THEORETICAL ANALYSES FOR A CLASS OF KERNELS WITH AN INVARIANT METRIC

Akira Tanaka and Masaaki Miyakoshi

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

ABSTRACT

One of central topics of kernel machines in the field of machine learning is a model selection, especially a selection of a kernel or its parameters. In our previous work, we discussed a class of kernels whose corresponding reproducing kernel Hilbert spaces have an invariant metric and proved that the kernel corresponding to the smallest reproducing kernel Hilbert space, including an unknown true function, gives the optimal model. However, discussions for properties that make the metrics of reproducing kernel Hilbert spaces invariant are insufficient. In this paper, we show a necessary and sufficient condition that makes the metrics of reproducing kernel Hilbert spaces invariant.

Index Terms— kernel machine, reproducing kernel Hilbert space, generalization ability, metric

1. INTRODUCTION

Learning based on kernel machines[1], represented by the support vector machine[2] and the kernel ridge regression[3, 4], is widely known as a powerful tool for various fields of information science such as pattern recognition, regression estimation, and density estimation. In general, an appropriate model selection is required in order to obtain a desirable learning result by kernel machines. There exists two classes of model selection. One is a selection of a model space to which a learning result belongs. The other is a selection of a learning machine in a fixed model space. The latter, such as a selection of a regularization parameter under a fixed kernel, is sufficiently investigated in terms of theoretical and practical senses (See [5, 6] for instance). On the other hand, the former, that is, a selection of a kernel or its parameters, is not discussed sufficiently in terms of theoretical sense although practical algorithms for a selection of a kernel (or its parameters), such as a cross-validation, are revealed. The difficulty of theoretical analyses for a selection of a kernel (or its parameters) lies on the fact that the metrics of two reproducing kernel Hilbert spaces (RKHS)[7, 8] corresponding to two different kernels may differ in general. In order to avoid this difficulty, we analyzed the properties of a class of RKHS's with an invariant metric and proved that the kernel corresponding to the smallest RKHS, including an unknown true function, gives the optimal model in our previous paper[9]. Although we have an example of a class of kernels, whose corresponding RKHS's have an invariant metric, such as the sinc kernel (sampling function used in Shannon's sampling theorem), discussions for properties that make the metrics of RKHS's invariant are insufficient. In this paper, we show a necessary and sufficient condition that makes the metrics of RKHS's invariant.

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

In this section, we prepare some mathematical tools concerned with the theory of RKHS's[7, 8].

Definition 1 *[7] 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. *For every fixed $\tilde{\boldsymbol{x}} \in \mathcal{D}$, $K(\cdot, \tilde{\boldsymbol{x}})$ is a function belonging to $\mathcal{H}$.*
2. *For every fixed $\tilde{\boldsymbol{x}} \in \mathcal{D}$ and every fixed $f \in \mathcal{H}$,*

$$f(\tilde{\boldsymbol{x}}) = \langle f(\cdot), K(\cdot, \tilde{\boldsymbol{x}}) \rangle_{\mathcal{H}}, \tag{1}$$

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

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 Eq.(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. Note that reproducing kernels are positive definite [7]:

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

for any integer 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[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].

Next, we introduce the Schatten product [10] that is a convenient tool to reveal the reproducing property of kernels.

Definition 2 *[10] Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces. The Schatten product of $g \in \mathcal{H}_2$ and $h \in \mathcal{H}_1$ is defined by*

$$(g \otimes h)f = \langle f, h \rangle_{\mathcal{H}_1} g, \quad f \in \mathcal{H}_1. \tag{3}$$

Note that $(g \otimes h)$ is a linear operator from $\mathcal{H}_1$ onto $\mathcal{H}_2$. It is easy to show that the following relations hold for $h, v \in \mathcal{H}_1$, $g, u \in \mathcal{H}_2$.

$$(h \otimes g)^* = (g \otimes h), \tag{4}$$

$$(h \otimes g)(u \otimes v) = \langle u, g \rangle_{\mathcal{H}_2} (h \otimes v), \tag{5}$$

where the superscript * denotes the adjoint operator.

3. FORMULATION OF LEARNING PROBLEMS AND KERNEL SPECIFIC GENERALIZATION ABILITY

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

$$y_k = f(\boldsymbol{x}_k) + n_k, \tag{6}$$

where f denotes the unknown true function and n_k denotes a zero-mean additive noise. In pattern recognition problems, y_k denotes a class label, and in regression or density estimation problems, it denotes a value of the function f at a point $\boldsymbol{x}_k$ with additive noise. The aim of machine learning is to estimate the unknown true function f by using the given training data set and statistical properties of the noise.

In this paper, we assume that the unknown true function f belongs to the RKHS $\mathcal{H}_K$ corresponding to a certain kernel function K. If $f \in \mathcal{H}_K$, then Eq.(6) is rewritten as

$$y_k = \langle f(\cdot), K(\cdot, \boldsymbol{x}_k) \rangle_{\mathcal{H}_K} + n_k, \tag{7}$$

on the basis of the reproducing property of kernels. Let $\boldsymbol{y} = [y_1, \ldots, y_\ell]'$ and $\boldsymbol{n} = [n_1, \ldots, n_\ell]'$ with the superscript $'$ denoting the transposition operator, then applying the Schatten product to Eq.(7) yields

$$\boldsymbol{y} = \left(\sum_{k=1}^{\ell} [\boldsymbol{e}_k^{(\ell)} \otimes K(\cdot, \boldsymbol{x}_k)] \right) f(\cdot) + \boldsymbol{n}, \tag{8}$$

where $\boldsymbol{e}_k^{(\ell)}$ denotes the k-th vector of the canonical basis of $\mathbf{R}^\ell$. For a convenience of description, we write

$$A_K = \left(\sum_{k=1}^{\ell} [\boldsymbol{e}_k^{(\ell)} \otimes K(\cdot, \boldsymbol{x}_k)] \right). \tag{9}$$

The operator A_K is a linear map from $\mathcal{H}_K$ onto $\mathbf{R}^\ell$ and Eq.(8) can be rewritten as

$$\boldsymbol{y} = A_K f(\cdot) + \boldsymbol{n}, \tag{10}$$

which represents the relationship between the unknown true function f and an output vector $\boldsymbol{y}$. The information of input vectors is integrated in the operator A_K. Therefore, a machine learning problem can be interpreted as an inversion problem of the linear equation Eq.(10)[11]. In general, an estimated function $\hat{f}$ is represented as

$$\hat{f}(\cdot) = L\boldsymbol{y}, \tag{11}$$

where L denotes a learning operator specified by a learning criterion such as that of the support vector machine and that of the kernel ridge regression.

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

$$\hat{f}(\cdot) = A_K^* \boldsymbol{\alpha} \tag{12}$$

$$= \left(\sum_{k=1}^{\ell} [K(\cdot, \boldsymbol{x}_k) \otimes \boldsymbol{e}_k^{(\ell)}] \right) \boldsymbol{\alpha} \tag{13}$$

$$= \sum_{k=1}^{\ell} \alpha_k K(\cdot, \boldsymbol{x}_k) \tag{14}$$

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

$$J(K, f) = ||f - P_K f||_{\mathcal{H}_K}^2, \tag{15}$$

where P_K denotes the orthogonal projector onto $\mathcal{R}(A_K^*)$ in $\mathcal{H}_K$ and $|| \cdot ||_{\mathcal{H}_K}$ denotes the induced norm in $\mathcal{H}_K$. Note that the orthogonality of P_K is also defined by the metric in $\mathcal{H}_K$. A selection of an element in $\mathcal{R}(A_K^*)$ as a learning result is out of the scope of this paper since this selection depends on learning criteria. We also ignore the observation noise in the following contents since the noise does not affect Eq.(15).

4. OPTIMAL KERNEL IN A CLASS OF REPRODUCING KERNEL HILBERT SPACES WITH AN INVARIANT METRIC

In [9], we discussed a class of kernels whose corresponding RKHS's have an invariant metric and proved that the kernel corresponding to the smallest RKHS including an unknown

true function gives the optimal model, that is, the best generalization ability. In this section, we review the discussions of [9].

Firstly, we give important theorems concerned with nested RKHS's shown in [7].

Theorem 1 *[7] 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 $f = f_1 + f_2$ with $f_i \in F_i$, and with the norm defined by*

$$||f||^2 = \min\left[||f_1||_1^2 + ||f_2||_2^2\right], \tag{16}$$

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

Theorem 2 *[7] If K is the reproducing kernel of the class F with the norm $||\cdot||$, and if the linear class $F_1 \subset F$ forms a Hilbert space with the norm $||\cdot||_1$, such that $||f||_1 \geq ||f||$ for any $f \in F_1$, then the class F_1 possesses a reproducing kernel K_1 such that $K^c = K - K_1$ is also a reproducing kernel.*

Theorem 3 *[7] If K and K_1 are the reproducing kernels of the classes of F and F_1 with the norms $||\cdot||$, $||\cdot||_1$, and if $K - K_1$ is a reproducing kernel, then $F_1 \subset F$ and $||f_1||_1 \geq ||f_1||$ for every $f_1 \in F_1$.*

Let us consider nested RKHS's $\mathcal{H}_{K_1}$ and $\mathcal{H}_{K_2}$ satisfying

$$\mathcal{H}_{K_1} \subset \mathcal{H}_{K_2}, \tag{17}$$

specified by a class of kernels $\{K_i \mid i \in \{1,2\}\}$. We assume that $\mathcal{H}_{K_i}$ has an invariant metric for any $i \in \{1,2\}$, that is,

$$\langle f, g\rangle_{\mathcal{H}_{K_1}} = \langle f, g\rangle_{\mathcal{H}_{K_2}} \tag{18}$$

for any $f, g \in \mathcal{H}_{K_1}$ and

$$||f||^2_{\mathcal{H}_{K_1}} = ||f||^2_{\mathcal{H}_{K_2}} \tag{19}$$

for any $f \in \mathcal{H}_{K_1}$. According to Theorem 2, there exists a kernel K^c such that

$$K_2 = K_1 + K^c. \tag{20}$$

The following theorem is the main result of [9].

Theorem 4 *[9] For any input vectors $\{\boldsymbol{x}_k \mid k \in \{1,\ldots,\ell\}\}$,*

$$||f - P_{K_1} f||^2_{\mathcal{H}_{K_2}} \leq ||f - P_{K_2} f||^2_{\mathcal{H}_{K_2}} \tag{21}$$

holds for any $f \in \mathcal{H}_{K_1}$.

According to Theorem 4, given a class of kernels that forms a nested class of RKHS's with an invariant metric, it is concluded that the kernel corresponding to the smallest RKHS, including the unknown true function, gives the best generalization ability among the given class of kernels. One of typical examples of a class of kernels whose corresponding RKHS's have an invariant metric is the sinc kernel (sampling function used in Shannon's sampling theorem). As is well known, the RKHS of the sinc kernel is a subspace of L^2. This theoretical result may play an important role in analyzing a generalization ability of kernel machines. However, the result in [9] does not contribute toward finding a class of kernels whose corresponding RKHS's have an invariant metric.

5. THEORETICAL ANALYSES FOR A CLASS OF KERNELS WITH AN INVARIANT METRIC

In this section, we discuss properties of a class of kernels whose corresponding RKHS's have an invariant metric.

The following theorem is the main result of this paper.

Theorem 5 *Let K_1 and $K_2 = K_1 + K^c$ be kernels whose corresponding RKHS's satisfy $\mathcal{H}_{K_1} \subset \mathcal{H}_{K_2}$. The following three statements are equivalent each other.*

1) *For any $f \in \mathcal{H}_{K_1}$, $||f||^2_{\mathcal{H}_{K_1}} = ||f||^2_{\mathcal{H}_{K_2}}$.*
2) *$\mathcal{H}_{K_1} \cap \mathcal{H}_{K^c} = \{0\}$.*
3) *For any $f_1 \in \mathcal{H}_{K_1}$ and $f_2 \in \mathcal{H}_{K^c}$, $\langle f_1, f_2\rangle_{\mathcal{H}_{K_2}} = 0$.*

Proof

1) $\rightarrow$ 2)

From Theorem 1,

$$||f||^2_{\mathcal{H}_{K_2}} = \min\left[||f_1||^2_{\mathcal{H}_{K_1}} + ||f_2||^2_{\mathcal{H}_{K^c}}\right], \tag{22}$$

holds for any $f \in \mathcal{H}_{K_2}$ with $f = f_1 + f_2$, $(f_1 \in \mathcal{H}_{K_1}, f_2 \in \mathcal{H}_{K^c})$. It is trivial that Eq.(22) holds for any $f \in \mathcal{H}_{K_1} \cap \mathcal{H}_{K^c}$ since $\mathcal{H}_{K_1} \cap \mathcal{H}_{K^c} \subset \mathcal{H}_{K_2}$. Thus, for any $f \in \mathcal{H}_{K_1} \cap \mathcal{H}_{K^c}$,

$$\begin{aligned} ||f||^2_{\mathcal{H}_{K_1}} = ||f||^2_{\mathcal{H}_{K_2}} &= \min\left[||f_1||^2_{\mathcal{H}_{K_1}} + ||f_2||^2_{\mathcal{H}_{K^c}}\right] \\ &\leq \min_{\alpha}\left[||\alpha f||^2_{\mathcal{H}_{K_1}} + ||(1-\alpha)f||^2_{\mathcal{H}_{K^c}}\right] \\ &= \frac{||f||^2_{\mathcal{H}_{K_1}}||f||^2_{\mathcal{H}_{K^c}}}{||f||^2_{\mathcal{H}_{K_1}} + ||f||^2_{\mathcal{H}_{K^c}}}. \end{aligned}$$

On the other hand, it is easy to show that

$$||f||^2_{\mathcal{H}_{K_1}} - \frac{||f||^2_{\mathcal{H}_{K_1}}||f||^2_{\mathcal{H}_{K^c}}}{||f||^2_{\mathcal{H}_{K_1}} + ||f||^2_{\mathcal{H}_{K^c}}} \geq 0$$

holds for any $f \in \mathcal{H}_{K_1} \cap \mathcal{H}_{K^c}$. Thus, it is concluded that

$$||f||^2_{\mathcal{H}_{K_1}} = \frac{||f||^2_{\mathcal{H}_{K_1}}||f||^2_{\mathcal{H}_{K^c}}}{||f||^2_{\mathcal{H}_{K_1}} + ||f||^2_{\mathcal{H}_{K^c}}}, \tag{23}$$

which trivially yields $||f||^2_{\mathcal{H}_{K_1}} = 0$ for any $f \in \mathcal{H}_{K_1} \cap \mathcal{H}_{K^c}$. Therefore, $\mathcal{H}_{K_1} \cap \mathcal{H}_{K^c} = \{0\}$ is obtained.

2) $\rightarrow$ 1)

If $\mathcal{H}_{K_1} \cap \mathcal{H}_{K^c} = \{0\}$ holds, then Theorem 1 yields

$$||f||^2_{\mathcal{H}_{K_2}} = ||f||^2_{\mathcal{H}_{K_1}} \tag{24}$$

since for any $f \in \mathcal{H}_{K_1}$, the decomposition

$$f = f_1 + f_2,\ (f_1 = f \in \mathcal{H}_{K_1}, f_2 = 0 \in \mathcal{H}_{K^c})$$

is unique.

2) $\rightarrow$ 3)

If $\mathcal{H}_{K_1} \cap \mathcal{H}_{K^c} = \{0\}$ holds, then for any $f \in \mathcal{H}_{K_2}$, the decomposition $f = f_1 + f_2,\ (f_1 \in \mathcal{H}_{K_1}, f_2 \in \mathcal{H}_{K^c})$ is unique. Thus, Theorem 1 yields

$$||f||^2_{\mathcal{H}_{K_2}} = ||f_1||^2_{\mathcal{H}_{K_1}} + ||f_2||^2_{\mathcal{H}_{K^c}}. \tag{25}$$

On the other hand, the uniqueness of the decomposition also yields

$$\begin{aligned} ||f||^2_{\mathcal{H}_{K_2}} &= ||f_1 + f_2||^2_{\mathcal{H}_{K_2}} \\ &= ||f_1||^2_{\mathcal{H}_{K_2}} + ||f_2||^2_{\mathcal{H}_{K_2}} + 2\langle f_1, f_2\rangle_{\mathcal{H}_{K_2}} \\ &= ||f_1||^2_{\mathcal{H}_{K_1}} + ||f_2||^2_{\mathcal{H}_{K^c}} + 2\langle f_1, f_2\rangle_{\mathcal{H}_{K_2}}. \end{aligned} \tag{26}$$

Thus, from Eqs.(25) and (26),

$$\langle f_1, f_2\rangle_{\mathcal{H}_{K_2}} = 0 \tag{27}$$

is obtained.

3) $\rightarrow$ 2)

If $\langle f_1, f_2\rangle_{\mathcal{H}_{K_2}} = 0$ for any $f_1 \in \mathcal{H}_{K_1}$ and $f_2 \in \mathcal{H}_{K^c}$, it is trivial that $\mathcal{H}_{K_1} \cap \mathcal{H}_{K^c} = \{0\}$. □

According to Theorem 5, it is concluded that the invariance of the metric of nested RKHS's is identical to the disjointness of the smaller RKHS and the added RKHS; and these two properties are identical to the orthogonality of these two RKHS's. Therefore, we can identify a class of kernels with an invariant metric by checking the disjointness or the orthogonality of these two RKHS's.

Combining Theorems 4 and 5 implies that if the unknown target function f belongs to $\mathcal{H}_{K_1}$, then the kernel $K_1 + K^c$ does not improve the (kernel specific) generalization ability when $\mathcal{H}_{K_1}$ and $\mathcal{H}_{K^c}$ are disjoint or orthogonal.

On the other hand, as mentioned in [9], the limitation of the invariant metric is quite severe. Thus, analyses for a class of RKHS's, whose metrics are not always invariant, are one of future works that should be resolved.

6. CONCLUSION

In this paper, we discussed a class of kernels that forms a nested class of RKHS's and proved that such a class of RKHS's has an invariant metric if and only if the smaller RKHS and the added one are disjoint or orthogonal. Analyses for a class of RKHS's whose metrics are not always invariant, are one of future works.

7. ACKNOWLEDGMENT

The authors would thank to Professor Masashi Sugiyama in Tokyo Institute of Technology for his valuable comments. This work was partially supported by Grant-in-Aid No.21700001 for Young Scientists (B) from the Ministry of Education, Culture, Sports and Technology of Japan.

8. 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. Sugiyama and H. Ogawa, "Subspace Information Criterion for Model Selection," *Neural Computation*, vol. 13, no. 8, pp. 1863–1889, 2001.

[6] M. Sugiyama, M. Kawanabe, and K. Muller, "Trading variance reduction with unbiasedness: The regularized subspace information criterion for robust model selection in kernel regression," *Neural Computation*, vol. 16, no. 5, pp. 1077–1104, 2004.

[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] A. Tanaka, , H. Imai, M. Kudo, and M. Miyakoshi, "Optimal kernel in a class of kernels with an invariant metric," in *Joint IAPR International Workshops SSPR 2008 and SPR 2008*. 2008, pp. 530–539, Springer.

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

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