IEICE TRANSACTIONS

on Fundamentals of Electronics, Communications and Computer Sciences

VOL. E100-A NO. 3
MARCH 2017

A PUBLICATION OF THE ENGINEERING SCIENCES SOCIETY
The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER

Theoretical Analyses on 2-Norm-Based Multiple Kernel Regressors

Akira TANAKA[†a)] ***and*** **Hideyuki IMAI**[†], ***Members***

SUMMARY The solution of the standard 2-norm-based multiple kernel regression problem and the theoretical limit of the considered model space are discussed in this paper. We prove that 1) The solution of the 2-norm-based multiple kernel regressor constructed by a given training data set does not generally attain the theoretical limit of the considered model space in terms of the generalization errors, even if the training data set is noise-free, 2) The solution of the 2-norm-based multiple kernel regressor is identical to the solution of the single kernel regressor under a noise free setting, in which the adopted single kernel is the sum of the same kernels used in the multiple kernel regressor; and it is also true for a noisy setting with the 2-norm-based regularizer. The first result motivates us to develop a novel framework for the multiple kernel regression problems which yields a better solution close to the theoretical limit, and the second result implies that it is enough to use the single kernel regressors with the sum of given multiple kernels instead of the multiple kernel regressors as long as the 2-norm based criterion is used.

key words: *multiple kernel regressor, reproducing kernel Hilbert space, generalization error, 2-norm criterion, 2-norm regularizer*

1. Introduction

The kernel method [1]–[3] is widely recognized as one of powerful tools in the field of machine learning, such as pattern recognition, regression estimation and density estimation; and it has been imported into the field of signal processing such as system identification problems [4], [5]. So far, many kernel-based learning machines have been proposed such as the support vector machines [3], [6] and the kernel ridge regressors [2]. Historically, the kernel-based learning started with a single kernel; and then it has been extended to the learning with multiple kernels [7]. In early multiple kernel learning methods, a kernel constructed by a linear (or convex) combination of kernels was used as an alternative single kernel, in which the main interest was how to construct a better kernel [2], [8]. Today, various frameworks of multiple kernel learning have been developed (See [7] and its references therein). The ensemble kernel learning [6] is a representative one, in which a learning result is constructed as a combination of the leaning results by each kernel through the boosting strategy for instance. As a more flexible model, we can consider a linear combination of all (n) kernels with all (ℓ) input training vectors, which necessarily has $n\ell$ coefficients to be determined [7]. This model is widely recognized to be more flexible than another models such as the single kernel regressor with a combined kernels.

Since many frameworks for the multiple kernel learning have been proposed, as mentioned above, theoretical analyses of their generalization capability and their advantages are needed to give a guideline in order to choose an appropriate method to each learning problem. In such theoretical analyses, statistical or asymptotic approaches were popular (See [9]–[11] for instance) in which a certain probabilistic structure was assumed in a training data set. On the other hand, it is also important to analyze the generalization capability of kernel machines which is specified only by a given training data set since we have to construct a learning result by a given training data set only in practical problems.

In general, a learning result of the kernel machines is modeled as a linear combination of the kernel functions whose one variable is fixed to each training input vector. It implies that the learning result belongs to the reproducing kernel Hilbert space (RKHS) which is uniquely specified by the kernel. Therefore, it is natural to assume that an unknown target for the estimation also belongs to the RKHS corresponding to the kernel in a mathematical analysis of the generalization capability of the kernel machines. In this paper, an unknown target for the estimation is assumed to be in an RKHS, which necessarily leads us to the analyses on the kernel-based regression problems.

The common objective of machine learning is to minimize the generalization error. Roughly speaking, the generalization error of a kernel machine is a difference between an unknown true function and an estimated one at all points (which may not be in a given training data set) in a domain. This implies that it is natural to adopt the (squared) norm of the difference between an unknown true function and an estimated one as the generalization error [12], [13] since the absolute difference of the unknown true function and an estimated one is upper bounded by the product of a certain constant (specified by an adopted kernel and a considered point) and the norm of the difference of them, which is a straightforward consequence of the reproducing property of kernels and the Schwarz's inequality.

As an optimization criterion in the kernel machines, we usually adopt the combination of an empirical error measure, such as the classical squared loss and a hinge loss [2], [3], and a regularizer, such as the ℓ_p-norm of coefficients used in a linear combination [2], [3] or the (squared) norm of the unknown true function itself evaluated in a cer-

Manuscript received November 16, 2015.
Manuscript revised October 21, 2016.

[†]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.E100.A.877

tain RKHS [11], [14]. Our generalization error (the norm of the difference between an unknown true function and an estimated one) partly supports the validity of adopting the (squared) norm of an estimated function as the regularizer since the minimum norm of an estimated function deeply related to the orthogonal projection of the unknown true function onto the linear subspace corresponding to the learning model, which is the optimal solution in terms of our generalization error. Thus, we adopt the squared norm of the estimated function evaluated in an RKHS as the target regularizer for the analysis. Although there exist many criteria for evaluating the empirical error, as mentioned above, we do not have sufficient theoretical knowledge about the generalization capability in terms of our generalization error even for the fundamental ℓ_2-based measure. Accordingly, we analyze the generalization capability of the fundamental multiple kernel regressor by the ℓ_2-norm-based empirical error measure with the squared-norm-based regularizer of an estimated function. Specifically, we discuss the following two issues in this paper. One is whether or not we can construct the orthogonal projection of the unknown true function onto the model space, that is the theoretically best solution, by a given training data set; and we give a negative conclusion for this issue, which gives us a motivation to develop a novel framework that gives a learning result which is closer to the theoretical limit. The other is the properties of the solution of the multiple kernel regressors by the 2-norm-based formulation. If we can not obtain the theoretical limit by the 2-norm-based formulation, we are interested in the relationship between the solution and the theoretical limit. Thus, we theoretically investigate their relationship and prove that the solution is identical to that by the single kernel regressor with the sum of the considered kernels, which reveals the superiority of a learning with a combined (single) kernel in terms of computational costs. We believe that our analyses provide useful tools for the analyses of the kernel machines based on another approaches.

The rest of this paper is organized as follows. In Sect. 2, we review the theory of reproducing kernel Hilbert spaces and give an overview of the single kernel regressors. In Sect. 3, we discuss the properties of the multiple kernel regressors and give some theoretical results. In Sect. 4, we give toy examples confirming our theoretical results. Finally, we give concluding remarks in Sect. 5.

2. Preliminaries

2.1 Theory of Reproducing Kernel Hilbert Spaces

Here, we prepare some mathematical tools concerned with the theory of reproducing kernel Hilbert spaces [15]–[17].

Definition 1: [15] Let $\mathbf{R}^d$ be a d-dimensional real vector space and let $\mathcal{H}$ be a class of functions defined on a domain $D \subset \mathbf{R}^d$, forming a Hilbert space of real-valued functions. The function $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$, $(\boldsymbol{x}, \tilde{\boldsymbol{x}} \in D)$ is called a reproducing kernel of $\mathcal{H}$, if the following two conditions hold.

1. For every fixed $\tilde{\boldsymbol{x}} \in D$,
$$K_{\tilde{\boldsymbol{x}}}(\cdot) = K(\cdot, \tilde{\boldsymbol{x}}) \in \mathcal{H}. \tag{1}$$
2. For every $\tilde{\boldsymbol{x}} \in D$ and every $f(\cdot) \in \mathcal{H}$,
$$f(\tilde{\boldsymbol{x}}) = \langle f(\cdot), K(\cdot, \tilde{\boldsymbol{x}}) \rangle_{\mathcal{H}}, \tag{2}$$
where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ stands for the inner product of the Hilbert space $\mathcal{H}$.

A Hilbert space that has a reproducing kernel K is called a reproducing kernel Hilbert space (RKHS), and it is denoted by $\mathcal{H}_K$. The reproducing property Eq. (2) enables us to treat a value of a function at a point in D, while we can not deal with a value of a function in a general Hilbert space such as $L^2(D)$ (the Hilbert space of all square-integrable functions defined on D). Note that reproducing kernels are positive definite:

$$\sum_{i,j=1}^{N} c_i c_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0, \tag{3}$$

for any positive integer $N \in \mathbf{N}$, $c_1, \ldots, c_N \in \mathbf{R}$, and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in D$ [15], where $\mathbf{N}$ stands for the set of natural numbers. In addition, $K(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = K(\tilde{\boldsymbol{x}}, \boldsymbol{x})$ for any $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in D$ is followed [15]. If a reproducing kernel $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ exists, it is unique [15]. Conversely, every positive definite function $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ has the unique corresponding RKHS [15]. In this paper, we assume that all RKHSs are separable [18] since many popular RKHSs are separable.

The following theorem, concerned with the sum of reproducing kernels, plays an important role in the analyses of the multiple kernel regressors.

Theorem 1: [15] 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

$$\|f(\cdot)\|^2 = \min \left[\|f_1(\cdot)\|_1^2 + \|f_2(\cdot)\|_2^2 \right], \tag{4}$$

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

This theorem shows that the sum of the kernels $K = K_1 + K_2$ has the unique corresponding RKHS with the norm defined by Eq. (4). Note that the original kernels K_1 and K_2 are no longer reproducing kernels of the RKHS corresponding to the kernel $K = K_1 + K_2$. This theorem is easily extended to the case of the sum of more than two reproducing kernels. In the following contents, we use the notation 'kernel' instead of 'reproducing kernel' for simplicity.

2.2 Overview of Single Kernel Regressor and Its Generalization Error

In this part, we review regression problems with a single

kernel and introduce the generalization error of the kernel regressor as preparations for the analyses on the multiple kernel regressors.

Let $\{(y_i, \boldsymbol{x}_i) \mid i \in \{1, \ldots, \ell\}\}$ be a given training data set with ℓ samples, where $y_i \in \mathbf{R}$ denotes an output value and $\boldsymbol{x}_i \in \mathbf{R}^d$ denotes the corresponding input vector, generated by the model:

$$y_i = f(\boldsymbol{x}_i) + n_i, \tag{5}$$

where $f(\cdot)$ denotes an unknown true function and n_i denotes an additive noise. The aim of regression problem is to estimate the unknown true function $f(\cdot)$ by using the given training data set and statistical properties of the noise (if available).

In the single kernel regressor using a kernel K, a learning result is modeled as a linear combination of $K(\cdot, \boldsymbol{x}_i), (i \in \{1, \ldots, \ell\})$, written as

$$\hat{f}(\cdot) = \sum_{i=1}^{\ell} c_i K(\cdot, \boldsymbol{x}_i) \tag{6}$$

with some coefficients $c_i \in \mathbf{R}, (i \in \{1, \ldots, \ell\})$. In general, the minimizer of the generalization error is adopted as the coefficients c_i. In this paper, we adopt

$$E_S(f(\cdot), \hat{f}(\cdot); \mathcal{H}_{K_e}) = \|f(\cdot) - \hat{f}(\cdot)\|^2_{\mathcal{H}_{K_e}} \tag{7}$$

as the generalization error of $\hat{f}(\cdot)$ [12], [13], [19], [20], where K_e is a certain kernel whose corresponding RKHS includes $f(\cdot)$ and $\hat{f}(\cdot)$. The subscript 'S' stands for 'Single kernel'. If $f(\cdot) \in \mathcal{H}_K$, it is natural to adopt $\mathcal{H}_K$ as $\mathcal{H}_{K_e}$ since $\hat{f}(\cdot) \in \mathcal{H}_K$ holds by Eqs. (1) and (6). The validity of using $E_S(f(\cdot), \hat{f}(\cdot); \mathcal{H}_K)$ is supported by the fact that

$$\begin{aligned} &|f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})| \\ &= |\langle f(\cdot) - \hat{f}(\cdot), K(\cdot, \boldsymbol{x})\rangle_{\mathcal{H}_K}| \\ &\le \|f(\cdot) - \hat{f}(\cdot)\|_{\mathcal{H}_K} \|K(\cdot, \boldsymbol{x})\|_{\mathcal{H}_K} \\ &= \|f(\cdot) - \hat{f}(\cdot)\|_{\mathcal{H}_K} K(\boldsymbol{x}, \boldsymbol{x})^{1/2}, \end{aligned} \tag{8}$$

holds for any $\boldsymbol{x} \in D$, by the reproducing property Eq. (2) of a kernel and the Schwarz's inequality. Therefore, minimizing $E_S(f(\cdot), \hat{f}(\cdot); \mathcal{H}_K)$ surely reduces the upper bound of the absolute error at any point in D.

Since we assume that $\mathcal{H}_K$ is separable, there exists a countable set Λ such that $\mathrm{span}\{K(\cdot, z_k) \mid z_k \in D, k \in \Lambda\}$ is dense in $\mathcal{H}_K$. Thus, the unknown true function $f(\cdot) \in \mathcal{H}_K$ can be represented by

$$f(\cdot) = \sum_{k \in \Lambda} \alpha_k K(\cdot, z_k), \tag{9}$$

with some (unknown) coefficients α_k, $(k \in \Lambda)$. Therefore, we have

$$\begin{aligned} &E_S(f(\cdot), \hat{f}(\cdot); \mathcal{H}_K) \\ &= \Big\| \sum_{k \in \Lambda} \alpha_k K(\cdot, z_k) - \sum_{i=1}^{\ell} c_i K(\cdot, \boldsymbol{x}_i) \Big\|^2_{\mathcal{H}_K} \\ &= \boldsymbol{\alpha}' G_K^{ZZ} \boldsymbol{\alpha} - 2\boldsymbol{\alpha}' G_K^{ZX} \boldsymbol{c} + \boldsymbol{c}' G_K^{XX} \boldsymbol{c}. \end{aligned} \tag{10}$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{|\Lambda|}]'$ with $|\Lambda|$ denoting the cardinality of Λ (or $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots]'$ in case of $|\Lambda| = \infty$), $\boldsymbol{c} = [c_1, \ldots, c_\ell]'$ and the superscript $'$ denoting the transposition operator. $G_K^{XX} = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))$, $G_K^{ZZ} = (K(z_i, z_j))$, and $G_K^{ZX} = (K(z_i, \boldsymbol{x}_j))$, satisfying $(G_K^{ZX})' = G_K^{XZ}$, are the kernel matrices of K with $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\ell\}$ and $Z = \{z_1, \ldots, z_{|\Lambda|}\}$ (or $Z = \{z_1, z_2, \ldots\}$ in case of $|\Lambda| = \infty$). We assume that all kernel matrices are of full rank since kernel matrices are usually of full rank except in the pathological cases, such as the cases with a training data set including duplicated training samples. Since Eq. (10) is a positive quadratic form, the minimizer of Eq. (10) is identified as its stationary point, which is obtained as

$$\frac{\partial E_S(f(\cdot), \hat{f}(\cdot); \mathcal{H}_K)}{\partial \boldsymbol{c}} = 2G_K^{XX} \boldsymbol{c} - 2G_K^{XZ} \boldsymbol{\alpha} = \boldsymbol{0}. \tag{11}$$

Therefore, the typical[†] optimal coefficient vector $\boldsymbol{c}$ is reduced to

$$\hat{\boldsymbol{c}}_S^{TL} = (G_K^{XX})^{-1} G_K^{XZ} \boldsymbol{\alpha} = (G_K^{XX})^{-1} \boldsymbol{f}, \tag{12}$$

since $\boldsymbol{f} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_\ell)]' = G_K^{XZ} \boldsymbol{\alpha}$ holds from Eq. (9). Note that the minimizer of $E_S(f(\cdot), \hat{f}(\cdot); \mathcal{H}_K)$ produces the orthogonal projection in $\mathcal{H}_K$ of the unknown true function $f(\cdot)$ onto the linear subspace

$$L_S = \mathrm{span}\{K(\cdot, \boldsymbol{x}_i) \mid i \in \{1, \ldots, \ell\}\}. \tag{13}$$

Hereafter, we call the function constructed by $\hat{\boldsymbol{c}}_S^{TL}$ the theoretical limit of the model space L_S since $\hat{\boldsymbol{c}}_S^{TL}$ is the theoretically best solution.

Here, we consider the following problem, which is a fundamental formulation of the kernel regressor with a single kernel.

Problem 1: Find the coefficient vector $\boldsymbol{c}$ of the model Eq. (6) that minimizes

$$J_S(\boldsymbol{c}) = \|\hat{f}(\cdot)\|^2_{\mathcal{H}_K} \tag{14}$$

subject to

$$y_i = \hat{f}(\boldsymbol{x}_i), \quad i \in \{1, \ldots, \ell\}. \tag{15}$$

The constraint Eq. (15) can be represented by

$$\boldsymbol{y} = G_K^{XX} \boldsymbol{c} \tag{16}$$

with $\boldsymbol{y} = [y_1, \ldots, y_\ell]'$ and the criterion Eq. (14) is reduced to

$$J_S(\boldsymbol{c}) = \boldsymbol{c}' G_K^{XX} \boldsymbol{c}. \tag{17}$$

It is well known [21] that the solution of Problem 1 is given by

$$\hat{\boldsymbol{c}}_1 = (G_K^{XX})^{-1} \boldsymbol{y}, \tag{18}$$

[†]The word 'typical' is used for the minimum-norm least-squares solution.

which agrees with the coefficient vector Eq. (12) of the theoretical limit with a noise-free training data set[†]. Accordingly, it is concluded that the empirical error minimization scheme (corresponding to Eq. (15)) achieves the minimum generalization error in the single kernel regressor under the conditions $f(\cdot) \in \mathcal{H}_K$ and $K_e = K$.

In case of a noisy training data set, a regularization scheme is usually adopted. The following problem is the formulation of the regression problem with a standard 2-norm-based regularizer.

Problem 2: Find the coefficient vector $\boldsymbol{c}$ of the model Eq. (6) that minimizes

$$J_{SR}(\boldsymbol{c}) = \lambda \|\hat{f}(\cdot)\|^2_{\mathcal{H}_K} + \sum_{i=1}^{\ell} (y_i - \hat{f}(\boldsymbol{x}_i))^2, \tag{19}$$

where λ is a positive regularization parameter.

The subscript 'SR' stands for 'Single kernel with Regularization'. The criterion Eq. (19) is reduced to

$$J_{SR}(\boldsymbol{c}) = \lambda \boldsymbol{c}' G_K^{XX} \boldsymbol{c} + \|\boldsymbol{y} - G_K^{XX} \boldsymbol{c}\|^2 \tag{20}$$

and its minimizer is easily obtained as

$$\hat{\boldsymbol{c}}_2 = (G_K^{XX} + \lambda I_\ell)^{-1} \boldsymbol{y}, \tag{21}$$

where I_ℓ denotes the identity matrix of degree ℓ.

3. Analyses on Multiple Kernel Regressors with 2-Norm-Based Criterion

In this section, we consider the class of kernels $\mathcal{K} = \{K_1, \ldots, K_n\}$ and discuss theoretical properties of the multiple kernel regressors using kernels in $\mathcal{K}$. A learning result of the multiple kernel regressor discussed in this paper is modeled as

$$\hat{f}(\cdot) = \sum_{p=1}^{n} \sum_{i=1}^{\ell} c_i^{(p)} K_p(\cdot, \boldsymbol{x}_i), \tag{22}$$

which is one of the most popular models in the multiple kernel scenario [7], [11], [14]. In contrast to the single kernel regressor, we face one problem, that is, a selection of an RKHS $\mathcal{H}_{K_e}$ which is used to evaluate the generalization error of the model Eq. (22). Since we have to evaluate the norm of individual kernel $K_p(\cdot, \boldsymbol{x}_i)$ in a certain RKHS $\mathcal{H}_{K_e}$, it must include the RKHSs corresponding to all the kernels in $\mathcal{K}$. One reasonable resolution is adopting the RKHS corresponding to the sum of all kernels in $\mathcal{K}$ as K_e, written as

$$K_u = \sum_{p=1}^{n} K_p, \tag{23}$$

whose validity is supported by Theorem 1[††]. As in [11], [14], the norm of an estimated function is generally evaluated in the RKHS corresponding to a certain positively weighted linear combination of given kernels in the multiple kernel regressors, while K_u seems to be an unweighted sum. However, K_u can deal with the evaluation by a weighted sum. In fact, given a set of kernels $\tilde{\mathcal{K}} = \{\tilde{K}_1, \ldots, \tilde{K}_n\}$ and their positive weights $\Theta = \{\theta_1, \ldots, \theta_n\}$, then $\mathcal{K} = \{K_1, \ldots, K_n\}$ with $K_p = \theta_p \tilde{K}_p$ surely produces a positively weighted linear combination of the kernels in $\tilde{\mathcal{K}}$. Note that our target for the analysis is the generalization capability with a fixed Θ and the selection of a better Θ is out of the scope of this paper (See [8] for the optimal selection of Θ).

Here, we give some preparations for the inner product of the kernels in $\mathcal{H}_{K_u}$.

Lemma 1: For every $\boldsymbol{x}, \boldsymbol{z} \in D$,

$$\sum_{p=1}^{n} \langle K_p(\cdot, \boldsymbol{x}), K_q(\cdot, \boldsymbol{z}) \rangle_{\mathcal{H}_{K_u}} = K_q(\boldsymbol{x}, \boldsymbol{z}), \tag{24}$$

$$\sum_{p=1}^{n} \sum_{q=1}^{n} \langle K_p(\cdot, \boldsymbol{x}), K_q(\cdot, \boldsymbol{z}) \rangle_{\mathcal{H}_{K_u}} = K_u(\boldsymbol{x}, \boldsymbol{z}) \tag{25}$$

hold.

Proof From the reproducing property Eq. (2), we have

$$\begin{aligned} &\sum_{p=1}^{n} \langle K_p(\cdot, \boldsymbol{x}), K_q(\cdot, \boldsymbol{z}) \rangle_{\mathcal{H}_{K_u}} \\ &= \Big\langle \sum_{p=1}^{n} K_p(\cdot, \boldsymbol{x}), K_q(\cdot, \boldsymbol{z}) \Big\rangle_{\mathcal{H}_{K_u}} \\ &= \langle K_u(\cdot, \boldsymbol{x}), K_q(\cdot, \boldsymbol{z}) \rangle_{\mathcal{H}_{K_u}} = K_q(\boldsymbol{x}, \boldsymbol{z}), \end{aligned}$$

since $K_q(\cdot, \boldsymbol{z}) \in \mathcal{H}_{K_u}$, which yields Eq. (24). Also we have,

$$\begin{aligned} &\sum_{p=1}^{n} \sum_{q=1}^{n} \langle K_p(\cdot, \boldsymbol{x}), K_q(\cdot, \boldsymbol{z}) \rangle_{\mathcal{H}_{K_u}} \\ &= \Big\langle \sum_{p=1}^{n} K_p(\cdot, \boldsymbol{x}), \sum_{q=1}^{n} K_q(\cdot, \boldsymbol{z}) \Big\rangle_{\mathcal{H}_{K_u}} \\ &= \langle K_u(\cdot, \boldsymbol{x}), K_u(\cdot, \boldsymbol{z}) \rangle_{\mathcal{H}_{K_u}} = K_u(\boldsymbol{x}, \boldsymbol{z}), \end{aligned}$$

which yields Eq. (25). □

We assume that the unknown true function $f(\cdot)$ belongs to $\mathcal{H}_{K_u}$ in the following contents. From the assumption of separable RKHSs, there exists a countable set Λ such that $\mathrm{span}\{K_u(\cdot, \boldsymbol{z}_k) \mid \boldsymbol{z}_k \in D,\ k \in \Lambda\}$ is dense in $\mathcal{H}_{K_u}$; and the unknown true function can be represented by

$$f(\cdot) = \sum_{k \in \Lambda} \alpha_k K_u(\cdot, \boldsymbol{z}_k) \tag{26}$$

[†]Since we assume that the kernel matrix is non-singular, the criterion Eq. (14) does not affect the solution.

[††]In our previous work [22], we adopted an alternative K_e specialized for the Gaussian kernel, which does not lead to the results in this paper shown below.

with certain (unknown) coefficients α_k, $(k \in \Lambda)$ as the same with the single kernel case.

First of all, we identify the theoretical limit of the model space, that is, the orthogonal projection of the unknown true function onto the model space

$$L_M = \mathrm{span}\{K_p(\cdot, \boldsymbol{x}_i) \mid p \in \{1, \ldots, n\},\ i \in \{1, \ldots, \ell\}\}. \tag{27}$$

Here, the subscript 'M' stands for 'Multiple kernel'. The generalization error of $\hat{f}(\cdot)$ evaluated in $\mathcal{H}_{K_u}$ is reduced to

$$\begin{aligned} E_M(f(\cdot), \hat{f}(\cdot); \mathcal{H}_{K_u}) &= \|f(\cdot) - \hat{f}(\cdot)\|^2_{\mathcal{H}_{K_u}} \\ &= \Big\| \sum_{k\in\Lambda} \alpha_k K_u(\cdot, \boldsymbol{z}_k) - \sum_{p=1}^{n} \sum_{i=1}^{\ell} c_i^{(p)} K_p(\cdot, \boldsymbol{x}_i) \Big\|^2_{\mathcal{H}_{K_u}} \\ &= \boldsymbol{\alpha}' G^{ZZ}_{K_u} \boldsymbol{\alpha} - 2 \sum_{p=1}^{n} \boldsymbol{\alpha}' G^{ZX}_{K_p} \boldsymbol{c}^{(p)} \\ &\quad + \sum_{p=1}^{n} \sum_{q=1}^{n} (\boldsymbol{c}^{(p)})' H^{XX}_{K_p, K_q} \boldsymbol{c}^{(q)}, \end{aligned} \tag{28}$$

where $\boldsymbol{c}^{(p)} = [c_1^{(p)}, \ldots, c_\ell^{(p)}]'$ and

$$\begin{aligned} G^{ZZ}_{K_u} &= (\langle K_u(\cdot, \boldsymbol{z}_i), K_u(\cdot, \boldsymbol{z}_j) \rangle_{\mathcal{H}_{K_u}}) = K_u(\boldsymbol{z}_i, \boldsymbol{z}_j), \\ G^{ZX}_{K_p} &= (\langle K_u(\cdot, \boldsymbol{z}_i), K_p(\cdot, \boldsymbol{x}_j) \rangle_{\mathcal{H}_{K_u}}) = K_p(\boldsymbol{z}_i, \boldsymbol{x}_j), \\ H^{XX}_{K_p, K_q} &= (\langle K_p(\cdot, \boldsymbol{x}_i), K_q(\cdot, \boldsymbol{x}_j) \rangle_{\mathcal{H}_{K_u}}), \end{aligned}$$

which are well-defined since $K_p(\cdot, \boldsymbol{x}_i) \in \mathcal{H}_{K_u}$. Let

$$H^{XX} = \begin{bmatrix} H^{XX}_{K_1,K_1} & \cdots & H^{XX}_{K_1,K_n} \\ \vdots & \ddots & \vdots \\ H^{XX}_{K_n,K_1} & \cdots & H^{XX}_{K_n,K_n} \end{bmatrix},$$

$$\boldsymbol{c} = \begin{bmatrix} \boldsymbol{c}^{(1)} \\ \vdots \\ \boldsymbol{c}^{(n)} \end{bmatrix}, \quad G^{XZ} = \begin{bmatrix} G^{XZ}_{K_1} \\ \vdots \\ G^{XZ}_{K_n} \end{bmatrix},$$

then Eq. (28) is rewritten as

$$\begin{aligned} &E_M(f(\cdot), \hat{f}(\cdot); \mathcal{H}_{K_u}) \\ &= \boldsymbol{\alpha}' G^{ZZ}_{K_u} \boldsymbol{\alpha} - 2\boldsymbol{\alpha}' (G^{XZ})' \boldsymbol{c} + \boldsymbol{c}' H^{XX} \boldsymbol{c}. \end{aligned} \tag{29}$$

Since $E_M(f(\cdot), \hat{f}(\cdot); \mathcal{H}_{K_u})$ is a positive quadratic form, the minimizer of $E_M(f(\cdot), \hat{f}(\cdot); \mathcal{H}_{K_u})$ is identified as its stationary point, which is obtained as

$$\frac{\partial E_M(f(\cdot), \hat{f}(\cdot); \mathcal{H}_{K_u})}{\partial \boldsymbol{c}} = 2H^{XX}\boldsymbol{c} - 2G^{XZ}\boldsymbol{\alpha} = \boldsymbol{0}, \tag{30}$$

which is reduced to the linear equation

$$H^{XX}\boldsymbol{c} = G^{XZ}\boldsymbol{\alpha}. \tag{31}$$

Therefore, the coefficient vector of the theoretical limit is obtained by

$$\hat{\boldsymbol{c}}^{TL}_M = (H^{XX})^{-1} G^{XZ} \boldsymbol{\alpha}. \tag{32}$$

Generally, we can not construct this theoretical limit only from the given training data set, even if it is noise-free on the contrary to the single kernel case. In fact, we only have $\boldsymbol{f} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_\ell)]'$, which can be represented by

$$\boldsymbol{f} = G^{XZ}_{K_u} \boldsymbol{\alpha} = \sum_{p=1}^{n} G^{XZ}_{K_p} \boldsymbol{\alpha} = (\boldsymbol{1}'_n \otimes I_\ell) G^{XZ} \boldsymbol{\alpha}, \tag{33}$$

by Eq. (26), which is a direct consequence of Lemma 1, where $\boldsymbol{1}_n$ denotes the n-dimensional vector whose all elements are unity and $\otimes$ denotes the Kronecker product [23] of two matrices. In order to obtain $G^{XZ}\boldsymbol{\alpha}$ in Eq. (32) from $\boldsymbol{f}$, we have to decompose $\boldsymbol{f}$ into the series of $G^{XZ}_{K_p}\boldsymbol{\alpha}$, $(p \in \{1, \ldots, n\})$. However, it is impossible since $\boldsymbol{\alpha}$ is unknown. We will discuss details of this issue later.

As mentioned before, the theoretical limit Eq. (32) gives the orthogonal projection of the unknown true function $f(\cdot)$ onto the linear subspace L_M. Thus, the obtained minimum generalization error is reduced to

$$\begin{aligned} E^{(min)}_M &= \min_{\boldsymbol{c}} E_M(f(\cdot), \hat{f}(\cdot); \mathcal{H}_{K_u}) \\ &= \boldsymbol{\alpha}' G^{ZZ}_{K_u} \boldsymbol{\alpha} - (\hat{\boldsymbol{c}}^{TL}_M)' (H^{XX}) \hat{\boldsymbol{c}}^{TL}_M \\ &= \boldsymbol{\alpha}' G^{ZZ}_{K_u} \boldsymbol{\alpha} - \boldsymbol{\alpha}' (G^{XZ})' (H^{XX})^{-1} G^{XZ} \boldsymbol{\alpha} \end{aligned} \tag{34}$$

by the Pythagorean theorem.

3.1 Analyses on Noise Free Cases

Now, we introduce the solution of the multiple kernel regressor defined by a typical 2-norm-based criterion for noise-free training data set; and discuss the relationship between the generalization errors of the estimated function by its solution and the theoretical limit, that is, the function constructed by Eq. (32).

The most popular and fundamental 2-norm-based multiple kernel regressor is formalized as the following problem.

Problem 3: Find the coefficient vector $\boldsymbol{c}$ of the model Eq. (22) that minimizes

$$J_M(\boldsymbol{c}) = \|\hat{f}(\cdot)\|^2_{\mathcal{H}_{K_u}} \tag{35}$$

subject to

$$y_i = \hat{f}(\boldsymbol{x}_i), \quad i \in \{1, \ldots, \ell\}. \tag{36}$$

The constraint Eq. (36) can be represented by

$$\boldsymbol{y} = \sum_{p=1}^{n} G^{XX}_{K_p} \boldsymbol{c}^{(p)} = (G^{XX})' \boldsymbol{c}, \tag{37}$$

where $G^{XX} = [G^{XX}_{K_1}, \ldots, G^{XX}_{K_n}]'$ and the criterion Eq. (35) is reduced to

$$J_M(\boldsymbol{c}) = \boldsymbol{c}' H^{XX} \boldsymbol{c}. \tag{38}$$

The solution of Problem 3 is easily obtained as

$$\hat{\boldsymbol{c}}_3 = (H^{XX})^{-1}G^{XX}\{(G^{XX})'(H^{XX})^{-1}G^{XX}\}^{-1}\boldsymbol{y} \tag{39}$$

as shown in [21]. From Eq. (33), this solution can be also represented by

$$\hat{\boldsymbol{c}}_3 = (H^{XX})^{-1}G^{XX}\{(G^{XX})'(H^{XX})^{-1}G^{XX}\}^{-1}G^{XZ}_{K_u}\boldsymbol{\alpha}, \tag{40}$$

when the training data set is noise-free.

Firstly, we investigate the properties of the learning result by the solution $\hat{\boldsymbol{c}}_3$ of Problem 3.

Theorem 2: Let $\hat{f}_3(\cdot)$ be the estimated function by $\hat{\boldsymbol{c}}_3$, then

$$\langle f(\cdot) - \hat{f}_3(\cdot), \hat{f}_3(\cdot)\rangle_{\mathcal{H}_{K_u}} = 0 \tag{41}$$

holds.

Proof Since $K_u = \sum_{p=1}^{n} K_p$ yields

$$\sum_{i=1}^{n} H^{XX}_{K_p,K_i} = \sum_{i=1}^{n} H^{XX}_{K_i,K_p} = G^{XX}_{K_p},$$

from Lemma 1, we have

$$(H^{XX})^{-1}G^{XX} = (H^{XX})^{-1}H^{XX}(\mathbf{1}_n \otimes I_\ell) = (\mathbf{1}_n \otimes I_\ell). \tag{42}$$

Therefore,

$$\begin{aligned}
&\langle f(\cdot), \hat{f}_3(\cdot)\rangle_{\mathcal{H}_{K_u}} = \boldsymbol{\alpha}'(G^{XZ})'\hat{\boldsymbol{c}}_3 \\
&= \boldsymbol{\alpha}'(G^{XZ})'(H^{XX})^{-1}G^{XX} \\
&\quad \times\{(G^{XX})'(H^{XX})^{-1}G^{XX}\}^{-1}G^{XZ}_{K_u}\boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}'(G^{XZ})'(\mathbf{1}_n \otimes I_\ell) \\
&\quad \times(\{G^{XX})'(H^{XX})^{-1}G^{XX}\}^{-1}G^{XZ}_{K_u}\boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}'(G^{XZ}_{K_u})'\{(G^{XX})'(H^{XX})^{-1}G^{XX}\}^{-1}G^{XZ}_{K_u}\boldsymbol{\alpha}
\end{aligned} \tag{43}$$

holds. On the other hand, we have

$$\begin{aligned}
&\|\hat{f}_3(\cdot)\|^2_{\mathcal{H}_{K_u}} = \hat{\boldsymbol{c}}_3'H^{XX}\hat{\boldsymbol{c}}_3 \\
&= \boldsymbol{\alpha}'(G^{XZ}_{K_u})'\{(G^{XX})'(H^{XX})^{-1}G^{XX}\}^{-1}(G^{XX})' \\
&\quad \times(H^{XX})^{-1}G^{XX}\{(G^{XX})'(H^{XX})^{-1}G^{XX}\}^{-1}G^{XZ}_{K_u}\boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}'(G^{XZ}_{K_u})'\{(G^{XX})'(H^{XX})^{-1}G^{XX}\}^{-1}G^{XZ}_{K_u}\boldsymbol{\alpha}.
\end{aligned} \tag{44}$$

Accordingly, Eqs. (43) and (44) imply

$$\begin{aligned}
&\langle f(\cdot) - \hat{f}_3(\cdot), \hat{f}_3(\cdot)\rangle_{\mathcal{H}_{K_u}} \\
&= \langle f(\cdot), \hat{f}_3(\cdot)\rangle_{\mathcal{H}_{K_u}} - \|\hat{f}^3(\cdot)\|^2_{\mathcal{H}_{K_u}} = 0,
\end{aligned}$$

which concludes the proof. □

According to Theorem 2, the estimated function of Problem 3 by $\hat{\boldsymbol{c}}_3$ given by Eq. (40) is the orthogonal projection of the unknown true function $f(\cdot)$ onto a certain linear subspace $\tilde{L}_M \subset L_M$. From Theorem 2 and the Pythagorean theorem, the generalization error of the estimated function by the solution of Problem 3 is reduced to

$$\begin{aligned}
E^{(est)}_M &= E_M(f(\cdot), \hat{f}_3(\cdot); \mathcal{H}_{K_u}) \\
&= \boldsymbol{\alpha}'G^{ZZ}_{K_u}\boldsymbol{\alpha} - \|\hat{f}_3(\cdot)\|^2_{\mathcal{H}_{K_u}} \\
&= \boldsymbol{\alpha}'G^{ZZ}_{K_u}\boldsymbol{\alpha} - \hat{\boldsymbol{c}}_3'H^{XX}\hat{\boldsymbol{c}}_3 \\
&= \boldsymbol{\alpha}'G^{ZZ}_{K_u}\boldsymbol{\alpha} \\
&\quad -\boldsymbol{\alpha}'(G^{XZ}_{K_u})'\{(G^{XX})'(H^{XX})^{-1}G^{XX}\}^{-1} \\
&\quad \times G^{XZ}_{K_u}\boldsymbol{\alpha}.
\end{aligned} \tag{45}$$

Secondary, we analyze the relationship between the generalization errors of the estimated function of Problem 3 constructed by $\hat{\boldsymbol{c}}_3$ and the theoretical limit, that is, the function constructed by $\hat{\boldsymbol{c}}^{TL}_M$.

Lemma 2: Let $H_{ij} \in \mathbf{R}^{m\times m}$, $i, j \in \{1, \ldots, n\}$) be positive definite symmetric matrices and assume that

$$H = \begin{bmatrix} H_{11} & \cdots & H_{1n} \\ \vdots & \ddots & \vdots \\ H_{n1} & \cdots & H_{nn} \end{bmatrix}$$

is also positive definite. Then

$$M = H^{-1} - (\mathbf{1}_n \otimes I_m)((\mathbf{1}'_n \otimes I_m)H(\mathbf{1}_n \otimes I_m))^{-1}(\mathbf{1}_n \otimes I_m) \tag{46}$$

is non-negative definite and $\mathcal{N}(M) = \mathcal{R}(H(\mathbf{1}_n \otimes I_m))$, where $\mathcal{R}(\cdot)$ and $\mathcal{N}(\cdot)$ denote the range space and the null space of a given matrix.

Proof Let $W = H^{1/2}(\mathbf{1}_n \otimes I_m)$, then we have,

$$M = H^{-1/2}(I_{mn} - W(W'W)^{-1}W')H^{-1/2}. \tag{47}$$

Since $W(W'W)^{-1}W'$ is the orthogonal projector onto $\mathcal{R}(W)$, $I_{mn} - W(W'W)^{-1}W'$ is the orthogonal projector onto its orthogonal complement, whose eigenvalues are trivially non-negative, which implies that M is a non-negative definite matrix by the Sylvester's law of inertia.

From Eq. (47), it is trivial that

$$\mathcal{N}(M) = \mathcal{R}(H^{1/2}W) = \mathcal{R}(H(\mathbf{1}_n \otimes I_m))$$

holds. □

Theorem 3:

$$E^{(est)}_M \geq E^{(min)}_M \tag{48}$$

holds and the equality is obtained if and only if

$$\hat{\boldsymbol{c}}^{TL}_M \in \mathcal{R}(\mathbf{1}_n \otimes I_\ell) \tag{49}$$

holds.

Proof Since

$$\begin{aligned}
&(G^{XX})'(H^{XX})^{-1}G^{XX} \\
&= (\mathbf{1}'_n \otimes I_\ell)H^{XX}(H^{XX})^{-1}H^{XX}(\mathbf{1}_n \otimes I_\ell) \\
&= (\mathbf{1}'_n \otimes I_\ell)H^{XX}(\mathbf{1}_n \otimes I_\ell)
\end{aligned} \tag{50}$$

and

$$(G^{XZ}_{K_u})' = (G^{XZ})'(\mathbf{1}_n \otimes I_\ell) \tag{51}$$

hold from Lemma 1, we have

$$\begin{aligned}
&E^{(est)}_M - E^{(min)}_M \\
&= \boldsymbol{\alpha}'(G^{XZ})'(H^{XX})^{-1}G^{XZ}\boldsymbol{\alpha} \\
&\quad -\boldsymbol{\alpha}'(G^{XZ}_{K_u})'\{(G^{XX})'(H^{XX})^{-1}G^{XX}\}^{-1}G^{XZ}_{K_u}\boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}'(G^{XZ})'(H^{XX})^{-1}G^{XZ}\boldsymbol{\alpha} \\
&\quad -\boldsymbol{\alpha}'(G^{XZ})'(\mathbf{1}_n \otimes I_\ell)\{(G^{XX})'(H^{XX})^{-1}G^{XX}\}^{-1} \\
&\quad \times(\mathbf{1}'_n \otimes I_\ell)G^{XZ}\boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}'(G^{XZ})'(H^{XX})^{-1}G^{XZ}\boldsymbol{\alpha} \\
&\quad -\boldsymbol{\alpha}'(G^{XZ})'(\mathbf{1}_n \otimes I_\ell)\{(\mathbf{1}'_n \otimes I_\ell)H^{XX}(\mathbf{1}_n \otimes I_\ell)\}^{-1} \\
&\quad \times(\mathbf{1}'_n \otimes I_\ell)G^{XZ}\boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}(G^{XZ})'\tilde{M}^{XX}G^{XZ}\boldsymbol{\alpha},
\end{aligned} \tag{52}$$

where

$$\begin{aligned}
&\tilde{M}^{XX} \\
&= (H^{XX})^{-1} \\
&\quad -(\mathbf{1}_n \otimes I_\ell)\{(\mathbf{1}'_n \otimes I_\ell)H^{XX}(\mathbf{1}_n \otimes I_\ell)\}^{-1}(\mathbf{1}'_n \otimes I_\ell).
\end{aligned}$$

Therefore, Lemma 2 immediately yields $E^{(est)}_M \geq E^{(min)}_M$.

From Lemma 2, $E^{(est)}_M = E^{(min)}_M$ holds only when

$$G^{XZ}\boldsymbol{\alpha} \in \mathcal{N}(\tilde{M}^{XX}) = \mathcal{R}(H^{XX}(\mathbf{1}_n \otimes I_\ell)), \tag{53}$$

which is identical to

$$\hat{\boldsymbol{c}}^{TL}_M = (H^{XX})^{-1}G^{XZ}\boldsymbol{\alpha} \in \mathcal{R}(\mathbf{1}_n \otimes I_\ell), \tag{54}$$

which concludes the proof. □

The former claim in Theorem 3 is easy to understand since $E^{(min)}_M$ is the theoretical lower bound of the generalization error of the model space L_M. In contrast, the latter claim in Theorem 3 is important since it implies that the estimated function by the optimal solution of Problem 3 can not always achieve the the theoretical lower bound of the generalization error. In fact, the condition Eq. (49) implies that $\hat{\boldsymbol{c}}^{TL}_M$ must lie on a certain ℓ-dimensional subspace (since $\mathrm{rank}(\mathbf{1}_n \otimes I_\ell) = \ell$) in (ℓn)-dimensional vector space. Note that Eqs. (8) and (48) imply that the upper bound of the absolute error of the theoretical limit is smaller than that of the estimated function obtained by Problem 3 at any point $\boldsymbol{x} \in D$, which may not be in X, when Eq. (49) does not hold.

Finally, we identify the linear subspace $\tilde{L}_M$ and its relation to the linear subspace $\mathcal{R}(\mathbf{1}_n \otimes I_\ell)$ given in Theorem 3. To this end, we consider the following problem.

Problem 4: Find the coefficient vector $\boldsymbol{c} = [c_1, \ldots, c_\ell]'$ for the model

$$\hat{f}(\cdot) = \sum_{i=1}^{\ell} c_i K_u(\cdot, \boldsymbol{x}_i) \tag{55}$$

that minimizes

$$J_C = \|\hat{f}(\cdot)\|^2_{\mathcal{H}_{K_u}} \tag{56}$$

subject to

$$y_i = \hat{f}(\boldsymbol{x}_i), \quad i \in \{1, \ldots, \ell\}. \tag{57}$$

The subscript 'C' stands for 'Combined kernel'. Although K_u involves multiple kernels in $\mathcal{K}$, it is reduced to a single kernel. Thus, the solution of Problem 4 is given by

$$\hat{\boldsymbol{c}}_4 = (G^{XX}_{K_u})^{-1}\boldsymbol{y} \tag{58}$$

as the same with Problem 1; and it is reduced to

$$\hat{\boldsymbol{c}}_4 = (G^{XX}_{K_u})^{-1}\boldsymbol{f} = (G^{XX}_{K_u})^{-1}G^{XZ}_{K_u}\boldsymbol{\alpha} \tag{59}$$

when the training data set is noise-free. We have the following important theorem for the relationship between the estimated functions by the solutions of Problems 3 and 4.

Theorem 4: The estimated functions by the solutions of Problems 3 and 4 are identical.

Proof From Eqs. (42) and (50),

$$(\mathbf{1}'_n \otimes I_\ell)H^{XX}(\mathbf{1}_n \otimes I_\ell) = G^{XX}_{K_u}, \tag{60}$$

holds, which is a direct consequence of by Lemma 1. Thus, we have

$$\begin{aligned}
&\hat{\boldsymbol{c}}_3 \\
&= (H^{XX})^{-1}G^{XX}((G^{XX})'(H^{XX})^{-1}G^{XX})^{-1}G^{XZ}_{K_u}\boldsymbol{\alpha} \\
&= (\mathbf{1}_n \otimes I_\ell)(G^{XX}_{K_u})^{-1}G^{XZ}_{K_u}\boldsymbol{\alpha} \\
&= (\mathbf{1}_n \otimes I_\ell)\hat{\boldsymbol{c}}_4.
\end{aligned}$$

Then the estimated function of Problem 3 is reduced to

$$\hat{f}_3(\cdot) = \sum_{p=1}^{n}\sum_{i=1}^{\ell}(\hat{\boldsymbol{c}}_4)_i K_p(\cdot, \boldsymbol{x}_i) = \sum_{i=1}^{\ell}(\hat{\boldsymbol{c}}_4)_i K_u(\cdot, \boldsymbol{x}_i), \tag{61}$$

where $(\hat{\boldsymbol{c}}_4)_i$ is the i-th component of $\hat{\boldsymbol{c}}_4$, which is identical with the estimated function $\hat{f}_4(\cdot)$ of Problem 4 by the optimal coefficient vector $\hat{\boldsymbol{c}}_4$.

□

According to Theorem 4, it is concluded that the linear subspace $\tilde{L}_M$ is identified as

$$\tilde{L}_M = \mathrm{span}\{K_u(\cdot, \boldsymbol{x}_i) \mid i \in \{1, \ldots, \ell\}\}. \tag{62}$$

Theorem 4 also claims that it is enough to consider the simpler Problem 4 instead of Problem 3 as long as the 2-norm-based criterion is used. Note that Theorem 4 is consistent with the representer theorem (See [8] and their references therein) since the representer theorem claims that the optimal solution of Problem 3 (and Problem 4), without any assumption on the model of a function, is specified by Eq. (55). Moreover, this theorem gives the intuitive interpretation of the linear subspace $\mathcal{R}(\mathbf{1}_n \otimes I_\ell)$ given in Theorem 3, that is, $\hat{\boldsymbol{c}}^{TL}_M$ satisfying the condition Eq. (49) yields the coefficient vectors $\hat{\boldsymbol{c}}^{(p)}$, $(p \in \{1, \ldots, n\})$ which is independent

from the kernel index p, and those vectors make the orthogonal projection of the unknown true function $f(\cdot)$ onto the linear subspace L_M also lying on $\tilde{L}_M$.

3.2 Analyses on Noisy Cases with 2-Norm Regularizer

When the training data set is noisy, a regularization scheme is usually adopted to stabilize the solution in regression problems. In this part, we discuss the multiple kernel regressor by a 2-norm-based criterion with a 2-norm-based regularizer; and prove that a similar result to Theorem 4, obtained for noise-free cases, also holds for a noisy case.

Let us consider the following problem, that is, a regularized version of Problem 3.

Problem 5: Find the coefficient vector $\boldsymbol{c}$ of the model Eq. (22) that minimizes

$$J_{MR}(\boldsymbol{c}) = \lambda \|\hat{f}(\cdot)\|^2_{\mathcal{H}_{K_u}} + \sum_{i=1}^{\ell} (y_i - \hat{f}(\boldsymbol{x}_i))^2, \tag{63}$$

where λ denotes a positive regularization parameter.

The subscript 'MR' stands for 'Multiple kernel with Regularization'. The criterion Eq. (63) can be also represented by

$$J_{MR}(\boldsymbol{c}) = \lambda \boldsymbol{c}' H^{XX} \boldsymbol{c} + \|\boldsymbol{y} - (G^{XX})' \boldsymbol{c}\|^2, \tag{64}$$

and the minimizer of Problem 5 is easily obtained as

$$\hat{\boldsymbol{c}}_5 = (G^{XX}(G^{XX})' + \lambda H^{XX})^{-1} G^{XX} \boldsymbol{y}. \tag{65}$$

Also, let us consider the following problem, that is, a regularized version of Problem 4.

Problem 6: Find the coefficient vector $\boldsymbol{c}$ of the model Eq. (55) that minimizes

$$J_{CR}(\boldsymbol{c}) = \lambda \|\hat{f}(\cdot)\|^2_{\mathcal{H}_{K_u}} + \sum_{i=1}^{\ell} (y_i - \hat{f}(\boldsymbol{x}_i))^2, \tag{66}$$

where λ denotes a positive regularization parameter.

The subscript 'CR' stands for 'Combined kernel with Regularization'. The optimal solution of Problem 6 is reduced to

$$\hat{\boldsymbol{c}}_6 = (G^{XX}_{K_u} + \lambda I_\ell)^{-1} \boldsymbol{y} \tag{67}$$

as the same with Problem 2.

Theorem 5: The estimated functions by the optimal solutions of Problems 5 and 6 are identical.

Proof We have to show that

$$\hat{\boldsymbol{c}}_5 = (\mathbf{1}_n \otimes I_\ell) \hat{\boldsymbol{c}}_6 \tag{68}$$

holds in order for the estimated functions in Problems 5 and 6 are identical.

Since Lemma 1 implies $G^{XX} = H^{XX}(\mathbf{1}_n \otimes I_\ell)$,

$$\begin{aligned}
&\hat{\boldsymbol{c}}_5 \\
&= \{H^{XX}(\mathbf{1}_n \otimes I_\ell)(\mathbf{1}_n' \otimes I_\ell) H^{XX} + \lambda H^{XX}\}^{-1} \\
&\quad \times H^{XX}(\mathbf{1}_n \otimes I_\ell)\boldsymbol{y} \\
&= \{(\mathbf{1}_n \otimes I_\ell)(\mathbf{1}_n' \otimes I_\ell) H^{XX} + \lambda I_{\ell n}\}^{-1} (\mathbf{1}_n \otimes I_\ell)\boldsymbol{y} \\
&= \{(\mathbf{1}_n \otimes I_\ell)(G^{XX})' + \lambda I_{\ell n}\}^{-1} (\mathbf{1}_n \otimes I_\ell)\boldsymbol{y}
\end{aligned} \tag{69}$$

holds. Then, we have

$$\{(\mathbf{1}_n \otimes I_\ell)(G^{XX})' + \lambda I_{\ell n}\} \hat{\boldsymbol{c}}_5 = (\mathbf{1}_n \otimes I_\ell)\boldsymbol{y}$$

and

$$\begin{aligned}
&\{(\mathbf{1}_n \otimes I_\ell)(G^{XX})' + \lambda I_{\ell n}\}(\mathbf{1}_n \otimes I_\ell)\hat{\boldsymbol{c}}_6 \\
&= (\mathbf{1}_n \otimes I_\ell)(G^{XX})'(\mathbf{1}_n \otimes I_\ell)(G^{XX}_{K_u} + \lambda I_\ell)^{-1}\boldsymbol{y} \\
&\quad + \lambda(\mathbf{1}_n \otimes I_\ell)(G^{XX}_{K_u} + \lambda I_\ell)^{-1}\boldsymbol{y} \\
&= (\mathbf{1}_n \otimes I_\ell) G^{XX}_{K_u}(G^{XX}_{K_u} + \lambda I_\ell)^{-1}\boldsymbol{y} \\
&\quad + \lambda(\mathbf{1}_n \otimes I_\ell)(G^{XX}_{K_u} + \lambda I_\ell)^{-1}\boldsymbol{y} \\
&= (\mathbf{1}_n \otimes I_\ell)(G^{XX}_{K_u} + \lambda I_\ell - \lambda I_\ell)(G^{XX}_{K_u} + \lambda I_\ell)^{-1}\boldsymbol{y} \\
&\quad + \lambda(\mathbf{1}_n \otimes I_\ell)(G^{XX}_{K_u} + \lambda I_\ell)^{-1}\boldsymbol{y} \\
&= (\mathbf{1}_n \otimes I_\ell)\boldsymbol{y} - \lambda(\mathbf{1}_n \otimes I_\ell)(G^{XX}_{K_u} + \lambda I_\ell)^{-1}\boldsymbol{y} \\
&\quad + \lambda(\mathbf{1}_n \otimes I_\ell)(G^{XX}_{K_u} + \lambda I_\ell)^{-1}\boldsymbol{y} \\
&= (\mathbf{1}_n \otimes I_\ell)\boldsymbol{y}.
\end{aligned}$$

Since $\{(\mathbf{1}_n \otimes I_\ell)(G^{XX})' + \lambda I_{\ell n}\}$ is non-singular, Eq. (68) is obtained, which concludes the proof. □

According to Theorem 5, it is enough to consider the simpler Problem 6 instead of Problem 5 as long as the 2-norm-based criterion with the 2-norm-based regularizer is adopted as the same with the noise-free cases. Note that Theorem 5 is consistent with the representer theorem as the same with Theorem 4.

4. Numerical Examples

In this section, we show toy examples which confirm the theoretical results obtained in the previous section numerically.

In order to construct the theoretical limit and the estimated function of multiple kernel regressors defined by Problems 3 and 5, we have to calculate the quadratic forms with the kernel matrix H^{XX} or its components $H^{XX}_{K_p;K_q}$, which can not always be obtained numerically except for finite dimensional RKHSs such as those corresponding to the polynomial kernels. Thus, we give a reasonable way to calculate them.

Since $\hat{f}(\cdot) \in \mathcal{H}_{K_u}$, Eq. (22) can be also represented by

$$\hat{f}(\cdot) = \sum_{p=1}^{n} \sum_{i=1}^{\ell} c_i^{(p)} K_p(\cdot, \boldsymbol{x}_i) = \sum_{j \in \Lambda} d_j K_u(\cdot, \boldsymbol{z}_j) \tag{70}$$

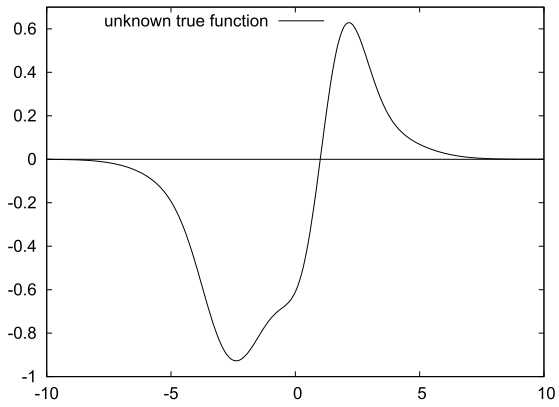


Fig. 1 The unknown true function.

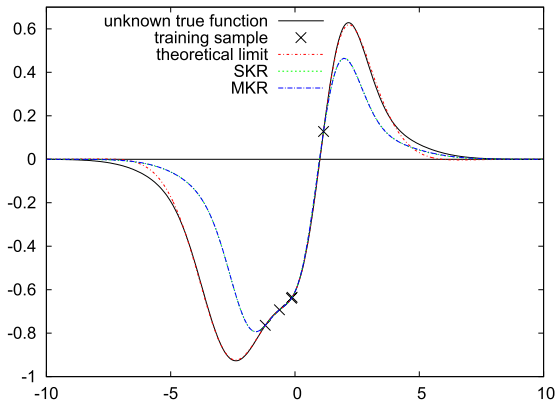


Fig. 2 Results for noise-free case. The estimated functions by MKR and SKR can not be discriminated as expected by our theoretical results.

with certain coefficients d_j. This equation implies

$$G^{ZZ}_{K_u}\boldsymbol{d} = (G^{XZ})'\boldsymbol{c}, \tag{71}$$

where $\boldsymbol{d} = [d_1, \ldots, d_{|\Lambda|}]'$ (or $\boldsymbol{d} = [d_1, d_2, \ldots]'$ in case of $|\Lambda| = \infty$). Therefore, we have

$$\begin{aligned} \|\hat{f}(\cdot)\|^2_{\mathcal{H}_{K_u}} &= \boldsymbol{c}' H^{XX} \boldsymbol{c} \\ &= \boldsymbol{d}' G^{ZZ}_{K_u} \boldsymbol{d} = \boldsymbol{c}' G^{XZ} (G^{ZZ}_{K_u})^{-1} (G^{XZ})' \boldsymbol{c}. \end{aligned} \tag{72}$$

Thus, adopting Eq. (72) as $\|\hat{f}(\cdot)\|^2_{\mathcal{H}_{K_u}}$ enables us to obtain actual solutions of the theoretical limit and the estimated functions of Problems 3 and 5. Note that $|\Lambda| \geq \ell n$ is needed to guarantee the non-singularity of the matrix $G^{XZ}(G^{ZZ}_{K_u})^{-1}(G^{XZ})'$. Also, $(G^{ZZ}_{K_u})^{-1}$ should be replaced by $(G^{ZZ}_{K_u} + \gamma I_{|\Lambda|})^{-1}$ in Eq. (72) with a small positive γ in order to stabilize the solutions. We believe that this approximation framework (especially for the theoretical limit) may be also useful for the performance evaluation of practical algorithms such as an online multiple kernel regressor with sparse dictionary [5] since comparison with the theoretical limit is critical rather than that with the unknown true function.

Let K_σ be the Gaussian kernel defined by

$$K_\sigma(x, y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-y)^2}{2\sigma^2}\right), \tag{73}$$

where σ denotes a positive shape parameter; and we adopted $\mathcal{K} = \{K_{\sigma_1}, K_{\sigma_2}\}$ with $\sigma_1 = 1.0$ and $\sigma_2 = 2.0$ as a class of kernels in the following examples, which implies

$$K_u(x, y) = K_{\sigma_1}(x, y) + K_{\sigma_2}(x, y). \tag{74}$$

We constructed the unknown true function $f(\cdot)$ by Eq. (26) with randomly generated 20 pairs of $z_k \in \mathbf{R}$ and $\alpha_k \in \mathbf{R}$ obtained by the standard normal distribution, which is shown in Fig. 1. As Z in Eqs. (71) and (72), we adopted $Z = \{0.05k \mid k \in [-200, 200] \subset \mathbf{N}\}$ so that it can cover the dominant part of the unknown true function with sufficient density.

4.1 Example for Noise-Free Case

We obtained randomly generated 5 training input points by the standard normal distribution and constructed the corresponding training output values by the unknown true function; and we calculated the theoretical limit of the model space and the estimated functions of Problems 3 and 4. Figure 2 shows the training data points and those functions together with the unknown true function. The estimated functions of Problems 3 and 4 are represented by 'MKR' and 'SKR' respectively in Fig. 2. Table 1 shows the actual generalization errors of the theoretical limit (TL), MKR, and SKR, where the squared norm of the unknown true function is 3.0284.

Table 1 Obtained generalization errors.

	TL	MKR	SKR
Generalization Error	0.0220	0.6709	0.6708

Table 1 numerically confirms the inequality in Theorem 3 and shows that the solutions of Problems 3 and 4 can not attain the theoretical limit. Also, Fig. 2 numerically confirms Theorem 4 since the solutions of Problems 3 and 4 are almost identical. As mentioned in the previous section, only the upper bound of the absolute error of the theoretical limit is guaranteed to be smaller than (or equal to) that of the estimated function obtained by Problems 3 and 4 at any point. Thus, there may exists an input point, where the actual absolute error of the theoretical limit is larger than that of the estimated function obtained by Problems 3 and 4 in general. However, it is confirmed by Fig. 2 that the actual absolute error of the theoretical limit is smaller than that of the estimated function obtained by Problems 3 and 4 at almost all of input points in this example.

4.2 Example for Noisy Case

We adopted the 5 training input points which are the same as the noise-free case and constructed the corresponding output values by the unknown true function with the zero-mean white Gaussian noise whose standard deviation is 0.1; and obtained the estimated functions of Problems 5 and 6 with $\lambda = 0.01$. Figure 3 shows the training data points and those functions together with the unknown true function and the

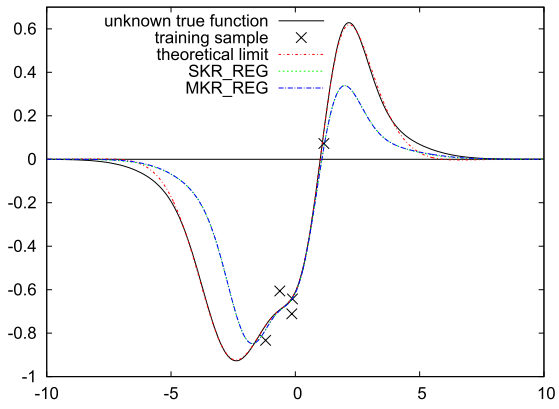


Fig. 3 Results for noisy case with regularizer. The estimated functions by MKR_REG and SKR_REG can not be discriminated as expected by our theoretical results.

theoretical limit which are the same with the noise-free case. The estimated functions of Problems 5 and 6 are represented by 'MKR_REG' and 'SKR_REG' respectively in Fig. 3.

According to this result, we can confirm that the solutions of Problems 5 and 6 are almost identical, which is claimed in Theorem 5.

5. Conclusion

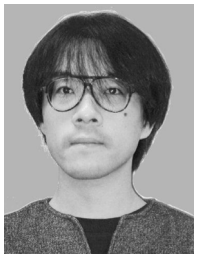
In this paper, we theoretically analyzed the 2-norm-based multiple kernel regressors and their regularized versions with the 2-norm-based regularizer. We also showed that the kernel-dependent coefficient model and the kernel-independent coefficient model are identical in both settings, which is consistent with the representer theorem. This result suggests the advantage of the kernel-independent model corresponding to the single kernel regressor with a combined kernel, in terms of computational efficiency. Moreover, we showed that the solution of these problems can not attain the theoretical limit of the model space, that is, the orthogonal projection of the unknown true function onto the model space, even if the training data set is noise-free. This result motivates us to develop a novel multiple kernel regression framework which is better than the empirical error minimization framework using the 2-norm-based formulation.

Acknowledgment

This work was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), 24500001 and 16K05264.

References

[1] K. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," IEEE Trans. Neural Netw., vol.12, pp.181–201, 2001.
[2] J. Shawe-Taylor and N. Cristianini, Kernel Methods for Pattern Recognition, Cambridge University Press, Cambridge, 2004.
[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] C. Richard, J.C.M. Bermudez, and P. Honeine, "Online prediction of time series data with kernels," IEEE Trans. Signal Process., vol.57, no.3, pp.1058–1067, 2009.
[5] M. Yukawa, "Multikernel adaptive filtering," IEEE Trans. Signal Process., vol.60, no.9, pp.4672–4682, 2012.
[6] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1999.
[7] M. Gonen and E. Alpaydin, "Multiple kernel learning algorithms," J. Mach. Learn. Res., vol.12, pp.2211–2268, 2011.
[8] C. Micchelli and M. Pontil, "Learning the kernel function via regularization," J. Mach. Learn. Res., vol.6, pp.1099–1125, 2005.
[9] C. Cortes, M. Mohri, and A. Rostamizadeh, "Generalization bounds for learning kernels," the 27th International Conference on Machine Learning (ICML2010), pp.247–254, 2010.
[10] V. Koltchinskii and D. Panchenko, "Empirical margin distributions and bounding the generalization error of combined classifiers," Ann. Statist., vol.30, no.1, pp.1–50, 2002.
[11] T. Suzuki, "Unifying framework for fast learning rate of non-sparse multiple kernel learning," Advances in Neural Information Processing Systems 24 (NIPS 2011), pp.1575–1583, 2011.
[12] M. Sugiyama and H. Ogawa, "Incremental active learning for optimal generalization," Neural Comput., vol.12, no.12, pp.2909–2940, 2000.
[13] 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, Sept. 2001.
[14] D. Sahoo, S. Hoi, and B. Li, "Online multiple kernel regression," the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.293–302, 2014.
[15] N. Aronszajn, "Theory of reproducing kernels," Trans. Am. Math. Soc., vol.68, no.3, pp.337–404, 1950.
[16] 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.
[17] S. Saitoh, Integral Transforms, Reproducing Kernels and Their Applications, Addison Wesley Longman Ltd, UK, 1997.
[18] M. Reed and B. Simon, Methods of Modern Mathematical Physics I: Functional Analysis (Revised and Enlarged Edition), Academic Press, San Diego, 1980.
[19] A. Tanaka, H. Imai, M. Kudo, and M. Miyakoshi, "Optimal kernel in a class of kernels with an invariant metric," Proc. S&SSPR2008, pp.530–539, 2008.
[20] 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.
[21] C.R. Rao and S.K. Mitra, Generalized Inverse of Matrices and its Applications, John Wiley & Sons, 1971.
[22] A. Tanaka, "Analyses on empirical error minimization in multiple kernel regressors," 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.2046–2050, 2015.
[23] J.R. Magnus and H. Neudecker, Matrix Differential Calculus with Applications in Statistics and Econometrics, Wiley, 1988.



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.

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.