JMLR: Workshop and Conference Proceedings 39:1–15, 2014 ACML 2014

Theoretical Analyses on Ensemble and Multiple Kernel Regressors

Akira Tanaka TAKIRA@MAIN.IST.HOKUDAI.AC.JP
Division of Computer Science and Information Technology
Hokkaido University
N14W9, Kita-ku, Sapporo, 060-0814 Japan

Ichigaku Takigawa TAKIGAWA@CRIS.HOKUDAI.AC.JP
Creative Research Institution
Hokkaido University
N21W10, Kita-ku, Sapporo, 001-0021 Japan

Hideyuki Imai IMAI@MAIN.IST.HOKUDAI.AC.JP
Mineichi Kudo MINE@MAIN.IST.HOKUDAI.AC.JP
Division of Computer Science and Information Technology
Hokkaido University
N14W9, Kita-ku, Sapporo, 060-0814 Japan

Editor: Dinh Phung and Hang Li

Abstract

For the last few decades, learning based on multiple kernels, such as the ensemble kernel regressor and the multiple kernel regressor, has attracted much attention in the field of machine learning. Although its efficacy was revealed numerically in many works, its theoretical ground is not investigated sufficiently. In this paper, we discuss regression problems with a class of kernels whose corresponding reproducing kernel Hilbert spaces have a common subspace with an invariant metric and show that the ensemble kernel regressor (the mean of kernel regressors with those kernels) gives a better learning result than the multiple kernel regressor (the kernel regressor with the sum of those kernels) in terms of the generalization ability of a model space.

Keywords: function estimation, ensemble kernel regressor, multiple kernel regressor, generalization ability,

1. Introduction

Learning based on kernel machines (Muller, Mika, Ratsch, Tsuda, and Scholkopf, 2001), represented by the support vector machine (Vapnik, 1999) and the kernel ridge regressor (Cristianini and Shawe-Taylor, 2000), 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 desirable learning result by kernel machines. Although the model selection in a fixed model space, such as selection of a regularization parameter, is sufficiently investigated in terms of the-

© 2014 A. Tanaka, I. Takigawa, H. Imai & M. Kudo.

oretical and practical senses (see (Sugiyama and Ogawa, 2001; Sugiyama, Kawanabe, and Muller, 2004) for instance), the selection of a model space itself is not sufficiently investigated in terms of a theoretical sense, while 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 (or its parameters) lies on the fact that the metrics of two reproducing kernel Hilbert spaces (Aronszajn, 1950; Mercer, 1909) 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. Recently, a novel framework for evaluating the generalization errors of model spaces specified by different kernels was introduced, in which the so-called invariant metric condition was imposed on the corresponding reproducing kernel Hilbert spaces; and some theoretical results for the selection of a kernel were obtained (Tanaka, Imai, Kudo, and Miyakoshi, 2008; Tanaka and Miyakoshi, 2010; Tanaka, Imai, Kudo, and Miyakoshi, 2011; Tanaka, Takigawa, Imai, and Kudo, 2012).

For the last few decades, learning based on multiple kernels has attracted much attention in this field, which can be regarded as one of model selection schemes. Learning machines using multiple kernels can be divided into two main streams. One is the ensemble kernel learning (see (Vapnik, 1999) for instance) that is a combination of kernel-based learning machines; and the other is the multiple kernel learning (see (Sonnenburg, Ratsch, Schafer, and Scholkopf, 2006) for instance) that is a learning machine based on a combination of kernels. Although their efficacy was revealed numerically in many works, their theoretical grounds were not discussed sufficiently. In this paper, we discuss regression problems with a class of kernels whose corresponding reproducing kernel Hilbert spaces have a common subspace with an invariant metric as the same with (Tanaka, Imai, Kudo, and Miyakoshi, 2008; Tanaka, Takigawa, Imai, and Kudo, 2012) and prove that the ensemble kernel regressor yields a better learning result than the multiple kernel regressor under the invariant metric condition in terms of the generalization ability of a model space.

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

In this section, we give some mathematical preliminaries concerned with the theory of reproducing kernel Hilbert spaces (Aronszajn, 1950; Mercer, 1909).

Definition 1 *(Aronszajn, 1950) Let $\mathbf{R}^d$ be a d-dimensional real vector space and let $\mathcal{H}$ be a class of functions defined on $\mathcal{D} \subset \mathbf{R}^d$, 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 $\tilde{\boldsymbol{x}} \in \mathcal{D}$, $K(\cdot, \tilde{\boldsymbol{x}}) \in \mathcal{H}$.*
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 $\mathcal{H}$.

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 (Aronszajn, 1950):

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

for any $N \in \mathbf{N}$, $c_1, \ldots, c_N \in \mathbf{R}$, and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathcal{D}$. In addition, $K(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = K(\tilde{\boldsymbol{x}}, \boldsymbol{x})$ holds for any $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathcal{D}$ (Aronszajn, 1950). If a reproducing kernel $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ exists, it is unique (Aronszajn, 1950). Conversely, every positive definite function $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ has the unique corresponding RKHS (Aronszajn, 1950). Hereafter, the RKHS corresponding to a reproducing kernel $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ is denoted by $\mathcal{H}_K$. In the following contents, we simply use the symbol K for a kernel by omitting $(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ except the cases where it is needed. In this paper, we assume that the RKHS is separable (Reed and Simon, 1980).

Next, we introduce the Schatten product (Schatten, 1960) that is a convenient tool to reveal the reproducing property Eq.(1) of kernels.

Definition 2 *(Schatten, 1960) 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 $h, v \in \mathcal{H}_1$, $g, u \in \mathcal{H}_2$.

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

where the superscript * denotes the adjoint operator.

We give some theorems concerned with sum and difference of reproducing kernels used in the following contents.

Theorem 3 *((Aronszajn, 1950), p.353) If K_i is the reproducing kernel of the class F_i with the norm $||\cdot||_i$, then $K = K_1 + K_2$ is the reproducing kernel of the class F of all functions $f(\cdot) = f_1(\cdot) + f_2(\cdot)$ with $f_i(\cdot) \in F_i$, and with the norm defined by*

$$||f(\cdot)||^2 = \min \left[||f_1(\cdot)||_1^2 + ||f_2(\cdot)||_2^2 \right], \tag{5}$$

the minimum taken for all the decompositions $f(\cdot) = f_1(\cdot) + f_2(\cdot)$ with $f_i(\cdot) \in F_i$.

Theorem 4 *((Aronszajn, 1950), Theorem II) 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 $||f(\cdot)||_1 \geq ||f(\cdot)||$ for any $f(\cdot) \in F_1$, then the class F_1 possesses a reproducing kernel K_1 such that $K^c = K - K_1$ is also a reproducing kernel.*

Theorem 5 *((Aronszajn, 1950), Theorem I) If K and K_1 are the reproducing kernels of the classes of F and F_1 with the norms $||\cdot||$, $||\cdot||_1$, and if $K - K_1$ is a reproducing kernel, then $F_1 \subset F$ and $||f_1(\cdot)||_1 \geq ||f_1(\cdot)||$ for every $f_1(\cdot) \in F_1$.*

Theorem 6 *((Saitoh, 1997), Theorem 6) Let K_1 and K_2 be kernels, then*

$$\mathcal{H}_{K_1} \subset \mathcal{H}_{K_2} \tag{6}$$

holds, if and only if there exists a positive constant γ such that

$$\gamma^2 K_2 - K_1 \tag{7}$$

is a kernel.

Theorem 3 guarantees that the RKHS corresponding to $K = K_1 + K_2$ includes $\mathcal{H}_{K_1}$ and $\mathcal{H}_{K_2}$ and Theorems 4, 5 and 6 reveal the relationship between the difference of two kernels and the corresponding RKHS's (and their norms).

3. Formulation of Regression Problems

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

$$y_i = f(\boldsymbol{x}_i) + n_i, \tag{8}$$

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

In this paper, we assume that the unknown function $f(\cdot)$ belongs to the RKHS $\mathcal{H}_K$ corresponding to a certain kernel K. If $f(\cdot) \in \mathcal{H}_K$, then Eq.(8) is rewritten as

$$y_i = \langle f(\cdot), K(\cdot, \boldsymbol{x}_i)\rangle_{\mathcal{H}_K} + n_i, \tag{9}$$

on the basis of the reproducing property of the 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.(9) yields

$$\boldsymbol{y} = \left(\sum_{k=1}^{\ell} [\boldsymbol{e}_k^{(\ell)} \otimes K(\cdot, \boldsymbol{x}_k)] \right) f(\cdot) + \boldsymbol{n}, \tag{10}$$

where $\boldsymbol{e}_k^{(\ell)}$ denotes the ℓ-dimensional unit vector whose k-th element is unity. For a convenience of description, we write

$$A_{K,X} = \left(\sum_{k=1}^{\ell} [\boldsymbol{e}_k^{(\ell)} \otimes K(\cdot, \boldsymbol{x}_k)] \right), \tag{11}$$

where $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\ell\}$. Note that $A_{K,X}$ is a linear operator from $\mathcal{H}_K$ onto $\mathbf{R}^\ell$ and Eq.(10) can be written by

$$\boldsymbol{y} = A_{K,X} f(\cdot) + \boldsymbol{n}, \tag{12}$$

which represents the relationship between the unknown true function $f(\cdot)$ and the output vector $\boldsymbol{y}$. Therefore, a regression problem can be interpreted as an inversion problem of the linear equation Eq.(12) (Ogawa, 1995).

4. Kernel Specific Generalization Ability and Some Known Results

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

$$\hat{f}(\cdot) = A^*_{K,X}\boldsymbol{\alpha} = \left(\sum_{k=1}^{\ell}[K(\cdot, \boldsymbol{x}_k) \otimes \boldsymbol{e}_k^{(\ell)}]\right)\boldsymbol{\alpha} = \sum_{k=1}^{\ell}\alpha_k K(\cdot, \boldsymbol{x}_k) \tag{13}$$

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

$$J(f(\cdot); K, X) = ||f(\cdot) - P_{K,X}f(\cdot)||^2_{\mathcal{H}_K}, \tag{14}$$

where $P_{K,X}$ denotes the orthogonal projector onto $\mathcal{R}(A^*_{K,X})$ and $||\cdot||_{\mathcal{H}_K}$ denotes the induced norm of $\mathcal{H}_K$. Note that the orthogonality of $P_{K,X}$ is also specified by the metric of $\mathcal{H}_K$. Selection of an element in $\mathcal{R}(A^*_{K,X})$ 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.(14).

Here, we give some propositions as preparations to evaluate Eq.(14).

Lemma 7 *(Tanaka, Imai, Kudo, and Miyakoshi, 2008)*

$$P_{K,X} = \sum_{i,j=1}^{\ell}(G^+_{K,X})_{ij}\left[K(\cdot, \boldsymbol{x}_i) \otimes K(\cdot, \boldsymbol{x}_j)\right], \tag{15}$$

where $G_{K,X}$ denotes the Grammian matrix of K with X and the superscript $^+$ denotes the Moore-Penrose generalized inverse(Rao and Mitra, 1971).

From Lemma 7, the orthogonal projection of $f(\cdot) \in \mathcal{H}_K$ onto $\mathcal{R}(A^*_{K,X})$ is given as

$$P_{K,X}f(\cdot) = \sum_{i,j=1}^{\ell} f(\boldsymbol{x}_i)(G^+_{K,X})_{ij}K(\cdot, \boldsymbol{x}_j), \tag{16}$$

and this formula immediately yields the following lemma.

Lemma 8 *(Tanaka, Imai, Kudo, and Miyakoshi, 2008) For any $f(\cdot) \in \mathcal{H}_K$,*

$$||P_{K,X}f(\cdot)||^2_{\mathcal{H}_K} = \boldsymbol{f}'G^+_{K,X}\boldsymbol{f} \tag{17}$$

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

Note that $\boldsymbol{f} \in \mathcal{R}(G_{K,X})$ holds, since $\mathcal{R}(A_{K,X}) = \mathcal{R}(A_{K,X}A^*_{K,X}) = \mathcal{R}(G_{K,X})$ trivially holds by Eq.(4).

Let K_1 and K^c be kernels, then $K_2 = K_1 + K^c$ is also a kernel whose corresponding RKHS includes $\mathcal{H}_{K_1}$ from Theorem 3. Since $K_1 = K_2 - K^c$ holds, we have

$$||f(\cdot)||^2_{\mathcal{H}_{K_1}} \geq ||f(\cdot)||^2_{\mathcal{H}_{K_2}} \tag{18}$$

for any $f(\cdot) \in \mathcal{H}_{K_1}$ from Theorem 5. In (Tanaka and Miyakoshi, 2010), the following theorem concerned with the equality in Eq.(18) was introduced, which plays a crucial role in the following contents.

Theorem 9 *(Tanaka and Miyakoshi, 2010) Let K_1 and K^c be kernels and let $K_2 = K_1 + K^c$. The following three statements are equivalent each other.*

1) For any $f(\cdot) \in \mathcal{H}_{K_1}$, $||f(\cdot)||^2_{\mathcal{H}_{K_1}} = ||f(\cdot)||^2_{\mathcal{H}_{K_2}}$,

2) $\mathcal{H}_{K_1} \cap \mathcal{H}_{K^c} = \{0\}$,

3) For any $f_1(\cdot) \in \mathcal{H}_{K_1}$ and for any $f_2(\cdot) \in \mathcal{H}_{K^c}$, $\langle f_1(\cdot), f_2(\cdot)\rangle_{\mathcal{H}_{K_2}} = 0$.

In the following contents, we omit the symbol X from Grammian matrices and projectors since we adopt an arbitrarily fixed X for all cases.

5. Analyses on Ensemble and Multiple Kernel Regressors

We consider a class of kernels $\mathcal{K} = \{K_1, \ldots, K_n\}$ and corresponding RKHS written as $\mathcal{H}_{K_i}$, $(i \in \{1, \ldots, n\})$. We assume that

$$L = \overline{\cap_{i=1}^n \mathcal{H}_{K_i}} \tag{19}$$

is a non-empty linear class and we discuss the regression problem for $f(\cdot) \in L$ in order for $P_{K_p} f(\cdot)$, $(p \in \{1, \ldots, n\})$ to be consistent in terms of the orthogonal projection [1]. Under these settings, we discuss two kernel regression schemes using all kernels in $\mathcal{K}$. One is the multiple kernel regressor based on a linear combination of given kernels with positive weights determined by some criterion. The other is the ensemble kernel regressor which is a convex combination of kernel regressors based on each kernels whose weights are specified by boosting strategy for instance. As mentioned above, the strategies of combination in the both schemes are quite different, which implies that it is difficult to analyze their generalization ability in a unified way. Therefore, we analyze quite simple and special cases in this paper as follows. Also note that we analyze the optimal results of both schemes in noise free case, that is, the orthogonal projection of the unknown true function onto the model space for simplicity of analysis, while the optimal result are not always achieved in practical problems.

1. If $f(\cdot) \notin L$, there may exist K_p by which the orthogonal projection can not be constructed from the training data set.

We define the multiple kernel regressor as that based on the sum of all kernels, written as

$$K_u = \sum_{i=1}^{n} K_i. \tag{20}$$

It is trivial that

$$K_u - K_i = \sum_{j=1, j\neq i}^{n} K_j. \tag{21}$$

is also a kernel from Theorem 3, which implies that for any fixed $i \in \{1, \ldots, n\}$,

$$\mathcal{H}_{K_i} \subset \mathcal{H}_{K_u}, \quad ||f(\cdot)||_{\mathcal{H}_{K_i}} \geq ||f(\cdot)||_{\mathcal{H}_{K_u}} \tag{22}$$

holds for any $f(\cdot) \in \mathcal{H}_{K_i}$ from Theorem 5. The learning result by the multiple kernel regressor is written as

$$\begin{aligned} \hat{f}_m(\cdot) &= P_{K_u} f(\cdot) \\ &= \sum_{i,j=1}^{\ell} f(\boldsymbol{x}_i)(G^+_{K_u})_{ij} K_u(\cdot, \boldsymbol{x}_j). \end{aligned} \tag{23}$$

We define the ensemble kernel regressor as the mean of the kernel regressors by the individual kernels $K_i, (i \in \{1, \ldots, n\})$. The learning result by the ensemble kernel regressor is written as

$$\begin{aligned} \hat{f}_e(\cdot) &= \frac{1}{n} \sum_{p=1}^{n} P_{K_p} f(\cdot) \\ &= \frac{1}{n} \sum_{p=1}^{n} \sum_{i,j=1}^{\ell} f(\boldsymbol{x}_i)(G^+_{K_p})_{ij} K_p(\cdot, \boldsymbol{x}_j). \end{aligned} \tag{24}$$

The generalization error, defined by Eq.(14), of the multiple kernel regressor Eq.(23) is straightforwardly obtained by

$$\begin{aligned} E_m &= J(f(\cdot); K_u, X) = ||f(\cdot) - P_{K_u} f(\cdot)||^2_{\mathcal{H}_{K_u}} \\ &= ||f||^2_{\mathcal{H}_{K_u}} - \boldsymbol{f}' G^+_{K_u} \boldsymbol{f} \end{aligned} \tag{25}$$

from Lemma 8 and the Pythagorean theorem. Note that the evaluation by the norm $||\cdot||_{\mathcal{H}_{K_u}}$ is the best choice for the multiple kernel regressor since the orthogonality of P_{K_u} is specified by the metric of $\mathcal{H}_{K_u}$.

Next, we evaluate the generalization error of the ensemble kernel regressor Eq.(24) with the same norm as Eq.(25), which is written as

$$E_e = \left\| f(\cdot) - \frac{1}{n} \sum_{p=1}^{n} P_{K_p} f(\cdot) \right\|^2_{\mathcal{H}_{K_u}}. \tag{26}$$

We give the following Lemmas to evaluate Eq.(26).

Lemma 10 *Let K be a kernel whose corresponding RKHS is separable, and let α be a positive real number, then*

$$\mathcal{H}_K = \mathcal{H}_{\alpha K} \tag{27}$$

as the class of functions. Moreover

$$\alpha ||f(\cdot)||^2_{\mathcal{H}_{\alpha K}} = ||f(\cdot)||^2_{\mathcal{H}_K} \tag{28}$$

holds for any $f(\cdot) \in \mathcal{H}_K$.

Proof Let α_1 and α_2 be a real positive numbers satisfying $\alpha_2 < \alpha < \alpha_1$, then,

$$\alpha_1 K - (\alpha K), \quad \frac{1}{\alpha_2}(\alpha K) - K$$

are also kernels. Therefore, Eq.(27) immediately holds from Theorem 6.

Since $\mathcal{H}_K$ is separable, there exists a countable set $\{(\beta_k, \boldsymbol{z}_k) \mid k \in \mathbf{N}\}$ for any $f(\cdot) \in \mathcal{H}_K$ such that

$$f(\cdot) = \sum_{k \in \mathbf{N}} \beta_k K(\cdot, \boldsymbol{z}_k).$$

Then, we have

$$||f(\cdot)||^2_{\mathcal{H}_K} = \sum_{i,j \in \mathbf{N}} \beta_i \beta_j K(\boldsymbol{z}_i, \boldsymbol{z}_j).$$

On the other hand, we have

$$\begin{aligned} ||f(\cdot)||^2_{\mathcal{H}_{\alpha K}} &= \left\| \frac{1}{\alpha} \sum_{k \in \mathbf{N}} \beta_k \alpha K(\cdot, \boldsymbol{z}_k) \right\|^2_{\mathcal{H}_{\alpha K}} \\ &= \frac{1}{\alpha^2} \sum_{i,j \in \mathbf{N}} \beta_i \beta_j \alpha K(\boldsymbol{z}_i, \boldsymbol{z}_j) = \frac{1}{\alpha} \sum_{i,j \in \mathbf{N}} \beta_i \beta_j K(\boldsymbol{z}_i, \boldsymbol{z}_j) \\ &= \frac{1}{\alpha} ||f(\cdot)||^2_{\mathcal{H}_K}, \end{aligned}$$

which concludes the proof. ■

Lemma 11 *Let K_i, $(i \in \{1, \ldots, n\})$ be kernels and let $K_u = \sum_{i=1}^n K_i$. For any function $f(\cdot) = \sum_{i=1}^n f_i(\cdot)$ with $f_i(\cdot) \in \mathcal{H}_{K_i}$,*

$$\left\| \sum_{i=1}^n f_i(\cdot) \right\|^2_{\mathcal{H}_{K_u}} \leq \sum_{i=1}^n \|f_i(\cdot)\|^2_{\mathcal{H}_{K_i}} \tag{29}$$

holds.

Proof This lemma is a trivial consequence of Theorem 3. ■

Here, we consider the following assumption.

Assumption 12 *There exists a linear class $S \subset L$ such that*

$$||f(\cdot)||_{\mathcal{H}_{K_i}} = ||f(\cdot)||_{\mathcal{H}_{K_j}}, \ (i, j \in \{1, \ldots, n\}) \tag{30}$$

for any $f(\cdot) \in S$.

When Assumption 12 holds, there exists a kernel K_S such that

$$K_i^c = K_i - K_S, \ (i \in \{1, \ldots, n\}) \tag{31}$$

is also a kernel from Theorem 4. Hereafter, we use $\mathcal{H}_{K_S}$ instead of S since K_S is guaranteed to be a kernel. Note that

$$\mathcal{H}_{K_S} \cap \mathcal{H}_{K_i^c} = \{0\} \tag{32}$$

holds from Theorem 9. The importance of Assumption 12 will be discussed in Section 6.

Lemma 13 *If Assumption 12 is satisfied,*

$$E_e \leq ||f(\cdot)||^2_{\mathcal{H}_{K_u}} - \frac{1}{n^2} \sum_{p=1}^{n} \boldsymbol{f}' G^+_{K_p} \boldsymbol{f} \tag{33}$$

holds for any $f(\cdot) \in \mathcal{H}_{K_S}$.

Proof From Lemma 10, Theorem 9 and Assumption 12, we have

$$||f(\cdot)||^2_{\mathcal{H}_{K_p}} = ||f(\cdot)||^2_{\mathcal{H}_{K_S}} = n||f(\cdot)||^2_{\mathcal{H}_{nK_S}} = n||f(\cdot)||^2_{\mathcal{H}_{K_u}} \tag{34}$$

for any $f(\cdot) \in \mathcal{H}_{K_S}$ since

$$K_u = nK_S + \sum_{p=1}^{n} K_p^c$$

and $\mathcal{H}_{nK_S} \cap \mathcal{H}_{K^c} = \{0\}$ hold, where $K^c = \sum_{p=1}^{n} K_p^c$. Therefore, from Lemma 11 and the Pythagorean theorem, we have

$$\begin{aligned}
E_e &= \left\| \frac{1}{n} \sum_{p=1}^{n} \left(f(\cdot) - P_{K_p} f(\cdot) \right) \right\|^2_{\mathcal{H}_{K_u}} \\
&\leq \frac{1}{n^2} \sum_{p=1}^{n} ||f(\cdot) - P_{K_p} f(\cdot)||^2_{\mathcal{H}_{K_p}} \\
&= \frac{1}{n^2} \sum_{p=1}^{n} (||f(\cdot)||^2_{\mathcal{H}_{K_p}} - \boldsymbol{f}' G^+_{K_p} \boldsymbol{f}) \\
&= ||f(\cdot)||^2_{\mathcal{H}_{K_u}} - \frac{1}{n^2} \sum_{p=1}^{n} \boldsymbol{f}' G^+_{K_p} \boldsymbol{f},
\end{aligned}$$

which concludes the proof. ■

Lemma 14 *Let $G_i \in \mathbf{R}^{d\times d}$, $(i \in \{1,\ldots,n\})$ be non-negative definite real symmetric matrices and let $\boldsymbol{v} \in \cap_{i=1}\mathcal{R}(G_i)$. Then,*

$$\boldsymbol{v}'\left(\frac{1}{n^2}\sum_{i=1}^{n} G_i^+ - \left(\sum_{i=1}^{n} G_i\right)^+\right)\boldsymbol{v} \geq 0 \tag{35}$$

holds.

Proof Let $S = \sum_{i=1}^n G_i$ and $T = \frac{1}{n^2}\sum_{i=1}^n G_i^+$, then $\mathcal{R}(S) = \mathcal{R}(S^+) = \mathcal{R}(T)$ holds since S and T are non-negative definite symmetric matrices. Therefore, we have

$$\begin{aligned}
&\boldsymbol{v}'(T - S^+)\boldsymbol{v} \\
&= \boldsymbol{v}'S^+S(T-S^+)SS^+\boldsymbol{v} = \boldsymbol{v}'S^+(STS - S)S^+\boldsymbol{v} \\
&= \boldsymbol{v}'S^+\left(\frac{1}{n^2}\sum_{i=1}^{n} SG_i^+S - S\right)S^+\boldsymbol{v}
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{i=1}^{n} SG_i^+S - n^2S \\
&= \sum_{i=1}^{n}(nG_i + S - nG_i)G_i^+(nG_i + S - nG_i) - n^2S \\
&= \sum_{i=1}^{n}((S-nG_i)G_i^+(S-nG_i) + nG_iG_i^+(S-nG_i) + n(S-nG_i)G_i^+G_i)) \\
&= \sum_{i=1}^{n}((S-nG_i)G_i^+(S-nG_i) + n(G_iG_i^+S + SG_i^+G_i - 2nG_i)) \\
&= \sum_{i=1}^{n}((S-nG_i)G_i^+(S-nG_i) + n(G_iG_i^+S + SG_i^+G_i)) - 2n^2S.
\end{aligned}$$

Let

$$\begin{aligned}
T_1 &= \sum_{i=1}^{n}(S-nG_i)G_i^+(S-nG_i), \\
T_2 &= n\sum_{i=1}^{n}(G_iG_i^+S + SG_i^+G_i) - 2n^2S,
\end{aligned}$$

then

$$\begin{aligned}
&\boldsymbol{v}'S^+T_2S^+\boldsymbol{v} \\
&= n\sum_{i=1}^{n}\boldsymbol{v}'(S^+G_iG_i^+ + G_i^+G_iS^+)\boldsymbol{v} - 2n^2\boldsymbol{v}'S^+\boldsymbol{v} \\
&= n\sum_{i=1}^{n}2\boldsymbol{v}'S^+\boldsymbol{v} - 2n^2\boldsymbol{v}'S^+\boldsymbol{v} \\
&= 0
\end{aligned}$$

holds, since $\boldsymbol{v} \in \mathcal{R}(G_i) \subset \mathcal{R}(S)$ for any $i \in \{1, \ldots, n\}$ and $G_i G_i^+ = G_i^+ G_i$ is the orthogonal projector onto $\mathcal{R}(G_i)$. Therefore, we have

$$\boldsymbol{v}'(T - S^+)\boldsymbol{v} = \frac{1}{n^2}\boldsymbol{v}' S^+ T_1 S^+ \boldsymbol{v} \geq 0,$$

since T_1 is a non-negative definite symmetric matrix, which concludes the proof. ■

The next theorem is the main result of this paper.

Theorem 15 *If Assumption 12 is satisfied,*

$$E_m - E_e \geq 0 \tag{36}$$

holds for any $f(\cdot) \in \mathcal{H}_{K_S}$.

Proof From the fact that $\boldsymbol{f} \in \mathcal{R}(G_{K_p})$ for any $p \in \{1, \ldots, n\}$ and Lemmas 13 and 14,

$$\begin{aligned} E_m - E_e &\geq (||f||^2_{\mathcal{H}_{K_u}} - \boldsymbol{f}' G^+_{K_u} \boldsymbol{f}) - \left(||f||^2_{\mathcal{H}_{K_u}} - \frac{1}{n^2} \sum_{p=1}^{n} \boldsymbol{f}' G^+_{K_p} \boldsymbol{f} \right) \\ &= \frac{1}{n^2} \sum_{p=1}^{n} \boldsymbol{f}' G^+_{K_p} \boldsymbol{f} - \boldsymbol{f}' G^+_{K_u} \boldsymbol{f} \geq 0 \end{aligned}$$

is obtained, which concludes the proof. ■

According to Theorem 15, it is concluded that the ensemble kernel regressor yields a better result than the multiple kernel regressor under Assumption 12. Note that Assumption 12 is a quite strong condition. In fact, popular kernels, such as the Gaussian kernels with various parameters, do not satisfy Assumption 12. Therefore, relaxation of Assumption 12 is one of important issues that should be undertaken.

6. Examples

In this section, we give simple examples confirming the importance of Assumption 12 in Theorem 15. Let

$$\begin{aligned} K_1(x, y) &= 1 + xy \\ K_2(x, y) &= (1 + xy)^2 = 1 + 2xy + x^2 y^2 \end{aligned}$$

be polynomial kernels defined on $\mathbf{R} \times \mathbf{R}$. Note that $\dim \mathcal{H}_{K_1} = 2$ and $\mathcal{H}_{K_1}$ is spanned by the functions $b_1(x) = 1$ and $b_2(x) = x$. Similarly, $\dim \mathcal{H}_{K_2} = 3$ and $\mathcal{H}_{K_2}$ is spanned by the functions $b_1(x)$, $b_2(x)$, and $b_3(x) = x^2$. Therefore, the linear class L is spanned by $b_1(x)$ and $b_2(x)$. Also note that $K_u(x, y) = 2 + 3xy + x^2 y^2$. We investigate the generalization errors E_m and E_e for $f(\cdot) \in L$. We adopt $X = \{1\}$ as the input training data set in the following contents.

6.1. Example with Assumption 12

Let us consider the linear class spanned by $b_1(x)$, then any function in the class is represented by $f(x) = \alpha b_1(x) = \alpha$ with $\alpha \in \mathbf{R}$. Since

$$\begin{aligned} f(x) &= \alpha K_1(x, 0), \\ f(x) &= \alpha K_2(x, 0), \end{aligned}$$

we have $||f(x)||^2_{\mathcal{H}_{K_1}} = ||f(x)||^2_{\mathcal{H}_{K_2}} = \alpha^2$, which implies that Assumption 12 is satisfied for the linear class $\mathcal{H}_{K_S}$ spanned by $b_1(x)$. From Eq.(16), we have

$$\begin{aligned} \hat{f}_1(x) &= P_{K_1} f(x) = \frac{\alpha}{2}(1 + x), \\ \hat{f}_2(x) &= P_{K_2} f(x) = \frac{\alpha}{4}(1 + 2x + x^2), \end{aligned}$$

as the learning results by K_1 and K_2, which implies that the learning result by the ensemble kernel regressor is reduced to

$$\hat{f}_e(x) = \frac{\alpha}{8}(3 + 4x + x^2). \tag{37}$$

Similarly, the learning result by the multiple kernel regressor is reduced to

$$\hat{f}_m(x) = \frac{\alpha}{6}(2 + 3x + x^2). \tag{38}$$

Note that

$$\begin{aligned} d_e(x) &= \hat{f}_e(x) - f(x) = \frac{\alpha}{8}(3 + 4x + x^2) - \alpha \\ &= \frac{\alpha}{8}(-5 + 4x + x^2) \\ &= -\frac{\alpha}{48} K_u(x, -1) - \frac{21\alpha}{48} K_u(x, 0) + \frac{7\alpha}{48} K_u(x, 1) \\ d_m(x) &= \hat{f}_m(x) - f(x) = \frac{\alpha}{6}(2 + 3x + x^2) - \alpha \\ &= \frac{\alpha}{6}(-4 + 3x + x^2) \\ &= -\frac{3\alpha}{6} K_u(x, 0) + \frac{\alpha}{6} K_u(x, 1) \end{aligned}$$

holds. By applying the fact that $\langle K_u(\cdot, x), K_u(\cdot, y) \rangle_{\mathcal{H}_{K_u}} = K_u(x, y)$, derived from Eq.(1), we have

$$\begin{aligned} E_e &= \frac{113}{384}\alpha^2 \simeq 0.294\alpha^2, \\ E_m &= \frac{1}{3}\alpha^2 \simeq 0.333\alpha^2. \end{aligned}$$

Accordingly, it is confirmed that the inequality Eq.(36) surely holds with Assumption 12 in these settings.

6.2. Example without Assumption 12

Let us consider the function $f(x) = x \in L$. Since

$$\begin{aligned} f(x) &= -K_1(x,0) + K_1(x,1), \\ f(x) &= -\frac{1}{4}K_2(x,-1) + \frac{1}{4}K_2(x,1), \end{aligned}$$

we have $||f(x)||^2_{\mathcal{H}_{K_1}} = 1$ and $||f(x)||^2_{\mathcal{H}_{K_2}} = 1/2$, which implies that Assumption 12 does not hold in this case. Since $f(1) = 1$, the learning results by the ensemble and the multiple kernel regressors are the same with Eqs.(37) and (38) with $\alpha = 1$. Note that

$$\begin{aligned} d_e(x) &= \hat{f}_e(x) - f(x) = \frac{1}{8}(3 + 4x + x^2) - x \\ &= \frac{1}{8}(3 - 4x + x^2) \\ &= \frac{7}{48}K_u(x,-1) + \frac{3}{48}K_u(x,0) - \frac{1}{48}K_u(x,1) \\ d_m(x) &= \hat{f}_m(x) - f(x) = \frac{1}{6}(2 + 3x + x^2) - x \\ &= \frac{1}{6}(2 - 3x + x^2) \\ &= \frac{1}{6}K_u(x,-1) \end{aligned}$$

holds, which yields

$$\begin{aligned} E_e &= \frac{65}{384} \simeq 0.169, \\ E_m &= \frac{1}{6} \simeq 0.167, \end{aligned}$$

and $E_e > E_m$. Accordingly, it is confirmed that the inequality Eq.(36) does not hold without Assumption 12 in such a simple setting with a popular polynomial kernel, which supports the importance of Assumption 12.

7. Conclusion

In this paper, we discussed a class of kernels whose corresponding RKHS's have a common subspace with an invariant metric and proved that the ensemble kernel regressor with those kernels gives a better result than the multiple kernel regressor with the sum of those kernels. Relaxation of the assumption in the main theorem, extending the obtained result to practical learning machines, such as the support vector machine and the kernel ridge regressor, and similar analyses for variance of additive noise are ones of our future works that should be undertaken.

Acknowledgments

This work was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), 24500001.

References

N. Aronszajn. Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

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

J. Mercer. Functions of Positive and Negative Type and Their Connection with The Theory of Integral Equations. *Transactions of the London Philosophical Society*, A(209):415–446, 1909.

K. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf. An Introduction to Kernel-Based Learning Algorithms. *IEEE Transactions on Neural Networks*, 12:181–201, 2001.

H. Ogawa. Neural Networks and Generalization Ability. *IEICE Technical Report*, NC95-8: 57–64, 1995.

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

M. Reed and B. Simon. *Methods of Modern Mathematical Physics I : Functional Analysis (Revised and Enlarged Edition)*. Academic Press, San Diego, 1980.

S. Saitoh. *Integral Transforms, Reproducing Kernels and Their Applications*. Addison Wesley Longman Ltd, UK, 1997.

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

S. Sonnenburg, G. Ratsch, C. Schafer, and B. Scholkopf. Large Scale Multiple Kernel Learning. *The Journal of Machine Learning Research*, 7:1531–1565, 2006.

M. Sugiyama and H. Ogawa. Subspace Information Criterion for Model Selection. *Neural Computation*, 13(8):1863–1889, 2001.

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*, 16(5):1077–1104, 2004.

A. Tanaka and M. Miyakoshi. Theoretical Analyses for a Class of Kernels with an Invariant Metric. In *2010 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2074–2077, 2010.

A. Tanaka, H. Imai, M. Kudo, and M. Miyakoshi. Optimal Kernel in a Class of Kernels with an Invariant Metric. In *Joint IAPR Internatioanl Workshops SSPR 2008 and SPR 2008*, pages 530–539. Springer, 2008.

A. Tanaka, , H. Imai, M. Kudo, and M. Miyakoshi. Theoretical Analyses on a Class of Nested RKHS's. In *2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2011)*, pages 2072–2075, 2011.

A. Tanaka, I. Takigawa, H. Imai, and M. Kudo. Extended Analyses for an Optimal Kernel in a Class of Kernels with an Invariant Metric. In *Joint IAPR Internatioanl Workshops SSPR 2012 and SPR 2012*, pages 345–353. Springer, 2012.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1999.