Analyses on Kernel-Specific Generalization Ability for Kernel Regressors with Training Samples

Akira Tanaka and Masaaki Miyakoshi
Division of Computer Science,
Graduate School of Information Science and Technology, Hokkaido University
Kita-14, Nishi-9, Kita-ku, Sapporo, 060-0814, Japan
{takira,miyakosi}@main.ist.hokudai.ac.jp

***Abstract*— *Theoretical analyses on generalization error of a model space for kernel regressors with respect to training samples are given in this paper. In general, the distance between an unknown true function and a model space tends to be small with a larger set of training samples. However, it is not clarified that a larger set of training samples achieves a smaller difference at each point of the unknown true function and the orthogonal projection of it onto the model space, compared with a smaller set of training samples. In this paper, we show that the upper bound of the squared difference at each point of these two functions with a larger set of training samples is not larger than that with a smaller set of training samples. We also give some numerical examples to confirm our theoretical result.*

Keywords*— *kernel regressor; reproducing kernel Hilbert space; generalization error; training samples;

I. Introduction

Learning based on kernel machines[1], represented by the support vector machine[2] and the kernel ridge regressor[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 model selection, such as the leave-one-out cross-validation, were proposed, it is important to analyze generalization error theoretically since it may be useful to improve the performance of the model selection methods.

Generalization error is usually defined as the distance between an unknown true function and an estimated one; and it can be decomposed into two components in kernel machines. One is the distance between the unknown true function and an adopted model space, which is the linear subspace spanned by kernel functions corresponding to points in a set of training samples. The other is the distance between the estimated function and the orthogonal projection of the unknown true function onto the model space. The former one was not sufficiently investigated so far, while the latter one was discussed in many articles (see [3], [4] for instance). In our previous work[5], we investigated the former one and showed that a kernel corresponding to the smallest reproducing kernel Hilbert space including an unknown true function gives the best model space among a class of kernels with an invariant metric.

In this paper, we theoretically analyze the generalization error of the model space with respect to training samples. It is intuitively trivial that a larger set of training samples achieves a better model space compared with a smaller set of training samples. However, it is not trivial that a larger set of training samples achieves a smaller difference at each point of the unknown true function and the orthogonal projection of it onto the model space compared with a smaller set of training samples. We show that the upper bound of the squared difference at each point of these two functions with a large set of training samples is not larger than that with a small set of training samples. We also give some numerical examples to confirm our theoretical result.

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[6], [7].

Definition 1: [6] 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 $f(\cdot) \in \mathcal{H}$,

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

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 positive definite [6]:

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

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[6]. If a reproducing kernel $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ exists, it is unique[6]. Conversely,

every positive definite function $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ has the unique corresponding RKHS [6].

Next, we introduce the Schatten product [8] that is a convenient tool to reveal the reproducing property of kernels.

Definition 2: [8] 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 $h, v \in \mathcal{H}_1$, $g, u \in \mathcal{H}_2$.

$$(h \otimes g)^* = (g \otimes h), \tag{4}$$

$$(h \otimes g)(u \otimes v) = \langle u, g \rangle_{\mathcal{H}_2} (h \otimes v), \tag{5}$$

where the superscript * denotes the adjoint operator.

III. Formulation of Learning Problems

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

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

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

In this paper, we assume that the unknown true function $f(\cdot)$ belongs to the RKHS $\mathcal{H}_K$ corresponding to a certain kernel function K. If $f(\cdot) \in \mathcal{H}_K$, 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 transposed matrix (or vector), 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

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

where $X = \{\boldsymbol{x}_k \mid k \in \{1, \ldots, \ell\}\}$. The operator $A_{K,X}$ is a linear operator that maps an element in $\mathcal{H}_K$ 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 $f(\cdot)$ and an output vector $\boldsymbol{y}$. Therefore, a machine learning problem can be interpreted as an inversion problem of the linear equation Eq.(10)[9].

IV. Kernel Specific Generalization Ability

In general, a learning result by kernel machines is represented by a linear combination of $K(\cdot, \boldsymbol{x}_i)$, which means that the learning result is an element in $\mathcal{R}(A_{K,X}^*)$ (the range space of the linear operator $A_{K,X}^*$) since

$$\begin{aligned} \hat{f}(\cdot) &= A_K^* \boldsymbol{\alpha} \\ &= \left(\sum_{k=1}^{\ell} [K(\cdot, \boldsymbol{x}_k) \otimes \boldsymbol{e}_k^{(\ell)}] \right) \boldsymbol{\alpha} \\ &= \sum_{k=1}^{\ell} \alpha_k K(\cdot, \boldsymbol{x}_k) \end{aligned} \tag{11}$$

holds, where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_\ell]'$ denotes an arbitrary vector in $\mathbf{R}^\ell$. The point at issue of this paper is selection of a model space, that is, the generalization ability of $\mathcal{R}(A_{K,X}^*)$ which is independent from criteria of learning machines. In order to discuss $\mathcal{R}(A_{K,X}^*)$, the orthogonal projector onto $\mathcal{R}(A_{K,X}^*)$ in $\mathcal{H}_K$, written as $P_{K,X}$, plays an important role.

Lemma 1: [5]

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

where $G_{K,X}$ denotes the Gramian matrix of K with X and the superscript $^+$ denotes the Moore-Penrose generalized inverse matrix[10].

Let $|| \cdot ||_{\mathcal{H}_K}$ be the induced norm in $\mathcal{H}_K$, then a generalization error of $\mathcal{R}(A_{K,X}^*)$ for $f(\cdot) \in \mathcal{H}_K$ can be defined as

$$J_{K,X}^{(N)} = ||f(\cdot) - P_{K,X} f(\cdot)||_{\mathcal{H}_K}^2, \tag{13}$$

which is called 'norm-based generalization error.' In [5], we discussed the norm-based generalization error of a class of kernels with an invariant metric; and obtained the following result concerned with a selection of RKHS itself.

Theorem 1: [5] Let K_1 and K_2 be kernels satisfying the following properties:

(1) $\mathcal{H}_{K_1} \subset \mathcal{H}_{K_2}$ (nested).
(2) $\langle f(\cdot), g(\cdot) \rangle_{\mathcal{H}_{K_1}} = \langle f(\cdot), g(\cdot) \rangle_{\mathcal{H}_{K_2}}$ for any $f(\cdot), g(\cdot) \in \mathcal{H}_{K_1}$ (invariant metric).

Then, for any $f(\cdot) \in \mathcal{H}_{K_1}$ and any set of input vectors X,

$$J_{K_1,X}^{(N)} \leq J_{K_2,X}^{(N)} \tag{14}$$

holds.

This theorem claims that the kernel corresponding to the smallest RKHS gives the best model among a class of kernels with an invariant metric whose corresponding RKHS's include the unknown true function.

On the other hand in this paper, we discuss the properties of $\mathcal{R}(A_{K,X}^*)$ with respect to a set of training samples. We show

the following lemmas as preparations.

Lemma 2: [5] For any $f(\cdot) \in \mathcal{H}_K$ and X,

$$||P_{K,X} f(\cdot)||^2_{\mathcal{H}_K} = \boldsymbol{f}' G^+_{K,X} \boldsymbol{f} \tag{15}$$

holds, where $\boldsymbol{f} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_\ell)]'$.

Lemma 3: Let

$$G = \left[\begin{array}{cc} A & B \\ B' & C \end{array} \right] \in \mathbf{R}^{(n+m)\times(n+m)} \tag{16}$$

be an *n.n.d.* symmetric matrix with $A \in \mathbf{R}^{n\times n}$, $C \in \mathbf{R}^{m\times m}$, and $B \in \mathbf{R}^{n\times m}$ and let $\boldsymbol{v} \in \mathcal{R}(G)$. Then,

$$\boldsymbol{v}' \left(G^+ - \left[\begin{array}{cc} A^+ & O_{n,m} \\ O_{m,n} & O_{m,m} \end{array} \right] \right) \boldsymbol{v} \geq 0 \tag{17}$$

holds, where $O_{m,n} \in \mathbf{R}^{m\times n}$ denotes the zero matrix.

Proof: Since G is *n.n.d.*, it is trivial that ZGZ' is also *n.n.d.* for any matrix $Z \in \mathbf{R}^{p\times(m+n)}$. Let

$$Z = \left[\begin{array}{cc} I_n - AA^+ & O_{n,m} \\ -B'A^+ & I_m \end{array} \right],$$

where I_n denotes the identity matrix of degree n. Then, using the facts that $I_n - A^+A$ is a symmetric orthogonal projector and $A^+A = AA^+$ holds for any symmetric matrix A yield that

$$\begin{aligned} & ZG'Z \\ &= \left[\begin{array}{cc} I_n - AA^+ & O_{n,m} \\ -B'A^+ & I_m \end{array} \right] \left[\begin{array}{cc} A & B \\ B' & C \end{array} \right] \\ &\quad \times \left[\begin{array}{cc} I_n - AA^+ & O_{n,m} \\ -B'A^+ & I_m \end{array} \right]' \\ &= \left[\begin{array}{cc} I_n - AA^+ & O_{n,m} \\ -B'A^+ & I_m \end{array} \right] \left[\begin{array}{cc} A & B \\ B' & C \end{array} \right] \\ &\quad \times \left[\begin{array}{cc} I_n - AA^+ & -A^+B \\ O_{m,n} & I_m \end{array} \right] \\ &= \left[\begin{array}{cc} O_{n,n} & (I_n - AA^+)B \\ B'(I_n - A^+A) & C - B'A^+B \end{array} \right] \end{aligned}$$

is also *n.n.d.*

Since $\boldsymbol{v} \in \mathcal{R}(G)$, there exist $\boldsymbol{z}$ such that $\boldsymbol{v} = G\boldsymbol{z}$. Thus,

$$\begin{aligned} & \boldsymbol{v}' \left(G^+ - \left[\begin{array}{cc} A^+ & O_{n,m} \\ O_{m,n} & O_{m,m} \end{array} \right] \right) \boldsymbol{v} \\ &= \boldsymbol{z}' G \left(G^+ - \left[\begin{array}{cc} A^+ & O_{n,m} \\ O_{m,n} & O_{m,m} \end{array} \right] \right) G\boldsymbol{z} \\ &= \boldsymbol{z}' \left(\left[\begin{array}{cc} A & B \\ B' & C \end{array} \right] - \left[\begin{array}{cc} A & AA^+B \\ B'A^+A & B'A^+B \end{array} \right] \right) \boldsymbol{z} \\ &= \boldsymbol{z}' \left[\begin{array}{cc} O_{n,n} & (I_n - AA^+)B \\ B'(I_n - A^+A) & C - B'A^+B \end{array} \right] \boldsymbol{z} \geq 0 \end{aligned}$$

is obtained, which concludes the proof. □

Let $X_S = \{\boldsymbol{x}_i \mid i \in \{1, \ldots, \ell_S\}\}$ and let $X_L = \{\boldsymbol{x}_i \mid i \in \{1, \ldots, \ell_L\}\}$ with $\ell_S < \ell_L$. Note that $X_S \subset X_L$ holds. It is intuitively trivial that $\mathcal{R}(A^*_{K,X_L})$ gives a better model space than $\mathcal{R}(A^*_{K,X_S})$ since $\mathcal{R}(A^*_{K,X_S}) \subset \mathcal{R}(A^*_{K,X_L})$. Its theoretical ground is given by the following theorem.

Theorem 2: Let $f(\cdot) \in \mathcal{H}_K$, then

$$J^{(N)}_{K,X_L} \leq J^{(N)}_{K,X_S} \tag{18}$$

holds.

Proof: From Lemma 2 and the Pythagorean theorem, we have

$$\begin{aligned} & J^{(N)}_{K,X} \\ &= ||f(\cdot) - P_{K,X} f(\cdot)||^2_{\mathcal{H}_K} \\ &= ||f(\cdot)||^2_{\mathcal{H}_K} - ||P_{K,X} f(\cdot)||^2_{\mathcal{H}_K} \\ &= ||f(\cdot)||^2_{\mathcal{H}_K} - \boldsymbol{f}' G^+_{K,X} \boldsymbol{f}, \end{aligned} \tag{19}$$

where $\boldsymbol{f} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_\ell)]'$. Thus,

$$\begin{aligned} & J^{(N)}_{K,X_S} - J^{(N)}_{K,X_L} \\ &= ||f(\cdot) - P_{K,X_S} f(\cdot)||^2_{\mathcal{H}_K} \\ &\quad - ||f(\cdot) - P_{K,X_L} f(\cdot)||^2_{\mathcal{H}_K} \\ &= (||f(\cdot)||^2_{\mathcal{H}_K} - \boldsymbol{f}'_S G^+_{K,X_S} \boldsymbol{f}_S) \\ &\quad - (||f(\cdot)||^2_{\mathcal{H}_K} - \boldsymbol{f}'_L G^+_{K,X_L} \boldsymbol{f}_L) \\ &= \boldsymbol{f}'_L G^+_{K,X_L} \boldsymbol{f}_L - \boldsymbol{f}'_S G^+_{K,X_S} \boldsymbol{f}_S \end{aligned} \tag{20}$$

is obtained, where

$$\begin{aligned} \boldsymbol{f}_S &= [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_{\ell_S})]', \\ \boldsymbol{f}_L &= [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_{\ell_L})]'. \end{aligned}$$

Note that

$$\boldsymbol{f}_L \in \mathcal{R}(A_{K,X_L}) = \mathcal{R}(A_{K,X_L} A^*_{K,X_L}) = \mathcal{R}(G_{K,X_L})$$

holds since $f(\cdot) \in \mathcal{H}_K$, $\boldsymbol{f}_L = A_{K,X_L} f(\cdot)$, and

$$\begin{aligned} & A_{K,X_L} A^*_{K,X_L} \\ &= \left(\sum_{k=1}^{\ell_L} [\boldsymbol{e}_k^{(\ell_L)} \otimes K(\cdot, \boldsymbol{x}_k)] \right) \left(\sum_{j=1}^{\ell_L} [\boldsymbol{e}_j^{(\ell_L)} \otimes K(\cdot, \boldsymbol{x}_j)] \right)^* \\ &= \left(\sum_{k=1}^{\ell_L} [\boldsymbol{e}_k^{(\ell_L)} \otimes K(\cdot, \boldsymbol{x}_k)] \right) \left(\sum_{j=1}^{\ell_L} [K(\cdot, \boldsymbol{x}_j)] \otimes \boldsymbol{e}_j^{(\ell_L)} \right) \\ &= \sum_{k=1}^{\ell_L} \sum_{j=1}^{\ell_L} K(\boldsymbol{x}_k, \boldsymbol{x}_j) (\boldsymbol{e}_k^{(\ell_L)} \otimes \boldsymbol{e}_j^{(\ell_L)}) \\ &= \sum_{k=1}^{\ell_L} \sum_{j=1}^{\ell_L} K(\boldsymbol{x}_k, \boldsymbol{x}_j) \boldsymbol{e}_k^{(\ell_L)} (\boldsymbol{e}_j^{(\ell_L)})' = G_{K,X_L} \end{aligned}$$

hold. Let

$$E = \left[\begin{array}{c} I_{\ell_S} \\ O_{\ell_D, \ell_S} \end{array} \right] \in \mathbf{R}^{\ell_L \times \ell_S}, \tag{21}$$

where $\ell_D = \ell_L - \ell_S$, then

$$E' G_{K,X_L} E = G_{K,X_S}.$$

Thus, applying Lemma 3 to Eq.(20) yields

$$\begin{aligned}
&\boldsymbol{f}'_L G^+_{K,X_L} \boldsymbol{f}_L - \boldsymbol{f}'_S G^+_{K,X_S} \boldsymbol{f}_S \\
&= \boldsymbol{f}'_L G^+_{K,X_L} \boldsymbol{f}_L - \boldsymbol{f}'_L E G^+_{K,X_S} E' \boldsymbol{f}_L \\
&= \boldsymbol{f}'_L G^+_{K,X_L} \boldsymbol{f}_L - \boldsymbol{f}'_L \begin{bmatrix} G^+_{K,X_S} & O_{\ell_S,\ell_D} \\ O_{\ell_D,\ell_S} & O_{\ell_D,\ell_D} \end{bmatrix} \boldsymbol{f}_L \\
&= \boldsymbol{f}'_L \left(G^+_{K,X_L} - \begin{bmatrix} G^+_{K,X_S} & O_{\ell_S,\ell_D} \\ O_{\ell_D,\ell_S} & O_{\ell_D,\ell_D} \end{bmatrix} \right) \boldsymbol{f}_L \\
&\geq 0
\end{aligned}$$

is obtained, which concludes the proof. □

As mentioned above, the statement in Theorem 2 is rather trivial since $\mathcal{R}(A^*_{K,X_S}) \subset \mathcal{R}(A^*_{K,X_L})$. However, it is not trivial that a larger set of training samples achieves smaller squared difference at each point $\boldsymbol{x} \in \mathcal{D}$ of $f(\cdot)$ and $P_{K,X} f(\cdot)$, that is, Eq.(18) does not always mean

$$(f(\boldsymbol{x}) - P_{K,X_L} f(\boldsymbol{x}))^2 \leq (f(\boldsymbol{x}) - P_{K,X_S} f(\boldsymbol{x}))^2 \tag{22}$$

at an arbitrarily fixed point $\boldsymbol{x} \in \mathcal{D}$.

In [11], we analyzed the difference of $f(\cdot)$ and $P_{K,X} f(\cdot)$ at a point $\boldsymbol{x} \in \mathcal{D}$.

Lemma 4: [11] Let $f(\cdot) \in \mathcal{H}_K$, then

$$(f(\boldsymbol{x}) - P_{K,X} f(\boldsymbol{x}))^2 \leq ||f||^2_{\mathcal{H}_K} E_{K,X}(\boldsymbol{x}) \tag{23}$$

holds for any $\boldsymbol{x} \in \mathcal{D}$, where

$$\begin{aligned}
&E_{K,X}(\boldsymbol{x}) \\
&= K(\boldsymbol{x},\boldsymbol{x}) \\
&\quad - \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} K(\boldsymbol{x},\boldsymbol{x}_j)(G^+_{K,X})_{i,j} K(\boldsymbol{x},\boldsymbol{x}_i).
\end{aligned} \tag{24}$$

Proof: [11] Using the property $P_{K,X} = P^*_{K,X}$, the reproducing property Eq.(1), the Schwarz's inequality, and the Pythagorean theorem, we have

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

which concludes the proof. □

This lemma claims that the upper bound of the squared difference between $f(\cdot)$ and $P_{K,X} f(\cdot)$ at a point $\boldsymbol{x} \in \mathcal{D}$ is proportional to $||f(\cdot)||^2_{\mathcal{H}_K}$ and $E_{K,X}(\boldsymbol{x})$. Since a kernel K is fixed in the context of this paper, we define the relative upper bound of the squared difference of $f(\cdot)$ and $P_{K,X} f(\cdot)$ at a point $\boldsymbol{x} \in \mathcal{D}$ by $E_{K,X}(\boldsymbol{x})$. Incorporating this lemma, we give the following theorem which is the main result of this paper.

Theorem 3: For any $\boldsymbol{x} \in \mathcal{D}$,

$$E_{K,X_L}(\boldsymbol{x}) \leq E_{K,X_S}(\boldsymbol{x}) \tag{25}$$

holds.

Proof: Let

$$\boldsymbol{k}_L = [K(\boldsymbol{x},\boldsymbol{x}_1), \ldots, K(\boldsymbol{x},\boldsymbol{x}_{\ell_L})]'$$

then

$$\boldsymbol{k}_L \in \mathcal{R}(A_{K,X_L}) = \mathcal{R}(A_{K,X_L} A^*_{K,X_L}) = \mathcal{R}(G_{K,X_L})$$

since $K(\boldsymbol{x},\cdot) \in \mathcal{H}_K$ and $\boldsymbol{k}_L = A_{K,X_L} K(\boldsymbol{x},\cdot)$ holds for an arbitrarily fixed $\boldsymbol{x} \in \mathcal{D}$. Let

$$\boldsymbol{k}_S = [K(\boldsymbol{x},\boldsymbol{x}_1), \ldots, K(\boldsymbol{x},\boldsymbol{x}_{\ell_S})]',$$

then $\boldsymbol{k}_S$ can be written as

$$\boldsymbol{k}_S = E' \boldsymbol{k}_L$$

with E in Eq.(21). Thus,

$$\begin{aligned}
&E_{K,X_S}(\boldsymbol{x}) - E_{K,X_L}(\boldsymbol{x}) \\
&= \left(K(\boldsymbol{x},\boldsymbol{x}) - \sum_{i=1}^{\ell_S} \sum_{j=1}^{\ell_S} K(\boldsymbol{x},\boldsymbol{x}_j)(G^+_{K,X_S})_{i,j} K(\boldsymbol{x},\boldsymbol{x}_i) \right) \\
&\quad - \left(K(\boldsymbol{x},\boldsymbol{x}) - \sum_{i=1}^{\ell_L} \sum_{j=1}^{\ell_L} K(\boldsymbol{x},\boldsymbol{x}_j)(G^+_{K,X_L})_{i,j} K(\boldsymbol{x},\boldsymbol{x}_i) \right) \\
&= \boldsymbol{k}'_L G^+_{K,X_L} \boldsymbol{k}_L - \boldsymbol{k}'_S G^+_{K,X_S} \boldsymbol{k}_S \\
&= \boldsymbol{k}'_L G^+_{K,X_L} \boldsymbol{k}_L - \boldsymbol{k}'_L E G^+_{K,X_S} E' \boldsymbol{k}_L \\
&= \boldsymbol{k}'_L \left(G^+_{K,X_L} - \begin{bmatrix} G^+_{K,X_S} & O_{\ell_S,\ell_D} \\ O_{\ell_D,\ell_S} & O_{\ell_D,\ell_D} \end{bmatrix} \right) \boldsymbol{k}_L \\
&\geq 0
\end{aligned}$$

is obtained from Lemma 3, which concludes the proof. □

According to Theorem 3, it is guaranteed that the upper bound of the squared difference at each point $\boldsymbol{x} \in \mathcal{D}$ of $f(\cdot)$ and $P_{K,X_L} f(\cdot)$ is not greater than that of $f(\cdot)$ and $P_{K,X_S} f(\cdot)$.

Since the range of $E_{K,X}(\boldsymbol{x})$ depends on $\boldsymbol{x} \in \mathcal{D}$, it is significant to consider the normalized version of it. $E_{K,X}(\boldsymbol{x})$ is non-negative since it is the squared norm of $K(\cdot,\boldsymbol{x}) - P_{K,X} K(\cdot,\boldsymbol{x})$. Moreover,

$$\sum_{i=1}^{\ell} \sum_{j=1}^{\ell} K(\boldsymbol{x},\boldsymbol{x}_j)(G^+_{K,X})_{i,j} K(\boldsymbol{x},\boldsymbol{x}_i)$$

in Eq.(24) is also non-negative since it is a quadratic form with the *n.n.d.* matrix $G^+_{K,X}$. Therefore, we have

$$K(\boldsymbol{x},\boldsymbol{x}) \geq \sum_{i=1}^{\ell}\sum_{j=1}^{\ell} K(\boldsymbol{x},\boldsymbol{x}_j)(G^+_{K,X})_{i,j}K(\boldsymbol{x},\boldsymbol{x}_i) \geq 0. \quad (26)$$

Note that when $K(\boldsymbol{x},\boldsymbol{x}) = 0$, $E_{K,X}(\boldsymbol{x})$ is necessarily reduced to 0. When $K(\boldsymbol{x},\boldsymbol{x}) \neq 0$, we also have

$$1 \geq \frac{1}{K(\boldsymbol{x},\boldsymbol{x})}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell} K(\boldsymbol{x},\boldsymbol{x}_j)(G^+_{K,X})_{i,j}K(\boldsymbol{x},\boldsymbol{x}_i) \geq 0 \quad (27)$$

from Eq.(26).

Accordingly, we define the normalized point-wise generalization error of $\mathcal{R}(A^*_{K,X})$ at a point $\boldsymbol{x} \in \mathcal{D}$ by

$$J^{(P)}_{K,X}(\boldsymbol{x}) = \begin{cases} \dfrac{E_{K,X}(\boldsymbol{x})}{K(\boldsymbol{x},\boldsymbol{x})} & : \quad (K(\boldsymbol{x},\boldsymbol{x}) \neq 0) \\ 0 & : \quad (K(\boldsymbol{x},\boldsymbol{x}) = 0) \end{cases} \quad (28)$$

and it satisfies

$$0 \leq J^{(P)}_{K,X}(\boldsymbol{x}) \leq 1 \quad (29)$$

for any $\boldsymbol{x} \in \mathcal{D}$. When $J^{(P)}_{K,X}(\boldsymbol{x})$ is close to 0, the generalization error at $\boldsymbol{x} \in \mathcal{D}$ is small; and when $J^{(P)}_{K,X}(\boldsymbol{x})$ is close to 1, the generalization error at $\boldsymbol{x} \in \mathcal{D}$ is close to its upper bound. For $J^{(P)}_{K,X}(\boldsymbol{x})$, we have the following corollary concerned with a set of training samples.

Corollary 1:

$$J^{(P)}_{K,X_L}(\boldsymbol{x}) \leq J^{(P)}_{K,X_S}(\boldsymbol{x}) \quad (30)$$

holds for any $\boldsymbol{x} \in \mathcal{D}$.

Proof: It is trivial in case of $K(\boldsymbol{x},\boldsymbol{x}) = 0$.
When $K(\boldsymbol{x},\boldsymbol{x}) \neq 0$, for an arbitrarily fixed $\boldsymbol{x} \in \mathcal{D}$,

$$\begin{aligned} & J^{(P)}_{K,X_S}(\boldsymbol{x}) - J^{(P)}_{K,X_L}(\boldsymbol{x}) \\ &= \frac{1}{K(\boldsymbol{x},\boldsymbol{x})}E_{K,X_S}(\boldsymbol{x}) - \frac{1}{K(\boldsymbol{x},\boldsymbol{x})}E_{K,X_L}(\boldsymbol{x}) \\ &= \frac{1}{K(\boldsymbol{x},\boldsymbol{x})}(E_{K,X_S}(\boldsymbol{x}) - E_{K,X_L}(\boldsymbol{x})) \end{aligned}$$

is obtained. Thus, Theorem 3 and the fact that $K(\boldsymbol{x},\boldsymbol{x}) > 0$ immediately concludes the proof. □

We can evaluate $J^{(P)}_{K,X}(\boldsymbol{x})$ at the point which yields the worst generalization error instead of each point $\boldsymbol{x} \in \mathcal{D}$ by

$$J^{(WP)}_{K,X} = \sup_{\boldsymbol{x}\in\mathcal{D}} J^{(P)}_{K,X}(\boldsymbol{x}). \quad (31)$$

We have the following corollary for $J^{(WP)}_{K,X}$ as the same with $J^{(P)}_{K,X}(\boldsymbol{x})$.

Corollary 2:

$$J^{(WP)}_{K,X_L} \leq J^{(WP)}_{K,X_S}. \quad (32)$$

Proof: Let

$$\boldsymbol{x}_L = \arg\sup_{\boldsymbol{x}\in\mathcal{D}} J^{(P)}_{K,X_L}(\boldsymbol{x}).$$

then Corollary 1 yields

$$J^{(P)}_{K,X_L}(\boldsymbol{x}_L) \leq J^{(P)}_{K,X_S}(\boldsymbol{x}_L).$$

On the other hand, it is trivial that

$$J^{(P)}_{K,X_S}(\boldsymbol{x}_L) \leq \sup_{\boldsymbol{x}\in\mathcal{D}} J^{(P)}_{K,X_S}(\boldsymbol{x}),$$

which concludes the proof. □

In the norm-based generalization error, the norm of the unknown true function plays a crucial role since

$$J^{(N)}_{K,X} = ||f||^2_{\mathcal{H}_K} - \boldsymbol{f}' G^+_{K,X}\boldsymbol{f}$$

as shown in the previous section. Thus, $J^{(N)}_{K,X}$ is rather meaningless for evaluation of the generalization ability of $\mathcal{R}(A^*_{K,X})$ itself.

On the other hand in $E_{K,X}(\boldsymbol{x})$, the norm of the unknown true function appears as a coefficient. Thus, its influence to the generalization error can be vanished by normalization as in $J^{(P)}_{K,X}(\boldsymbol{x})$ (and in $J^{(WP)}_{K,X}$), which enables us to have a quantitative goodness of $\mathcal{R}(A^*_{K,X})$ which is in the interval between the worst case (upper bound) and the best case. This is an essential value of considering $J^{(P)}_{K,X}(\boldsymbol{x})$ or $J^{(WP)}_{K,X}$.

V. Numerical Examples

In this section, we confirm the theoretical results in the previous section by numerical examples for 1-dimensional function estimation. We adopt the Gaussian kernel given by

$$K(x,y) = \exp\left(-\frac{(x-y)^2}{4}\right), \quad x,y \in [0,16] \in \mathbf{R} \quad (33)$$

as a kernel and

$$f(\cdot) = \sum_{k=1}^{100} c_i K(\cdot, x_i) \in \mathcal{H}_K \quad (34)$$

as an unknown true function, where x_i is randomly generated from the uniform distribution on $[0,16]$ written as $U(0,16)$ and c_i is randomly generated from the standard normal distribution $N(0,1)$. Coefficient c_i is normalized so that $||f(\cdot)||^2_{\mathcal{H}_K} = 1$. As a smaller and a larger sets of training samples, we adopt

$$\begin{aligned} X_S &= \{x_i \sim U(0,16) \mid i \in \{1,\ldots,10\}\}, \\ X_L &= X_S \cup X_E \end{aligned}$$

where $X_E = \{x_i \sim U(0,16) \mid i \in \{1,\ldots,5\}\}$. Figure 1 shows $f(\cdot)$, $P_{K,X_S}f(\cdot)$, $P_{K,X_L}f(\cdot)$, and training samples in which points in X_S are denoted by '$\times$' and those in X_E are denoted by '+'.

Figure 2 shows $J^{(P)}_{K,X_S}$ and $J^{(P)}_{K,X_L}$ with respect to $x \in [0,16]$.

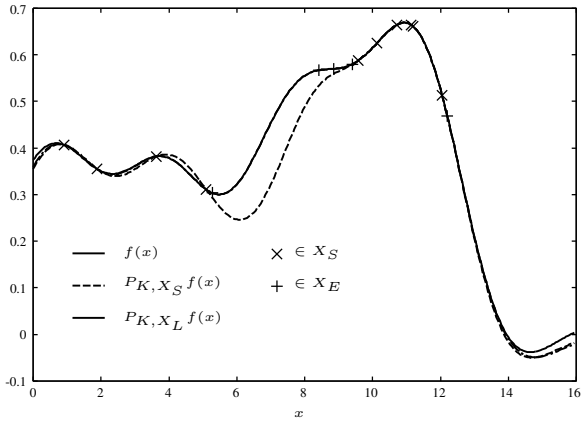

Fig. 1. An unknown true function, sampling points, and its orthogonal projections by P_{K,X_S} and P_{K,X_L}.

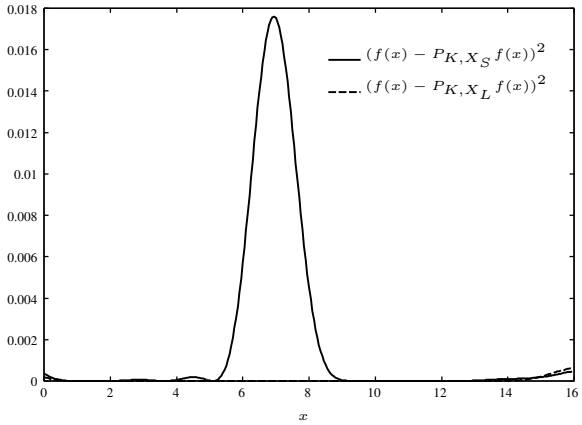

Fig. 3. Squared difference of $f(x)$ and $P_{K,X} f(x)$ with X_S and X_L.

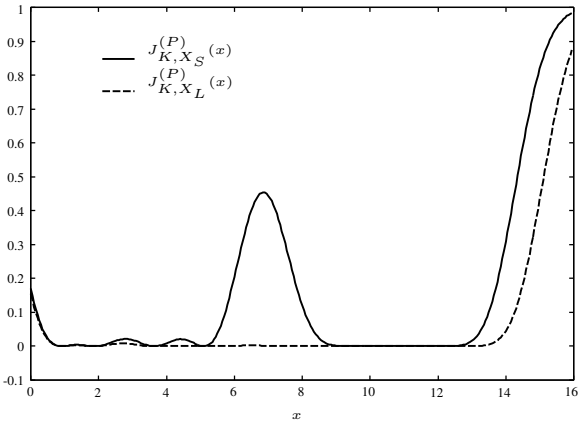

Fig. 2. $J^{(P)}_{K,X_L}$ and $J^{(P)}_{K,X_S}$ with respect to x.

According to this result, it is confirmed that the upper bound of the squared difference at each point in $\mathcal{D}$ with a larger set of training samples is smaller than that with a smaller set of training samples, which agrees with the theoretical results obtained in the previous section.

Figure 3 shows the squared difference of $f(\cdot)$ and $P_{K,X} f(\cdot)$ with X_S and X_L.

According to this result, there exist points $\boldsymbol{x} \in \mathcal{D}$, where the squared difference of $f(\boldsymbol{x})$ and $P_{K,X_L} f(\boldsymbol{x})$ is larger than that of $f(\boldsymbol{x})$ and $P_{K,X_S} f(\boldsymbol{x})$. This is a counter example for Eq.(22). Thus, our evaluation based on upper bound of the squared difference seems to be crucial.

VI. Conclusions

In this paper, we theoretically analyzed the generalization ability of a model space in kernel machines with respect to a set of training samples; and showed that a larger set of training samples achieves a smaller upper bound of the squared difference of the unknown true function and the orthogonal projection of it onto the model space compared with a smaller set of training samples. We also gave numerical examples to confirm our theoretical results. Applying our result to practical model selection is one of future works that should be resolved.

Acknowledgments

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] M. Sugiyama and H. Ogawa, "Subspace Information Criterion for Model Selection," *Neural Computation*, vol. 13, no. 8, pp. 1863–1889, 2001.

[4] M. Sugiyama, M. Kawanabe, and K. Muller, "Trading variance reduction with unbiasedness: The regularized subspace information criterion for robust model selection in kernel regression," *Neural Computation*, vol. 16, no. 5, pp. 1077–1104, 2004.

[5] 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.

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

[7] J. Mercer, "Functions of Positive and Negative Type and Their Connection with The Theory of Integral Equations," *Transactions of the London Philosophical Society*, vol. A, no. 209, pp. 415–446, 1909.

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

[9] H. Ogawa, "What Can we See behind Sampling Theorems?" *IEICE Transactions on Fundamentals*, vol. E92-A, no. 3, pp. 688–707, 2009.

[10] C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and its Applications*. John Wiley & Sons, 1971.

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