A Relationship Between Generalization Error and Training Samples in Kernel Regressors

Akira Tanaka, Hideyuki Imai, Mineichi Kudo, and Masaaki Miyakoshi
Division of Computer Science
Hokkaido University
Sapporo, Japan
e-mail: {takira,imai,mine,miyakosi}@main.ist.hokudai.ac.jp

***Abstract*—A relationship between generalization error and training samples in kernel regressors is discussed in this paper. The generalization error can be decomposed into two components. One is a distance between an unknown true function and an adopted model space. The other is a distance between an estimated function and the orthogonal projection of the unknown true function onto the model space. In our previous work, we gave a framework to evaluate the first component. In this paper, we theoretically analyze the second one and show that a larger set of training samples usually causes a larger generalization error.**

***Keywords*-kernel regressor; reproducing kernel Hilbert space; generalization error; sample points;**

I. Introduction

Learning based on kernel machines[1], represented by the support vector machine[2] and the kernel ridge regression[2], is widely known as a powerful tool for various fields of information science such as pattern recognition, regression estimation, and density estimation. In general, an appropriate model selection is required in order to obtain a small generalization error in kernel machines. Although many methods for the model selection, such as the leave-one-out cross-validation, are proposed, it is important to analyze the generalization error theoretically since it may be useful to improve the performance of the model selection methods. In kernel machines, a model space is specified by a linear space spanned by kernel functions corresponding to points in training data set. Theoretical analyses of the generalization error with respect to a kernel or its parameters is usually difficult since the metrics of the corresponding reproducing kernel Hilbert spaces may differ. Thus, we focus on theoretical analyses of the generalization error with respect to a training data set. There are two kinds of generalization error. One is a distance between an unknown true function and an adopted model space. The other is a distance between an estimated function and the orthogonal projection of the unknown true function onto the model space. In our previous work[3], we discussed the first generalization error and gave an upper bound of the absolute difference between the unknown true function and its orthogonal projection onto the model space at each point. According to the results, it immediately follows that the first generalization error decreases when the number of training samples increases. In this paper, we theoretically analyze the second generalization error and show that a larger set of training samples usually causes a larger generalization error. Numerical examples are also given to confirm our theoretical results.

II. Mathematical Preliminaries for The Theory of Reproducing Kernel Hilbert Spaces

In this section, we prepare some mathematical tools concerned with the theory of reproducing kernel Hilbert spaces[4].

Definition 1: [4] Let $\mathbf{R}^n$ be an n-dimensional real vector space and let $\mathcal{H}$ be a class of functions defined on $\mathcal{D} \subset \mathbf{R}^n$, forming a Hilbert space of real-valued functions. The function $K(\boldsymbol{x}, \tilde{\boldsymbol{x}}),\ (\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathcal{D})$ is called a reproducing kernel of $\mathcal{H}$, if

1) For every $\tilde{\boldsymbol{x}} \in \mathcal{D}$, $K(\cdot, \tilde{\boldsymbol{x}})$ is a function belonging to $\mathcal{H}$.
2) For every $\tilde{\boldsymbol{x}} \in \mathcal{D}$ and every $f \in \mathcal{H}$,

$$f(\tilde{\boldsymbol{x}}) = \langle f(\cdot), K(\cdot, \tilde{\boldsymbol{x}}) \rangle_{\mathcal{H}}, \tag{1}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the inner product of the Hilbert space $\mathcal{H}$.

The Hilbert space $\mathcal{H}$ that has a reproducing kernel is called a reproducing kernel Hilbert space (RKHS). The reproducing property Eq.(1) enables us to treat a value of a function at a point in $\mathcal{D}$. Note that reproducing kernels are positve definite (*p.d.*) [4]:

$$\sum_{i,j=1}^{N} c_i c_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0, \tag{2}$$

for any N, $c_1, \ldots, c_N \in \mathbf{R}$, and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathcal{D}$. In addition, $K(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = K(\tilde{\boldsymbol{x}}, \boldsymbol{x})$ for any $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathcal{D}$ is followed[4]. If a reproducing kernel $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ exists, it is unique[4]. Conversely, every *p.d.* function $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ has the unique corresponding RKHS [4].

Next, we introduce the Schatten product [5] that is a convenient tool to reveal the reproducing property of kernels.

Definition 2: [5] Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces. The Schatten product of $g \in \mathcal{H}_2$ and $h \in \mathcal{H}_1$ is defined by

$$(g \otimes h)f = \langle f, h\rangle_{\mathcal{H}_1} g, \quad f \in \mathcal{H}_1. \tag{3}$$

Note that $(g \otimes h)$ is a linear operator from $\mathcal{H}_1$ onto $\mathcal{H}_2$. It is easy to show that the following relations hold for $h, v \in \mathcal{H}_1$, $g, u \in \mathcal{H}_2$.

$$\begin{aligned} (h \otimes g)^* &= (g \otimes h), & (4)\\ (h \otimes g)(u \otimes v) &= \langle u, g\rangle_{\mathcal{H}_2}(h \otimes v), & (5) \end{aligned}$$

where the superscript * denotes the adjoint operator.

III. Problem Formulation of Learning and Kernel Regressors

Let $\{(y_i, \boldsymbol{x}_i) \mid i \in \{1, \ldots, \ell\}\}$ be a given training data set with $y_i \in \mathbf{R}$, $\boldsymbol{x}_i \in \mathcal{D} \subset \mathbf{R}^n$, satisfying

$$y_i = f(\boldsymbol{x}_i) + n_i, \tag{6}$$

where f denotes the unknown function and n_i denotes a zero-mean additive noise. The aim of machine learning is to estimate the unknown function f by using the given training data set and statistical properties of the noise.

In this paper, we assume that the unknown function f belongs to the RKHS $\mathcal{H}_K$ corresponding to a certain kernel function K. If $f \in \mathcal{H}_K$, then Eq.(6) is rewritten as

$$y_i = \langle f(\cdot), K(\cdot, \boldsymbol{x}_i)\rangle_{\mathcal{H}_K} + n_i, \tag{7}$$

on the basis of the reproducing property of kernels. Let $\boldsymbol{y} = [y_1, \ldots, y_\ell]'$ and $\boldsymbol{n} = [n_1, \ldots, n_\ell]'$ with the superscript $'$ denoting the transposition operator, then applying the Schatten product to Eq.(7) yields

$$\boldsymbol{y} = \left(\sum_{k=1}^{\ell} [\boldsymbol{e}_k^{(\ell)} \otimes K(\cdot, \boldsymbol{x}_k)]\right) f(\cdot) + \boldsymbol{n}, \tag{8}$$

where $\boldsymbol{e}_k^{(\ell)}$ denotes the k-th vector of the canonical basis of $\mathbf{R}^\ell$. For a convenience of description, we write

$$A_{K,X} = \left(\sum_{k=1}^{\ell} [\boldsymbol{e}_k^{(\ell)} \otimes K(\cdot, \boldsymbol{x}_k)]\right) \tag{9}$$

since $A_{K,X}$ is specified by the kernel K and the set of input vectors $X = \{\boldsymbol{x}_i \in \mathcal{D} \mid i \in \{1, \ldots, \ell\}\}$. $A_{K,X}$ is a linear operator that maps an element of $\mathcal{H}_K$ onto $\mathbf{R}^\ell$ and Eq.(8) can be written by

$$\boldsymbol{y} = A_{K,X} f(\cdot) + \boldsymbol{n}, \tag{10}$$

which represents the relation between the unknown true function f and an output vector $\boldsymbol{y}$. Therefore, a machine learning problem can be interpreted as an inversion problem of the linear equation Eq.(10)[6].

The minimum norm least squares solution for Eq.(10) is given by

$$\begin{aligned} \hat{f}(\cdot) &= A_{K,X}^+ \boldsymbol{y} = A_{K,X}^* G_{K,X}^+ \boldsymbol{y} \\ &= \sum_{k=1}^{\ell} \boldsymbol{y}' G_{K,X}^+ \boldsymbol{e}_k^{(\ell)} K(\cdot, \boldsymbol{x}_k), \end{aligned} \tag{11}$$

where $G_{K,X}$ denotes the Gramian matrix of K with X and the superscript $^+$ denotes the Moore-Penrose generalized inverse. Note that $A_{K,X}^+ A_{K,X}$ is the orthogonal projector onto $\mathcal{R}(A_{K,X}^*)$ (the range space of $A_{K,X}^*$, that is, the linear subspace spanned by $\{K(\cdot, \boldsymbol{x}_i) \mid i \in \{1, \ldots, \ell\}\}$) and its closed form is given by

$$P_{K,X} = \sum_{i,j=1}^{\ell} (G_{K,X}^+)_{i,j} K(\cdot, \boldsymbol{x}_i) \otimes K(\cdot, \boldsymbol{x}_j) \tag{12}$$

as shown in [7].

In practical problems, a solution by the kernel ridge regressor, which is a regularized version of Eq.(11) and is defined as

$$\hat{f}(\cdot) = \sum_{k=1}^{\ell} \boldsymbol{y}' (G_{K,X} + \mu I)^{-1} \boldsymbol{e}_k^{(\ell)} K(\cdot, \boldsymbol{x}_k) \tag{13}$$

with $\mu > 0$ denoting a regularization parameter, is used instead of Eq.(11). However, theoretical analyses of a solution based on Eq.(11) can be an important basis of all other kernel machines including Eq.(13). Thus, we theoretically analyze the generalization error of the solution Eq.(11) in the following contents.

IV. Generalization Error of a Model Space

In [3], we gave a framework to evaluate the generalization error of a model space, that is, $\mathcal{R}(A_{K,X}^*)$. In this section, we review results of [3] related to this paper.

Let $f \in \mathcal{H}_K$ be an unknown true function, then for any $\boldsymbol{x} \in \mathcal{D}$,

$$\begin{aligned} &|f(\boldsymbol{x}) - P_{K,X} f(\boldsymbol{x})| \\ &\quad = |\langle f(\cdot), K(\cdot, \boldsymbol{x})\rangle_{\mathcal{H}_K} - \langle P_{K,X} f(\cdot), K(\cdot, \boldsymbol{x})\rangle_{\mathcal{H}_K}| \\ &\quad = |\langle f(\cdot), K(\cdot, \boldsymbol{x})\rangle_{\mathcal{H}_K} - \langle f(\cdot), P_{K,X} K(\cdot, \boldsymbol{x})\rangle_{\mathcal{H}_K}| \\ &\quad = |\langle f(\cdot), K(\cdot, \boldsymbol{x}) - P_{K,X} K(\cdot, \boldsymbol{x})\rangle_{\mathcal{H}_K}| \\ &\quad \leq ||f||_{\mathcal{H}_K} E_{K,X}^{1/2}(\boldsymbol{x}), \end{aligned}$$

holds, where

$$\begin{aligned} E_{K,X}(\boldsymbol{x}) &= ||K(\cdot, \boldsymbol{x}) - P_{K,X} K(\cdot, \boldsymbol{x})||_{\mathcal{H}_K}^2 \\ &= \left(K(\boldsymbol{x}, \boldsymbol{x}) - \sum_{i,j=1}^{\ell} K(\boldsymbol{x}, \boldsymbol{x}_i)(G^+)_{i,j} K(\boldsymbol{x}, \boldsymbol{x}_j)\right), \end{aligned}$$

which implies that the absolute difference between f and $P_{K,X} f$ at a point $\boldsymbol{x} \in \mathcal{D}$ is proportional to $||f||_{\mathcal{H}_K}$ and $E_{K,X}^{1/2}(\boldsymbol{x})$[3]. Let

$$J_{K,X}^{(1)} = \sup_{\boldsymbol{x} \in \mathcal{D}} \left(E_{K,X}(\boldsymbol{x}) / K(\boldsymbol{x}, \boldsymbol{x})\right)^{1/2}, \tag{14}$$

then $0 \leq J^{(1)}_{K,X} \leq 1$ holds; and if $J^{(1)}_{K,X}$ is sufficiently close to zero, the model space $\mathcal{R}(A^*_{K,X})$ has a sufficient ability to represent any $f \in \mathcal{H}_K$. Note that it is trivial that when $\tilde{X} \subset X$, $J^{(1)}_{K,X} \leq J^{(1)}_{K,\tilde{X}}$ holds. Thus, it is concluded that a larger set of training samples implies a smaller generalization error of an adopted model space with a fixed kernel.

V. Generalization Error in a Model Space

The minimum norm least squares solution for Eq.(10) given in Section III can be decomposed as

$$\hat{f}(\cdot) = A^+_{K,X}\boldsymbol{y} = A^+_{K,X}A_{K,X}f(\cdot) + A^+_{K,X}\boldsymbol{n}. \tag{15}$$

The first term is the orthogonal projection of f onto $\mathcal{R}(A^*_{K,X})$ and its generalization error was analyzed in the previous section. The second term is the generalization error in $\mathcal{R}(A^*_{K,X})$, coming from the additive noise, whose closed form is given as

$$\begin{aligned}\hat{f}_n(\cdot) &= A^+_{K,X}\boldsymbol{n} \\ &= \left(\sum_{k=1}^{\ell} K(\cdot,\boldsymbol{x}_i)\otimes \boldsymbol{e}^{(\ell)}_k\right) G^+_{K,X}\boldsymbol{n} \\ &= \sum_{k=1}^{\ell} \boldsymbol{n}' G^+_{K,X}\boldsymbol{e}^{(\ell)}_k K(\cdot,\boldsymbol{x}_k),\end{aligned} \tag{16}$$

and its squared norm is given as

$$\begin{aligned}J^{(2)}_{K,X} &= ||\hat{f}_n||^2_{\mathcal{H}_K} = \langle \hat{f}_n, \hat{f}_n\rangle_{\mathcal{H}_K} \\ &= \boldsymbol{n}' G^+_{K,X}G_{K,X}G^+_{K,X}\boldsymbol{n} = \boldsymbol{n}' G^+_{K,X}\boldsymbol{n}.\end{aligned}$$

Lemma 1: Let

$$G = \begin{bmatrix} A & B \\ B' & C \end{bmatrix} \in \mathbf{R}^{(n+m)\times(n+m)} \tag{17}$$

be a *p.d.* matrix with $A \in \mathbf{R}^{n\times n}$, $C \in \mathbf{R}^{m\times m}$, and $B \in \mathbf{R}^{n\times m}$, then

$$Z = G^{-1} - \begin{bmatrix} A^{-1} & O_{n,m} \\ O_{m,n} & O_{m,m} \end{bmatrix} \tag{18}$$

is non-negative definite (*n.n.d.*), where $O_{m,n}$ denotes the zero matrix in $\mathbf{R}^{m\times n}$.

Proof: As shown in [8], G^{-1} can be represented as

$$G^{-1} = \begin{bmatrix} A^{-1}+FE^{-1}F' & -FE^{-1} \\ -E^{-1}F' & E^{-1} \end{bmatrix},$$

where $E = C - B'A^{-1}B$ and $F = A^{-1}B$. Thus, Eq.(18) is reduced to

$$Z = \begin{bmatrix} FE^{-1}F' & -FE^{-1} \\ -E^{-1}F' & E^{-1} \end{bmatrix}. \tag{19}$$

It is trivial that E^{-1} is *p.d.* since G, G^{-1}, and all their principal minors are *p.d.* Let $\boldsymbol{v}_1 \in \mathbf{R}^n$ and $\boldsymbol{v}_2 \in \mathbf{R}^m$ be arbitrary vectors and let $\boldsymbol{v} = [\boldsymbol{v}_1' \;\; \boldsymbol{v}_2']'$, then

$$\begin{aligned}&\boldsymbol{v}'Z\boldsymbol{v} \\ &= [\boldsymbol{v}_1' \;\; \boldsymbol{v}_2'] \begin{bmatrix} FE^{-1}F' & -FE^{-1} \\ -E^{-1}F' & E^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{v_1} \\ \boldsymbol{v_2} \end{bmatrix} \\ &= \boldsymbol{v}_1'FE^{-1}F'\boldsymbol{v}_1 - \boldsymbol{v}_1'FE^{-1}\boldsymbol{v}_2 \\ &\quad -\boldsymbol{v}_2'E^{-1}F'\boldsymbol{v}_1 + \boldsymbol{v}_2'E^{-1}\boldsymbol{v}_2 \\ &= (F'\boldsymbol{v}_1 - \boldsymbol{v}_2)'E^{-1}(F'\boldsymbol{v}_1 - \boldsymbol{v}_2) \geq 0\end{aligned}$$

holds, which concludes the proof. ■

Note that the non-singularity of G in this lemma is crucial. In fact, the singular matrix

$$G = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

gives a simple counter example since $G^+ - [1\;0]'([1\;0]G[1\;0]')^{-1}[1\;0]$ is not *n.n.d.*

Theorem 1: Let $X = \{\boldsymbol{x}_i \mid i \in \{1,\ldots,\ell\}\}$ be a set of training input vectors and let $\tilde{X} \subset X$ be a subset of X with $|\tilde{X}| = m < \ell$. Let $\boldsymbol{n} \in \mathbf{R}^\ell$ be the noise vector for X and let $\tilde{\boldsymbol{n}} \in \mathbf{R}^m$ be the noise vector for $\tilde{X}$ defined by $\tilde{\boldsymbol{n}} = P_1'\boldsymbol{n}$, where $P_1' \in \mathbf{R}^{m\times\ell}$ denotes the full row-rank matrix, such that all components are zero except for one component being unity in each row, that extracts components corresponding to $\tilde{X}$ from those of X. Then, if $G_{K,X}$ is non-singular,

$$J^{(2)}_{K,X} - J^{(2)}_{K,\tilde{X}} \geq 0 \tag{20}$$

holds.

Proof: Let $P_2' \in \mathbf{R}^{(\ell-m)\times\ell}$ be a full row-rank matrix, such that all components are zero except for one components being unity in each row, satisfying $P_1'P_2 = O_{m,\ell-m}$. It is obvious that $[P_1\;P_2]$ is a permutation matrix, which implies that it is also an orthogonal matrix. Then,

$$\begin{aligned}&J^{(2)}_{K,X} - J^{(2)}_{K,\tilde{X}} \\ &= \boldsymbol{n}'G^{-1}_{K,X}\boldsymbol{n} - \tilde{\boldsymbol{n}}'G^{-1}_{K,\tilde{X}}\tilde{\boldsymbol{n}} \\ &= \boldsymbol{n}'[P_1\;P_2][P_1\;P_2]'G^{-1}_{K,X}[P_1\;P_2][P_1\;P_2]'\boldsymbol{n} \\ &\quad -\boldsymbol{n}'P_1(P_1'G_{K,X}P_1)^{-1}P_1'\boldsymbol{n} \\ &= \boldsymbol{n}'[P_1\;P_2]([P_1\;P_2]'G_{K,X}[P_1\;P_2])^{-1}[P_1\;P_2]'\boldsymbol{n} \\ &\quad -\boldsymbol{n}'P_1(P_1'G_{K,X}P_1)^{-1}P_1'\boldsymbol{n} \\ &= \boldsymbol{n}'[P_1\;P_2]\begin{bmatrix} P_1'G_{K,X}P_1 & P_1'G_{K,X}P_2 \\ P_2'G_{K,X}P_1 & P_2'G_{K,X}P_2 \end{bmatrix}^{-1} \\ &\quad \times[P_1\;P_2]'\boldsymbol{n} \\ &\quad -\boldsymbol{n}'[P_1\;P_2]\begin{bmatrix} (P_1'G_{K,X}P_1)^{-1} & O_{m,\ell-m} \\ O_{\ell-m,m} & O_{\ell-m,\ell-m} \end{bmatrix} \\ &\quad \times[P_1\;P_2]'\boldsymbol{n} \geq 0\end{aligned}$$

holds from Lemma 1. ■

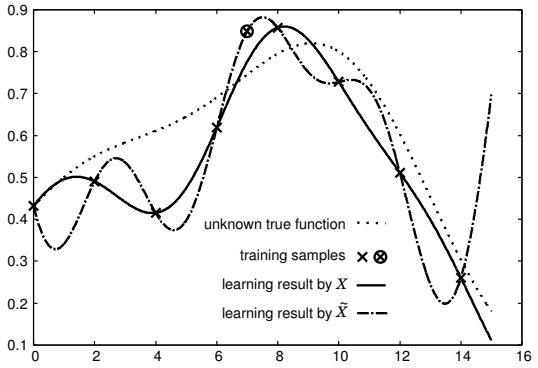

Figure 1. The unknown true function, the training samples, and the learning results by $\tilde{X}$ and X.

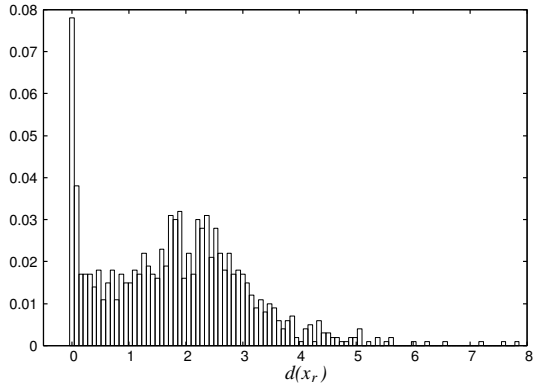

Figure 2. Normalized histogram of $d(x_r)$.

According to Theorem 1, given a realization of an additive noise vector $\boldsymbol{n}$ for X, the norm of the noise component in a learning result is larger than that based on $\tilde{\boldsymbol{n}}$ for $\tilde{X}$. It trivially holds that

$$E_{\boldsymbol{n}} J^{(2)}_{K,X} - E_{\tilde{\boldsymbol{n}}} J^{(2)}_{K,\tilde{X}} \geq 0 \tag{21}$$

when components of $\boldsymbol{n}$ and $\tilde{\boldsymbol{n}}$ are *i.i.d.* with the variance σ^2 since $E_{\boldsymbol{n}} J^{(2)}_{K,X} = E_{\boldsymbol{n}}(\boldsymbol{n}' G^{+}_{K,X} \boldsymbol{n}) = \sigma^2 \mathrm{tr}(G^{-1}_{K,X})$.

VI. Numerical Example

In this section, we verify the behavior of the norm of the noise component in the solution Eq.(11) with an artificial data. We adopt the Gaussian kernel given by

$$K(x, y) = \exp\left(-\frac{(x-y)^2}{16}\right) \tag{22}$$

as a kernel.

Figure 1 shows the unknown true function in the corresponding RKHS, training data set with $\sigma = 0.1$ where $\tilde{X} = \{0, 2, 4, 6, 8, 10, 12, 14\}$ denoted by '$\times$' and $X = \tilde{X} \cup X_e$, where $X_e = \{7\}$ denoted by '$\otimes$', and the learning results based on $\tilde{X}$ and X. $J^{(1)}_{K,\tilde{X}} = 0.139$ and $J^{(1)}_{K,X} = 0.116$, which implies that the model space $\mathcal{R}(A^{*}_{K,X})$ is slightly better than $\mathcal{R}(A^{*}_{K,\tilde{X}})$. On the other hand, $J^{(2)}_{K,\tilde{X}} = 0.408$ and $J^{(2)}_{K,X} = 58.227$, which agrees with the theoretical analyses given in the previous section.

Figure 2 shows the normalized histogram of

$$d(x_r) = \log\left(1 + J^{(2)}_{K,X} - J^{(2)}_{K,\tilde{X}}\right) \tag{23}$$

over $1,000$ trials with $X_e = \{x_r\}$ in which x_r is randomly selected from $[0, 15]$. According to Fig. 2, it is confirmed that $J^{(2)}_{K,X}$ is larger than $J^{(2)}_{K,\tilde{X}}$ in all cases.

VII. Conclusion

In this paper, we investigated the relationship between the generalization error and training samples in kernel regressors and showed that a larger set of training samples causes a larger generalization error corresponding to noise components. Extending our result for other practical kernel machines is one of future works that should be resolved.

Acknowledgment

This work was partially supported by Grant-in-Aid No.21700001 for Young Scientists (B) from the Ministry of Education, Culture, Sports and Technology of Japan.

References

[1] K. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," *IEEE Transactions on Neural Networks*, vol. 12, pp. 181–201, 2001.

[2] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge: Cambridge University Press, 2000.

[3] A. Tanaka, H. Imai, and M. Miyakoshi, "Kernel-induced sampling theorem," *IEEE Transactions on Signal Processing (in printing)*.

[4] N. Aronszajn, "Theory of Reproducing Kernels," *Transactions of the American Mathematical Society*, vol. 68, no. 3, pp. 337–404, 1950.

[5] R. Schatten, *Norm Ideals of Completely Continuous Operators*. Springer-Verlag, Berlin, 1960.

[6] H. Ogawa, "Neural Networks and Generalization Ability," *IEICE Technical Report*, vol. NC95-8, pp. 57–64, 1995 (in Japanese).

[7] A. Tanaka, H. Imai, M. Kudo, and M. Miyakoshi, "Optimal kernel in a class of kernels with an invariant metric," in *Proc. S&SSPR2008*, 2008, pp. 530–539.

[8] H. Yanai and K. Takeuchi, *Projection Matrices, Generalized Inverses, and Singular Value Decomposition (in Japanese)*. Tokyo: University of Tokyo Press, 1983.