PAPER

Ensemble and Multiple Kernel Regressors: Which Is Better?

Akira TANAKA[†a)], ***Member***, **Hirofumi TAKEBAYASHI**[†], **Ichigaku TAKIGAWA**[†], ***Nonmembers***, **Hideyuki IMAI**[†], ***and*** **Mineichi KUDO**[†], ***Members***

SUMMARY For the last few decades, learning with multiple kernels, represented by the ensemble kernel regressor and the multiple kernel regressor, has attracted much attention in the field of kernel-based machine learning. Although their efficacy was investigated numerically in many works, their theoretical ground is not investigated sufficiently, since we do not have a theoretical framework to evaluate them. In this paper, we introduce a unified framework for evaluating kernel regressors with multiple kernels. On the basis of the framework, we analyze the generalization errors of the ensemble kernel regressor and the multiple kernel regressor, and give a sufficient condition for the ensemble kernel regressor to outperform the multiple kernel regressor in terms of the generalization error in noise-free case. We also show that each kernel regressor can be better than the other without the sufficient condition by giving examples, which supports the importance of the sufficient condition.

key words: *kernel regression, ensemble kernel regressor, multiple kernel regressor, generalization error, reproducing kernel Hilbert spaces*

1. Introduction

Kernel-based learning machines [1], represented by the support vector machine [2] and the kernel ridge regressor [3], are widely recognized as powerful tools 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 (fixed kernel), such as selection of a regularization parameter, is sufficiently discussed in terms of both theoretical and practical senses (see [4], [5] for instance), the selection of a model space, specified by a kernel and training input vectors, is not sufficiently discussed in terms of theoretical sense, while practical algorithms for selection of a kernel (or its parameters), such as cross-validation, are widely used. The difficulty of the theoretical analyses on the selection of a kernel (or its parameters) is due to the fact that the metrics of two reproducing kernel Hilbert spaces 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 on the basis of the condition [6]–[9].

For the last few decades, learning based on multiple kernels has attracted much attention in the field of kernel-based machine learning, which can be regarded as one of model selection schemes. There exist two representative learning machines with multiple kernels. One is the ensemble kernel learning (see [2] for instance) that is a convex combination of kernel-based learning machines; and the other is the multiple kernel learning (see [10] for instance) that is a learning machine based on a convex combination of kernels. Although their efficacy was revealed numerically in many works, their theoretical grounds were not discussed sufficiently. Its difficulty is similar to that of the selection of a kernel mentioned above. In this paper, we introduced a unified framework for evaluating the generalization errors of kernel regressors with multiple kernels, and analyzed the generalization errors of the ensemble kernel regressor and the multiple kernel regressor. As a result, we obtained a sufficient condition for the ensemble kernel regressor to outperform the multiple kernel regressor in terms of the theoretical limit of the generalization error, that is, the attainable minimum generalization error achieved by the orthogonal projection of the unknown true function onto the solution subspace in the noise-free case. The sufficient condition is deeply related to the invariant metric condition given in [6] and the superiority of the ensemble kernel regressor against the multiple kernel regressor is deeply related to the relationship between the arithmetic mean and the harmonic mean. We also showed that each kernel regressor can be better than the other without the sufficient condition, depending on an unknown true function, which is the theoretical knowledge that has not been revealed in the past literatures and supports the importance of our sufficient condition.

Note that we discussed a similar problem in our previous work [11] in which the unweighted sum of kernels and the unweighted sum of kernel machines are discussed only. This paper is an extension of the result obtained in [11] to an arbitrary convex combination. Also note that this paper includes detailed descriptions of our previous works [12], [13], which gave an overview of some parts of this paper.

Manuscript received May 21, 2015.
Manuscript revised July 24, 2015.

[†]The authors are with the Division of Computer Science and Information Technology, Graduate School of Information Science and Technology, Hokkaido University, Sapporo-shi, 060-0814 Japan.
a) E-mail: takira@main.ist.hokudai.ac.jp

DOI: 10.1587/transfun.E98.A.2315

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

In this section, we give mathematical preliminaries concerned with the theory of reproducing kernel Hilbert spaces [14], [15].

Definition 1: [14] 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 the following two conditions hold.

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 $\mathcal{D}$ in contrast to ordinary Hilbert spaces such as $L^2(\mathcal{D})$, the Hilbert space consisting of all square integrable functions defined on $\mathcal{D}$. Note that reproducing kernels are positive definite [14]:

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

for any $N \in \mathbf{N}$, $c_1, \ldots, c_N \in \mathbf{R}$, and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathcal{D}$, where $\mathbf{N}$ stands for the set of natural numbers. In addition, $K(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = K(\tilde{\boldsymbol{x}}, \boldsymbol{x})$ holds for any $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathcal{D}$ [14]. If a reproducing kernel $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ exists, it is unique [14]. Conversely, every positive definite function $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ has the unique corresponding RKHS [14]. 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 [16] since popular RKHS's are separable [17].

Next, we introduce the Schatten product [18] that is a convenient tool to represent an inner product of two elements in Hilbert spaces as an operator for one of them.

Definition 2: [18] 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 $h \in \mathcal{H}_1$, $g, u \in \mathcal{H}_2$. $v \in \mathcal{H}_3$,

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

where the superscript * denotes the adjoint operator.

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

Theorem 1: [14] 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 2: [14] 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 3: [14] 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 4: [19] 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 1 guarantees that the RKHS corresponding to $K = K_1 + K_2$ includes all functions in $\mathcal{H}_{K_1}$ and those in $\mathcal{H}_{K_2}$; and Theorems 2, 3 and 4 reveal the relationship between the difference of two kernels and the corresponding RKHS's (and their norms). Note that Theorem 1 can be easily extended to more than two kernels.

3. Formulation of Regression Problems

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

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

where $f(\cdot)$ denotes an unknown true function and n_i denotes an observation noise. The aim of the regression problem considered in this paper is to estimate the unknown true function $f(\cdot)$ by using the given training data set and statistical properties of the 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 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 a kernel. 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 $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\ell\}$ is the set of the training input vectors. 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, the regression problem can be interpreted as the inversion problem of the linear equation Eq. (12) [20].

4. Generalization Error of Kernel Regressor and Some Known Results

In general, a learning result by a kernel machine is represented by a linear combination of $K(\cdot, \boldsymbol{x}_i)$, $(i \in \{1, \ldots, \ell\})$, 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

$$\begin{aligned} \hat{f}(\cdot) &= A_{K,X}^* \boldsymbol{\alpha} = \left(\sum_{i=1}^{\ell} [K(\cdot, \boldsymbol{x}_i) \otimes \boldsymbol{e}_i^{(\ell)}] \right) \boldsymbol{\alpha} \\ &= \sum_{i=1}^{\ell} \alpha_i K(\cdot, \boldsymbol{x}_i) \end{aligned} \tag{13}$$

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. Roughly speaking, the generalization error is the difference between the unknown true function and an estimated one at any point $\boldsymbol{x} \in \mathcal{D}$, which may not be in X. 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 $\mathcal{R}(A_{K,X}^*)$ [6], [21], [22] written as

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

where $P_{K,X}$ denotes the orthogonal projector onto $\mathcal{R}(A_{K,X}^*)$ and $\|\cdot\|_{\mathcal{H}_K}$ denotes the induced norm of $\mathcal{H}_K$. The validity of $J(f(\cdot); K, X)$ as the generalization error is supported by the fact that

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

holds for any $\boldsymbol{x} \in \mathcal{D}$, which is a trivial consequence of the reproducing property of a kernel and the Schwarz's inequality. 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). Note that ignoring the noise and adopting the orthogonal projection imply the analyses on the theoretical limit of the generalization error. Here, we give some propositions in order to evaluate Eq. (14).

Lemma 1: [6]

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

where $G_{K,X}$ denotes the Gram matrix of K with X, defined by $G_{K,X} = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))$, and the superscript $^+$ denotes the Moore-Penrose generalized inverse [23].

From Lemma 1, 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 2: [6] For any $f(\cdot) \in \mathcal{H}_K$,

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

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 $\boldsymbol{f} \in \mathcal{R}(A_{K,X}) = \mathcal{R}(A_{K,X} A_{K,X}^*) = \mathcal{R}(G_{K,X})$ trivially holds from 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 1. Since $K_1 = K_2 - K^c$ holds, we have

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

for any $f(\cdot) \in \mathcal{H}_{K_1}$ from Theorem 3. In [7], the following theorem, concerned with the equality in Eq. (18), was introduced, which plays a crucial role in the following contents.

Theorem 5: [7] 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 $f(\cdot) \in \mathcal{H}_{K_1}$, $\|f(\cdot)\|_{\mathcal{H}_{K_1}}^2 = \|f(\cdot)\|_{\mathcal{H}_{K_2}}^2$,
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 Gram matrices and projectors except the cases where it is needed.

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 $\mathcal{H}_{K_p}$, $(p \in \{1, \ldots, n\})$. We consider the linear class

$$L = \cap_{p=1}^{n} \mathcal{H}_{K_p}, \tag{19}$$

and 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[†]. Note that L always exists for any $\mathcal{K}$ since at least $f(\boldsymbol{x}) = 0 \in L$ holds[††]. Under these settings, we discuss two kernel-based regression schemes using all kernels in $\mathcal{K}$. One is the multiple kernel regressor, which is a kernel regressor based on a convex combination of given kernels. The other is the ensemble kernel regressor, which is a convex combination of kernel regressors by each kernel. 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, since our main interest is in the theoretical limit of the generalization error.

We define the multiple kernel regressor as the regressor based on a convex combinations of the kernels in $\mathcal{K}$, written as

$$K_u = \sum_{p=1}^{n} \alpha_p K_p, \quad \left(\alpha_p > 0, \ \sum_{p=1}^{n} \alpha_p = 1 \right). \tag{20}$$

Note that K_u is trivially a kernel from Theorem 1. The learning result by the multiple kernel regressor is written as

$$\begin{aligned} \hat{f}_m(\cdot) &= P_{K_u} f(\cdot) \\ &= \sum_{i,j=1}^{\ell} f(\boldsymbol{x}_i)(G^{+}_{K_u})_{ij} K_u(\cdot, \boldsymbol{x}_j), \end{aligned} \tag{21}$$

from Eq. (16).

We define the ensemble kernel regressor as the convex combination of the kernel regressors by each kernel $K_p, (p \in \{1, \ldots, n\})$. The learning result by the ensemble kernel regressor is written as

$$\begin{aligned} \hat{f}_e(\cdot) &= \sum_{p=1}^{n} \alpha_p P_{K_p} f(\cdot) \\ &= \sum_{p=1}^{n} \alpha_p \sum_{i,j=1}^{\ell} f(\boldsymbol{x}_i)(G^{+}_{K_p})_{ij} K_p(\cdot, \boldsymbol{x}_j). \end{aligned} \tag{22}$$

In general, the optimal coefficients α_p may differ in the both regressors. However, we adopt the same coefficients in the following analyses since if one regressor outperforms the other with the same coefficients, the former always outperforms the latter with their optimal coefficients.

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

[††]However, $L = \{0\}$ is a meaningless case. Thus, we are interested in the case of $\dim L \geq 1$.

The generalization error, defined by Eq. (14), of the multiple kernel regressor Eq. (21) is straightforwardly obtained by

$$\begin{aligned} E_m &= J(f(\cdot); K_u, X) = \|f(\cdot) - P_{K_u} f(\cdot)\|^2_{\mathcal{H}_{K_u}} \\ &= \|f\|^2_{\mathcal{H}_{K_u}} - \boldsymbol{f}' G^{+}_{K_u} \boldsymbol{f} \end{aligned} \tag{23}$$

from Lemma 2 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. (22) with the same norm as Eq. (23), which is written as

$$E_e = \left\| f(\cdot) - \sum_{p=1}^{n} \alpha_p P_{K_p} f(\cdot) \right\|^2_{\mathcal{H}_{K_u}}. \tag{24}$$

We give the following Lemmas to evaluate Eq. (24).

Lemma 3: [11] Let K be a kernel whose corresponding RKHS is separable, and let α be a positive real number, then

$$\mathcal{H}_K = \mathcal{H}_{\alpha K} \tag{25}$$

holds as the class of functions[†††]. Moreover

$$\alpha \|f(\cdot)\|^2_{\mathcal{H}_{\alpha K}} = \|f(\cdot)\|^2_{\mathcal{H}_K} \tag{26}$$

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

Proof Let α_1 and α_2 be 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. (25) immediately holds from Theorem 4.

Since $\mathcal{H}_K$ is separable, there exists a countable set $\{(\beta_k, \boldsymbol{z}_k) \mid k \in \mathbf{N}, \ \beta_k \in \mathbf{R}, \ \boldsymbol{z}_k \in \mathcal{D}\}$ 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. □

[†††]The metrics of the two RKHS's are necessarily different except the case of $\alpha = 1$.

Lemma 4: [11] Let K_p, $(p \in \{1, \ldots, n\})$ be kernels and let $K_u = \sum_{p=1}^{n} K_p$. For any function $f(\cdot) = \sum_{p=1}^{n} f_p(\cdot)$ with $f_p(\cdot) \in \mathcal{H}_{K_p}$,

$$\left\| \sum_{p=1}^{n} f_p(\cdot) \right\|_{\mathcal{H}_{K_u}}^2 \leq \sum_{p=1}^{n} \left\| f_p(\cdot) \right\|_{\mathcal{H}_{K_p}}^2 \tag{27}$$

holds.

Proof From Theorem 1, we have

$$\begin{aligned} \left\| \sum_{p=1}^{n} f_p(\cdot) \right\|_{\mathcal{H}_{K_u}}^2 &= \min_{\sum_{p=1}^{n} \tilde{f}_p = \sum_{p=1}^{n} f_p,\ \tilde{f}_p \in \mathcal{K}_{K_p}} \sum_{p=1}^{n} \left\| \tilde{f}_p(\cdot) \right\|_{\mathcal{H}_{K_p}}^2 \\ &\leq \sum_{i=p}^{n} \left\| f_p(\cdot) \right\|_{\mathcal{H}_{K_p}}^2, \end{aligned}$$

which concludes the proof. □

Lemma 5: For any $f(\cdot) \in L$,

$$E_e \leq \sum_{p=1}^{n} \alpha_p (\|f(\cdot)\|_{\mathcal{H}_{K_p}}^2 - \boldsymbol{f}' G_{K_p}^{+} \boldsymbol{f}) \tag{28}$$

holds.

Proof From Theorem 1, Lemmas 2, 3 and 4, and the fact that $f(\cdot) = \sum_{p=1}^{n} \alpha_p f(\cdot)$, we have

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

which concludes the proof. □

Accordingly, we have

$$\begin{aligned} &E_m - E_e \\ &\geq (\|f\|_{\mathcal{H}_{K_u}}^2 - \boldsymbol{f}' G_{K_u}^{+} \boldsymbol{f}) \\ &\quad - \sum_{p=1}^{n} \alpha_p (\|f(\cdot)\|_{\mathcal{H}_{K_p}}^2 - \boldsymbol{f}' G_{K_p}^{+} \boldsymbol{f}) \\ &= \left(\|f\|_{\mathcal{H}_{K_u}}^2 - \sum_{p=1}^{n} \alpha_p \|f(\cdot)\|_{\mathcal{H}_{K_p}}^2 \right) \\ &\quad + \left(\sum_{p=1}^{n} \alpha_p \boldsymbol{f}' G_{K_p}^{+} \boldsymbol{f} - \boldsymbol{f}' G_{K_u}^{+} \boldsymbol{f} \right). \end{aligned} \tag{29}$$

Let T_1 and T_2 be the first and the second terms in Eq. (29), and we analyze them in the following.

Here, we consider the linear class $S \subset L$ such that

$$\|f(\cdot)\|_S = \|f(\cdot)\|_{\mathcal{H}_{K_p}},\ (p \in \{1, \ldots, n\}) \tag{30}$$

for any $f(\cdot) \in S$. Note that such a linear class always exists since the norm of $f(\boldsymbol{x}) = 0 \in L$ is identical to zero in any Hilbert space[†]. For S, there exists a kernel K_S such that

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

is also a kernel from Theorem 2. Hereafter, we use $\mathcal{H}_{K_S}$ instead of S since K_S is guaranteed to be a kernel. Note that

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

holds from Theorem 5, which immediately yields

$$\mathcal{H}_{K_S} \cap \mathcal{H}_{K^c} = \{0\}, \tag{33}$$

where $K^c = \sum_{p=1}^{n} \alpha_p K_p^c$. Therefore, we have

$$\begin{aligned} &\sum_{p=1}^{n} \alpha_p \|f(\cdot)\|_{\mathcal{H}_{K_p}}^2 \\ &= \sum_{p=1}^{n} \alpha_p \|f(\cdot)\|_{\mathcal{H}_{K_S}}^2 = \sum_{p=1}^{n} \alpha_p \|f(\cdot)\|_{\mathcal{H}_{K_u}}^2 = \|f(\cdot)\|_{\mathcal{H}_{K_u}}^2 \end{aligned} \tag{34}$$

for any $f(\cdot) \in \mathcal{H}_{K_S}$ from Theorem 5, and $T_1 = 0$ is obtained.

Lemma 6: Let $G_p \in \mathbf{R}^{m \times m}$, $(p \in \{1, \ldots, n\})$ be non-negative definite symmetric matrices and $\boldsymbol{v} \in \cap_{p=1}^{n} \mathcal{R}(G_p)$ and let α_p, $(p \in \{1, \ldots, n\})$ be positive constants satisfying $\sum_{p=1}^{n} \alpha_p = 1$. Then,

$$\boldsymbol{v}' \left(\sum_{p=1}^{n} \alpha_p G_p^{+} - \left(\sum_{p=1}^{n} \alpha_p G_p \right)^{+} \right) \boldsymbol{v} \geq 0 \tag{35}$$

holds.

Proof Let $S = \sum_{p=1}^{n} \alpha_p G_p$ and $T = \sum_{p=1}^{n} \alpha_p G_p^{+}$, then $\mathcal{R}(S) = \mathcal{R}(S^{+}) = \mathcal{R}(T)$ trivially holds. Thus, we have

$$\begin{aligned} &\boldsymbol{v}'(T - S^{+})\boldsymbol{v} \\ &= \boldsymbol{v}' S^{+} S (T - S^{+}) S S^{+} \boldsymbol{v} = \boldsymbol{v}' S^{+} (STS - S) S^{+} \boldsymbol{v}, \end{aligned}$$

since $\boldsymbol{v} \in \cap_{p=1}^{n} \mathcal{R}(G_p) \subset \mathcal{R}(S^{+}) = \mathcal{R}(T)$. Here, the matrix $STS - S$ can be represented as

$$\begin{aligned} &STS - S \\ &= \sum_{p=1}^{n} \alpha_p (S - G_p + G_p) G_p^{+} (S - G_p + G_p) - S \\ &= \sum_{p=1}^{n} \alpha_p (S - G_p) G_p^{+} (S - G_p) \\ &\quad + \sum_{p=1}^{n} \alpha_p (S - G_p) G_p^{+} G_p + \sum_{p=1}^{n} \alpha_p G_p G_p^{+} (S - G_p) \end{aligned}$$

[†]As the same with L, $S = \{0\}$ is a meaningless linear class. Thus, we are interested in the case of $\dim S \geq 1$.

$$= \sum_{p=1}^{n} \alpha_p (S - G_p) G_p^+ (S - G_p) + \sum_{p=1}^{n} \alpha_p S G_p^+ G_p + \sum_{p=1}^{n} \alpha_p G_p G_p^+ S - 2S.$$

Let

$$U_1 = \sum_{p=1}^{n} \alpha_p (S - G_p) G_p^+ (S - G_p),$$

$$U_2 = \sum_{p=1}^{n} \alpha_p S G_p^+ G_p + \sum_{p=1}^{n} \alpha_p G_p G_p^+ S - 2S.$$

Since $\boldsymbol{v} \in \mathcal{R}(G_p) \subset \mathcal{R}(S)$ yields

$$\boldsymbol{v}' S^+ \left(\sum_{p=1}^{n} \alpha_p S G_p^+ G_p \right) S^+ \boldsymbol{v} = \sum_{p=1}^{n} \alpha_p \boldsymbol{v}' S^+ S G_p^+ G_p S^+ \boldsymbol{v} = \sum_{p=1}^{n} \alpha_p \boldsymbol{v}' S^+ \boldsymbol{v} = \boldsymbol{v}' S^+ \boldsymbol{v},$$

$$\boldsymbol{v}' S^+ \left(\sum_{p=1}^{n} \alpha_p G_p G_p^+ S \right) S^+ \boldsymbol{v} = \sum_{p=1}^{n} \alpha_p \boldsymbol{v}' S^+ G_p G_p^+ S S^+ \boldsymbol{v} = \sum_{p=1}^{n} \alpha_p \boldsymbol{v}' S^+ \boldsymbol{v} = \boldsymbol{v}' S^+ \boldsymbol{v},$$

we have

$$\boldsymbol{v}' S^+ U_2 S^+ \boldsymbol{v} = 2\boldsymbol{v}' S^+ \boldsymbol{v} - 2\boldsymbol{v}' S^+ S S^+ \boldsymbol{v} = 0. \tag{36}$$

Since U_1 is non-negative definite, it is concluded that

$$\boldsymbol{v}'(T - S^+)\boldsymbol{v} = \boldsymbol{v}' S^+ U_1 S^+ \boldsymbol{v} \geq 0 \tag{37}$$

holds, which concludes the proof. □

Note that Lemma 6 is an extension of the relationship between the reciprocals of the arithmetic mean and the harmonic mean [24] to quadratic forms with non-negative definite symmetric matrices.

The next theorem is the main result of this paper.

Theorem 6: If $f(\cdot) \in \mathcal{H}_{K_S}$,

$$E_m - E_e \geq 0 \tag{38}$$

holds.

Proof As mentioned above, $T_1 = 0$ holds for any $f(\cdot) \in \mathcal{H}_{K_S}$. Since $G_{K_u} = \sum_{p=1}^{m} \alpha_p G_{K_p}$, and $\boldsymbol{f} \in \mathcal{R}(G_{K_p})$ holds for any $p \in \{1, \ldots, n\}$, we have $T_2 \geq 0$ from Lemma 6. Therefore,

$$E_m - E_e \geq T_1 + T_2 = T_2 \geq 0$$

is immediately obtained for any $f(\cdot) \in \mathcal{H}_{K_S}$, which concludes the proof. □

According to Theorem 6, it is concluded that the ensemble kernel regressor yields a better result than the multiple kernel regressor for any $f(\cdot) \in \mathcal{H}_{K_S}$ in terms of the theoretical limit of the generalization error.

Note that the condition Eq. (30) may be too strong except for trivial (and meaningless) cases, such as $S = \{0\}$. Thus, we analyze the lower bound of $E_m - E_e$ for $f(\cdot) \notin \mathcal{H}_{K_S}$ next.

Since we assume that RKHS's are separable, there exists a countable set $Z_p = \{z_k^{(p)} \mid k \in \mathbf{N},\ z_k \in \mathcal{D}\}$ for each K_p, $(p \in \{1, \ldots, n\})$ by which the set $\{K_p(\cdot, z_k^{(p)}) \mid k \in \mathbf{N}\}$ is dense in $\mathcal{H}_{K_p}$; and there also exists a countable set $Z_u = \{z_k^{(u)} \mid k \in \mathbf{N},\ z_k \in \mathcal{D}\}$ by which the set $\{K_u(\cdot, z_k^{(u)}) \mid k \in \mathbf{N}\}$ is dense in $\mathcal{H}_{K_u}$. Let $Z = (\cup_{p=1}^{n} Z_p) \cup Z_u = \{z_k \mid k \in \mathbf{N}\}$, then the set $\{K_p(\cdot, z_k) \mid k \in \mathbf{N}\}$ is dense in $\mathcal{H}_{K_p}$ and the set $\{K_u(\cdot, z_k) \mid k \in \mathbf{N}\}$ is dense in $\mathcal{H}_{K_u}$. Since we assume $f(\cdot) \in L$, there exists a countable set $\{\beta_k^{(p)} \mid k \in \mathbf{N},\ \beta_k^{(p)} \in \mathbf{R}\}$ such that

$$f(\cdot) = \sum_{k \in \mathbf{N}} \beta_k^{(p)} K_p(\cdot, z_k) \tag{39}$$

for each $p \in \{1, \ldots, n\}$ and there exists a countable set $\{\beta_k \mid k \in \mathbf{N},\ \beta_k \in \mathbf{R}\}$ such that

$$f(\cdot) = \sum_{k \in \mathbf{N}} \beta_k K_u(\cdot, z_k). \tag{40}$$

The coefficient vectors $\boldsymbol{\beta}^{(p)} = [\beta_1^{(p)}, \ldots, \beta_k^{(p)}, \ldots]'$ and $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_k, \ldots]'$ are connected by

$$G_{K_u,Z} \boldsymbol{\beta} = G_{K_p,Z} \boldsymbol{\beta}^{(p)}, \tag{41}$$

where $G_{K_u,Z}$ and $G_{K_p,Z}$ denote the Gram matrices of K_u and K_p with Z. On the basis of these preparations, we have

$$\begin{aligned} T_1 &= \|f(\cdot)\|_{\mathcal{H}_{K_u}}^2 - \sum_{p=1}^{n} \alpha_p \|f(\cdot)\|_{\mathcal{H}_{K_p}}^2 \\ &= \boldsymbol{\beta}' G_{K_u,Z} \boldsymbol{\beta} - \sum_{p=1}^{n} \alpha_p (\boldsymbol{\beta}^{(p)})' G_{K_p,Z} \boldsymbol{\beta}^{(p)} \\ &= \boldsymbol{\beta}' G_{K_u,Z} G_{K_u,Z}^+ G_{K_u,Z} \boldsymbol{\beta} - \sum_{p=1}^{n} \alpha_p (\boldsymbol{\beta}^{(p)})' G_{K_p,Z} G_{K_p,Z}^+ G_{K_p,Z} \boldsymbol{\beta}^{(p)} \\ &= \boldsymbol{\beta}' G_{K_u,Z} G_{K_u,Z}^+ G_{K_u,Z} \boldsymbol{\beta} - \sum_{p=1}^{n} \alpha_p \boldsymbol{\beta}' G_{K_u,Z} G_{K_p,Z}^+ G_{K_u,Z} \boldsymbol{\beta} \\ &= \boldsymbol{\beta}' G_{K_u,Z} \left(G_{K_u,Z}^+ - \sum_{p=1}^{n} \alpha_p G_{K_p,Z}^+ \right) G_{K_u,Z} \boldsymbol{\beta}, \end{aligned}$$

which implies that $T_1 \leq 0$ for $f(\cdot) \notin \mathcal{H}_{K_S}$ from Lemma 6. Also, we have

$$\boldsymbol{f} = G_{K_u,(X,Z)} \boldsymbol{\beta} \tag{42}$$

from Eq. (40), where $G_{K_u,(X,Z)} = (K_u(\boldsymbol{x}_i, \boldsymbol{z}_j))$. Let

$$M_1 = G_{K_u,Z}^+ - \sum_{p=1}^{n} \alpha_p G_{K_p,Z}^+,$$

$$M_2 = \sum_{p=1}^{m} \alpha_p G_{K_p}^+ - G_{K_u}^+,$$

and

$$M = \left(G_{K_u,Z} M_1 G_{K_u,Z} + G_{K_u,(Z,X)} M_2 G_{K_u,(X,Z)}\right),$$

where $G_{K_u,(Z,X)} = G'_{K_u,(Z,X)}$. Then, we have a lower bound of $E_m - E_e$ for $f(\cdot) \notin \mathcal{H}_{K_S}$ written as

$$E_m - E_e \geq \boldsymbol{\beta}' M \boldsymbol{\beta}. \tag{43}$$

Since $T_1 \leq 0$ and $T_2 \geq 0$, the lower bound of $E_m - E_e$ with a fixed M (specified by X and the kernels) in Eq. (43) could be both positive and negative in general, depending on $\boldsymbol{\beta}$ that specifies the unknown true function $f(\cdot)$. Accordingly, when $f(\cdot) \notin \mathcal{H}_{K_S}$, $E_m - E_e \geq 0$ is guaranteed only for the unknown true function specified by $\boldsymbol{\beta}$ that makes the quadratic form Eq. (43) non-negative[†]. These analyses reveals the importance of the condition Eq. (30) and it implies that Eq. (30) is not too strong condition to obtain Eq. (38) in terms of the lower bound. However, we can not eliminate the possibility of $E_m - E_e \geq 0$ for all $f(\cdot) \notin \mathcal{H}_{K_S}$ even if the obtained lower bound $\boldsymbol{\beta}' M \boldsymbol{\beta}$ is negative with a certain $\boldsymbol{\beta}$ since our lower bound may be too loose. Thus, we give an example of $f(\cdot) \notin \mathcal{H}_{K_S}$ that makes actual value of $E_m - E_e$ negative in the next section.

Note that when $\mathrm{tr}(M) > 0$, $E_m > E_e$ is expected for many functions in L together with the fact that $\boldsymbol{\beta}' M \boldsymbol{\beta}$ is a lower bound of $E_m - E_e$. Thus, $\mathrm{tr}(M)$ can be used as a measure for deciding whether or not the ensemble kernel regressor is better (in some sense) than the multiple kernel regressor.

6. Numerical Examples

In this section, we give some numerical examples confirming our theoretical results obtained in the previous section with a simple polynomial kernel defined by

$$K_p(x, y) = (1 + xy)^p, \quad x, y \in \mathbf{R}, \tag{44}$$

where p denote a positive integer. We consider $\mathcal{K} = \{K_1, K_2\}$ as a class of kernels. 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 $b_2(x)$, that is, $L = \mathrm{span}\{1, x\} = \{a + bx \mid a, b \in \mathbf{R}\}$. Note that since $\dim\mathcal{H}_{K_1} = 2$ and $\dim\mathcal{H}_{K_2} = \dim\mathcal{H}_{K_u} = 3$, we can adopt $Z = \{-1, 0, 1\}$ that yields a dense set for each RKHS, such as $\{K_u(x, -1), K_u(x, 0), K_u(x, 1)\}$ for $\mathcal{H}_{K_u}$. Since

$$f(x) = a + bx = (a - b)K_1(x, 0) + bK_1(x, 1), \tag{45}$$

we have

$$\|f(x)\|^2_{\mathcal{H}_{K_1}} = a^2 + b^2. \tag{46}$$

Similarly, since

$$\begin{aligned} f(x) &= a + bx \\ &= -\frac{b}{4}K_2(x, -1) + aK_2(x, 0) + \frac{b}{4}K_2(x, 1), \end{aligned} \tag{47}$$

we have

$$\|f(x)\|^2_{\mathcal{H}_{K_2}} = a^2 + \frac{b^2}{2}. \tag{48}$$

Therefore, Eq. (30) holds if and only if $b = 0$, which implies that $\mathcal{H}_{K_S} = \mathrm{span}\{1\} = \{a \mid a \in \mathbf{R}\}$.

We adopt $\alpha_1 = 2/3$ and $\alpha_2 = 1/3$ as the coefficients for convex combinations. Then, we have

$$K_u(x, y) = \frac{2}{3}K_1(x, y) + \frac{1}{3}K_2(x, y) = 1 + \frac{4}{3}xy + \frac{1}{3}x^2y^2.$$

We investigate the generalization errors E_m and E_e for $f(\cdot) = ax + b \in L$ which is or is not included in $\mathcal{H}_{K_S}$. We adopt $X = \{1\}$ as the input training data set.

Since $f(1) = a + b$, learning results by K_1 and K_2 are reduced to

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

from Eq. (16). Therefore, the learning result by ensemble kernel regressor is given as

$$\hat{f}_e(x) = \frac{2}{3}\hat{f}_1(x) + \frac{1}{3}\hat{f}_2(x) = \frac{a + b}{12}(5 + 6x + x^2). \tag{49}$$

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

$$\hat{f}_m(x) = \frac{a + b}{8}(3 + 4x + x^2). \tag{50}$$

Note that

$$\begin{aligned} d_e(x) = \hat{f}_e(x) - f(x) &= \frac{a + b}{12}(5 + 6x + x^2) - (a + bx) \\ &= \frac{1}{12}((-7a + 5b) + 6(a - b)x + (a + b)x^2) \\ &= \frac{-a + 5b}{16}K_u(x, -1) + \frac{-5a + b}{6}K_u(x, 0) \\ &\quad + \frac{5a - b}{16}K_u(x, 1), \\ d_m(x) = \hat{f}_m(x) - f(x) &= \frac{a + b}{8}(3 + 4x + x^2) - (a + bx) \\ &= \frac{1}{8}((-5a + 3b) + 4(a - b)x + (a + b)x^2) \\ &= \frac{3b}{8}K_u(x, -1) - aK_u(x, 0) + \frac{3a}{8}K_u(x, 1) \end{aligned}$$

hold. Therefore, we have

$$E_e = \|d_e(x)\|^2_{\mathcal{H}_{K_u}} = \frac{79a^2 - 118ab + 55b^2}{144} \tag{51}$$

$$E_m = \|d_m(x)\|^2_{\mathcal{H}_{K_u}} = \frac{5a^2 - 6ab + 3b^2}{8}, \tag{52}$$

and the actual value of $E_m - E_e$ is reduced to

[†]Since $f(\cdot)$ is unknown, $\boldsymbol{\beta}$ is also unknown, which implies that we can not identify the better regressor only from the training data set when $f(\cdot) \notin \mathcal{H}_{K_S}$.

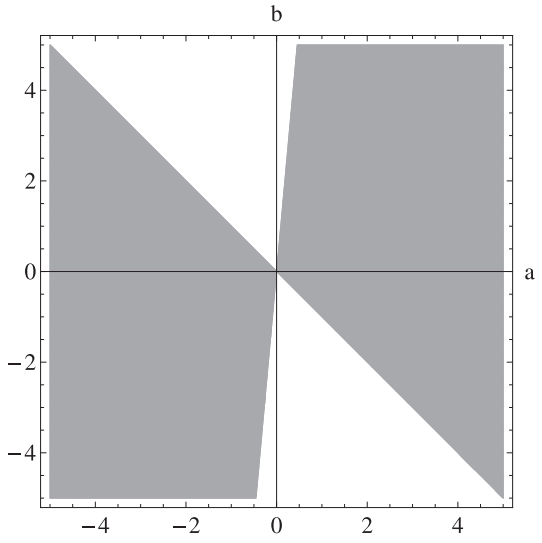


Fig. 1 The region of positive $E_m - E_e$ (gray) and the region of negative $E_m - E_e$ (white).

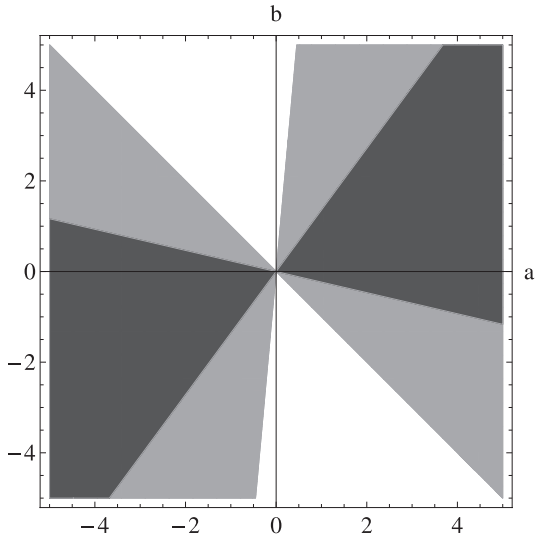


Fig. 2 The region of positive lower bound given in Eq. (43) (dark gray) superimposed on Fig. 1.

$$E_m - E_e = \frac{11a^2 + 10ab - b^2}{144} = \frac{(11a - b)(a + b)}{144}. \quad (53)$$

6.1 Example for $f(\cdot) \in \mathcal{H}_{K_S}$

As mentioned above, Eq. (30) holds if and only if $b = 0$ and then $\mathcal{H}_{K_S} = \mathrm{span}\{1\}$ hold. Therefore, we have

$$E_m - E_e = \frac{11}{144}a^2 > 0$$

for any $f(x) = a \in \mathcal{H}_{K_S}$. Accordingly, it is confirmed that the inequality Eq. (38) surely holds for any function in $\mathcal{H}_{K_S}$ in these settings.

6.2 Example for $f(\cdot) \notin \mathcal{H}_{K_S}$

From Eq. (53), $E_m - E_e \geq 0$ is satisfied only when $(11a - b)(a + b) \geq 0$. Figure 1 shows the regions in the a-b plane, where the actual value of $E_m - E_e$ is positive (gray region) or negative (white region).

According to this result, it is concluded that there exists $f(\cdot) \notin \mathcal{H}_{K_S}$, say $f(x) = 1 + x$, satisfying $E_m - E_e \geq 0$ (corresponding to gray region in Fig. 1), which implies that $f(\cdot) \in \mathcal{H}_{K_S}$ is not a necessary condition for $E_m - E_e \geq 0$. Also, the existence of $f(\cdot) \notin \mathcal{H}_{K_S}$, say $f(x) = x$, that makes $E_m - E_e$ negative is confirmed (white region in Fig. 1), which numerically supports the importance of Eq. (30) to obtain Eq. (38), while it is only supported in terms of a lower bound by the analyses given in the previous section. Therefore, it is concluded that the ensemble (or multiple) kernel regressor can be better than the other in terms of the theoretical limit of the generalization error when $f(\cdot) \notin \mathcal{H}_{K_S}$.

Finally, we investigate the tightness of the lower bound obtained in Eq. (43). Since

$$f(x) = \beta_1 K_u(x, -1) + \beta_2 K_u(x, 0) + \beta_3 K_u(x, 1)$$
$$= \left(\sum_{i=1}^{3} \beta_i\right) + \frac{3x}{4}(-\beta_1 + \beta_3) + \frac{x^2}{3}(\beta_1 + \beta_3),$$

and $f(\cdot) \in L, \beta_1 + \beta_3 = 0$ is required, Therefore, the unknown true function can be rewritten as

$$f(x) = \beta_2 - \frac{3\beta_1}{2}x, \quad (54)$$

which implies

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} -\frac{2}{3}b \\ a \\ \frac{2}{3}b \end{bmatrix}. \quad (55)$$

Substituting Eq. (55) to the lower bound $\boldsymbol{\beta}' M \boldsymbol{\beta}$ in Eq. (43) yields

$$\boldsymbol{\beta}' M \boldsymbol{\beta} = \frac{a^2}{24} + \frac{4ab}{27} - \frac{32b^2}{243}. \quad (56)$$

Figure 2 shows the region in the a-b plane, where the lower bound given in Eq. (43) is positive (dark gray region), which is superimposed on Fig. 1.

According to this result, it is confirmed that the lower bound obtained in Eq. (43) is not so tight, and there exist a function, say $f(x) = 1 + 2x$, whose actual $E_m - E_e$ is positive while the corresponding lower bound is negative. This is caused by the inequality given in Lemma 4. Thus, improvement of Lemma 4 is one of future works that should be undertaken in order to make $\mathrm{tr}(M)$ to be more accurate as the measure for ensuring the advantage of the ensemble kernel regressor.

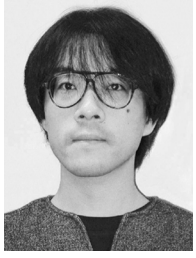
7. Conclusion

In this paper, we introduced a unified framework for evaluating the generalization errors of the kernel regressors with

multiple kernels, and analyzed the ensemble kernel regressor and the multiple kernel regressor, which are the representative kernel regressors using multiple kernels. As a result, we obtained a sufficient condition for the ensemble kernel regressor to outperform the multiple kernel regressor in terms of the theoretical limit of the generalization errors. We also clarified that the superiority of the ensemble kernel regressor was deeply related to the relationship between the arithmetic mean and the harmonic mean. Moreover, we further analyzed their generalization errors in the cases where the sufficient condition was not satisfied, and clarified that each regressor could be better than the other in such cases, depending on the unknown true function.

References

[1] K. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," IEEE Trans. Neural Networks, vol.12, pp.181–201, 2001.
[2] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1999.
[3] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, Cambridge, 2000.
[4] M. Sugiyama and H. Ogawa, "Subspace information criterion for model selection," Neural Computation, vol.13, no.8, pp.1863–1889, 2001.
[5] 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.
[6] A. Tanaka, H. Imai, M. Kudo, and M. Miyakoshi, "Optimal kernel in a class of kernels with an invariant metric," Joint IAPR Internatioanl Workshops SSPR 2008 and SPR 2008, vol.5342, pp.530–539, Springer, 2008.
[7] A. Tanaka and M. Miyakoshi, "Theoretical analyses for a class of kernels with an invariant metric," 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.2074–2077, 2010.
[8] A. Tanaka, H. Imai, M. Kudo, and M. Miyakoshi, "Theoretical analyses on a class of nested RKHS's," 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2011), pp.2072–2075, 2011.
[9] A. Tanaka, I. Takigawa, H. Imai, and M. Kudo, "Extended analyses for an optimal kernel in a class of kernels with an invariant metric," Joint IAPR Internatioanl Workshops SSPR 2012 and SPR 2012, vol.7626, pp.345–353, Springer, 2012.
[10] S. Sonnenburg, G. Ratsch, C. Schafer, and B. Scholkopf, "Large scale multiple kernel learning," J. Machine Learning Research, vol.7, pp.1531–1565, 2006.
[11] A. Tanaka, I. Takigawa, H. Imai, and M. Kudo, "Theoretical analyses on ensemble and multiple kernel regressors," 6th Asian Conference on Machine Learning (ACML2014), 2014.
[12] A. Tanaka, I. Takigawa, H. Imai, and M. Kudo, "Analyses on generalizaion errors of multiple and ensemble kernel regressors," 29th Signal Processing Symposium, pp.425–426, 2014.
[13] H. Takebayashi and A. Tanaka, "A consideration on generalization ability of multiple and ensemble kernel regressors," 29th Signal Processing Symposium, pp.120–123, 2014.
[14] N. Aronszajn, "Theory of reproducing kernels," Trans. Am. Math. Soc., vol.68, no.3, pp.337–404, 1950.
[15] J. Mercer, "Functions of positive and negative type and their connection with the theory of integral equations," Trans. London Philosophical Society, vol.A, no.209, pp.415–446, 1909.
[16] M. Reed and B. Simon, Methods of Modern Mathematical Physics I: Functional Analysis (Revised and Enlarged Edition), Academic Press, San Diego, 1980.
[17] A. Tanaka, H. Imai, and M. Kudo, "A sufficient condition for reproducing kernel hilbert spaces being separable," 28th Signal Processing Symposium, pp.350–354, 2013.
[18] R. Schatten, Norm Ideals of Completely Continuous Operators, Springer-Verlag, Berlin, 1960.
[19] S. Saitoh, Integral Transforms, Reproducing Kernels and Their Applications, Addison Wesley Longman, UK, 1997.
[20] H. Ogawa, "Neural networks and generalization ability," IEICE Technical Report, NC95-8, 1995.
[21] M. Sugiyama and H. Ogawa, "Incremental active learning for optimal generalization," Neural Computation, vol.12, no.12, pp.2909–2940, 2000.
[22] M. Sugiyama and H. Ogawa, "Active learning for optimal generalization in trigonometric polynomial models," IEICE Trans. Fundamentals, vol.E84-A, no.9, pp.2319–2329, 2001.
[23] C.R. Rao and S.K. Mitra, Generalized Inverse of Matrices and its Applications, John Wiley & Sons, 1971.
[24] G.H. Hardy, J.E. Littlewood, and G. Pólya, Inequalities, 2nd Ed., Cambridge University Press, 1952.



Akira Tanaka received the D.E. from Hokkaido University in 2000. He joined the Graduate School of Information Science and technology, Hokkaido University. His research interests include image processing, acoustic signal processing, and machine learning theory.

Hirofumi Takebayashi received the B.E. from Hokkaido University in 2013. He is a master course student in the Graduate School of Information Science and technology, Hokkaido University. His research interests include machine learning theory.

Ichigaku Takigawa received the D.E. from Hokkaido University in 2004. He joined the Graduate School of Information Science and Technology, Hokkaido University. His research interests include machine learning and data mining.

Hideyuki Imai received the D.E. from Hokkaido University in 1999. He joined the Graduate School of Information Science and Technology, Hokkaido University. His research interests include statistical inference.

Mineichi Kudo received his Dr. Eng. degree in Information Engineering from the Hokkaido University in 1988. Starting from an instructor, since 2001, he is a a professor (2001-) in Hokkaido University. In 2001 he received with professor Jack Sklansky the twenty-seventh annual pattern recognition society award. He was elected to a fellow of the International Association for Pattern Recognition on December 10, 2008. His current research interests include design of pattern recognition systems, image processing, data mining and computational learning theory. He is a member of the Pattern Recognition Society and the IEEE.