A Simple Proof for The Equivalence of Multiple Kernel Regressors and Single Kernel Regressors with Sum of Kernels

Akira Tanaka*
* Division of Computer Science and Information Technology, Hokkaido University,
Sapporo, 060-0814 Japan
E-mail: takira@main.ist.hokudai.ac.jp Tel: +81-11-706-6809

***Abstract*—It is widely recognized that the kernel-based learning scheme is one of powerful tools in the field of machine learning. Recently, learning with multiple kernels, instead of a single kernel, attracts much attention in this field. Although their efficacy was investigated in terms of practical sense, their theoretical grounds were not sufficiently discussed in the past studies. In our previous work, we theoretically analyzed the standard 2-norm-based multiple-kernel regressor, and proved that the solution of the multiple kernel regressor obtained by 2-norm-based criterion reduces to the solution of the single kernel regressor with the sum of the kernels. However, the proof was hard to understand intuitively. In this work, we give a simple proof for the theorem in which the roles of the 2-norm-based criteria are intuitively convincing.**

I. Introduction

Kernel-based learning machines [1], represented by the support vector machines [2], [3] and kernel ridge regressors [4], are widely recognized as ones of powerful tools in the field of information science such as pattern recognition, regression estimation and density estimation. Recently, learning with multiple kernels [5], instead of a single kernel, attracts much attention in this field. Although their performance was investigated by experimental results in many works, their theoretical grounds were not sufficiently discussed. In our previous work [6], we theoretically analyzed the standard 2-norm-based multiple-kernel regressor, which is defined as the minimizer of the squared norm of an estimated function with the 2-norm-based empirical error minimization constraint, and proved that the solution of the multiple kernel regressor obtained by the 2-norm-based criteria reduces to the solution of the single kernel regressor with the sum of all the kernels used in the multiple kernel regressor, which gave a negative conclusion for the advantage of the multiple kernel regressors. In [6], we directly compared the solutions of the multiple and the single kernel regressors to obtain the above mentioned theorem. Therefore, it did not lead the intuitive interpretation of the theorem since mathematical structures specified by the corresponding criteria were vanished in the solutions. In this work, we focus on the optimization criterion of the multiple kernel regressor, and analyze the role of the empirical error minimization scheme and minimization of the squared norm of the estimated function in order to reveal the reason why the multiple kernel regressor reduces to the single kernel regressor with the sum of the kernels.

The rest of this paper is organized as follows. In Section II, we review the theory of reproducing kernel Hilbert spaces and give a summary for regression problems with a single kernel. In Section III, we show an overview of the result obtained in our previous work [6], concerned with the multiple kernel regressors. In Section IV, we analyze the optimization criterion of the multiple kernel regressor and give a simple proof for the equivalency of the multiple kernel regressor and the single kernel regressor with the sum of the kernels. Finally, we give concluding remarks in Section V.

II. Mathematical Preliminaries

A. Theory of Reproducing Kernel Hilbert spaces

In this part, we prepare some mathematical tools concerned with the theory of reproducing kernel Hilbert spaces [7], [8], [9].

Definition 1: [7] 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 $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes 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 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 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in D$ [7], 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 [7]. If a reproducing kernel $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ exists, it is unique [7]. Conversely, every positive definite function $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ has the unique corresponding RKHS [7]. In this paper, we assume that all RKHS's are separable [10] since many popular RKHS's 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: [7] If K_i, $(i \in \{1, 2\})$ 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 $f_i(\cdot) \in F_i$.

Note that this theorem can be easily extended to more than two reproducing kernels. In the following contents, we use the notation 'kernel' instead of 'reproducing kernel' for simplicity.

B. Overview of Kernel-based Regression Problems

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 $\boldsymbol{x}_i \in \mathbf{R}^n$ 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).

We assume that the unknown true function $f(\cdot)$ belongs to a certain RKHS $\mathcal{H}_K$; and adopt the estimation model written as

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

which implies that the estimated function belongs to the linear subspace

$$L_S = \mathrm{span}\{K(\cdot, \boldsymbol{x}_i) \mid i \in \{1, \ldots, \ell\}\}. \tag{7}$$

In general, the coefficient c_i that minimizes the so-called generalization error is adopted in the kernel regressors. Roughly speaking, the generalization error is the difference between the unknown true function and an estimated one at any point $\boldsymbol{x}$ in D, which may not be in the set of training input vectors $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\ell\}$. Since we have

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

for any $\boldsymbol{x} \in D$, which is a trivial consequence of the reproducing property Eq.(2) of a kernel and the Schwarz inequality, it is natural to adopt

$$E_S(f(\cdot), \hat{f}(\cdot); K) = ||f(\cdot) - \hat{f}(\cdot)||_{\mathcal{H}_K}^2 \tag{8}$$

as the generalization error [11], [12], [13]. Therefore, the kernel regressor is formalized as the problem to find the coefficient c_i that makes $E_S(f(\cdot), \hat{f}(\cdot); K)$ as small as possible only from the training data set. From the reproducing property Eq.(2), we have

$$\begin{aligned} &E_S(f(\cdot), \hat{f}(\cdot); K) \\ &= \left\| f(\cdot) - \sum_{i=1}^{\ell} c_i K(\cdot, \boldsymbol{x}_i) \right\|_{\mathcal{H}_K}^2 \\ &= ||f(\cdot)||_{\mathcal{H}_K}^2 + \boldsymbol{c}' G_K^{XX} \boldsymbol{c} - 2\boldsymbol{c}'\boldsymbol{f}, \end{aligned} \tag{9}$$

where $G_K^{XX} = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))$ denotes the Gram matrix of the kernel K with the set of training input vectors X, the superscript $'$ stands for the transposition operator, $\boldsymbol{f} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_\ell)]'$, and $\boldsymbol{c} = [c_1, \ldots, c_\ell]'$. Thus, the minimizer of $E_S(f(\cdot), \hat{f}(\cdot); K)$ is obtained by the linear equation

$$\frac{\partial E_S(f(\cdot), \hat{f}(\cdot); K)}{\partial \boldsymbol{c}} = 2(G_K^{XX} \boldsymbol{c} - \boldsymbol{f}) = \boldsymbol{0}, \tag{10}$$

and its solution is reduced to

$$\hat{\boldsymbol{c}}_S^{TL} = (G_K^{XX})^{-1} \boldsymbol{f}, \tag{11}$$

which gives the theoretical limit of the model space L_S*.

On the other hand, the kernel regressor based on the training data set is formalized as follows.

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

$$J_S(\boldsymbol{c}) = ||\hat{f}(\cdot)||_{\mathcal{H}_K}^2 \tag{12}$$

subject to

$$y_i = \hat{f}(\boldsymbol{x}_i), \ i \in \{1, \ldots, \ell\}. \tag{13}$$

The constraint Eq.(13) can be represented by

$$\boldsymbol{y} = G_K^{XX} \boldsymbol{c} \tag{14}$$

with $\boldsymbol{y} = [y_1, \ldots, y_\ell]'$ and the criterion Eq.(12) is reduced to

$$J_S(\boldsymbol{c}) = \boldsymbol{c}' G_K^{XX} \boldsymbol{c}. \tag{15}$$

It is well known [14] that the solution of Problem 1 is given by

$$\hat{\boldsymbol{c}}_1 = (G_K^{XX})^{-1} \boldsymbol{y}, \tag{16}$$

which agrees with the theoretical limit Eq.(11) in noise-free case. Thus, it is concluded that Problem 1 surely achieves the theoretical limit in the kernel regressor with a single kernel.

*In this paper, we assume that the Gram matrix is non-singular.

III. Known Results for The Multiple Kernel Regressors

In this section, we introduce the important results for the multiple kernel regressors, which were obtained in our previous work [6].

We consider the class of kernels $\mathcal{K} = \{K_1, \ldots, K_n\}$ and discuss the regression problem by the multiple kernel regressor using all kernels in $\mathcal{K}$. The estimation model of the multiple kernel regressor, considered in this paper, is given as

$$\hat{f}(\cdot) = \sum_{p=1}^{n}\sum_{i=1}^{\ell} c_i^{(p)} K_p(\cdot, \boldsymbol{x}_i), \tag{17}$$

which implies that the estimated function belongs to the linear subspace

$$L_M = \mathrm{span}\{K_p(\cdot, \boldsymbol{x}_i) \mid p \in \{1, \ldots, n\},\ i \in \{1, \ldots, \ell\}\}. \tag{18}$$

We assume that the unknown true function $f(\cdot)$ belongs to the RKHS corresponding to

$$K_u = \sum_{p=1}^{n} K_p \tag{19}$$

since $\hat{f}(\cdot)$ in Eq.(17) is guaranteed to be in $\mathcal{H}_{K_u}$ from Theorem 1. Thus, we evaluate the generalization error by the norm of $\mathcal{H}_{K_u}$. Firstly, we identify the theoretical limit of the model space L_M. Since we assume that an RKHS is separable, there exists a countable set $\{\boldsymbol{z}_k \mid k \in \mathbf{N}, \boldsymbol{z}_k \in D\}$ such that the set $\{K_u(\cdot, \boldsymbol{z}_k) \mid k \in \mathbf{N}\}$ is dense in $\mathcal{H}_{K_u}$. Thus, there also exists the coefficients $\alpha_k,\ (k \in \mathbf{N})$ satisfying

$$f(\cdot) = \sum_{k \in \mathbf{N}} \alpha_k K_u(\cdot, \boldsymbol{z}_k). \tag{20}$$

The generalization error of $\hat{f}(\cdot)$ evaluated in $\mathcal{H}_{K_u}$ is reduced to

$$\begin{aligned} E_M(f(\cdot), \hat{f}(\cdot); K_u) &= \|f(\cdot) - \hat{f}(\cdot)\|^2_{\mathcal{H}_{K_u}} \\ &= \Big\| \sum_{k \in \mathbf{N}} \alpha_k K_u(\cdot, \boldsymbol{z}_k) - \sum_{p=1}^{n}\sum_{i=1}^{\ell} c_i^{(p)} K_p(\cdot, \boldsymbol{x}_i) \Big\|^2_{\mathcal{H}_{K_u}} \\ &= \boldsymbol{\alpha}' G^{ZZ}_{K_u} \boldsymbol{\alpha} - 2\sum_{p=1}^{n} \boldsymbol{\alpha}' G^{ZX}_{K_p} \boldsymbol{c}^{(p)} \\ &\quad + \sum_{p=1}^{n}\sum_{q=1}^{n} (\boldsymbol{c}^{(p)})' H^{XX}_{K_p,K_q} \boldsymbol{c}^{(q)}, \end{aligned} \tag{21}$$

where $\boldsymbol{c}^{(p)} = [c_1^{(p)}, \ldots, c_\ell^{(p)}]'$, $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots]'$ 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

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

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

then Eq.(21) is rewritten as

$$\begin{aligned} & E_M(f(\cdot), \hat{f}(\cdot); K_u) \\ &= \boldsymbol{\alpha}' G^{ZZ}_{K_u} \boldsymbol{\alpha} - 2\boldsymbol{\alpha}'(G^{XZ})'\boldsymbol{c} + \boldsymbol{c}' H^{XX} \boldsymbol{c}. \end{aligned} \tag{22}$$

Therefore, the minimizer of $E_M(f(\cdot), \hat{f}(\cdot); K_u)$ is obtained by

$$\frac{\partial E_M(f(\cdot), \hat{f}(\cdot); \mathcal{H}_{K_u})}{\partial \boldsymbol{c}} = 2H^{XX}\boldsymbol{c} - 2G^{XZ}\boldsymbol{\alpha} = \boldsymbol{0}, \tag{23}$$

which is reduced to the linear equation

$$H^{XX}\boldsymbol{c} = G^{XZ}\boldsymbol{\alpha}. \tag{24}$$

Therefore, the coefficient vector of the theoretical limit of the model space L_M is reduced to

$$\hat{\boldsymbol{c}}_M^{TL} = (H^{XX})^{-1} G^{XZ} \boldsymbol{\alpha}. \tag{25}$$

Generally, we can not construct this theoretical limit from the given training data set even if it is noise-free on the contrary to the single kernel case, since $\boldsymbol{\alpha}$ is unknown.

On the other hand, the multiple kernel regressor based on the training data set is formalized as follows.

Problem 2: Find the coefficient vector $\boldsymbol{c}$ of the model Eq.(17) that minimizes

$$J_M(\boldsymbol{c}) = \|\hat{f}(\cdot)\|^2_{\mathcal{H}_{K_u}} \tag{26}$$

subject to

$$y_i = \hat{f}(\boldsymbol{x}_i),\ i \in \{1, \ldots, \ell\}. \tag{27}$$

Note that the constraint Eq.(27) can be represented by

$$\boldsymbol{y} = \sum_{p=1}^{n} G^{XX}_{K_p} \boldsymbol{c}^{(p)} = (G^{XX})'\boldsymbol{c}, \tag{28}$$

where $G^{XX} = [G^{XX}_{K_1}, \ldots, G^{XX}_{K_n}]'$ and the criterion Eq.(26) is reduced to

$$J_M(\boldsymbol{c}) = \boldsymbol{c}' H^{XX} \boldsymbol{c}. \tag{29}$$

The solution of Problem 2 is easily obtained as

$$\hat{\boldsymbol{c}}_2 = (H^{XX})^{-1} G^{XX} \{(G^{XX})'(H^{XX})^{-1} G^{XX}\}^{-1} \boldsymbol{y} \tag{30}$$

as shown in [14], [6].

The following theorems are the important results obtained in [6].

Theorem 2: [6] Let $\hat{f}_2(\cdot)$ be the estimated function by $\hat{\boldsymbol{c}}_2$, then

$$\langle f(\cdot) - \hat{f}_2(\cdot), \hat{f}_2(\cdot)\rangle_{\mathcal{H}_{K_u}} = 0 \tag{31}$$

holds.

Theorem 3: [6] The estimated function by the solution of Problems 2 is identical to the estimated function by the solution of Problem 1 with $K = K_u$.

Theorem 2 implies that the estimated function by the coefficient vector obtained by Problem 2 is the orthogonal projection of the unknown true function $f(\cdot)$ onto a certain linear subspace $\tilde{L}_M \subset L_M$; and Theorem 3 implies that $\tilde{L}_M$ is reduced to

$$\tilde{L}_M = \mathrm{span}\{K_u(\cdot, \boldsymbol{x}_i) \mid i \in \{1, \ldots, \ell\}\}, \tag{32}$$

which gives the negative conclusion for the advantage of the model Eq.(17) of the multiple kernel regressor.

In [6], we proved Theorem 3 by comparing the solutions Eqs.(16) and (30). Therefore, it did not lead the intuitive interpretation of the theorem since mathematical structures specified by the corresponding criteria were vanished in the solutions.

IV. A Simple Proof for Theorem 3

In this section, we give a simple proof for Theorem 3 in which the roles of 2-norm-based criteria are revealed.

Since $K_u = \sum_{p=1}^{n} K_p$, we have $\sum_{p=1}^{n} H^{XX}_{K_p;K_q} = G^{XX}_{K_q}$ and $\sum_{p=1}^{n} G^{XZ}_{K_p} = G^{XZ}_{K_u}$. Therefore, the constraint Eq.(28) can be represented by

$$(\boldsymbol{1}_n' \otimes I_\ell) G^{XZ} \boldsymbol{\alpha} = (\boldsymbol{1}_n' \otimes I_\ell) H^{XX} \boldsymbol{c}, \tag{33}$$

where $\boldsymbol{1}_n$, I_ℓ, and $\otimes$ denote the n dimensional vector whose elements are unity, the identity matrix of degree ℓ, and the Kronecker product of two matrices [15] defined by

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix} \in \mathbf{R}^{mp \times nq}$$

for $A = (a_{ij}) \in \mathbf{R}^{m \times n}$ and $B \in \mathbf{R}^{p \times q}$, which implies that the constraint Eq.(28) for the minimum empirical error can be regarded as the linear equation Eq.(24) that $\hat{\boldsymbol{c}}^{TL}_M$ must satisfy, degenerated by pre-multiplying $(\boldsymbol{1}_n' \otimes I_\ell)$. Thus, we can rewrite the criterion of Problem 2 as

$$J_M(\boldsymbol{c}) = \frac{1}{2}\boldsymbol{c}' H^{XX} \boldsymbol{c} - \boldsymbol{\mu}'(\boldsymbol{1}_n' \otimes I_\ell)(G^{XZ}\boldsymbol{\alpha} - H^{XX}\boldsymbol{c}), \tag{34}$$

with the Lagrange multiplier $\boldsymbol{\mu}$. Then, we have

$$\frac{J_M(\boldsymbol{c})}{\partial \boldsymbol{c}} = H^{XX}\boldsymbol{c} - H^{XX}(\boldsymbol{1}_n \otimes I_\ell)\boldsymbol{\mu} = \boldsymbol{0}.$$

Therefore, we have

$$\boldsymbol{c} = (\boldsymbol{1}_n \otimes I_\ell)\boldsymbol{\mu}, \tag{35}$$

which trivially implies $\boldsymbol{c}^{(1)} = \cdots = \boldsymbol{c}^{(n)}$. Also, substituting Eq.(35) to the constraint Eqs.(28),(29), and (17) yields the constraint

$$\boldsymbol{y} = G^{XX}_{K_u}\boldsymbol{\mu}, \tag{36}$$

the criterion

$$J_M(\boldsymbol{\mu}) = \boldsymbol{\mu}' G^{XX}_{K_u}\boldsymbol{\mu}, \tag{37}$$

and the model

$$\hat{f}(\cdot) = \sum_{p=1}^{n}\sum_{i=1}^{\ell} c_i^{(p)} K_p(\cdot, \boldsymbol{x}_i) = \sum_{i=1}^{\ell} \mu_i K_u(\cdot, \boldsymbol{x}_i) \tag{38}$$

with $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_\ell]'$, which are identical to those in Problem 1 with $K = K_u$, which concluded the proof of Theorem 3.

Accordingly, the above proof reveals that

(1) the constraint for the empirical error minimization is the degenerated version of the linear equation appeared in obtaining the theoretical limit of the model space L_M, and

(2) together with the above constraint, the criterion for the minimum squared norm necessarily produces the coefficient vector $\boldsymbol{c}^{(i)}$ which does not depend on the index $p \in \{1, \ldots, n\}$,

in the 2-norm-based multiple kernel regressor.

V. Conclusion

In this paper, we gave a simple proof for the equivalency of the 2-norm-based multiple kernel regressor and the 2-norm-based single kernel regressor adopting the sum of the kernels, which was discussed in our previous work. Our proof obtained in this paper revealed the roles of the constraint for the empirical error minimization and the criterion for the squared-norm of the estimated function.

References

[1] K. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," *IEEE Transactions on Neural Networks*, vol. 12, pp. 181–201, 2001.

[2] 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] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Recognition*, Cambridge University Press, Cambridge, 2004.

[5] M. Gonen and E Alpaydin, "Multiple Kernel Learning Algorithms," *Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.

[6] A. Tanaka and H. Imai, "Theoretical Analyses on 2-Norm-Based Multiple Kernel Regressors," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences (submitted).*

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

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

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

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

[11] M. Sugiyama and H. Ogawa, "Incremental Active Learning for Optimal Generalization," *Neural Computation*, vol. 12, no. 12, pp. 2909–2940, 2000.

[12] M. Sugiyama and H. Ogawa, "Active Learning for Optimal Generalization in Trigonometric Polynomial Models," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E84-A, no. 9, pp. 2319–2329, 2001.

[13] A. Tanaka, H. Imai, M. Kudo, and M. Miyakoshi, "Optimal kernel in a class of kernels with an invariant metric," in *Proc. S&SSPR2008*, 2008, pp. 530–539.

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

[15] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Wiley, 1988.