Optimal Kernel in a Class of Kernels with an Invariant Metric

Akira Tanaka, Hideyuki Imai, Mineichi Kudo, and Masaaki Miyakoshi

Division of Computer Science,
Graduate School of Information Science and Technology, Hokkaido University,
Sapporo, 060-0814, Japan
{takira,imai,mine,miyakosi}@main.ist.hokudai.ac.jp

Abstract. Learning based on kernel machines is widely known as a powerful tool for various fields of information science such as pattern recognition and regression estimation. One of central topics of kernel machines is model selection, especially selection of a kernel or its parameters. In this paper, we consider a class of kernels that forms a monotonic classes of reproducing kernel Hilbert spaces with an invariant metric and show that the kernel corresponding to the smallest reproducing kernel Hilbert space including an unknown true function gives the optimal model for the unknown true function.

1 Introduction

Learning based on kernel machines[1], represented by the support vector machine [2] and the kernel ridge regression[3,4], is widely known as a powerful tool for various fields of information science such as pattern recognition, regression estimation, and density estimation. In general, appropriate model selection is required in order to obtain a desirable learning result by kernel machines. There exists two classes of model selection. One is selection of a model space to which a learning machine belongs. The other is selection of a learning machine in a fixed model space. The latter, such as selection of a regularization parameter under a fixed kernel, is sufficiently investigated in terms of theoretical and practical senses (See [5,6] for instance). On the other hand, the former, that is, selection of a kernel or its parameters, is not discussed sufficiently in terms of theoretical sense although practical algorithms for selection of a kernel (or its parameters) such as cross-validation are revealed. The difficulty of the theoretical analyses for selection of a kernel lies on the fact that the metrics of two reproducing kernel Hilbert spaces (RKHS)[7,8] corresponding to two different kernels may differ in general, which means that we do not have a unified framework to evaluate learning results obtained by different kernels. In order to avoid this difficulty, we considered a class of kernels that forms a monotonic classes of functions with an invariant metric and introduced a method for selecting an optimal kernel from the class of kernels in the previous work[9]. However, mathematical properties of the selected kernel or the corresponding RKHS are not discussed sufficiently in [9].

In this paper, we consider a class of kernels that forms a monotonic classes of functions with an invariant metric as the same with [9] and show that the kernel corresponding to the smallest RKHS including an unknown true function gives the optimal model for the true function among the class of kernels.

2 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[7,8].

Definition 1. [7] 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 $K(\mathbf{x}, \tilde{\mathbf{x}})$, $(\mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{D})$ is called a reproducing kernel of $\mathcal{H}$, if

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

$$f(\tilde{\mathbf{x}}) = \langle f(\mathbf{x}), K(\mathbf{x}, \tilde{\mathbf{x}}) \rangle_{\mathcal{H}}, \quad (1)$$

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 [7]:

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

for any N , $c_1, \dots, c_N \in \mathbf{R}$, and $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathcal{D}$. In addition, $K(\mathbf{x}, \tilde{\mathbf{x}}) = K(\tilde{\mathbf{x}}, \mathbf{x})$ for any $\mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{D}$ is followed[7]. If a reproducing kernel $K(\mathbf{x}, \tilde{\mathbf{x}})$ exists, it is unique[7]. Conversely, every positive definite function $K(\mathbf{x}, \tilde{\mathbf{x}})$ has the unique corresponding RKHS [7].

Next, we introduce the Schatten product [10] that is a convenient tool to reveal the reproducing property of kernels.

Definition 2. [10] 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

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

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 $h, v \in \mathcal{H}_1$, $g, u \in \mathcal{H}_2$.

$$(h \otimes g)^* = (g \otimes h), \quad (h \otimes g)(u \otimes v) = \langle u, g \rangle_{\mathcal{H}_2} (h \otimes v), \quad (4)$$

where the superscript $*$ denotes the adjoint operator.

3 Formulation of Learning Problems

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

$$y_i = f(\mathbf{x}_i) + n_i, \quad (5)$$

where f denotes the unknown true 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 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.(5) is rewritten as

$$y_i = \langle f(\mathbf{x}), K(\mathbf{x}, \mathbf{x}_i) \rangle_{\mathcal{H}_K} + n_i, \quad (6)$$

on the basis of the reproducing property of kernels. Let $\mathbf{y} = [y_1, \dots, y_\ell]'$ and $\mathbf{n} = [n_1, \dots, n_\ell]'$ with the superscript $'$ denoting the transposed matrix (or vector), then applying the Schatten product to Eq.(6) yields

$$\mathbf{y} = \left(\sum_{k=1}^{\ell} [\mathbf{e}_k^{(\ell)} \otimes K(\mathbf{x}, \mathbf{x}_k)] \right) f(\mathbf{x}) + \mathbf{n}, \quad (7)$$

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

$$A_K = \left(\sum_{k=1}^{\ell} [\mathbf{e}_k^{(\ell)} \otimes K(\mathbf{x}, \mathbf{x}_k)] \right). \quad (8)$$

The operator A_K is linear one that maps an element of $\mathcal{H}_K$ onto $\mathbf{R}^\ell$ and Eq.(7) can be written by

$$\mathbf{y} = A_K f(\mathbf{x}) + \mathbf{n}, \quad (9)$$

which represents the relation between the unknown true function f and an output vector $\mathbf{y}$. The information of input vectors is integrated in the operator A_K . Therefore, a machine learning problem can be interpreted as an inversion problem of the linear equation Eq.(9)[11]. In general, an estimated function $\hat{f}(\mathbf{x})$ is represented as

$$\hat{f}(\mathbf{x}) = L\mathbf{y}, \quad (10)$$

where L denotes a learning operator specified by a learning criterion such as those of the support vector machines and kernel ridge regression.

4 Kernel Specific Generalization Ability

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

$$\hat{f}(\mathbf{x}) = A_K^* \boldsymbol{\alpha} = \left(\sum_{k=1}^{\ell} [K(\mathbf{x}, \mathbf{x}_k) \otimes \mathbf{e}_k^{(\ell)}] \right) \boldsymbol{\alpha} = \sum_{k=1}^{\ell} \alpha_k K(\mathbf{x}, \mathbf{x}_k) \quad (11)$$

holds, where $\alpha = [\alpha_1, \dots, \alpha_\ell]'$ denotes an arbitrary vector in $\mathbf{R}^\ell$. The point at issue of this paper is to discuss selection of a model space, that is, the generalization ability of $\mathcal{R}(A_K^*)$ which is independent from criteria of learning machines. Therefore, we define the generalization ability of kernel machines specified by a kernel K as the distance between the unknown true function f and $\mathcal{R}(A_K^*)$ written as

$$J(K, f) = \|f - P_K f\|_{\mathcal{H}_K}^2, \quad (12)$$

where P_K denotes the orthogonal projector onto $\mathcal{R}(A_K^*)$ and $\|\cdot\|_{\mathcal{H}_K}$ denotes the induced norm of $\mathcal{H}_K$. Note that the orthogonality of P_K is also defined by the metric of $\mathcal{H}_K$. Selection of an element in $\mathcal{R}(A_K^*)$ as a learning result is out of the scope of this paper since the selection depends on learning criteria. We also ignore the observation noise in the following contents since the noise does not affect Eq.(12).

Here, we give some propositions as preparations to evaluate Eq.(12).

Lemma 1

$$P_K = \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} (G_K^+)_{i,j} [K(\mathbf{x}, \mathbf{x}_i) \otimes K(\mathbf{x}, \mathbf{x}_j)], \quad (13)$$

where G_K denotes Gram's matrix of K with input vectors $\mathbf{x}_i$, ($i = 1, \dots, \ell$) and the superscript $+$ denotes the Moore-Penrose generalized inverse[12].

Proof. The orthogonal projector onto $\mathcal{R}(A_K^*)$ is given by $P_K = A_K^+ A_K$ and $A_K^+ = A_K^* (A_K A_K^*)^+$ holds. Since

$$\begin{aligned} & A_K A_K^* \\ &= \left(\sum_{i=1}^{\ell} [e_i^{(\ell)} \otimes K(\mathbf{x}, \mathbf{x}_i)] \right) \left(\sum_{i=1}^{\ell} [e_i^{(\ell)} \otimes K(\mathbf{x}, \mathbf{x}_i)] \right)^* \\ &= \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} [e_i^{(\ell)} \otimes K(\mathbf{x}, \mathbf{x}_i)] [K(\mathbf{x}, \mathbf{x}_j) \otimes e_j^{(\ell)}] \\ &= \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \langle K(\mathbf{x}, \mathbf{x}_j), K(\mathbf{x}, \mathbf{x}_i) \rangle_{\mathcal{H}_K} (e_i^{(\ell)} \otimes e_j^{(\ell)}) \\ &= \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} K(\mathbf{x}_i, \mathbf{x}_j) (e_i^{(\ell)} e_j^{(\ell)*}) \end{aligned}$$

is reduced to G_K , the concrete form of P_K is written as

$$\begin{aligned} P_K &= A_K^+ A_K = A_K^* (A_K A_K^*)^+ A_K \\ &= \left(\sum_{i=1}^{\ell} [e_i^{(\ell)} \otimes K(\mathbf{x}, \mathbf{x}_i)] \right)^* G_K^+ \left(\sum_{i=1}^{\ell} [e_i^{(\ell)} \otimes K(\mathbf{x}, \mathbf{x}_i)] \right) \\ &= \left(\sum_{i=1}^{\ell} [K(\mathbf{x}, \mathbf{x}_i) \otimes e_i^{(\ell)}] \right) G_K^+ \left(\sum_{i=1}^{\ell} [e_i^{(\ell)} \otimes K(\mathbf{x}, \mathbf{x}_i)] \right) \end{aligned}$$

$$= \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} (G_K^+)_{i,j} [K(\mathbf{x}, \mathbf{x}_i) \otimes K(\mathbf{x}, \mathbf{x}_j)],$$

which concludes the proof. $\square$

Lemma 2. For any $f \in \mathcal{H}_K$ and input vectors $\mathbf{x}_i$, ($i \in \{1, \dots, \ell\}$),

$$\|P_K f\|_{\mathcal{H}_K}^2 = \mathbf{f}' G_K^+ \mathbf{f} \quad (14)$$

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

Proof. Since $f \in \mathcal{H}_K$,

$$\begin{aligned} P_K f &= \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} (G_K^+)_{i,j} [K(\mathbf{x}, \mathbf{x}_i) \otimes K(\mathbf{x}, \mathbf{x}_j)] f \\ &= \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} (G_K^+)_{i,j} f(\mathbf{x}_j) K(\mathbf{x}, \mathbf{x}_i) \end{aligned}$$

holds. Thus

$$\begin{aligned} \|P_K f\|_{\mathcal{H}_K}^2 &= \left\langle \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} (G_K^+)_{i,j} f(\mathbf{x}_j) K(\mathbf{x}, \mathbf{x}_i), \sum_{k=1}^{\ell} \sum_{l=1}^{\ell} (G_K^+)_{k,l} f(\mathbf{x}_l) K(\mathbf{x}, \mathbf{x}_k) \right\rangle_{\mathcal{H}_K} \\ &= \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \sum_{k=1}^{\ell} \sum_{l=1}^{\ell} (G_K^+)_{i,j} (G_K^+)_{k,l} f(\mathbf{x}_j) f(\mathbf{x}_l) \langle K(\mathbf{x}, \mathbf{x}_i), K(\mathbf{x}, \mathbf{x}_k) \rangle_{\mathcal{H}_K} \\ &= \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \sum_{k=1}^{\ell} \sum_{l=1}^{\ell} f(\mathbf{x}_j) (G_K^+)_{j,i} (G_K^+)_{i,k} (G_K^+)_{k,l} f(\mathbf{x}_l) \\ &= \mathbf{f}' G_K^+ G_K G_K^+ \mathbf{f} = \mathbf{f}' G_K^+ \mathbf{f} \end{aligned}$$

holds since G_K is symmetric. $\square$

5 Optimal Kernel in a Class of Kernels Forming a Monotonic Classes of Functions with an Invariant Metric

Let us consider a monotonic classes of functions $\mathcal{H}_{K_i}$, ($i \in \{1, \dots, N\}$) written as

$$\mathcal{H}_{K_1} \subset \mathcal{H}_{K_2} \subset \dots \subset \mathcal{H}_{K_N}, \quad (15)$$

forming Hilbert spaces specified by a class of kernels K_i , ($i \in \{1, \dots, N\}$) with an invariant metric satisfying

$$\langle f, g \rangle_{\mathcal{H}_{K_1}} = \langle f, g \rangle_{\mathcal{H}_{K_2}} = \dots = \langle f, g \rangle_{\mathcal{H}_{K_N}} \quad (16)$$

for any $f, g \in \mathcal{H}_{K_1}$. Note that

$$\|f\|_{\mathcal{H}_{K_1}}^2 = \|f\|_{\mathcal{H}_{K_2}}^2 = \cdots = \|f\|_{\mathcal{H}_{K_N}}^2 \quad (17)$$

trivially holds for any $f \in \mathcal{H}_{K_1}$ from Eq.(16).

Theorem 1. [7] *If K is the reproducing kernel of the class F with the norm $\|\cdot\|$, and if the linear class $F_1 \subset F$ forms a Hilbert space with the norm $\|\cdot\|_1$, such that $\|f\|_1 \geq \|f\|$ for any $f \in F_1$, then the class F_1 possesses a reproducing kernel K_1 such that $K^c = K - K_1$ is also a reproducing kernel.*

According to Theorem 1, there exists a kernel K_j^c such that

$$K_j = K_1 + K_j^c,$$

for any $j > 1$ since $\|f\|_{\mathcal{H}_{K_j}} = \|f\|_{\mathcal{H}_{K_1}}$ for any $f \in \mathcal{H}_{K_1}$. Note that corresponding Gram's matrices satisfy

$$G_{K_j} = G_{K_1} + G_{K_j^c}$$

for a set of input vectors.

Lemma 3. [12] *Let H_1 and H_2 be non-negative definite (n.n.d.) Hermitian matrices, then there exists a non-singular matrix T that jointly diagonalizes H_1 and H_2 written as*

$$TH_1T^* = \Lambda_1, \quad TH_2T^* = \Lambda_2,$$

where Λ_1 and Λ_2 denote diagonal matrices.

Lemma 4. *Let H_1 and H_2 be n.n.d. Hermitian matrices and let T be a non-singular joint diagonalizer of H_1 and H_2 written as*

$$TH_1T^* = \Lambda_1, \quad TH_2T^* = \Lambda_2,$$

where Λ_1 and Λ_2 denote diagonal matrices. The matrix $T^*(\Lambda_1 + \Lambda_2)^+T$ is generalized inverse matrix of $H_1 + H_2$.

Proof. Since $H_1 + H_2 = T^{-1}(\Lambda_1 + \Lambda_2)(T^*)^{-1}$,

$$\begin{aligned} & (H_1 + H_2)(T^*(\Lambda_1 + \Lambda_2)^+T)(H_1 + H_2) \\ &= (T^{-1}(\Lambda_1 + \Lambda_2)(T^*)^{-1})(T^*(\Lambda_1 + \Lambda_2)^+T)(T^{-1}(\Lambda_1 + \Lambda_2)(T^*)^{-1}) \\ &= T^{-1}(\Lambda_1 + \Lambda_2)(\Lambda_1 + \Lambda_2)^+(\Lambda_1 + \Lambda_2)(T^*)^{-1} \\ &= T^{-1}(\Lambda_1 + \Lambda_2)(T^*)^{-1} = H_1 + H_2 \end{aligned}$$

holds, which concludes the proof. $\square$

Lemma 5. *Let H_1 and H_2 be n.n.d. Hermitian matrices, then $H_1(H_1 + H_2)^-H_2$ is invariant for any $(H_1 + H_2)^-$, where the superscript $-$ denotes a generalized inverse[12].*

Proof. Let $B = H_1 + H_2$ and let B_1^- and B_2^- be different generalized inverse matrices of B , then

$$B_2^- = B_1^- + Y - B_1^- BYBB_1^- \quad (18)$$

holds with a certain matrix Y [12]. Thus,

$$\begin{aligned} & H_1 B_1^- H_2 - H_1 B_2^- H_2 \\ &= H_1 B_1^- H_2 - H_1 (B_1^- + Y - B_1^- BYBB_1^-) H_2 \\ &= -H_1 Y H_2 + H_1 B_1^- BYBB_1^- H_2 \\ &= -H_1 Y H_2 + H_1 Y H_2 = O \end{aligned}$$

holds since $\mathcal{R}(H_i) \subset \mathcal{R}(B)$, ($i = 1, 2$). $\square$

Theorem 2. Let H_1 and H_2 be n.n.d. Hermitian matrices and let $\mathbf{y} \in \mathcal{R}(H_1)$, then

$$J = \mathbf{y}^* (H_1^+ - (H_1 + H_2)^+) \mathbf{y} \geq 0 \quad (19)$$

holds.

Proof. Since $\mathbf{y} \in \mathcal{R}(H_1)$, there exists $\mathbf{z}$ such that $\mathbf{y} = H_1 \mathbf{z}$. Thus,

$$\begin{aligned} J &= \mathbf{y}^* (H_1^+ - (H_1 + H_2)^+) \mathbf{y} \\ &= \mathbf{z}^* H_1 (H_1^+ - (H_1 + H_2)^+) H_1 \mathbf{z} \\ &= \mathbf{z}^* H_1 \mathbf{z} - \mathbf{z}^* H_1 (H_1 + H_2)^+ H_1 \mathbf{z} \end{aligned} \quad (20)$$

$$\begin{aligned} &= \mathbf{z}^* H_1 \mathbf{z} - \mathbf{z}^* (H_1 + H_2 - H_2) (H_1 + H_2)^+ (H_1 + H_2 - H_2) \mathbf{z} \\ &= \mathbf{z}^* H_1 \mathbf{z} - \mathbf{z}^* (H_1 + H_2) \mathbf{z} - \mathbf{z}^* H_2 (H_1 + H_2)^+ H_2 \mathbf{z} + 2\mathbf{z}^* H_2 \mathbf{z} \\ &= \mathbf{z}^* H_2 \mathbf{z} - \mathbf{z}^* H_2 (H_1 + H_2)^+ H_2 \mathbf{z}. \end{aligned} \quad (21)$$

From Eqs.(20) and (21), we have

$$\begin{aligned} & 2J \\ &= \mathbf{z}^* (H_1 + H_2) \mathbf{z} \\ &\quad - \mathbf{z}^* H_1 (H_1 + H_2)^+ H_1 \mathbf{z} - \mathbf{z}^* H_2 (H_1 + H_2)^+ H_2 \mathbf{z} \\ &= \mathbf{z}^* (H_1 + H_2) \mathbf{z} \\ &\quad - \mathbf{z}^* [(H_1 + H_2)(H_1 + H_2)^+ (H_1 + H_2) \\ &\quad \quad - H_1 (H_1 + H_2)^+ H_2 - H_2 (H_1 + H_2)^+ H_1] \mathbf{z} \\ &= \mathbf{z}^* (H_1 + H_2) \mathbf{z} \\ &\quad - \mathbf{z}^* [(H_1 + H_2) - H_1 (H_1 + H_2)^+ H_2 - H_2 (H_1 + H_2)^+ H_1] \mathbf{z} \\ &= \mathbf{z}^* [H_1 (H_1 + H_2)^+ H_2 + H_2 (H_1 + H_2)^+ H_1] \mathbf{z}. \end{aligned} \quad (22)$$

From Lemma 3, there exists a non-singular matrix T such that $TH_1T^* = \Lambda_1$ and $TH_2T^* = \Lambda_2$. From Lemmas 4 and 5, Eq.(22) is reduced to

$$\begin{aligned} 2J &= \mathbf{z}^* [H_1 (H_1 + H_2)^+ H_2 + H_2 (H_1 + H_2)^+ H_1] \mathbf{z} \\ &= 2\mathbf{z}^* [T^{-1} \Lambda_1 (\Lambda_1 + \Lambda_2)^+ \Lambda_2 (T^*)^{-1}] \mathbf{z} \\ &= 2\mathbf{z}^* T^{-1} [\Lambda_1 (\Lambda_1 + \Lambda_2)^+ \Lambda_2] (T^{-1})^* \mathbf{z}. \end{aligned}$$

On the basis of Sylvester's law of inertia, Λ_1 and Λ_2 are *n.n.d.*, which immediately yields $J \geq 0$. $\square$

The following theorem is the main result of this paper.

Theorem 3. *For any $f \in \mathcal{H}_{K_1}$ and input vectors $\mathbf{x}_i$, ($i \in \{1, \dots, \ell\}$),*

$$\|f - P_{K_1} f\|_{\mathcal{H}_{K_N}}^2 \leq \|f - P_{K_j} f\|_{\mathcal{H}_{K_N}}^2 \quad (23)$$

holds for any $1 < j \leq N$.

Proof. For any $j \in \{1, \dots, N\}$ and any $f \in \mathcal{H}_{K_1}$,

$$\begin{aligned} \|f - P_{K_j} f\|_{\mathcal{H}_{K_N}}^2 &= \|f - P_{K_j} f\|_{\mathcal{H}_{K_j}}^2 \\ &= \langle f, f \rangle_{\mathcal{H}_{K_j}} + \langle P_{K_j} f, P_{K_j} f \rangle_{\mathcal{H}_{K_j}} - 2\langle f, P_{K_j} f \rangle_{\mathcal{H}_{K_j}} \\ &= \langle f, f \rangle_{\mathcal{H}_{K_j}} + \langle P_{K_j} f, P_{K_j} f \rangle_{\mathcal{H}_{K_j}} - 2\langle f, P_{K_j}^2 f \rangle_{\mathcal{H}_{K_j}} \\ &= \langle f, f \rangle_{\mathcal{H}_{K_j}} + \langle P_{K_j} f, P_{K_j} f \rangle_{\mathcal{H}_{K_j}} - 2\langle P_{K_j} f, P_{K_j} f \rangle_{\mathcal{H}_{K_j}} \\ &= \langle f, f \rangle_{\mathcal{H}_{K_j}} - \langle P_{K_j} f, P_{K_j} f \rangle_{\mathcal{H}_{K_j}} \\ &= \langle f, f \rangle_{\mathcal{H}_{K_N}} - \mathbf{f}' G_{K_j}^+ \mathbf{f}, \end{aligned}$$

from Lemma 2 and the fact that P_{K_j} is an orthogonal projector. It is obvious that $\mathbf{f} \in \mathcal{R}(A_{K_1}) = \mathcal{R}(A_{K_1} A_{K_1}^*) = \mathcal{R}(G_{K_1})$. Therefore, Theorems 1 and 2 yield

$$\begin{aligned} \|f - P_{K_j} f\|_{\mathcal{H}_{K_N}}^2 - \|f - P_{K_1} f\|_{\mathcal{H}_{K_N}}^2 &= (\langle f, f \rangle_{\mathcal{H}_{K_N}} - \mathbf{f}' G_{K_j}^+ \mathbf{f}) - (\langle f, f \rangle_{\mathcal{H}_{K_N}} - \mathbf{f}' G_{K_1}^+ \mathbf{f}) \\ &= \mathbf{f}' (G_{K_1}^+ - G_{K_j}^+) \mathbf{f} \\ &= \mathbf{f}' (G_{K_1}^+ - (G_{K_1} + G_{K_j^c})^+) \mathbf{f} \geq 0, \end{aligned}$$

which concludes the proof. $\square$

According to Theorem 3, given a class of kernels that forms a monotonic classes of functions with an invariant metric, it is concluded that the kernel corresponding to the smallest class of functions gives the best generalization ability among the given class of kernels.

6 Discussions

One of classes of kernels whose corresponding RKHSs have an invariant metric is the sinc kernel written as

$$K^\alpha(\mathbf{x}, \tilde{\mathbf{x}}) = \frac{\sin \alpha \|\mathbf{x} - \tilde{\mathbf{x}}\|}{\pi \|\mathbf{x} - \tilde{\mathbf{x}}\|},$$

where $\alpha > 0$ denotes the parameter. In [9], we proposed a method of selecting the best kernel from a class of kernels forming a monotonic classes of functions

with an invariant metric and we observed that the RKHS corresponding to the selected kernel by the method was the smallest class of functions including the unknown true function in the numerical experiment using the sinc kernel. This result is given its theoretical ground in this paper.

As is well known, the best kernel that perfectly reconstructs the unknown true function f with a finite (at least one) number of training samples, is

$$K^P(\mathbf{x}, \tilde{\mathbf{x}}) = f(\mathbf{x})f(\tilde{\mathbf{x}}).$$

In fact,

$$\begin{aligned} P_{K^P} f(\mathbf{x}) &= \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} (G_K^+)_{i,j} [K^P(\mathbf{x}, \mathbf{x}_i) \otimes K^P(\mathbf{x}, \mathbf{x}_j)] f(\mathbf{x}) \\ &= \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} (G_K^+)_{i,j} K^P(\mathbf{x}, \mathbf{x}_i) f(\mathbf{x}_j) \\ &= \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} (G_K^+)_{i,j} f(\mathbf{x}) f(\mathbf{x}_i) f(\mathbf{x}_j) \\ &= f(\mathbf{x}) \mathbf{f}'(\mathbf{f} \mathbf{f}')^+ \mathbf{f} = f(\mathbf{x}) \mathbf{f}^+ \mathbf{f} = f(\mathbf{x}) \end{aligned}$$

holds since $\mathbf{f}^+ \mathbf{f}$ is the one-dimensional projector with $\text{rank}(\mathbf{f}^+ \mathbf{f}) = 1$, that is, $\mathbf{f}^+ \mathbf{f} = 1$. It is obvious that K^P corresponds to the smallest class of functions that includes f , which agrees with the analytical result of this paper.

As already mentioned, the scope of this paper is in selection of $\mathcal{R}(A_K^*)$. However, selection of $\mathcal{R}(A_K^*)$ does not always have precedence over selection of an element in $\mathcal{R}(A_K^*)$ since the total generalization ability, that is, minimization of $\|f - \hat{f}\|$ is required in practical problems.

In this paper, we discussed a class of kernels whose corresponding RKHSs have an invariant metric. However, this limitation is quite severe. In fact, this analysis cannot be applied to popular kernels such as Gaussian kernels and polynomial kernels. Thus, one of important future works is an extension of our discussions to a class of kernels with variant metrics.

7 Conclusion

In this paper, we considered a class of kernels that forms a monotonic classes of functions with an invariant metric and showed that the kernel corresponding to the smallest class of functions including the unknown true function gives the best model space for the true function. An extension of our result to a class of kernels corresponding to classes of functions with variant metrics is one of important future works that should be resolved.

Acknowledgement. This work was partially supported by Grant-in-Aid No. 18700001 for Young Scientists (B) from the Ministry of Education, Culture, Sports and Technology of Japan.

References

1. Muller, K., Mika, S., Ratsch, G., Tsuda, K., Scholkopf, B.: An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks* 12, 181–201 (2001)
2. Vapnik, V.N.: *The Nature of Statistical Learning Theory*. Springer, New York (1999)
3. Shawe-Taylor, J., Cristianini, N.: *Kernel Methods for Pattern Recognition*. Cambridge University Press, Cambridge (2004)
4. Cristianini, N., Shawe-Taylor, J.: *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, Cambridge (2000)
5. Sugiyama, M., Ogawa, H.: Subspace Information Criterion for Model Selection. *Neural Computation* 13, 1863–1889 (2001)
6. Sugiyama, M., Kawanage, M., Muller, K.: Trading variance reduction with unbiasedness: The regularized subspace information criterion for robust model selection in kernel regression. *Neural Computation* 16, 1077–1104 (2004)
7. Aronszajn, N.: Theory of Reproducing Kernels. *Transactions of the American Mathematical Society* 68, 337–404 (1950)
8. Mercer, J.: Functions of Positive and Negative Type and Their Connection with The Theory of Integral Equations. *Transactions of the London Philosophical Society A*, 415–446 (1909)
9. Tanaka, A., Sugiyama, M., Imai, H., Kudo, M., Miyakoshi, M.: Model selection using a class of kernels with an invariant metric. In: *Joint IAPR International Workshops SSPR 2006 and SPR 2006*, pp. 862–870. Springer, Heidelberg (2006)
10. Schatten, R.: *Norm Ideals of Completely Continuous Operators*. Springer, Berlin (1960)
11. Ogawa, H.: Neural Networks and Generalization Ability. *IEICE Technical Report NC95-8*, 57–64 (1995)
12. Rao, C.R., Mitra, S.K.: *Generalized Inverse of Matrices and its Applications*. John Wiley & Sons, Chichester (1971)