ANALYSES ON EMPIRICAL ERROR MINIMIZATION IN MULTIPLE KERNEL REGRESSORS

Akira Tanaka

Division of Computer Science and Information Technology, Hokkaido University,
N14W9, Kita-ku, Sapporo, 060-0814 Japan.

ABSTRACT

Theoretical validity of empirical error minimization in multiple kernel regressors is discussed in this paper. Generalization error of a kernel machine is usually evaluated by the induced norm of the difference between an unknown true function and an estimated one in an appropriate reproducing kernel Hilbert space. It is well known that empirical error minimization also achieves the minimum generalization error in single kernel regressors. However, it is not clarified whether or not that is true for multiple kernel regressors. Moreover, possibility of constructing the minimizer of the generalization error by a given training date set is not also clarified. In this paper, we give negative conclusions for these problems through theoretical analyses on the generalization error of multiple kernel regressors and also give an example by popular Gaussian kernels.

Index Terms— multiple kernel regressor, reproducing kernel Hilbert space, generalization error, empirical error minimization

1. INTRODUCTION

Learning based on kernel machines [1, 2, 3, 4] is widely known as a powerful tool for various fields of information science such as pattern recognition, regression estimation, and density estimation. For the last few decades, kernel machines with multiple kernels attract much attention in these fields and various learning methods have been developed (see [5] and its references for instance). In this work, we concentrate on a kernel-based regression problem due to its simple structure for analyses. The aim of the regression problem is to minimize the generalization error, that is, the error for arbitrary input vectors. However, many kernel regressors are formalized as minimization of empirical error, that is, the error for training data set, with some constraints or regularization terms. In kernel regressors with a single kernel, it is clarified that the empirical error minimization also achieves the minimum generalization error which is evaluated by the induced norm of the difference between an unknown true function and an estimated one in the reproducing kernel Hilbert space corresponding to the adopted kernel (see [6, 7] for instance), On the other hand, it is not clarified whether or not that is true for multiple kernel regressors. Moreover, possibility of constructing the minimizer of the generalization error by a given training data set is not also clarified. In this paper, we discuss theoretical properties of the optimal learning result of the multiple kernel regressor and show that (a) the empirical error minimization can not always minimize the generalization error; and (b) the optimal learning result, in terms of minimum generalization error, can not always obtained from a given training data set. We also show an example with popular Gaussian kernels with various shape parameters in order to confirm our theoretical results.

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

2. MATHEMATICAL PRELIMINARIES FOR THE THEORY OF REPRODUCING KERNEL HILBERT SPACES

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

Definition 1 *[8] Let $\mathbf{R}^n$ be an n-dimensional real vector space and let $\mathcal{H}$ be a class of functions defined on $\mathcal{D} \subset \mathbf{R}^n$, forming a Hilbert space of real-valued functions. The function $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$, $(\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathcal{D})$ is called a reproducing kernel of $\mathcal{H}$, if*

1. $\forall \tilde{\boldsymbol{x}} \in \mathcal{D};\ K(\cdot, \tilde{\boldsymbol{x}}) \in \mathcal{H}$,
2. $\forall \tilde{\boldsymbol{x}} \in \mathcal{D},\ \forall f(\cdot) \in \mathcal{H};\ f(\tilde{\boldsymbol{x}}) = \langle f(\cdot), K(\cdot, \tilde{\boldsymbol{x}}) \rangle_{\mathcal{H}}$,

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the inner product of $\mathcal{H}$.

The Hilbert space that has a reproducing kernel K is called a reproducing kernel Hilbert space (RKHS), denoted by $\mathcal{H}_K$. The reproducing property (the second condition in Definition 1) enables us to treat a value of a function at a point in $\mathcal{D}$, while we can not deal with a value of a function in a general Hilbert space such as L^2 (the Hilbert space consisting of all square-integrable functions). Note that reproducing kernels are positive definite [8]:

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

for any $N \in \mathbf{N}$, $c_1, \ldots, c_N \in \mathbf{R}$, and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathcal{D}$. In addition, $K(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = K(\tilde{\boldsymbol{x}}, \boldsymbol{x})$ for any $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathcal{D}$ is followed [8]. If a reproducing kernel $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ exists, it is unique [8]. Conversely, every positive definite function $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ has the unique corresponding RKHS [8]. We assume that an RKHS is separable [10] in this paper.

3. PROBLEM FORMULATION AND OVERVIEW OF SINGLE KERNEL REGRESSOR

Let $\{(y_i, \boldsymbol{x}_i) \mid i \in \{1, \ldots, \ell\}\}$ be a given training data set with an output value $y_i \in \mathbf{R}$ and the corresponding input vector $\boldsymbol{x}_i \in \mathbf{R}^n$, satisfying

$$y_i = f(\boldsymbol{x}_i) + n_i, \tag{2}$$

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). Note that we ignore the additive noise in the following contents since analyses on the additive noise are out of the scope of this paper.

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

$$\hat{f}(\cdot) = \sum_{i=1}^{\ell} c_i K(\cdot, \boldsymbol{x}_i) \tag{3}$$

with some coefficients $c_i \in \mathbf{R}, (i \in \{1, \ldots, \ell\})$. In general, these coefficients are obtained by minimization of the empirical error defined by

$$J_1^{(emp)} = \sum_{j=1}^{\ell} \left(y_j - \sum_{i=1}^{\ell} c_i K(\boldsymbol{x}_j, \boldsymbol{x}_i) \right)^2, \tag{4}$$

which can be also represented by

$$J_1^{(emp)} = ||\boldsymbol{y} - G_X^{(K)} \boldsymbol{c}||^2, \tag{5}$$

where $\boldsymbol{y} = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_\ell)]'$ (since we ignore the noise), $\boldsymbol{c} = [c_1, \ldots, c_\ell]'$ with the superscript $'$ denoting the transposition operator and $G_X^{(K)} = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))$ with $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\ell\}$. It is trivial that a typical[1] solution that minimizes Eq.(5) is given as

$$\hat{\boldsymbol{c}}_1^{(emp)} = (G_X^{(K)})^+ \boldsymbol{y}, \tag{6}$$

where the superscript $^+$ denotes the Moore-Penrose generalized inverse [11]. Note that it is not so trivial that the learning result by Eq.(6) gives a small error for an input vector $\boldsymbol{x} \notin X$.

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

Under the assumption $f(\cdot) \in \mathcal{H}_K$, the generalization error of $\hat{f}(\cdot)$ is defined as

$$J_1^{(gen)}(\hat{f}(\cdot)) = ||f(\cdot) - \hat{f}(\cdot)||_{\mathcal{H}_K}^2, \tag{7}$$

where $|| \cdot ||_{\mathcal{H}_K}$ denotes the induced norm of $\mathcal{H}_K$. The validity of Eq.(7) as the measure of the generalization error is supported by

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

obtained by the reproducing property of kernels and the Schwarz's inequality, where $\boldsymbol{x}$ is an arbitrary input vector in $\mathcal{D}$ which may not be in X. Note that if $f(\cdot) \notin \mathcal{H}_K$, Eq.(7) is meaningless since Eq.(8) does not make sense, which implies that we can not discuss generalization capability of kernel machines when $f(\cdot) \notin \mathcal{H}_K$ from the theoretical point of view.

Next, we confirm that the solution Eq.(6) is identical to the optimal solution in terms of the generalization error defined by Eq.(7). Since we assume that $\mathcal{H}_K$ is separable, there exists a countable set $\{(\alpha_k, \boldsymbol{z}_k) \mid k \in \{1, \ldots, N\}\}$ with $\alpha_k \in \mathbf{R}$ and $\boldsymbol{z}_k \in \mathcal{D}$ such that

$$f(\cdot) = \sum_{k=1}^{N} \alpha_k K(\cdot, \boldsymbol{z}_k). \tag{9}$$

Therefore, Eq.(7) is reduced to

$$\begin{aligned} J_1^{(gen)}(\hat{f}(\cdot)) &= \left\| \sum_{k=1}^{N} \alpha_k K(\cdot, \boldsymbol{z}_k) - \sum_{i=1}^{\ell} c_i K(\cdot, \boldsymbol{x}_i) \right\|_{\mathcal{H}_K}^2 \\ &= \boldsymbol{\alpha}' G_Z^{(K)} \boldsymbol{\alpha} + \boldsymbol{c}' G_X^{(K)} \boldsymbol{c} - 2\boldsymbol{c}' G_{XZ}^{(K)} \boldsymbol{\alpha}, \end{aligned} \tag{10}$$

where $Z = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$, $G_Z^{(K)} = (K(\boldsymbol{z}_i, \boldsymbol{z}_j))$ and $G_{XZ}^{(K)} = (K(\boldsymbol{x}_i, \boldsymbol{z}_j))$. Since $J_1^{(gen)}(\hat{f}(\cdot))$ is a positive quadratic form with respect to $\boldsymbol{c}$,

$$\frac{\partial J_1^{(gen)}(\hat{f}(\cdot))}{\partial \boldsymbol{c}} = 2G_X^{(K)} \boldsymbol{c} - 2G_{XZ}^{(K)} \boldsymbol{\alpha} = \boldsymbol{0} \tag{11}$$

yields the optimal solution and a typical optimal solution is given as

$$\hat{\boldsymbol{c}}_1^{(gen)} = (G_X^{(K)})^+ G_{XZ}^{(K)} \boldsymbol{\alpha} = (G_X^{(K)})^+ \boldsymbol{y}, \tag{12}$$

since $\boldsymbol{y} = G_{XZ}^{(K)} \boldsymbol{\alpha}$ holds from Eq.(9), that agrees with Eq.(6) which is obtained by the empirical error minimization. Therefore, it is confirmed that the empirical error minimization yields the optimal solution in terms of the generalization error Eq.(7) and the solution can be constructed by the given training date set since Eq.(12) does not contain the unknown parameter $\boldsymbol{\alpha}$. Note that these fasts can be also confirmed by the optimal solution of a linear inversion problem of the sampling process [6, 7]. Also note that the minimizer of the generalization error gives the orthogonal projection of $f(\cdot)$ onto the linear subspace spanned by $\{K(\cdot, \boldsymbol{x}_i) \mid i \in \{1, \ldots, \ell\}\}$, whose orthogonality is specified by the metric of $\mathcal{H}_K$.

4. ANALYSES ON MULTIPLE KERNEL REGRESSOR

In this section, we discuss the multiple kernel regressor with the class of kernels $\mathcal{K} = \{K_1, \ldots, K_m\}$. The model of a learning result is given as

$$\hat{f}(\cdot) = \sum_{p=1}^{m} \sum_{i=1}^{\ell} c_i^{(p)} K_p(\cdot, \boldsymbol{x}_i), \tag{13}$$

with some coefficients $c_i^{(p)}$, $(i \in \{1, \ldots, \ell\},\ p \in \{1, \ldots, m\})$ [5]. We assume that $\mathcal{H}_{K_p} \subset \mathcal{H}_K$ for any $p \in \{1, \ldots, m\}$ so that the evaluation of the generalization error by the norm of $\mathcal{H}_K$ makes sense [2].

The empirical error of the model Eq.(13) is defined as

$$J_2^{(emp)} = \sum_{j=1}^{\ell} \left(y_j - \sum_{p=1}^{m} \sum_{i=1}^{\ell} c_i^{(p)} K_p(\boldsymbol{x}_j, \boldsymbol{x}_i) \right)^2, \tag{14}$$

which can be also represented by

$$J_2^{(emp)} = \left\| \boldsymbol{y} - \sum_{p=1}^{m} G_X^{(K_p)} \boldsymbol{c}^{(p)} \right\|^2, \tag{15}$$

where $\boldsymbol{c}^{(p)} = [c_1^{(p)}, \ldots, c_\ell^{(p)}]'$. Since $J_2^{(emp)}$ is a positive quadratic form with respect to each $\boldsymbol{c}^{(p)}$, the minimizer of $J_2^{(emp)}$ is obtained by $\partial J_2^{(emp)} / \partial \boldsymbol{c}^{(p)} = \boldsymbol{0}$ for all $p \in \{1, \ldots, m\}$, which is reduced to the linear equation:

$$M_1 \boldsymbol{c} = \boldsymbol{b}_1, \tag{16}$$

where

$$M_1 = \begin{bmatrix} (G_X^{(K_1)})^2 & \cdots & G_X^{(K_1)} G_X^{(K_m)} \\ \vdots & \ddots & \vdots \\ G_X^{(K_m)} G_X^{(K_1)} & \cdots & (G_X^{(K_m)})^2 \end{bmatrix},$$

$\boldsymbol{c} = [(\boldsymbol{c}^{(1)})', \ldots, (\boldsymbol{c}^{(m)})']'$ and $\boldsymbol{b}_1 = [G_X^{(K_1)}, \ldots, G_X^{(K_m)}]' \boldsymbol{y}$. Therefore, a typical optimal solution in terms of the empirical error is obtained by

$$\hat{\boldsymbol{c}}_2^{(emp)} = M_1^{+} \boldsymbol{b}_1. \tag{17}$$

Next, we investigate the minimizer of the generalization error of the model Eq.(13), which is defined as

$$\begin{aligned} & J_2^{(gen)}(\hat{f}(\cdot)) \\ &= \left\| \sum_{k=1}^{N} \alpha_k K(\cdot, \boldsymbol{z}_k) - \sum_{p=1}^{m} \sum_{i=1}^{\ell} c_i^{(p)} K_p(\cdot, \boldsymbol{x}_i) \right\|_{\mathcal{H}_K}^2 \\ &= \boldsymbol{\alpha}' G_Z^{(K)} \boldsymbol{\alpha} - 2 \sum_{p=1}^{m} (\boldsymbol{c}^{(p)})' G_{XZ}^{(K_p)} \boldsymbol{\alpha} \\ &\quad + \sum_{p=1}^{m} \sum_{q=1}^{m} (\boldsymbol{c}^{(p)})' H_X^{(K_p;K_q)} \boldsymbol{c}^{(q)}, \end{aligned} \tag{18}$$

[2] As shown in the next section, this assumption is natural in the light of the way of using the popular Gaussian kernels in multiple kernel regressors.

where $H_X^{(K_p;K_q)} = (\langle K_p(\cdot, \boldsymbol{x}_i), K_q(\cdot, \boldsymbol{x}_j) \rangle_{\mathcal{H}_K})$. Since $J_2^{(gen)}(\hat{f}(\cdot))$ is a positive quadratic form with respect to each $\boldsymbol{c}^{(p)}$, the minimizer of $J_2^{(gen)}(\hat{f}(\cdot))$ is obtained by $\partial J_2^{(gen)}(\hat{f}(\cdot)) / \partial \boldsymbol{c}^{(p)} = \boldsymbol{0}$ for all $p \in \{1, \ldots, m\}$, which can be represented by the linear equation:

$$M_2 \boldsymbol{c} = \boldsymbol{b}_2, \tag{19}$$

where

$$M_2 = \begin{bmatrix} H_X^{(K_1;K_1)} & H_X^{(K_1;K_2)} & \cdots & H_X^{(K_1;K_m)} \\ H_X^{(K_2;K_1)} & H_X^{(K_2;K_2)} & \cdots & H_X^{(K_2;K_m)} \\ \vdots & \vdots & \ddots & \vdots \\ H_X^{(K_m;K_1)} & H_X^{(K_m;K_2)} & \cdots & H_X^{(K_m;K_m)} \end{bmatrix},$$

and $\boldsymbol{b}_2 = [G_{ZX}^{(K_1)}, \ldots, G_{ZX}^{(K_m)}]' \boldsymbol{\alpha}$. Therefore, a typical optimal solution in terms of the generalization error is obtained by

$$\hat{\boldsymbol{c}}_2^{(gen)} = M_2^{+} \boldsymbol{b}_2. \tag{20}$$

In general, solutions (17) and (20) are different, which implies that the empirical error minimization can not always yield the optimal solution of the multiple kernel regressor in terms of the generalization error, while it gives the optimal solution in the single kernel regressor as shown in the previous section. Moreover, it is also confirmed that the optimal solution in terms of the generalization error can not be constructed from the training data set, since we can not obtain $G_{XZ}^{(K_p)} \boldsymbol{\alpha}$ in $\boldsymbol{b}_2$ from $\boldsymbol{y} = G_{XZ}^{(K)} \boldsymbol{\alpha}$ unless we have a priori information about Z and $\boldsymbol{\alpha}$. In the following contents, we call the function Eq.(13) with $\hat{\boldsymbol{c}}_2^{(gen)}$ 'theoretical limit' of the multiple kernel regressors.

5. NUMERICAL EXAMPLE

In this section, we show an example in order to confirm that the learning result by Eq.(17) and the theoretical limit by Eq.(20) actually differ.

Generally, we can not always calculate the matrices $H_X^{(K_p;K_q)}$ in Eq.(19). On the other hand, we developed the way to calculate them for the popular Gaussian kernels in [12], which is defined as

$$K_\sigma(x, y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-y)^2}{2\sigma^2} \right), \tag{21}$$

where σ denote the positive shape parameter. We show some important results obtained in [12], which is used in our numerical examples given below.

Theorem 1 *[12] For any positive parameters σ_1 and σ_2, satisfying $0 < \sigma_2 < \sigma_1$,*

$$\mathcal{H}_{K_{\sigma_1}} \subset \mathcal{H}_{K_{\sigma_2}}$$

holds.

Table 1. Empirical error, generalization error, and L^2 error by the EMP and the LIM with respect to σ.

	$J_2^{(emp)}$		$J_2^{(gen)}$		$J_2^{(L^2)}$	
σ	EMP	LIM	EMP	LIM	EMP	LIM
0.2	$\mathbf{1.88 \times 10^{-17}}$	1.29×10^{1}	9.05×10^{8}	$\mathbf{7.38 \times 10^{2}}$	3.23×10^{3}	$\mathbf{1.45 \times 10^{1}}$
0.4	$\mathbf{2.73 \times 10^{-18}}$	2.27×10^{0}	5.11×10^{7}	$\mathbf{8.83 \times 10^{1}}$	7.51×10^{2}	$\mathbf{3.01 \times 10^{0}}$
0.6	$\mathbf{3.26 \times 10^{-18}}$	1.19×10^{-1}	2.98×10^{5}	$\mathbf{6.01 \times 10^{0}}$	4.68×10^{1}	$\mathbf{4.02 \times 10^{-1}}$
0.8	$\mathbf{4.71 \times 10^{-18}}$	8.13×10^{-4}	2.03×10^{3}	$\mathbf{1.57 \times 10^{-2}}$	1.81×10^{0}	$\mathbf{6.77 \times 10^{-2}}$
1.0	$\mathbf{1.45 \times 10^{-18}}$	1.96×10^{-21}	5.80×10^{1}	$\mathbf{1.56 \times 10^{-3}}$	1.57×10^{-1}	$\mathbf{1.00 \times 10^{-1}}$

Theorem 2 *[12] For any positive parameter σ_3, satisfying $0 < \sigma_3 < \min\{\sigma_1, \sigma_2\}$,*

$$\langle K_{\sigma_1}(\cdot, x), K_{\sigma_2}(\cdot, y)\rangle_{\mathcal{H}_{K_{\sigma_3}}} = K_{\sqrt{\sigma_1^2+\sigma_2^2-\sigma_3^2}}(x, y) \quad (22)$$

holds.

According to Theorem 1, K_σ with

$$\sigma < \min\{\sigma_1, \ldots, \sigma_m\} \quad (23)$$

can be a kernel of the host RKHS for a class of kernels $\mathcal{K} = \{K_{\sigma_1}, \ldots, K_{\sigma_m}\}$. Moreover, Theorem 2 enables us to calculate $H_X^{(K_{\sigma_p}; K_{\sigma_q})}$.

In multiple kernel regressors with the Gaussian kernels, we try to estimate the unknown true function by the Gaussian kernels which are steeper as much as the steepest part of the unknown true function and by more gradual ones corresponding to larger shape parameters for gradual part of the unknown true function. Therefore, selection of σ satisfying Eq.(23) seems to be natural in terms of a practical sense.

We consider a class of kernels $\mathcal{K} = \{K_{\sigma_1}, K_{\sigma_2}\}$ with $\sigma_1 = 1.0$ and $\sigma_2 = 1.2$; and adopt K_σ with a σ satisfying Eq.(23). We constructed an unknown true function $f(\cdot)$ by Eq.(9) with K_σ, $N = 20$ and randomly generated $\alpha_k \in \mathbf{R}$ and $z_k \in \mathbf{R}$ by the standard normal distribution; and also generated training input points randomly (by the standard normal distribution) with $\ell = 5$ and training output values by these points and $f(\cdot)$.

Figure 1 shows an example of the unknown true function $f(\cdot)$, the learning result by the empirical error minimizer $\hat{\boldsymbol{c}}_2^{(emp)}$, and the theoretical limit obtained by $\hat{\boldsymbol{c}}_2^{(gen)}$ with $\sigma = 0.8$, in which the points in the training data set are denoted by '+'. According to Fig.1, it is confirmed that the learning result by the empirical error minimization is surely different from the theoretical limit.

Table 1 shows $J_2^{(emp)}$, $J_2^{(gen)}$, and L^2 norm of $f(\cdot) - \hat{f}(\cdot)$ (as a reference and denoted by $J_2^{(L^2)}$) of the learning results by the empirical error minimizer (denoted by 'EMP') and the theoretical limit (denoted by 'LIM') for the same setting with Fig.1 with respect to $\sigma = 0.2, 0.4, 0.6, 0.8, 1.0$. According to Table 1, it is confirmed that the EMP surely obtains a smaller empirical error, while it fails to obtain a small generalization error evaluated in $\mathcal{H}_{K_\sigma}$, which is a consequence of our theoretical analyses, and also in L^2, which is only supported numerically. Note that $\hat{\boldsymbol{c}}_2^{(emp)}$ in this example is a stable solution since the training data set is noise-free and $1/\mathrm{cond}(M_1) = 2.01 \times 10^{-9}$, which is much larger than the machine epsilon.

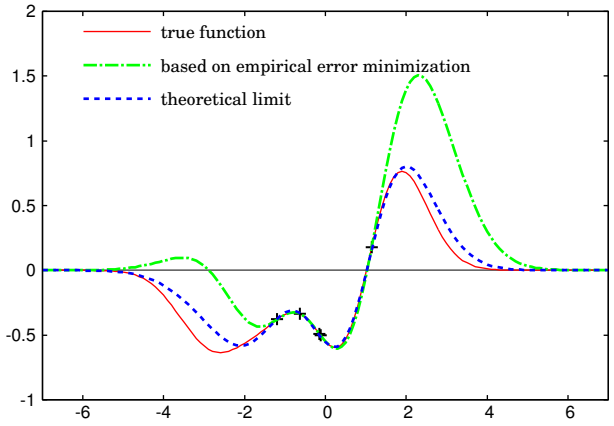

Fig. 1. An example of a true function, the learning result by empirical error minimization, and the theoretical limit.

6. CONCLUSION

In this paper, we theoretically analyzed a typical multiple kernel regressor and its model space; and revealed the theoretical limit of the model space in terms of the generalization error, defined by the norm of the difference between an unknown true function and an estimated one, evaluated in an appropriate reproducing kernel Hilbert space. On the basis of the analyses, we showed that the empirical error minimization does not always attain the minimum generalization error in contrast to single kernel regressors; and also showed that the theoretical limit can not always be constructed from a given training data set. These results give us a motivation of developing a novel framework of multiple kernel regressors that yields a learning result closer to the theoretical limit. Analyses for additive noise and properties of the solution obtained by a regularization scheme are ones of our future works that should be undertaken.

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

[4] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, Cambridge, 2000.

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

[6] H. Ogawa, "Neural Networks and Generalization Ability," *IEICE Technical Report*, vol. NC95-8, pp. 57–64, 1995.

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

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

[9] 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.

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

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

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