with a common χ_1^2 distribution, $V_1,\ldots,V_{\underline{M}}$ are independent random variables with a common χ_2^2 distribution. Hence, when we define

$$\overline{V}_{\underline{M}}=\max_{1\leq n\leq\underline{M}}V_n, \tag{66}$$

we can apply Lemma 2 to $\overline{V}_{\underline{M}}$ and obtain

$$E\left[\overline{V}_{\underline{M}}\right]=C^*(2,\underline{M}). \tag{67}$$

On the other hand, since, $\tilde{c}_{l_1}^2\geq\tilde{c}_{l_2}^2\geq\tilde{c}_l^2$ for $l\neq l_1\neq l_2$,

$$(\tilde{c}_{l_1}^2+\tilde{c}_{l_2}^2)/\sigma_*^2\geq V_n \tag{68}$$

holds for any $n<\underline{M}$. Thus,

$$\overline{U}_2=(\tilde{c}_{l_1}^2+\tilde{c}_{l_2}^2)/\sigma_*^2\geq\overline{V}_{\underline{M}}. \tag{69}$$

From this inequality, (67) and Lemma 1, we obtain

$$C^{ODV}(2,N)=E\left[\overline{U}_2\right]\geq E\left[\overline{V}_{\underline{M}}\right]=C^*(2,\underline{M}). \tag{70}$$

By setting $\underline{C}^{ODV}(2,N)=C^*(2,\underline{M})$, we obtain (64). □

Next, we give the following lemma.

Lemma 4 *For* $s\geq 1$,

$$C^{ODV}(s,N)\leq s\cdot C^{ODV}(1,N). \tag{71}$$

Proof When $s=1$, (71) obviously holds with equality. When $s\geq 2$,

$$\overline{U}_s=\frac{1}{\sigma_*^2}\sum_{n=1}^{s}\tilde{c}_{l_n}^2\leq\frac{1}{\sigma_*^2}\cdot s\cdot\tilde{c}_{l_1}^2=s\cdot\overline{U}_1 \tag{72}$$

holds because $\tilde{c}_{l_1}^2\geq\tilde{c}_{l_n}^2$ for any $n<N$. By this inequality and Lemma 1, we obtain (71). □

Lemma 5 *For $s > 2$,*

$$C^{ODV}(s,N) \geq \underline{C}^{ODV}(s,N) = 2 \cdot \{\gamma + \log \underline{M} - \delta(\underline{M})\}. \tag{73}$$

And

$$C^{ODV}(1,N) \geq \underline{C}^{ODV}(1,N) = \gamma + \log \underline{M} - \delta(\underline{M}). \tag{74}$$

Proof For $s > 2$,

$$\overline{U}_s = \frac{1}{\sigma_*^2} \sum_{n=1}^{s} \tilde{c}_{l_n}^2 = \overline{U}_2 + \frac{1}{\sigma_*^2} \sum_{n=3}^{s} \tilde{c}_{l_n}^2. \tag{75}$$

holds. Because $\tilde{c}_{l_n}^2 \geq 0$, $\overline{U}_2 \leq \overline{U}_s$ in the above equation. Hence, we obtain (73) by Lemma 1 and 3. On the other hand, when we set $s = 2$ in Lemma 4,

$$C^{ODV}(1,N) \geq \frac{1}{2} C^{ODV}(2,N) \tag{76}$$

holds. Thus, by putting

$$\underline{C}^{ODV}(1,N) = \frac{1}{2} \underline{C}^{ODV}(2,N), \tag{77}$$

we obtain (74) from Lemma 3. □

As shown in Lemma 2, $\underline{C}^{ODV}(s,N)$ goes to infinity as $N \to \infty$. Consequently, $C^{ODV}(s,N)$ goes to infinity as $N \to \infty$ while $C^{OF}(s,N)$ does not depend on N, which is given in Theorem 1. This is a remarkable property which is induced by the selectability of basis functions. In the following, we will give an upper bound for $C^{ODV}(s,N)$, which is denoted by $\overline{C}^{ODV}(s,N)$.

Lemma 6 *For $s \geq 1$,*

$$C^{ODV}(s,N) \leq \overline{C}^{ODV}(s,N) = s \cdot 2 \cdot \{\gamma + \log N + \delta(N)\}. \tag{78}$$

Proof When we write $U_l = \tilde{c}_l^2/\sigma_*^2$; $l = 1,\ldots,N$, U_l ; $l = 1,\ldots,N$ are independent random variables with a common χ_1^2 distribution. We choose $V_1,\ldots,V_N$ with a common χ_1^2 distribution so as to hold the independency of the random variables $W_l = U_l + V_l$; $l = 1,\ldots,N$. Then, W_l ; $l = 1 \ldots N$ are independent random variables with a common χ_2^2 distribution. By defining

$$\overline{W}_N = \max_{1 \leq l \leq N} W_l, \tag{79}$$

$\overline{U}_1 \leq \overline{W}_N$ holds because $V_l \geq 0$. Thus, by Lemma 1 and 3,

$$C^{ODV}(s,N) = E\left[\overline{U}_s\right] \leq s \cdot E\left[\overline{U}_1\right] \leq s \cdot E\left[\overline{W}_N\right] = s \cdot C^*(2,N). \tag{80}$$

holds for any $s \geq 1$. Hence, we immediately obtain (78) by Lemma 2. □

Taking into account the fact that $\underline{C}^{ODV}(s,N)$ and $\overline{C}^{ODV}(s,N)$ are of the order of $\log N$ if s is fixed, the following theorem is obvious from Theorem 2, Lemma 3, 5 and 6.

Theorem 3

$$\sigma_*^2 - \frac{\sigma_*^2}{N} \overline{C}^{ODV}(s,N) \leq R_{\text{emp}}^{ODV}(s,N) \leq \sigma_*^2 - \frac{\sigma_*^2}{N} \underline{C}^{ODV}(s,N) \tag{81}$$

$$\sigma_*^2 + \frac{\sigma_*^2}{N} \underline{C}^{ODV}(s,N) \leq R^{ODV}(s,N) \leq \sigma_*^2 + \frac{\sigma_*^2}{N} \overline{C}^{ODV}(s,N) \tag{82}$$

and, under the fixed number of basis functions,

$$C^{ODV}(s,N) = O\ (\log N)\,. \tag{83}$$

8. Discussion

In this section, based on the previous results, we discuss about the expectations of the minimum of empirical loss and the expected loss of $\mathcal{M}^{ODV}(s)$. Furthermore, we compare the statistical properties of $\mathcal{M}^{ODV}(s)$ with that of $\mathcal{M}^{OF}(s)$. In the following, when we mention the behaviour of $\mathcal{M}^{ODV}(s)$ and $\mathcal{M}^{OF}(s)$ with the increase of N, we keep the number of basis functions fixed.

When we define $B^{OF}(s,N) \equiv \sigma_*^2 C^{OF}(s,N)/N$ in Theorem 1, $R^{OF}_{\rm emp}(s,N)$ is smaller than the true variance σ_*^2 by $B^{OF}(s,N)$ and, conversely, $R^{OF}(s,N)$ is larger than σ_*^2 by $B^{OF}(s,N)$. $B^{OF}(s,N)$ arises obviously from the parameter estimation. On the other hand, in Theorem 2, when we define $B^{ODV}(s,N) \equiv \sigma_*^2 C^{ODV}(s,N)/N$, $R^{ODV}_{\rm emp}(s,N)$ is smaller than the true variance σ_*^2 by $B^{ODV}(s,N)$ and, conversely, $R^{ODV}(s,N)$ is larger than σ_*^2 by $B^{ODV}(s,N)$. It is clear that $B^{OF}(s,N)$ is of $O(1/N)$ by Theorem 1 while $B^{ODV}(s,N)$ is of $O(\log N/N)$ by Theorem 3. Thus, we immediately obtain the following corollary.

Corollary 1 *Under the fixed number of basis functions,*

$$\lim_{N\to\infty} R^{OF}_{\rm emp}(s,N) = \lim_{N\to\infty} R^{ODV}_{\rm emp}(s,N) = \sigma_*^2 \tag{84}$$

$$\lim_{N\to\infty} R^{OF}(s,N) = \lim_{N\to\infty} R^{ODV}(s,N) = \sigma_*^2. \tag{85}$$

Since the estimated variance $\widehat{\sigma}_s^2$ is defined as the empirical loss, (84) asserts that the estimated variance of $\mathcal{M}^{ODV}(s)$, as well as that of $\mathcal{M}^{OF}(s)$, is an asymptotically unbiased estimate for the true variance. Hence, in both of $\mathcal{M}^{ODV}(s)$ and $\mathcal{M}^{OF}(s)$, if the number of samples is sufficiently large, the estimated output is close to the output of the true function ; i.e. $h(x_i) = 0$; $i = 1,\ldots,N$. However, because the convergence rate of $R^{ODV}_{\rm emp}(s,N)$ is slower than that of $R^{OF}_{\rm emp}(s,N)$, the sample size needed for $\mathcal{M}^{ODV}(s)$ is larger compared to the sample size needed for $\mathcal{M}^{OF}(s)$ if we keep the same deviation between the true output and the estimated output for both of $\mathcal{M}^{ODV}(s)$ and $\mathcal{M}^{OF}(s)$. Furthermore, by Theorem 2 and Corollary 1, we can conclude that $R^{ODV}(s,N)$ and $R^{OF}(s,N)$ asymptotically attain their minima because $B^{ODV}(s,N), B^{OF}(s,N) \geq 0$.

$C^{ODV}(s,N)$ increases monotonically with the increase of N by Theorem 3. By the definition of the DVFR, the basis parameter space is spread as N increases, which is regarded as increasing in variety of the FR. However, by Corollary 1, the difference between $R^{ODV}_{\rm emp}(s,N)$ and σ_*^2 decreases in the order of $\log N/N$. Consequently, the FR cannot pursue the large size samples although the variety of the FR increases with the increase of the number of samples. This is due to fixate the number of basis functions.

By Theorem 1 and 3, there exists the N such that $C^{OF}(s,N) < C^{ODV}(s,N)$ if the number of basis functions is fixed. Thus, the following is true.

Corollary 2 *When the number of basis functions s is fixed and $s \ll N$,*

$$R^{OF}_{\rm emp}(s,N) > R^{ODV}_{\rm emp}(s,N) \tag{86}$$

$$R^{OF}(s,N) < R^{ODV}(s,N) \tag{87}$$

holds.

That is, the minimum of the average squared error given by the ODVFR may be smaller than the minimum given by the OFFR while the expected loss at the estimated parameter vector for the ODVFR may be larger than the expected loss for the OFFR. In the above corollary, (86) agrees with the result in function approximation[5][12], in which the capability of neural networks brought about by the selectability of basis functions has been shown. Generally speaking, good fitting to the given samples induces bad generalization for unseen samples. The above corollary tells us that this is ture in our situation and the phenomenon is more remarkable for the ODVFR compared to the OFFR. Thus, the number of samples needed for the ODVFR is larger compared to the number of samples needed for the OFFR to identify the true function buried under noise, which is mentioned above.

Furthermore, let us devote a little more space to discussing the model selection problem of $\mathcal{M}^{ODV}(s)$. The following relations hold by Theorem 1 and 2.

$$R^{OF}(s,N) = R^{OF}_{\rm emp}(s,N) + 2\frac{\sigma_*^2}{N} C^{OF}(s,N) \tag{88}$$

$$R^{ODV}(s,N) = R^{ODV}_{\rm emp}(s,N) + 2\frac{\sigma_*^2}{N} C^{ODV}(s,N). \tag{89}$$

As mentioned in the last part of section 4., the relation for $\mathcal{M}^{OF}(s)$ is exactly the same as the relation employed in PSE[4] since $C^{OF}(s,N) = s$. The PSE is a model selection criterion in which $R^{OF}_{\rm emp}(s,N)$ and σ_*^2 are substituted by the empirical loss and the proper estimate of the variance respectively. The second term of PSE is determined by $C^{OF}(s,N)$, which penalizes the model complexity. By Theorem 1, $R^{OF}_{\rm emp}(s,N)$ decreases as $C^{OF}(s,N)$ increases. The same is true for $\mathcal{M}^{ODV}(s)$ by Theorem 2. Under the least square error method, the average squared error decreases as the complexity of the FR increases. Thus, $C^{ODV}(s,N)$ and $C^{OF}(s,N)$ can be regarded as the complexity of each FR. Consequently, we should notice that, under the fixed number of basis functions and the fixed number of samples, the ODVFR is more complex than the OFFR by Corollary 2. Moreover, it is found that we can not directly apply the PSE for $\mathcal{M}^{ODV}(s)$ by Theorem 3. However, unfortunately, the above discussion holds provided only that the set of samples is a Gaussian noise sequence. Therefore, we need more general results to discuss about the adequate model selection criterion for $\mathcal{M}^{ODV}(s)$ and this is a part of the future work.

Finally, we consider the variable basis type FR with the continuous basis parameters. We denote the family of the ODVFR by $\mathcal{G}_s^{ODV}$. We define an orthonormal continuous variable basis type FR as follows. We call a FR an orthonormal continuous variable basis type, which satisfies that $\boldsymbol{B}_j = \boldsymbol{R}^t$; $j = 1,\ldots,s$ and $\mathcal{G}_s^{ODV} \subseteq \mathcal{G}_s^{OCV}$, where $\mathcal{G}_s^{OCV}$ is a family of orthonormal continuous variable basis type FRs. The orthonormal continuous variable basis type FR is denoted by OCVFR. In the above, t corresponds to the number of basis parameters in a basis function. The regression model using the OCVFR is denoted by $\mathcal{M}^{OCV}(s)$. The $R_{\rm emp}(s,N)$ of $\mathcal{M}^{OCV}(s)$ is denoted by $R^{OCV}_{\rm emp}(s,N)$. The $r_{\rm emp}(\widehat{\omega}_s)$ of $\mathcal{M}^{ODV}(s)$ and that of $\mathcal{M}^{OCV}(s)$ are denoted by $r^{ODV}_{\rm emp}(\widehat{\omega}_s)$ and $r^{OCV}_{\rm emp}(\widehat{\omega}_s)$ respectively. Because $0 \leq R^{OCV}_{\rm emp}(s,N)$, $R^{ODV}_{\rm emp}(s,N) \leq \sigma_*^2$ holds provided that the set of samples is a Gaussian noise sequence, $R^{OCV}_{\rm emp}(s,N)$ and $R^{ODV}_{\rm emp}(s,N)$ are bounded. On the other hand, because $\mathcal{G}_s^{ODV} \subseteq \mathcal{G}_s^{OCV}$ holds, $r^{OCV}_{\rm emp}(\widehat{\omega}_s) \leq r^{ODV}_{\rm emp}(\widehat{\omega}_s)$ under the least square error method. Thus, by Lemma 1 and Theorem 3, we obtain the following corollary.

Corollary 3 *When the set of samples is a Gaussian noise sequence with the mean* 0 *and the variance* σ_*^2,

$$R_{\mathrm{emp}}^{OCV}(s,N) \leq \sigma_*^2 - \frac{\sigma_*^2}{N}\underline{C}^{ODV}(s,N). \tag{90}$$

Generally speaking, in the parameter estimation of $\mathcal{M}^{OCV}(s)$, there is no guarantee to be obtained the global minimum of the empirical loss. So, we should remark that the above corollary holds only if the global minimum is attained in the parameter estimation algorithm.

Example Let the range of inputs be $[0, 2\pi)$ and determine the N input points so as to satisfy $x_i = 2\pi(i-1)/N \ ; \ i = 1,\ldots,N$. In this situation, it can be shown that the basis described below satisfies the orthonormality condition.

In the case that N is even :

$$\left\{\frac{1}{\sqrt{N}}, \frac{\cos x}{\sqrt{N/2}}, \frac{\sin x}{\sqrt{N/2}}, \ldots, \frac{\sin(\frac{N}{2}-1)x}{\sqrt{N/2}}, \frac{\cos\frac{N}{2}x}{\sqrt{N}}\right\} \tag{91}$$

In the case that N is odd :

$$\left\{\frac{1}{\sqrt{N}}, \frac{\cos x}{\sqrt{N/2}}, \frac{\sin x}{\sqrt{N/2}}, \ldots, \frac{\cos(\lceil\frac{N}{2}\rceil)x}{\sqrt{N/2}}, \frac{\sin(\lceil\frac{N}{2}\rceil)x}{\sqrt{N/2}}\right\}. \tag{92}$$

Let us consider the case that N is odd. We choose $b'_1 = (0,\kappa)$ so as to satisfy $\sin(\kappa) = 1/\sqrt{2}$ and determine $b'_2 = (1,\pi/2)$, $b'_3 = (1,0)$, $b'_4 = (2,\pi/2)$, $b'_5 = (2,0)$, ..., $b'_{N-1} = (\lceil\frac{N}{2}\rceil,\pi/2)$, $b'_N = (\lceil\frac{N}{2}\rceil,0)$. Then the basis parameter space is given by $\boldsymbol{B}_j = \{b'_1,\ldots,b'_N\}$. The basis is constructed as follows.

$$(\sin(b_1\cdot\overline{x})/\sqrt{N/2}, \sin(b_2\cdot\overline{x})/\sqrt{N/2}, \ldots, \sin(b_s\cdot\overline{x})/\sqrt{N/2}) \ , \ b_j \in \boldsymbol{B}_j \ ; \ j = 1,\ldots,s \tag{93}$$

where $\overline{x} = (x,1)$ and $\cdot$ denotes the inner product. The FR prescribed by the above basis is the ODVFR. Thus, Theorem 2, 3 and Corollary 1, 2 hold for the FR. Now, we consider a 3-layered neural network with one-input and one-output, which has threshold to hidden layer. And the networks has the sinusoidal function in hidden layer and linear function in input and output layers. Then the network with s hidden units is the OCVFR. Hence, if the global minimum of the average squared error can be obtained by learning, Corollary 3 holds for the network. □

9. Conclusion

In this paper, to focus on the selectability of basis functions in the context of the regression model, we analyzed the statistical properties of the orthonormal discrete variable basis type function representations by taking the squared error as a loss function. We showed that the calculation of the expectations of the minimum of the empirical loss and the expected loss with respect to the joint distribution of the given samples is reduced to that of the expectation of the sum of the s largest values of the sequence of independent random variables with a common χ^2 distribution with one degree of freedom, where s corresponds to the number of basis functions. Along this line, we obtained lower and upper bounds for the expectations. Based on these results, we obtain the fact that the minimum of the empirical loss is asymptotically unbiased and the expected loss attains its minimum in the limit. Furthermore, the results tell

us that the expectation of the minimum of the empirical loss given by the orthonormal discrete variable basis type function representations is smaller than the expectation of the minimum given by the orthonormal fixed basis type function representations and the converse relation holds for the expectation of the expected loss. The results also tell us that the model selection criterion for the regression models using the orthonormal discrete variable basis type function representations should be different from the PSE. Furthermore, by using these bounds, we obtained an upper bound of the expectation of the minimum of the empirical loss given by the orthonormal continuous variable basis type function representations. And we obtained an upper bound on the expectation of the minimum of the empirical loss for a specific type of 3-layered neural network.

Our results obtained here are restricted in terms of the function representation. Thus, we should extend the results for the case of discrete variable basis type function representations and function representations with continuous variable basis functions, which include 3-layered neural networks. And, although we dealt with the case that the set of samples is a Gaussian noise sequence, we will consider the case that the set of samples in which the true function is not 0 or the true function is not in a class of function representations. Furthermore, we will consider a reasonable model selection criterion for the variable basis type function representations.

References

[1] Akaike H. : "Information Theory and an Extension of the Maximum Likelihood Principle", In *2nd International Symposium on Information Theory*, B.N.Petrov and F.Csáki eds., Akadémia Kiado, Budapest, pp.267–281, (1973).

[2] Akaike H. : "A New Look at the Statistical Model Identification", *IEEE Trans. Automat. Contr.*, **AC-19**, 6, pp.716–722, (1974).

[3] Amemiya T. : "Advanced econometrics", Basil Blackwell Ltd., (1985).

[4] Barron A. R. : "Predicted Squared Error : A Criterion for Automatic Model Selection", In *Self-Organizing Methods in Modeling*, S. Farlow, ed., Marcel Dekker, New York, pp.87–103, (1984).

[5] Barron A. R. : "Universal Approximation Bounds for Superposition of a Sigmoidal Function", *IEEE Trans. on Information Theory*, **39**, 3, pp.930–945, (1993).

[6] Fogel D.B. : "Criterion for Optimal Neural Network Selection", *IEEE Trans. on Neural Networks*, **2**, 5, pp.490–497, (1991).

[7] Geman S., Bienenstock E. and Doursat R. : "Neural Networks and the Bias/Variance Dilemma", *Neural Computation*, **4**, pp.1–58, (1992).

[8] Hagiwara K., Toda N. and Usui S. : "On the Problem of Applying AIC to Determine the Structure of a Layered Feed-Forward Neural Network", *Proceedings of International Joint Conference on Neural Networks*, Nagoya, Japan, **Vol.III**, pp.2263–2266, (1993).

[9] Kurita T. : "A Method to Determine the Number of Hidden Units of Three Layered Neural Networks by Information Criteria", *Trans. IEICE*, **Vol. J-73-D-II**, pp.1872–1878, (1990), in Japanese.

[10] Moody J. E. : "The effective Number of Parameters : An Analysis of Generalization and Regularization in Nonlinear Learning Systems", In *Advances in Neural Information Processing Systems* **4**, pp.598–605, Moody J. E., Hanson S. J. and Lippmann R. P. eds., Morgan Kaufmann, (1992).

[11] Murata N., Yoshizawa S. and Amari S. : "Network Information Criterion – Determining

the Number of Hidden Units for an Artificial Neural Network Model", *IEEE Trans. on Neural Networks*, **5**, 6, pp.865–872, (1994).

[12] Murata N. : "Function Approximation by Three-Layered Networks and Its Error Bounds — An Integral Representation Theorem", *Technical Reports, Mathematical Engineering Section*, **METR 94-19**, Univ. of Tokyo, (1994).

[13] Rumelhart D.E. and McClelland J.L. (Eds.) : "Parallel Distributed Processing", MIT Press, (1986).