Integrated Kernels and Their Properties

Akira Tanaka, Hideyuki Imai, Mineichi Kudo, and Masaaki Miyakoshi
Division of Computer Science,
Graduate School of Information Science and Technology,
Hokkaido University, N14W9, Kita-ku, Sapporo 060-0814, Japan.

abstract

Kernel machines are widely considered to be powerful tools in various fields of information science. By using a kernel, an unknown target is represented by a function that belongs to a reproducing kernel Hilbert space (RKHS) corresponding to the kernel. The application area is widened by enlarging the RKHS such that it includes a wide class of functions. In this study, we demonstrate a method to perform this by using parameter integration of a parameterized kernel. Some numerical experiments show that the unresolved problem of finding a good parameter can be neglected.

keyword

kernel, reproducing kernel Hilbert space, projection learning, parameter integration

1 Introduction

Learning based on kernel machines [1, 2, 3] has been widely considered to be a powerful tool in various fields of information science such as pattern recognition, regression estimation, density estimation, etc. In several existing approaches, the adequacy of kernel machines is measured by the difference between the predictive output of an assumed model with the training input data set and a training output data set (e.g., class labels in a pattern recognition problem, function values in a density or regression estimation problem, etc.). In these approaches, a kernel, whose mathematical properties are described in detail in [4], is recognized as a useful tool for calculating the inner product in a certain feature space.

On the other hand, Ogawa formulated a learning problem as an estimation of the unknown function in a certain function space. In this approach, the adequacy of learning is measured by the difference between the unknown true function and the estimated one in the function space determined by a kernel. Therefore, a kernel plays a crucial role in the determination of a function space to which the unknown target function belongs. This scheme is referred to as

(parametric) projection learning [5, 6, 7, 8] and it has been yielding several interesting results, primarily in the field of neural networks (see [9] for instance). This approach seems reasonable since it is widely known that the mathematical essence of using a kernel is that the unknown target (a classifier in a pattern recognition problem, function in a density or regression estimation problem, etc.) can be represented by a function that belongs to a reproducing kernel Hilbert space (RKHS) [4] that corresponds to an adopted kernel.

In the real world, however, targets to be learned are not always functions as one found in a pattern recognition problem in which some classes essentially overlap. The application of kernel machines to such problems requires methods such as the "soft margin" technique in a support vector machine (SVM) [10]. On the other hand, it is also important to analyze the performance and properties of a kernel machine when the unknown target can be represented by a function. In such a case, the use of kernel is theoretically validated. Here, one of the primary topics that requires theoretical clarification is the question: What constitutes a good kernel? As already mentioned, the condition that the unknown true function belongs to the RKHS corresponding to an adopted kernel imparts theoretical consistency to kernel machines. This understanding provides a suggestion for designing a kernel. In general, information about the unknown true function is limited. Therefore, we have to construct the RKHS to be as large as possible. This imparts consistency to kernel machines for a wide class of functions. In this study, we show that a kernel corresponding to such a large RKHS is realized by the parameter integration of a parameterized kernel. Further, some numerical examples are given in order to validate our theory.

2 Mathematical Preliminaries for the RKHS Theory

In this section, we prepare some mathematical tools to deal with the RKHS theory.

Definition 1 *[4] 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 $K(\boldsymbol{x}, \boldsymbol{y})$ $(\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D})$ is referred to as a reproducing kernel of $\mathcal{H}$, if*

1. *for every $\boldsymbol{y} \in \mathcal{D}$, $K(\boldsymbol{x}, \boldsymbol{y})$ is a function of $\boldsymbol{x}$ belonging to $\mathcal{H}$ and*
2. *for every $\boldsymbol{y} \in \mathcal{D}$ and every $f \in \mathcal{H}$,*

$$f(\boldsymbol{y}) = \langle f(\boldsymbol{x}), K(\boldsymbol{x}, \boldsymbol{y}) \rangle, \tag{1}$$

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

The Hilbert space $\mathcal{H}$ is referred to as an RKHS when it has a reproducing kernel. The reproducing property, Eq.(1), enables the realization of the value

of the function at a point in $\mathcal{D}$. Note that the reproducing kernels are positive definite [4]:

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

for any N, $c_1, \ldots, c_N \in \mathbf{R}$, and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathcal{D}$. In addition, $K(\boldsymbol{x}, \boldsymbol{y}) = K(\boldsymbol{y}, \boldsymbol{x})$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}$ follows [4]. If a reproducing kernel $K(\boldsymbol{x}, \boldsymbol{y})$ exists, it is unique [4]. Conversely, every positive definite function $K(\boldsymbol{x}, \boldsymbol{y})$ has a unique corresponding RKHS [4]. Therefore, it is guaranteed that any positive definite function $K(\boldsymbol{x}, \boldsymbol{y})$ is always a reproducing kernel.

Next, we introduce the Schatten product [11] that distinctly reveals the reproducing property of kernels.

Definition 2 *[11] 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 as*

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

Note that $(g \otimes h)$ is a linear operator from $\mathcal{H}_1$ onto $\mathcal{H}_2$. It can be easily shown that the following relations hold for $h, v \in \mathcal{H}_1$ and $g, u \in \mathcal{H}_2$:

$$(h \otimes g)^* = (g \otimes h), \quad (h \otimes g)(u \otimes v) = \langle u, g \rangle (h \otimes v), \tag{4}$$

where X^* denotes the adjoint operator of X.

3 Interpretation of Learning as a Linear Inverse Problem

Let $(y_i, \boldsymbol{x}_i)$, $(y_i \in \mathbf{R},\ \boldsymbol{x}_i \in \mathbf{R}^n,\ i = 1, \ldots, \ell)$ be a given training data set of ℓ samples that satisfies

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

where f and $n_i \in \mathbf{R}$ denote a real-valued function and additive noise, respectively. In regression estimation or density estimation problems, y_i takes a real value, whereas y_i takes a class label in pattern recognition problems. In this study, the aim of the machine learning is assumed to be the estimation of the unknown function f using a training data set, *a priori* knowledge about the function space, and statistical properties of additive noise. In this study, we assume that f belongs to $\mathcal{H}_K$, the RKHS corresponding to a certain kernel K. Based on the reproducing property of kernels, the value of a function $f \in \mathcal{H}_K$ at a point $\boldsymbol{x}_i$ is written as

$$f(\boldsymbol{x}_i) = \langle f(\boldsymbol{x}), K(\boldsymbol{x}, \boldsymbol{x}_i) \rangle. \tag{6}$$

Therefore, Eq.(5) is rewritten as

$$y_i = \langle f(\boldsymbol{x}), K(\boldsymbol{x}, \boldsymbol{x}_i) \rangle + n_i. \tag{7}$$

Let $\boldsymbol{y} := [y_1, \ldots, y_\ell]'$ and $\boldsymbol{n} := [n_1, \ldots, n_\ell]'$, where X' denotes the transposed matrix (or vector) of X. By applying the Schatten product to Eq.(7), we have

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

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

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

It is important to note that operator A is linear, irrespective of whether $f(\boldsymbol{x})$s are linear or non-linear. Now, the simplest form of Eq.(8) can be expressed as

$$\boldsymbol{y} = Af(\boldsymbol{x}) + \boldsymbol{n}. \tag{10}$$

This equation represents the relation between the unknown target function $f(\boldsymbol{x})$ and output $\boldsymbol{y}$. All the information about the input vectors is integrated in operator A. Therefore, a machine learning problem can be interpreted as an inverse problem of Eq.(10) [5, 6].

Based on the model described by Eq.(10), Ogawa proposed a novel learning framework referred to as (parametric) projection learning [5, 6]. Projection learning yields a minimum variance unbiased estimator of the orthogonal projection of the unknown function $f(\boldsymbol{x})$ onto $\mathcal{R}(A^*)$, the range of A^*; on the other hand, parametric projection learning results in an improvement in the unknown function by incorporating the relaxation of the unbiasedness of projection learning to suppress the influence of noise. Parametric projection learning includes projection learning as a special case. In the framework of (parametric) projection learning, the solution is the learning operator B; by using it, the estimated function is expressed as

$$\hat{f}(\boldsymbol{x}) = B\boldsymbol{y}. \tag{11}$$

Parametric projection learning is defined as follows:

Definition 3 *[7, 8] The learning operator B_{PPL} of parametric projection learning is given as*

$$\begin{aligned} B_{PPL}(\gamma) := \mathrm{argmin}_B [\mathrm{tr}[(BA - P_{\mathcal{R}(A^*)})(BA - P_{\mathcal{R}(A^*)})^*] \\ + \gamma E_{\boldsymbol{n}} ||B\boldsymbol{n}||^2], \end{aligned} \tag{12}$$

where $P_{\mathcal{R}(A^*)}$ denotes the orthogonal projector onto $\mathcal{R}(A^*)$ and γ denotes a real positive parameter that controls the trade-off between the two terms.

As shown in [7, 8], one of the solutions of the parametric projection learning is given as

$$B_{PPL}(\gamma) = A^*(AA^* + \gamma Q)^+, \tag{13}$$

where X^+ denotes the Moore-Penrose generalized inverse [12] of X, and Q denotes the noise correlation operator defined by

$$Q := E_{\boldsymbol{n}}[\boldsymbol{n}\boldsymbol{n}^*]. \tag{14}$$

Finally, the solution of the parametric projection learning is written as

$$\hat{f}(\boldsymbol{x}) = B_{PPL}\boldsymbol{y}, \tag{15}$$

or specifically

$$\begin{aligned} \hat{f}(\boldsymbol{x}) &= \left(\sum_{i=1}^{\ell}\left[K(\boldsymbol{x},\boldsymbol{x_i})\otimes \boldsymbol{e}_i^{(\ell)}\right]\right)(G+\gamma Q)^+\boldsymbol{y} \\ &= \sum_{i=1}^{\ell}\boldsymbol{y}'(G+\gamma Q)^+\boldsymbol{e}_i^{(\ell)}K(\boldsymbol{x},\boldsymbol{x_i}), \end{aligned} \tag{16}$$

where $G = AA^*$ is the Gram's matrix of K expressed as $G = (g_{ij})$, $g_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, which is easily confirmed by the properties of the Schatten product (Eq.(4)). The appropriate parameter γ can be selected by using a criterion such as the subspace information criterion [13]. Note that the assumption $Q = O$ (zero matrix) yields the solution based on the Moore-Penrose generalized inverse of A, while the assumption $Q = I_\ell$ yields a solution identical to the kernel ridge regression [2, 3] and the Gaussian process [3].

4 Integrated Kernel and its Properties

Considering the learning problems in terms of Eq.(10), the only assumed condition is $f(\boldsymbol{x}) \in \mathcal{H}_K$, which is crucial for the theoretical consistency of the learning; further, it yields an important suggestion for kernel design or kernel selection. If *a priori* knowledge about the function space to which the unknown target function belongs is available, it is sufficient to adopt the corresponding kernel as long as such a kernel exists. However, in general, this knowledge is unavailable. Therefore, the second best method is adopting a kernel whose corresponding RKHS is as large as possible; this guarantees theoretical consistency for a wide class of functions. One resolution for this is to adopt the sum of a number of kernels since it is shown in [4] that the corresponding RKHS includes the RKHSs of the summed kernels. However, some difficulties such as the selection of kernels to be summed up and the cost of calculations still remain. To overcome these difficulties, we introduce an integrated kernel that actually resolves these difficulties.

The inner product $\langle\cdot,\cdot\rangle$, as introduced in Section 2, is defined by integration on the Lebesgue measurable space. Similarly, the following integrations are on the Lebesgue measurable space.

Definition 4 *Let $\mathcal{Q} \subset \mathbf{R}$ be a Borel set and let $K_\theta(\boldsymbol{x},\boldsymbol{y})$ be a kernel with parameter $\theta \in \mathcal{Q}$. We assume that $K_\theta(\boldsymbol{x},\boldsymbol{y})$ is an integrable function with*

respect to the Lebesgue measure for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}$*, that is,*

$$\int_{\mathcal{Q}} |K_\theta(\boldsymbol{x}, \boldsymbol{y})| d\theta < \infty. \tag{17}$$

The integrated kernel is defined as

$$K_{\mathcal{Q}}(\boldsymbol{x}, \boldsymbol{y}) := \int_{\mathcal{Q}} K_\theta(\boldsymbol{x}, \boldsymbol{y}) d\theta. \tag{18}$$

Note that $K_\theta(\boldsymbol{x}, \boldsymbol{y})$ *is also integrable on an arbitrary Borel set* $\Xi \subseteq \mathcal{Q}$ *since*

$$\int_{\Xi} |K_\theta(\boldsymbol{x}, \boldsymbol{y})| d\theta \leq \int_{\mathcal{Q}} |K_\theta(\boldsymbol{x}, \boldsymbol{y})| d\theta < \infty \tag{19}$$

holds.

Theorem 1 $K_\Xi(\boldsymbol{x}, \boldsymbol{y})$ *is a kernel for an arbitrary Borel set* $\Xi \subseteq \mathcal{Q}$*.*

Proof It is trivial that

$$\sum_{i,j=1}^{N} c_i c_j K_\Xi(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{20}$$

is finite for any N, $c_1, \ldots, c_N \in \mathbf{R}$, and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathcal{D}$ since $K_\Xi(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is finite for any $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{D}$. Therefore, it is sufficient to confirm that

$$\sum_{i,j=1}^{N} c_i c_j K_\Xi(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0. \tag{21}$$

On the basis of the facts that

1. $\sum_{i,j=1}^{N} c_i c_j K_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is a measurable function for any N, $c_1, \ldots, c_N \in \mathbf{R}$, and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathcal{D}$ since $K_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is measurable for any $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{D}$,
2. the integral of the sum of a finite number of measurable functions is identical to the sum of the integral of those measurable functions, and
3. $\sum_{i,j=1}^{N} c_i c_j K_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$ holds for any $\theta \in \Xi \subseteq \mathcal{Q}$,

it immediately follows that

$$\begin{aligned}
&\sum_{i,j=1}^{N} c_i c_j K_\Xi(\boldsymbol{x}_i, \boldsymbol{x}_j) \\
&= \sum_{i,j=1}^{N} c_i c_j \left(\int_{\Xi} K_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j) d\theta \right) \\
&= \int_{\Xi} \left(\sum_{i,j=1}^{N} c_i c_j K_\theta(\boldsymbol{x}_i, \boldsymbol{x}_j) \right) d\theta \geq 0,
\end{aligned} \tag{22}$$

which concludes the proof. □

Next, we investigate the relationships between $K_{\mathcal{Q}}(\boldsymbol{x}, \boldsymbol{y})$ and $K_\theta(\boldsymbol{x}, \boldsymbol{y})$. The following theorem is useful for investigating the inclusion relation of two RKHSs.

Theorem 2 *[14] Let the kernels be $K_1(\boldsymbol{x}, \boldsymbol{y})$ and $K_2(\boldsymbol{x}, \boldsymbol{y})$.*

$$\mathcal{H}_{K_1} \subset \mathcal{H}_{K_2} \tag{23}$$

holds, if and only if there exists a real positive number γ that makes

$$\gamma K_2(\boldsymbol{x}, \boldsymbol{y}) - K_1(\boldsymbol{x}, \boldsymbol{y}) \tag{24}$$

a kernel, that is, a positive definite function.

Definition 5 *Let $f(x)$ be a continuous function defined on a closed interval $[a, b]$, and let δ be a real positive constant satisfying*

$$\delta < \frac{b-a}{2}. \tag{25}$$

A function $f(x)$ is called a δ-monotone continuous function if $f(x)$ is monotonously increasing or decreasing on $[\xi - \delta, \xi + \delta]$ for any $\xi \in [a + \delta, b - \delta]$.

The concept of such functions is introduced in order to describe a continuous function without any extremal points. Note that monotonously increasing and decreasing continuous functions are δ-monotone continuous functions for any $0 < \delta < (b - a)/2$.

Lemma 1 *Let δ be a real positive constant satisfying*

$$0 < \delta < \frac{b-a}{2}, \tag{26}$$

and let $f(x)$ be a δ-monotone continuous function defined on $[a, b]$; then, there exists $c \in [a + \delta/2, b - \delta/2]$ such that

$$\int_{c-\delta/2}^{c+\delta/2} f(x)dx = \delta f(\xi) \tag{27}$$

for any $\xi \in [a + \delta, b - \delta]$.

Proof Let

$$F(t) = \int_{t-\delta/2}^{t+\delta/2} f(x)dx. \tag{28}$$

Note that $F(t)$ is a continuous function and $[t - \delta/2, t + \delta/2] \subset [a, b]$ for any $t \in [a + \delta/2, b - \delta/2]$. It is trivial that

$$F(\xi - \delta/2) \leq \delta f(\xi) \leq F(\xi + \delta/2) \tag{29}$$

holds when $f(x)$ is monotonously increasing on $[\xi - \delta, \xi + \delta]$. On the other hand, when $f(x)$ is monotonously decreasing on $[\xi - \delta, \xi + \delta]$,

$$F(\xi + \delta/2) \leq \delta f(\xi) \leq F(\xi - \delta/2) \tag{30}$$

holds. By the mean value theorem, the existence of $c \in [\xi - \delta/2, \xi + \delta/2]$ that satisfies

$$F(c) = \delta f(\xi) \tag{31}$$

is guaranteed. □

Theorem 3 *Let $\mathcal{Q}$ and $\mathcal{Q}_\delta$ be closed intervals expressed as*

$$\mathcal{Q} := [\theta_s, \theta_e] \tag{32}$$
$$\mathcal{Q}_\delta := [\theta_s + \delta, \theta_e - \delta] \tag{33}$$

with $\theta_s < \theta_e$, where δ denotes a real positive constant satisfying

$$\delta < \frac{\theta_e - \theta_s}{2}; \tag{34}$$

then, the RKHS corresponding to $K_{\mathcal{Q}}(\boldsymbol{x}, \boldsymbol{y})$ includes the RKHS corresponding to $K_\theta(\boldsymbol{x}, \boldsymbol{y})$ for any $\theta \in \mathcal{Q}_\delta$ if $K_\theta(\boldsymbol{x}, \boldsymbol{y})$ is a δ-monotone continuous function with respect to θ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}$.

Proof According to Lemma 1, for any $\xi \in [\theta_s + \delta, \theta_e - \delta]$, there exists $c \in [\theta_s + \delta/2, \theta_e - \delta/2]$ satisfying

$$\int_{c-\delta/2}^{c+\delta/2} K_\theta(\boldsymbol{x}, \boldsymbol{y}) d\theta = \delta K_\xi(\boldsymbol{x}, \boldsymbol{y}), \tag{35}$$

if $K_\theta(\boldsymbol{x}, \boldsymbol{y})$ is a δ-monotone continuous function with respect to θ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}$. Note that $\mathcal{Q}_c := [c - \delta/2, c + \delta/2] \subset \mathcal{Q}$ holds. Therefore,

$$\begin{aligned} &\frac{1}{\delta} K_{\mathcal{Q}}(\boldsymbol{x}, \boldsymbol{y}) - K_\xi(\boldsymbol{x}, \boldsymbol{y}) \\ &= \frac{1}{\delta} \int_{\mathcal{Q}} K_\theta(\boldsymbol{x}, \boldsymbol{y}) d\theta - \frac{1}{\delta} \int_{\mathcal{Q}_c} K_\theta(\boldsymbol{x}, \boldsymbol{y}) d\theta \\ &= \frac{1}{\delta} \int_{\mathcal{Q} - \mathcal{Q}_c} K_\theta(\boldsymbol{x}, \boldsymbol{y}) d\theta \end{aligned} \tag{36}$$

is a kernel by Theorem 1 since $K_\theta(\boldsymbol{x}, \boldsymbol{y})$ is a kernel for any $\theta \in \mathcal{Q} - \mathcal{Q}_c$, and $\mathcal{Q} - \mathcal{Q}_c \subset \mathcal{Q}$ is a Borel set. From Theorem 2 and the fact that $1/\delta$ is a positive real bounded constant, for any $\theta \in \mathcal{Q}_\delta$

$$\mathcal{H}_{K_\theta} \subset \mathcal{H}_{K_{\mathcal{Q}}} \tag{37}$$

follows. □

Note that a δ-monotone continuous function is also a δ'-monotone continuous function for any $\delta' \in (0, \delta)$. Thus, if $K_\theta(\boldsymbol{x}, \boldsymbol{y})$ is a δ-monotone continuous function with a certain δ, then Theorem 3 holds for any $\delta' \in (0, \delta)$. On the other hand, there exists $\delta' \in (0, \delta)$ that satisfies $\theta \in \mathcal{Q}_\delta$ for any $\theta \in (\theta_s, \theta_e)$. Thus,

$$\mathcal{H}_{K_\theta} \subset \mathcal{H}_{K_{\mathcal{Q}}} \tag{38}$$

holds for any $\theta \in (\theta_s, \theta_e)$. This implies that, in practice, the RKHS corresponding to the integrated kernel $K_{\mathcal{Q}}(\boldsymbol{x}, \boldsymbol{y})$ includes the RKHS corresponding to $K_\theta(\boldsymbol{x}, \boldsymbol{y})$ for any $\theta \in \mathcal{Q}$. Note that if a kernel has an extremal point at a certain $\theta_m \in \mathcal{Q}$ and thus the kernel is not a δ-monotone continuous function,

then the RKHS corresponding to $K_{\theta_m}(\boldsymbol{x}, \boldsymbol{y})$ may not be included in the RKHS corresponding to $K_{\mathcal{Q}}(\boldsymbol{x}, \boldsymbol{y})$.

The concept of integrated kernels can be easily extended to the cases in which the dimension of the parameter space is greater than unity as long as the conditions in Theorem 3 are satisfied for each parameter.

5 Numerical Examples

In this section, we illustrate some numerical examples in a pattern recognition problem and function estimation problem. As already mentioned, only the case in which the target to be learned can be expressed as a function is considered. Therefore, we do not consider the case in which misclassification is permitted for training samples, even in a pattern recognition problem. We adopt the Gaussian kernel

$$K_\sigma(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{y}||^2}{2\sigma^2}\right) \tag{39}$$

with various variances σ^2 as the kernel. Then, its integrated version is given as

$$\begin{aligned} & K_{\left[0,\sqrt{2}\sigma_0\right]}(\boldsymbol{x}, \boldsymbol{y}) \\ &= \int_0^{\sqrt{2}\sigma_0} \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{y}||^2}{\sigma^2}\right) d\sigma \\ &= \sqrt{2}\sigma_0 \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{y}||^2}{2\sigma_0^2}\right) \\ &\quad -\sqrt{\pi}||\boldsymbol{x} - \boldsymbol{y}||\mathrm{erfc}\left(\frac{||\boldsymbol{x} - \boldsymbol{y}||}{\sqrt{2}\sigma_0}\right), \end{aligned} \tag{40}$$

where erfc(x) denotes the complementary error function defined by

$$\mathrm{erfc(x)} = \frac{2}{\sqrt{\pi}} \int_x^\infty \exp\left(-t^2\right) dt. \tag{41}$$

It is empirically known that a Gaussian kernel with an extremely small variance overfits the training data set, while Eq.(40) does not exist for $\sigma_0 = \infty$. Therefore, we consider the maximum eigenvalue of the covariance matrix of the input vectors for σ_0^2 as a sufficiently large value. Note that considering a value greater than the value considered by us does not create any problem as long as the integral in Eq.(40) exists. Further, note that the RKHS corresponding to $K_{\left[0,\sqrt{2}\sigma_0\right]}(\boldsymbol{x}, \boldsymbol{y})$ includes the RKHSs corresponding to $K_\sigma(\boldsymbol{x}, \boldsymbol{y})$, $(\sigma \in (0, \sigma_0))$ since a Gaussian kernel is a monotonously increasing continuous function with respect to the parameter $\sigma > 0$. Therefore, if the problem can be represented by $K_\sigma(\boldsymbol{x}, \boldsymbol{y})$ with a certain $\sigma \in (0, \sigma_0)$, it is represented by $K_{\left[0,\sqrt{2}\sigma_0\right]}(\boldsymbol{x}, \boldsymbol{y})$. Figure 1 shows a graph of the integrated kernel with $\sigma_0^2 = 1$.

We assume $Q = O$ for all the following examples. Therefore, we adopt the Moore-Penrose generalized inverse of the operator A as the learning machine. It

should be noted that the issue under consideration is not the learning machine but the kernel itself.

5.1 Example of the Pattern Recognition Problem

First, we deal with a pattern recognition problem. Figure 2 shows the scatter diagram of the training data set, referred to as "SPIRAL," of two classes. In the diagram, the samples belonging to class #1 are indicated by "+" and those belonging to class #2, "∘." Each class has one hundred samples. Here, $y_i \in \{-1, 1\}$. Figures $3 \sim 5$ show the estimated separating hyperplanes[1] using Gaussian kernels with $\sigma^2 = 0.0003,\ 0.02,\ 1000.0$, respectively. Here, the parameter $\sigma^2 = 0.02$ for a Gaussian kernel is chosen so that the misclassification ratio in leave-one-out cross-validation is almost the smallest. The parameter $\sigma^2 = 0.0003$ (or 1000.0) is used as too small (or large) one which is an example far from $\sigma^2 = 0.02$. Figs. 6 and 7 show the estimated separating hyperplanes based on the proposed integrated kernel using $\sigma_0 = 2.67$ (the maximum eigenvalue of the covariance matrix of the input vectors) as our suggestion and $\sigma_0 = 2.67 \times 10^4$ as a possibly large value, respectively. Note that the range of display is limited to $[-8, 8] \times [-8, 8]$. Table 1 shows the misclassification ratio in leave-one-out cross-validation of each condition.

It is confirmed that a Gaussian kernel with an extremely large (or small) variance underfits (or overfits) the given data, while the proposed kernel yields an appropriate separating hyperplane. Further, the proposed kernel with a larger $\sigma_0 (= 2.67 \times 10^4)$ causes neither overfitting nor underfitting.

5.2 Example of Function Estimation

Next, we consider a function estimation problem. In Fig. 8, the solid line denotes the unknown target function ($\mathrm{sinc}(x)$) and "×" denotes the sample points (training data set). Figures $9 \sim 11$ show the estimated functions using Gaussian kernels with $\sigma^2 = 0.001,\ 10.0,\ 5000.0$, respectively. Here, the parameter $\sigma^2 = 10.0$ for a Gaussian kernel is chosen so as to attain almost the minimum squared-error between the unknown target function and the estimated one. The parameter $\sigma^2 = 0.001$ (or 5000.0) is used as too small (or large) one which is an example far from $\sigma^2 = 10.0$. Figures 12 and 13 show the estimated functions based on the proposed kernel using $\sigma_0 = 5.99$ (the variance of the input values) as our suggestion and $\sigma_0 = 5.99 \times 10^4$ as a possibly large value, respectively. Table 2 shows the squared-error between the unknown target function and the estimated one in each condition.

Similar to the pattern recognition problem, a Gaussian kernel with an extremely large (or small) variance does not represent the target function, while the proposed kernel can; further, adopting a larger σ_0 for the proposed kernel causes neither overfitting nor underfitting.

[1] The boundary of the two regions is given by $\{\boldsymbol{x} | \hat{f}(\boldsymbol{x}) = 0\}$.

5.3 Remarks

It is often considered that complex models tend to overfit a training data set. However, this is not always true. The above numerical examples illustrate the counterexamples. Indeed, the RKHS corresponding to the integrated kernel is larger than those corresponding to the integrand and the complexity of the larger RKHS is higher than that of the smaller RKHS on the basis of the Rademacher averages (see [2] for instance).

In this paper, we assumed that what to be estimated is a function, which means that our framework is only applicable to the classification problem in which the misclassification is not allowed for the given training data set. Moreover, our discussion is only on the kernel (model) selection, and does not reach to the learning machine construction. Thus, in order to apply our kernel to real-world problems, we have to resolve these problems theoretically.

6 Conclusion

In this study, on the basis of the framework of projection learning, we have proposed a new method of constructing a kernel that incorporates the parameter integration of parameterized kernels. The RKHS that corresponds to the proposed integrated kernel includes almost all the RKHSs that correspond to the integrand with the parameters belonging to the intervals of integration. The kernel used in this study does not require an optimization of parameter value, although a parameterized kernel often requires this optimization. The validity of the proposed kernel is confirmed by some artificial numerical examples.

Acknowledgments

This work was partially supported by Grant-in-Aid No.18700001 for Young Scientists (B) from the Ministry of Education, Culture, Sports and Technology of Japan. The authors would like to thank Dr. Sugiyama and Dr. Takigawa for their valuable comments. The authors would also like to thank Dr. Kitamura for his useful comments regarding the style of the revised manuscript.

References

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

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

[3] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, Cambridge, 2000.

[4] N. Aronszajn, Theory of Reproducing Kernels, Transactions of the American Mathematical Society 68 (3) (1950) 337–404.

[5] H. Ogawa, Neural Networks and Generalization Ability, IEICE Technical Report NC95-8 (1995) 57–64.

[6] M. Sugiyama, H. Ogawa, Incremental Projection Learning for Optimal Generalization, Neural Networks 14 (1) (2001) 53–66.

[7] E. Oja, H. Ogawa, Parametric Projection Filter for Image and Signal Restoration, IEEE Transactions on Acoustics, Speech and Signal Processing ASSP–34 (6) (1986) 1643–1653.

[8] H. Imai, A. Tanaka, M. Miyakoshi, The family of parametric projection filters and its properties for perturbation, The IEICE Transactions on Information and Systems E80–D (8) (1997) 788–794.

[9] M. Sugiyama, H. Ogawa, Incremental construction of projection generalizing neural networks, IEICE Transactions on Information and Systems E85-D (9) (2002) 1433–1442.

[10] V. N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1999.

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

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

[13] M. Sugiyama, H. Ogawa, Subspace Information Criterion for Model Selection, Neural Computation 13 (8) (2001) 1863–1889.

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

Akira Tanaka

received his B.E. degree in 1994; M.E. degree, 1996; and D.E. degree, 2000 from Hokkaido University, Japan. He joined the Graduate School of Information Science and Technology, Hokkaido University. His research interests include image processing, acoustic signal processing, and learning theory.

Hideyuki Imai

received his D.E. degree in 1999 from Hokkaido University, Japan. He joined the Graduate School of Information Science and Technology, Hokkaido University. His research interests include statistical inferences.

Mineichi Kudo

received his B.E. degree in 1983; M.E. degree, 1985; and D.E. degree, 1988 from Hokkaido University, Japan. He joined the Graduate School of Information Science and Technology, Hokkaido University. His current research interests include the design of pattern classifiers, machine learning, data analysis/representation, and image processing.

Masaaki Miyakoshi

received his D.E. degree in 1985 from Hokkaido University, Japan. He joined the Graduate School of Information Science and Technology, Hokkaido University. His research interests include fuzzy theory.

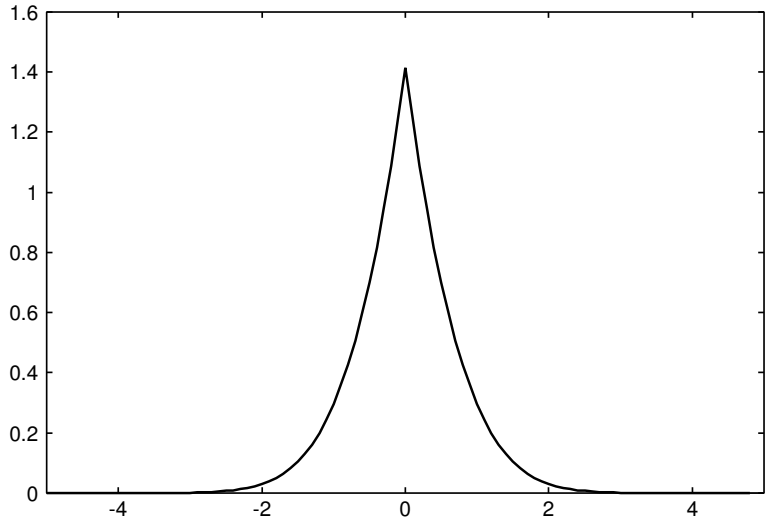

Figure 1: The graph of the integrated kernel $K_{[0,\sqrt{2}\sigma_0]}$ with $\sigma_0^2 = 1$. The horizontal axis denotes $||\boldsymbol{x} - \boldsymbol{y}||$.

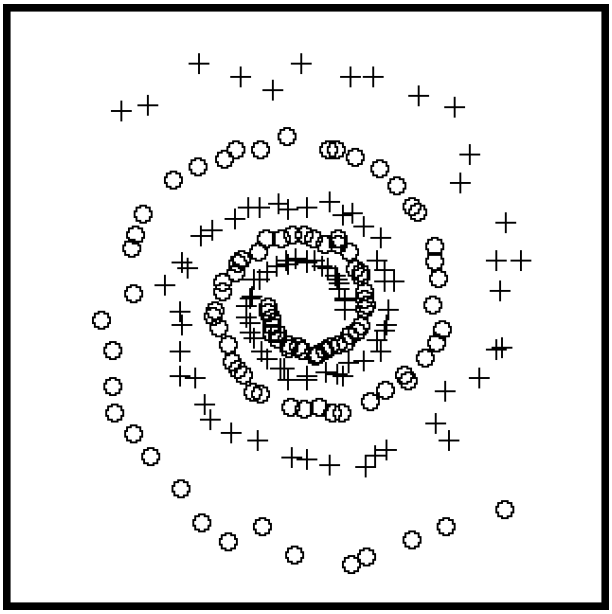

Figure 2: The scatter diagram of the training data set SPIRAL.

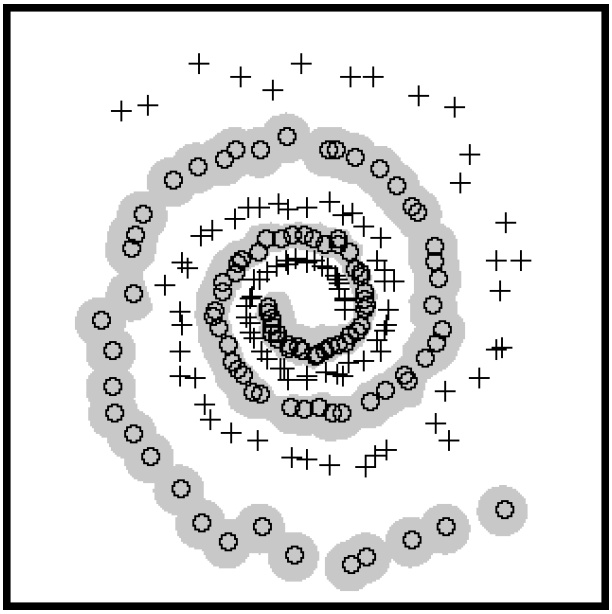

Figure 3: The result based on a Gaussian kernel with $\sigma^2 = 0.0003$ for SPIRAL.

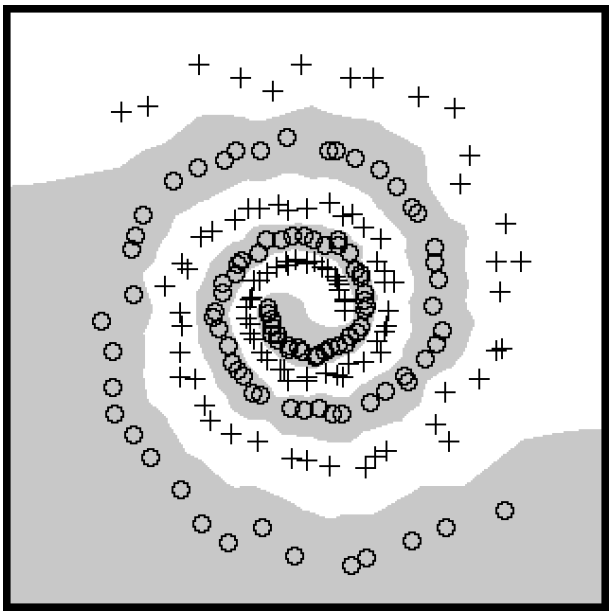

Figure 4: The result based on a Gaussian kernel with $\sigma^2 = 0.02$ for SPIRAL.

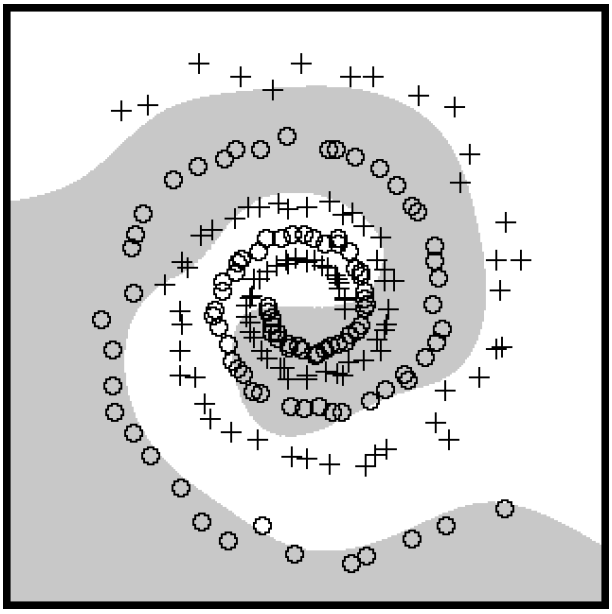

Figure 5: The result based on a Gaussian kernel with $\sigma^2 = 1000$ for SPIRAL.

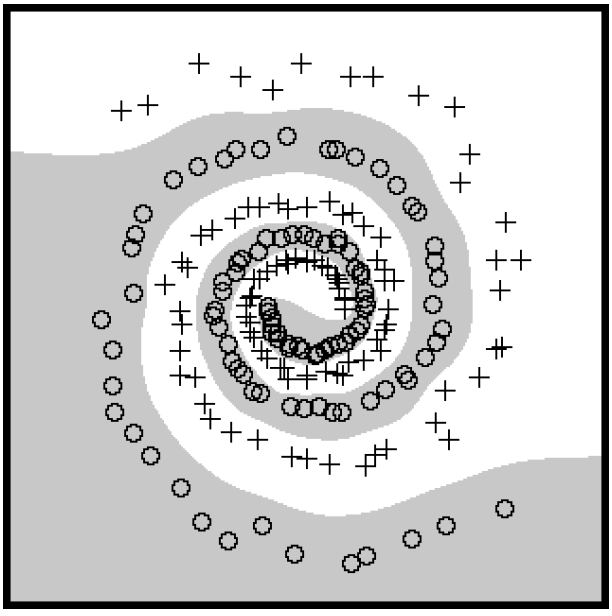

Figure 6: The result based on the proposed integrated kernel for SPIRAL with $\sigma_0 = 2.67$.

Table 1: Misclassification ratio in leave-one-out cross-validation.

Condition	Misclassification ratio
Gaussian kernel with $\sigma^2 = 0.0003$	9.5%
Gaussian kernel with $\sigma^2 = 0.02$	0.5%
Gaussian kernel with $\sigma^2 = 1000$	35.0%
Integrated kernel with $\sigma_0 = 2.67$	0.0%
Integrated kernel with $\sigma_0 = 2.67 \times 10^4$	1.5%

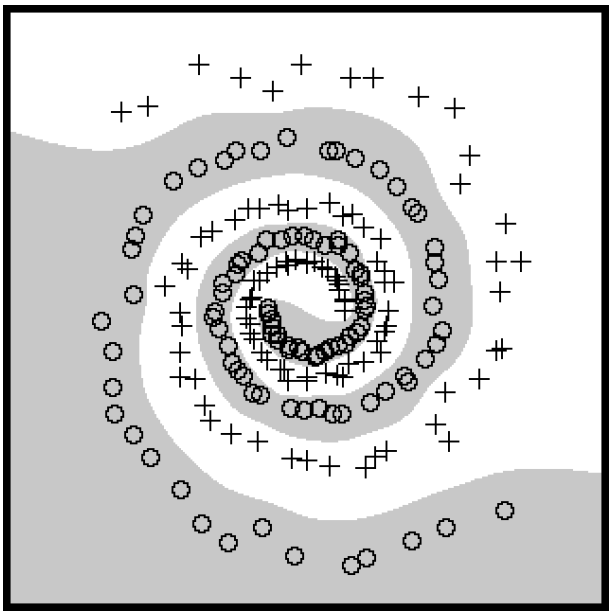

Figure 7: The result based on the proposed integrated kernel for SPIRAL with $\sigma_0 = 2.67 \times 10^4$.

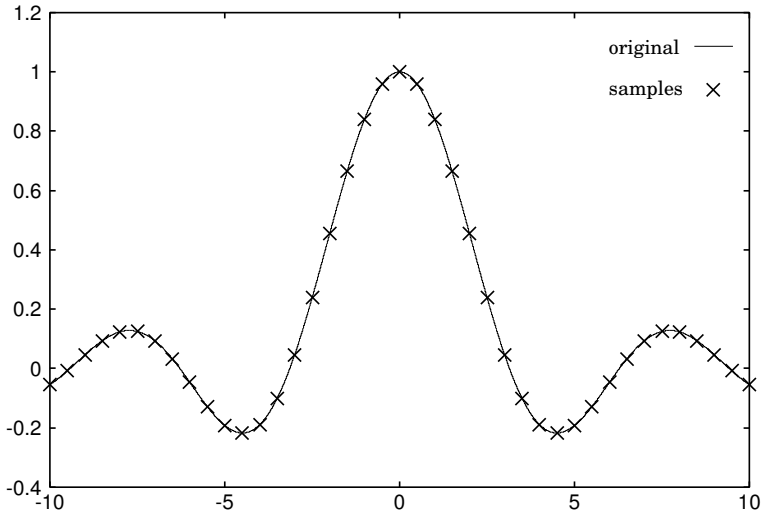

Figure 8: Target function and training samples.

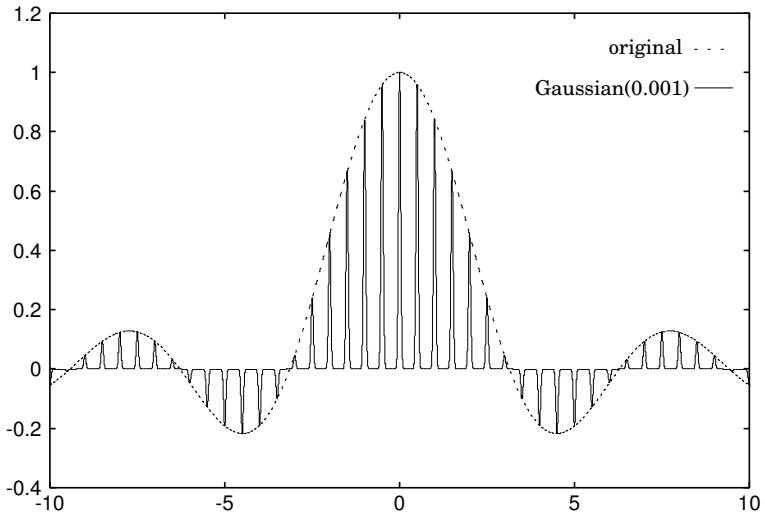

Figure 9: The result based on a Gaussian kernel with $\sigma^2 = 0.001$.

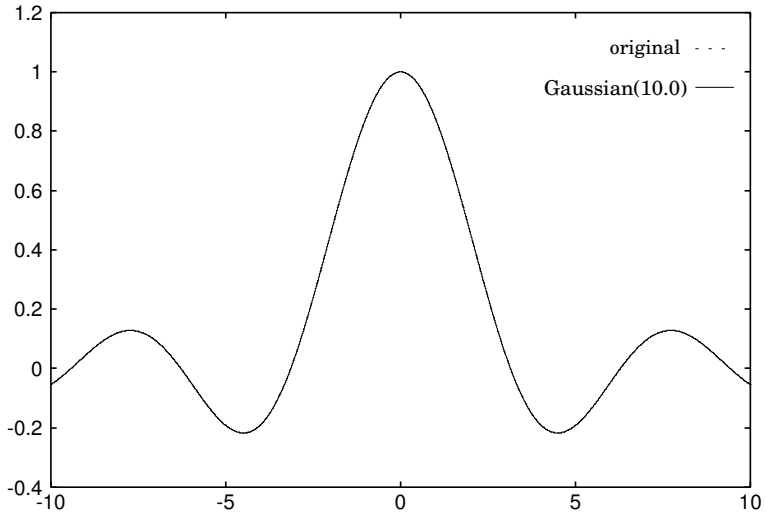

Figure 10: The result based on a Gaussian kernel with $\sigma^2 = 10.0$.

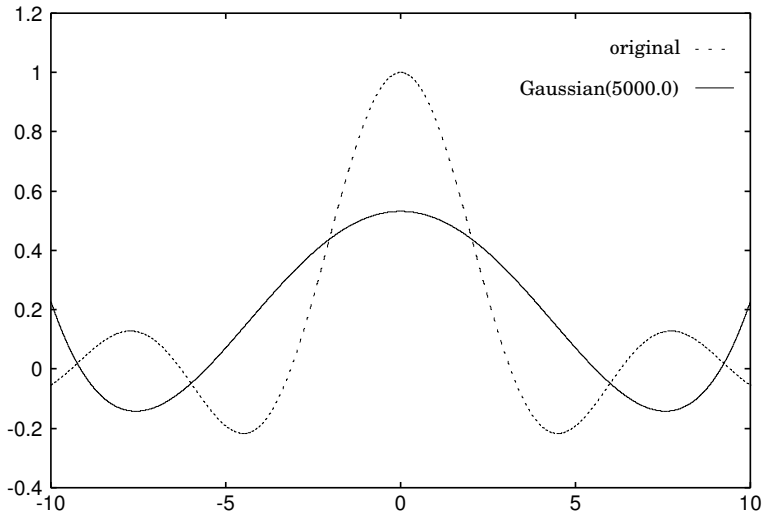

Figure 11: The result based on a Gaussian kernel with $\sigma^2 = 5000.0$.

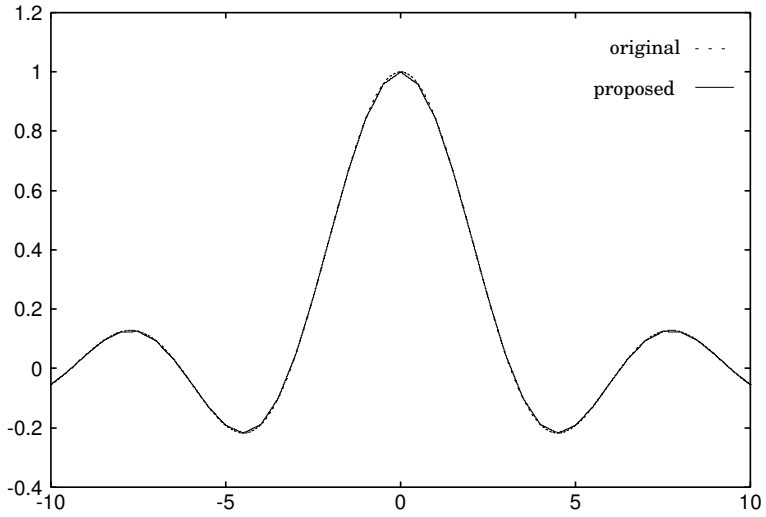

Figure 12: The result based on the proposed kernel with $\sigma_0 = 5.99$.

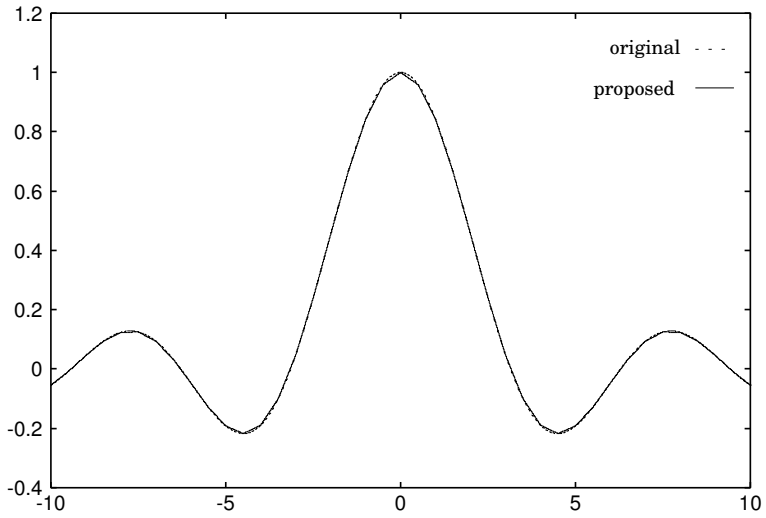

Figure 13: The result based on the proposed kernel with $\sigma_0 = 5.99 \times 10^4$.

Table 2: The squared-error between the unknown true function and the estimated one.

Condition	Squared-error
Gaussian kernel with $\sigma^2 = 0.001$	2.41
Gaussian kernel with $\sigma^2 = 10.0$	5.39×10^{-13}
Gaussian kernel with $\sigma^2 = 5000$	1.34
Integrated kernel with $\sigma_0 = 5.99$	2.74×10^{-4}
Integrated kernel with $\sigma_0 = 5.99 \times 10^4$	2.70×10^{-4}