Learning the Value of a Function by Using Hypercircle Inequality for Data Error

B. Novaprateep, K. Khompurngson, D. Poltem

***Abstract*—In this paper, we briefly review Hypercircle inequality for data error ($Hide$) measured with square loss. We provide it in the case that the unit ball B is replaced by δB where δ is any positive number. Moreover, we also discuss some important facts of $Hide$ for practical computation and study the problem in learning the value of a function in reproducing kernel Hilbert space (RKHS) by using the available material from $Hide$ with different values of δ. We compare our numerical experiment to the method of regularization, which is the standard method for learning problem.

***Keywords*—Hypercircle inequality, Reproducing Kernel Hilbert space, Regularization, Convex Optimization and Noise Data.

I. Introduction

In this paper, we briefly review Hypercircle inequality for data error ($Hide$) measured with square loss [2], [8]. We provide it in the case that the unit ball B is replaced by δB where δ is any positive number. Moreover, we also discuss some important facts of $Hide$ for practical computation and study the problem in learning the value of a function in reproducing kernel Hilbert space (RKHS) by using the available material from $Hide$ with different values of δ. We compare our numerical experiment to the method of regularization, which is the standard method for learning problem.

Given an input-output examples $\{(t_j, d_j) : j \in \mathbb{N}_n\} \subseteq \mathcal{T} \times \mathbb{R}$ where $\mathcal{T}$ is an input set, and we use the notation $\mathbb{N}_n = \{1, 2, ..., n\}$. The basic idea in learning problem is to determine a functional representation from data. Let the hypothesis space H be a reproducing kernel Hilbert space (RKHS) of real value function on a set $\mathcal{T}$. That is, $f : \mathcal{T} \to \mathbb{R}$ is the functional in the hypothesis space H, and d_j is a data representation of $f(t_j)$ for all $j \in \mathbb{N}_n$. The real function K of t and s in $\mathcal{T}$ is called a *reproducing kernel* of H if the following property is satisfied for every $t \in \mathcal{T}$ and every $f \in H$

$$f(t) = \langle f, K_t \rangle$$

where K_t is the function of $s \in \mathcal{T}$ and $K_t(s) = K(t, s)$. Therefore, for any $s, t \in \mathcal{T}$

$$K(s, t) = \langle K_s, K_t \rangle$$

B. Novaprateep is with the Department of Mathematics, Faculty of Science, Mahidol University, and Centre of Excellence in Mathematics, PERDO, CHE, THAILAND e-mail: scbnv@mahidol.ac.th (corresponding author).

K. Khompurngson is with the Faculty of Science, University of Phayao, and Centre of Excellence in Mathematics, PERDO, CHE, THAILAND e-mail: kannikak@nu.ac.th.

D. Poltem is with the School of Interdisciplinary, Kanchanaburi Campus, Mahidol University, and Centre of Excellence in Mathematics, PERDO, CHE, THAILAND e-mail: kadpt@mahidol.ac.th.

By the obove relations, for $t \in \mathcal{T}$ we also obatin that

$$||K_t||^2 = \langle K_t, K_t \rangle = K(t, t)$$

The Aronszajn and Moore theorem [1] states that a function $K : \mathcal{T} \times \mathcal{T} \longrightarrow \mathbb{R}$ is a reproducing kernel for some RKHS if and only if for any inputs $T = \{t_j : j \in \mathbb{N}_n\} \subseteq \mathcal{T}$ the $n \times n$ matrix $G = (K(t_i, t_j) : i, j \in \mathbb{N}_n)$ is a positive semi-definite. That is, for any $T = \{t_j : j \in \mathbb{N}_n\} \subseteq \mathcal{T}$ and any $a = (a_1, ..., a_n) \in \mathbb{R}^n$

$$\sum_{i,j \in \mathbb{N}_n} a_j a_i K(t_j, t_i) \geq 0.$$

Moreover, for any kernel K there is a unique RKHS with K as its reproducing kernel. These important and useful facts allow us to specify a hypothesis space by choosing K.

Alternatively, we consider here the following point of view. Given $t_0 \in \mathcal{T}$, we want to estimate $f(t_0)$ knowing that $||f||_K \leq \delta$ and $|d - Qf|_2^2 \leq \varepsilon$ where $Qf := (f(t_j) = \langle f, K_{t_j} \rangle : j \in \mathbb{N}_n)$ and $| \cdot |_2$ is a Euclidean norm on $\mathbb{R}^n$. The standard method for learning $f(t_0)$ is the method of regularization, [5], [6]. Given $\rho > 0$, we choose the function which minimize from the R_ρ functional defined for $f \in H$ as

$$R_\rho(f) := |d - Qf|_2^2 + \rho ||f||_K^2. \tag{1}$$

According to Representer Theorem [3], [10], [11], [12], the function which minimizes (1) has the form

$$f_\rho(t) = \sum_{j \in \mathbb{N}_n} c(\rho)_j K(t_j, t), \quad t \in \mathcal{T} \tag{2}$$

for some real vector $c(\rho) = (G + \rho I)^{-1} d$ where I is $n \times n$ identity matrix and $G = (K(t_i, t_j) : i, j \in \mathbb{N}_n)$. We choose $f_\rho(t_0)$ as our estimator. Consequently, we let $\varepsilon_\rho^2 := |d - Qf_\rho|_2^2$ and $\delta_\rho^2 := ||f_\rho||_K^2$. Next, we want to compare this method to the midpoint algorithm. We then define the interval of uncertainty

$$I(t_0, \varepsilon_\rho, \delta_\rho) = \{f(t_0) : |d - Qf|_2 \leq \varepsilon_\rho, ||f||_K \leq \delta_\rho\}.$$

Hence, the best choice for this number is a function whose values at t_0 is midpoint of the interval $I(t_0, \varepsilon_\rho, \delta_\rho)$. To compare both methods, regularization method and midpoint algorithm, we need to show that the regularization estimator $f_\rho(t_0)$ can be view as an element in the interval $I(t_0, \varepsilon_\rho, \delta_\rho)$. According to our previous work, we found that there is only one element, namely, $f_\rho(t_0)$ in $I(t_0, \varepsilon_\rho, \delta_\rho)$. Therefore, our strategy to compare the regularization and midpoint estimator must consider a bigger value of ε_ρ and δ_ρ. For this reason, we shall discuss and continue report some results from numerical experiments of learning the value of function in RKHS by midpoint algorithm with different values of δ

The organization of this paper is as follows. In Section II, we briefly review Hypercircle inequality for data error measured with square loss and discuss what we need for Section III. Moreover, we will give some important fact of $Hide$ for practical computation. In section III contains some results of numerical experiments of learing the value of a function in RKHS. Specificalily, we consider the data error measured with square loss.

II. Hypercircle inequality for data error

In this section we begin with Hilbert space H over the real number with inner product $\langle \cdot, \cdot \rangle$. We choose a finite set of linearly independent elements $\mathcal{X} = \{x_j : j \in \mathbb{N}_n\}$ in H. We shall denote by M the $n-$dimensional linear subspace of H spanned by the vectors in $\mathcal{X}$. Let $Q : H \to \mathbb{R}^n$ be a linear operator H onto $\mathbb{R}^n$, which is defined for any $x \in H$ as

$$Qx = (\langle x, x_j \rangle : j \in \mathbb{N}_n). \quad (3)$$

Alternatively, the adjoint map $Q^T : \mathbb{R}^n \longrightarrow H$ is given at $a = (a_j : j \in \mathbb{N}_n) \in \mathbb{R}^n$ as

$$Q^T a = \sum_{j \in \mathbb{N}_n} a_j x_j \quad (4)$$

and the Gram matrix of the vectors in $\mathcal{X}$ is

$$G = QQ^T = \begin{bmatrix} \langle x_1, x_1 \rangle & \langle x_1, x_2 \rangle & ... & \langle x_1, x_n \rangle \\ \langle x_2, x_1 \rangle & \langle x_2, x_2 \rangle & ... & \langle x_2, x_n \rangle \\ \vdots & \vdots & \vdots & \vdots \\ \langle x_n, x_1 \rangle & \langle x_n, x_2 \rangle & ... & \langle x_n, x_n \rangle \end{bmatrix}$$

which is a symmetric and positive definite. To prove this, we let $0 \neq a \in \mathbb{R}^n$ and we have that

$$\begin{aligned} a^T G a &= a^T Q Q^T a \\ &= (a, QQ^T a) \\ &= \langle Q^T a, Q^T a \rangle = ||Q^T a||^2 > 0. \end{aligned}$$

Therefore, G is a positive definite matrix. Next, let us review basic facts about Hi, [2], [9], and discuss what we need for $Hide$.

Theorem 1: Let $\mathcal{X} = \{x_j : j \in \mathbb{N}_n\}$ be set of linearly independent elements in H. Then given any $d \in \mathbb{R}^n$ we can find an element $x(d) \in \mathcal{R}(Q^T)$ such that

$$Qx(d) = d.$$

Moreover, we have that $x(d) = Q^T G^{-1} d$.

Proof. We refer the reader to [2] for the proof.

Moreover, from this formula we obtain the useful equation

$$\begin{aligned} \min\{||x|| : x \in H, Qx = d\} &= ||x(d)|| \\ &= \sqrt{(d, G^{-1}d)}. \quad (5) \end{aligned}$$

Definition 2: Let H be the Hilbert space over the real number and $\mathcal{X} = \{x_j : j \in \mathbb{N}_n\}$ be a finite set of linearly independent elements in H. Let d be a given vector in $\mathbb{R}^n$, and δ be a positive number. The hypercircle, $\mathcal{H}(d, \delta)$ is a subset of H, which is defined by

$$\mathcal{H}(d, \delta) = \{x : x \in \delta B, Qx = d\}.$$

where $B := \{x : x \in H, ||x|| \leq 1\}$ is the $unit\ ball$ in H. We point out that the hypercircle $\mathcal{H}(d, \delta)$ is a convex subset of H which is sequentially compact in the $weak$ topology on H.

Theorem 3: If $d \in \mathbb{R}^n$ then $\mathcal{H}(d, \delta) \neq \emptyset$ if and only if

$$||x(d)|| = \sqrt{(d, G^{-1}d)} \leq \delta.$$

Moreover, in this case $x(d) \in \mathcal{H}(d, \delta)$.

Proof. Suppose that $\mathcal{H}(d, \delta) \neq \emptyset$, contains an element x. Let $e = x - x(d)$ and we observe that

$$\begin{aligned} ||x - x(d)||^2 &= ||x||^2 - 2\langle x, x(d) \rangle + ||x(d)||^2 \\ &= ||x||^2 - 2\langle x - e, x(d) \rangle + ||x(d)||^2 \\ &= ||x||^2 - ||x(d)||^2. \end{aligned}$$

Therefore, $||x(d)|| \leq \delta$ and $||x(d)|| = \sqrt{(d, G^{-1}d)} \leq \delta$. Conversely, suppose that $||x(d)|| = \sqrt{(d, G^{-1}d)} \leq \delta$. By Theorem 1, we obtain that $x(d) \in \mathcal{H}(d, \delta)$. □

Theorem 4: If $H \neq M$ then $\mathcal{H}(d, \delta)$ consists of exactly one point if and only if $||x(d)|| = \delta$.

Proof. Suppose that $\mathcal{H}(d, \delta)$ consists of exactly one point x. Choose any $w \in H$ such that $Qw = 0$. Therefore, for any $t \in \mathbb{R} \setminus \{0\}$ we conclude that $||x + tw|| > \delta$. This inequality implies that $||x|| = \delta$. We see that

$$\begin{aligned} ||x + tw||^2 &= ||x||^2 + 2t\langle x, w \rangle + t^2||w||^2 \\ &= ||x||^2 + t(2\langle x, w \rangle + t||w||^2). \end{aligned}$$

If we choose $t = -\frac{2\langle x, w \rangle}{||w||^2}$, then $||x|| > \delta$ which is a contradiction. Thus, we have that $\langle x, w \rangle = 0$. From the second conclusion we learn that $x \in M$. Consequently, we see that $x = x(d)$, so that $||x(d)|| = \delta$. Conversely, suppose that $||x(d)|| = \delta$ and $x \in \mathcal{H}(d, \delta)$. Then the vector $e := x - x(d)$ has the property that $Qe = 0$. As a result we conclude that $\langle e, x(d) \rangle = 0$. Consequently, we get that $\langle x, x(d) \rangle = \langle x - e, x(d) \rangle = ||x(d)||^2$ and so $||e||^2 = ||x||^2 - ||x(d)||^2 \leq 0$. This confirms the fact that $x = x(d)$. □

We add one final remark before providing the material of Hypercircle inequality for data error. If $H = M$ then $\mathcal{H}(d, \delta)$ consists of at most one point, namely $x(d)$.

Now we ready to describe Hypercircle inequality for data error ($Hide$). We provided it in the case that the data error is measured with Euclidean norm. We refer the reader to the paper [8] for more information about the proof of $Hide$ measured with any norm on $\mathbb{R}^n$.

Definition 5: Let H be the Hilbert space over the real number and $\mathcal{X} = \{x_j : j \in \mathbb{N}_n\}$ be a finite set of linearly independent elements in H. Let $E = \{e : e \in \mathbb{R}^n, |e|_2 \leq \varepsilon\}$ where $|\cdot|_2 : \mathbb{R}^n \to \mathbb{R}_+$ is a Euclidean norm on $\mathbb{R}^n$ and ε is some prescribed positive number. Let d be a given vector in $\mathbb{R}^n$, and δ be a positive number. The $hyperellipse$, $\mathcal{H}_2(d|E(\delta))$ is a subset of H, which is defined by

$$\mathcal{H}_2(d|E(\delta)) = \{x : x \in \delta B, Qx - d \in E\}.$$

Moreover, we observe that

$$\mathcal{H}_2(d|E(\delta)) = \bigcup_{e \in E} \mathcal{H}(d+e, \delta). \tag{6}$$

Since the vector $d \in \mathbb{R}^n$ and ε are prescribed, we start out by giving the formula for choosing the value of δ such that $\mathcal{H}_2(d|E(\delta)) \neq \emptyset$. We then begin with the following lemma.

Lemma 6: For any $d \in \mathbb{R}^n$,

$$\min\{||x|| : |d - Qx|_2 \leq \varepsilon\} = \min\{\sqrt{(d+\varepsilon c, G^{-1}(d+\varepsilon c))} : c \in \mathbb{R}^n, |c|_2 \leq 1\}.$$

Proof. From the equation (5) and (6), we obtain that

$$\min\{||x|| : |d - Qx|_2 \leq \varepsilon\} = \min\{\sqrt{(d+\varepsilon c, G^{-1}(d+\varepsilon c))} : c \in \mathbb{R}^n, |c|_2 \leq 1\}.$$

□

Lemma 7: $\mathcal{H}_2(d|E(\delta)) \neq \emptyset$ if and only if

$$\min_{|c|_2 \leq 1} (d+\varepsilon c, G^{-1}(d+\varepsilon c)) \leq \delta^2. \tag{7}$$

Proof. Let $x \in \mathcal{H}_2(d|E(\delta))$. Then there is $e \in E$ such that $x \in \mathcal{H}(d+e, \delta)$ and $x = x(d+e) = Q^T G^{-1}(d+e)$. Thus, we see that

$$\begin{aligned} ||x(d+e)||^2 &= \langle x(d+e), Q^T G^{-1}(d+e)\rangle \\ &= (d+e, G^{-1}(d+e)) \leq \delta^2. \end{aligned} \tag{8}$$

Hence, $\min\{(d+\varepsilon e, G^{-1}(d+\varepsilon e)) : |e|_2 \leq 1\} \leq \delta^2$. Conversely, (7) and (8) certainly implies $\mathcal{H}_2(d|E(\delta)) \neq \emptyset$. □

Next, we will give the formula for checking when $\mathcal{H}_2(d|E(\delta)) \neq \emptyset$. We then begin with the following definition.

Definition 8: Let A be an $n \times n$ symmetric matrix and $d \in \mathbb{R}^n$. The *spectrum* of the pair (A, d) is defined to be the set of all real numbers Λ for which there exists an $x \in \mathbb{R}^n$ with Euclidean norm one such that

$$A(x - d) = \Lambda x. \tag{9}$$

$0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$ be eigenvalue of G^{-1}, $\{u^j : j \in \mathbb{N}_n\}$ be a corresponding orthonormal set of eigenvector , write the vector d in the form $d = \sum_{j \in \mathbb{N}_n} \gamma_j u^j$ for some constants $\gamma_j \in \mathbb{R}$ and define the subset I of $\mathbb{N}_n$ by $I := \{j : \lambda_j d_j = 0\}$.

Lemma 9: If Λ is the least value in the spectrum of the pair $(\varepsilon^2 G^{-1}, \frac{d}{\varepsilon})$ then

$$\min_{|c|_2 \leq 1} (d+\varepsilon c, G^{-1}(d+\varepsilon c)) = \Lambda + \Lambda \sum_{j \notin I} \frac{\lambda_j |\gamma_j|^2}{\Lambda - \varepsilon^2 \lambda_j}.$$

Proof. We refer the reader to the paper [4] for the proof.

Alternatively, we can conclude that $\mathcal{H}_2(d|E(\delta)) \neq \emptyset$ if $\delta \geq \Lambda + \Lambda \sum_{j \notin I} \frac{\lambda_j |\gamma_j|^2}{\Lambda - \varepsilon^2 \lambda_j}$.

As we want to find the best estimator to optimally estimate one feature of $x \in \mathcal{H}_2(d|E(\delta))$ when we define a feature of $x \in H$ as the value a linear functional F_{x_0} defined at x as $F_{x_0}(x) = \langle x, x_0\rangle$. We then define the uncertainty set by $I(x_0, d|E(\delta)) = \{F_{x_0}(x) : x \in \mathcal{H}_2(d|E(\delta))\}$. Since $\mathcal{H}_2(d|E(\delta))$ is convex subset of H which is sequentially compact in the weak topology on H, we obtain that the uncertainty set is a closed and bounded interval in $\mathbb{R}$. Consequently, we have

$$I(x_0, d|E(\delta)) = [m_-(x_0, d|E(\delta)), m_+(x_0, d|E(\delta))]$$

where

$$m_+(x_0, d|E(\delta)) = \max\{F_{x_0}(x) : x \in \mathcal{H}_2(d|E(\delta))\}$$

and

$$m_-(x_0, d|E(\delta)) = \min\{F_{x_0}(x) : x \in \mathcal{H}_2(d|E(\delta))\}.$$

Hence, the best estimator is the midpoint of this interval. According to our pervious work, we have the important theorem which obtain the midpoint of this interval.

Theorem 10: If $\mathcal{H}_2(d|E(\delta)) \neq \emptyset$ then there is an element $e_0 \in E$ such that $x(d+e_0)$ is the best estimator for the feature F_{x_0}.

Proof. We refer the reader to the paper [8] for the proof.

Remark that $x(d+e_0) = Q^T G^{-1}(d+e_0) \in M$ for some $e_0 \in E$. That is, we can see that the best estimator has the form of Representer Theorem .

Next, let us point out some important facts that

$$\begin{aligned} m_-(x_0, d|E(\delta)) &= \min\{F_{x_0}(x) : x \in \mathcal{H}_2(d|E(\delta))\} \\ &= \min\{-F_{x_0}(x) : -x \in \mathcal{H}_2(-d|E(\delta))\} \\ &= -\max\{F_{x_0}(x) : x \in \mathcal{H}_2(-d|E(\delta))\} \\ &= -m_+(x_0, -d|E(\delta)). \end{aligned}$$

To obtain the midpoint, we then only need to evaluate the two numbers Let $m_+(x_0, \pm d \mid E((\delta)))$ and then compute the midpoint

$$\begin{aligned} m(x_0, d|E(\delta)) &= \frac{1}{2}\Big(m_+(x_0, d \mid E(\delta)) - \\ &\quad m_+(x_0, -d \mid E(\delta))\Big). \end{aligned} \tag{10}$$

Next, we will describe a duality formula for the right hand side of the interval of uncertainty. We start out by introducing the convex function $V_\delta : \mathbb{R}^n \to \mathbb{R}$ defined for $c \in \mathbb{R}^n$

$$V_\delta(c) := \delta||x_0 - Q^T c|| + \varepsilon|c|_2 + (c, d). \qquad (11)$$

Next, let us describe the sufficient condition on $\mathcal{H}_2(d|E(\delta))$ such that V_δ has a minimum.

Theorem 11: If $\mathcal{H}_2(d|E(\delta))$ contains more than one point then the function V_δ has a minimum and every minimizing sequence is bounded.

Proof. We refer the reader to the paper [8] for the proof.

In our theorem below, we shall provide the conditions such that the function V_δ achieves its minimums at 0.

Theorem 12: If $x_0 \neq 0$ then the following statement are equivalent:
(i) $0 = \arg\ \min\{V_\delta(c) : c \in \mathbb{R}^n\}$.
(ii) $\dfrac{\delta x_0}{||x_0||} \in \mathcal{H}_2(d|E(\delta))$.
(iii) $\dfrac{\delta x_0}{||x_0||} = \arg\ \max\{\langle x, x_0\rangle : x \in \mathcal{H}_2(d|E(\delta))\}$.

Proof. The equation (i) holds if and only if $||x_0|| = V_\delta(0) \leq V_\delta(c)$ for all $c \in \mathbb{R}^n$. Since the function V_δ is a convex this inequality holds if and only if for all $c \in \mathbb{R}^n$

$$-\varepsilon|c|_* - (c, d) \leq \inf\{\frac{\delta||x_0 - \lambda Q^T c|| - \delta||x_0||}{\lambda} : \lambda > 0\}$$

which means for all $c \in \mathbb{R}^n$ that

$$-\varepsilon|c|_* - (c, d) \leq -(\delta\frac{Qx_0}{||x_0||}, c).$$

That is , we have that

$$(\delta\frac{Qx_0}{||x_0||} - d, c) \leq \varepsilon|c|_*.$$

which is equivalent to saying that $|\delta\frac{Qx_0}{||x_0||} - d|_2 \leq \varepsilon$. This establishes that $\dfrac{\delta x_0}{||x_0||} \in \mathcal{H}_2(d|E(\delta))$ and the equivalence of (i) and (ii). Now, we establish that (ii) and (iii) are also the same. Certainly, by the definition of "$\arg\ \max$" (iii) implies (ii). Conversely, if (ii) is true then we have that

$$||\delta x_0|| = \langle x_0, \delta\frac{x_0}{||x_0||}\rangle \leq m_+(x_0, d|E) \leq \delta||x_0||$$

where the last inequality follows from the fact that $\mathcal{H}_2(d|E(\delta)) \subseteq \delta B$. Hence (iii) holds. □

Now we ready to state the sufficient condition on $\mathcal{H}_2(d|E(\delta))$ which ensure that the minimum $0 \neq c^* \in \mathbb{R}^n$ is unique solution of the function V_δ.

Theorem 13: If $\mathcal{H}_2(d|E(\delta))$ contains more than one point, $x_0 \notin M$, and $\frac{\delta x_0}{||x_0||} \notin \mathcal{H}_2(d|E(\delta))$ then

$$m_+(x_0, d|E(\delta)) = \min_{c\in\mathbb{R}^n} V_\delta(c).$$

Moreover, the minimum $c^* \in \mathbb{R}^n$ is the unique solution of the nonlinear equation

$$-\delta Q(\frac{x_0 - Q^T c^*}{||x_0 - Q^T c^*||}) + \varepsilon\frac{c^*}{|c^*|_2} + d = 0 \qquad (12)$$

and

$$x_+(d) := \delta\frac{x_0 - Q^T c^*}{||x_0 - Q^T c^*||} \qquad (13)$$

satisfies

$$x_+(d) = \arg\ \max\{F_{x_0}(x) : x \in \mathcal{H}_2(d|E(\delta))\}. \qquad (14)$$

Proof. We refer the reader to the paper [8] for the proof.

According to our previous work [8], we found that if $\mathcal{H}_2(d|E(\delta))$ contains only one element then the convex function V_δ above does not assume its minimum.

Example. Let $H = \mathbb{R}^2$, $x_1 = (1, 0)$, $d = 1 + \varepsilon$, $x_0 = (1, 1)$, $\delta = 1$ and $\varepsilon > 0$ then $\mathcal{H}_2(d|E(\delta)) = \{(1, 0)\}$ and $m_+(x_0, d) = 1$. On the other hand, we see that for all $c \in \mathbb{R}^n$

$$V_\delta(c) = \sqrt{1 + (1 - c)^2} + (1 + \varepsilon)c + \varepsilon|c|.$$

We observe that the map $c \to \sqrt{1 + (1 - c)^2} + (1 + \varepsilon)c + \varepsilon|c|$ is increasing. Next, we claim that $\inf\{\sqrt{1 + (1 - c)^2} + (1 + \varepsilon)c + \varepsilon|c| : c \in \mathbb{R}\} = 1$. This follows from the fact that

$$\lim_{c\to-\infty} \sqrt{1 + (1 - c)^2} + (1 + \varepsilon)c + \varepsilon|c| \quad = \quad 1.$$

Therefore, $\inf\{\sqrt{1 + (1 - c)^2} + (1 + \varepsilon)c + \varepsilon|c| : c \in \mathbb{R}\} = 1$ and the infimum is not achieved.

To this end, let us point out the condition which ensure that $\mathcal{H}_2(d|E(\delta))$ contains more than one point.

Theorem 14: . If $H \neq M$ and there exists $e \in E$ such that $x(d + e) \in \mathcal{H}_2(d|E(\delta))$ such that $||x(d + e)|| < \delta$ then there is an infinite number of vectors in $\mathcal{H}_2(d|E(\delta))$.

Proof. Since $H \neq M$, we choose any $w \in H$ such that $Qw = 0$ and $||w|| = 1$. Using our assumption, we have that there exists $e \in E$ such that $x(d + e) \in \mathcal{H}_2(d|E(\delta))$ and $||x(d + e)|| < \delta$. Moreover, we know that $x(d + e) \in M$. Next we define $y = x(d + e) + tw$ for some $t \in \mathbb{R}$. We observe that $Qy - d = Qx(d + e) - d \in E$. Next, we see that

$$\begin{aligned} ||y||^2 &= ||x(d + e)||^2 - 2\langle x(d), w\rangle + t^2||w||^2 \\ &= ||x(d + e)||^2 + t^2||w||^2 \\ &< ||x(d + e)||^2 + t. \end{aligned}$$

Since $||x(d + e)|| < \delta$, there is an infinite number of t such that $||y||^2 < ||x(d + e)||^2 + t < \delta^2$. Therefore, $y \in \mathcal{H}_2(d|E(\delta))$. Hence, there is an infinite number of vectors in $\mathcal{H}_2(d|E(\delta))$. □

Theorem 15: If $H \neq M$ then $\mathcal{H}_2(d|E(\delta))$ consists of exactly one point if and only if

$$\min\{(d + \varepsilon e, G^{-1}(d + \varepsilon e)) : e \in \mathbb{R}^n,\ |e|_2 \leq 1\} = \delta^2.$$

Proof. Suppose that $\mathcal{H}_2(d|E(\delta))$ consists of exactly the one point x. From Theorem 4, we obtain that $x = x(d+\hat{e})$ for some $\hat{e} \in E$ and $||x(d+\hat{e})|| = \delta$. Therefore, we obtain that

$$\min\{(d+\varepsilon e, G^{-1}(d+\varepsilon e)) : e \in \mathbb{R}^n, |e|_2 \leq 1\} = \delta^2.$$

Conversely, suppose that $\min\{(d+\varepsilon e, G^{-1}(d+\varepsilon e)) : e \in \mathbb{R}^n, |e|_2 \leq 1\} = \delta^2$. Since the map $e \to (d+\varepsilon e, G^{-1}(d+\varepsilon e))$ is strictly convex, there is a unique $\hat{e} \in \mathbb{R}^n$ with $|\hat{e}|_2 \leq 1$ such that $(d+\varepsilon\hat{e}, G^{-1}(d+\varepsilon\hat{e})) = \delta^2$ and $\mathcal{H}_2(d|E(\delta)) = \mathcal{H}(d+\hat{e}, \delta)$. Next, we claim that $\mathcal{H}(d+\hat{e}, \delta)$ consists of one point. Let $x \in \mathcal{H}(d+\hat{e}, \delta)$ and the vector $y := x - x(d+\hat{e})$ has the property that $Qy = 0$. As a result we conclude that $\langle y, x(d+\hat{e})\rangle = 0$. Consequently, we get that

$$\langle x, x(d+\hat{e})\rangle = \langle x-y, x(d+\hat{e})\rangle = ||x(d+\hat{e})||^2$$

By Theorem 4 again, we obtain that $||y||^2 = ||x||^2 - ||x(d+\hat{e})||^2 \leq 0$. This confirms the fact that $x = x(d+\hat{e})$. Hence,we can conclude that $\mathcal{H}_2(d|E(\delta))$ consists of exactly one point. □

To this end let us now consider the special case that $\{x_0\} \cup \mathcal{X}$ is an *orthonormal* set of vector. We found that the midpoint of the uncertainty interval is zero, see example [8]. That is, the right hand endpoint of the interval is an even function of $d \in \mathbb{R}^n$. That is, we have that

$$m_+(x_0, d|E(\delta)) = m_+(x_0, -d|E(\delta)).$$

III. Numerical Experiments

In this section, we shall continue to report some results from numerical experiments in learning the value of a function in RKHS by the midpoint algorithm with different values of δ. Let H be a reproducing kernel Hilbert space over real numbers (RKHS). Given any set of points $T = \{t_j : j \in \mathbb{N}_n\} \subseteq \mathcal{T}$ where $\mathcal{T}$ is an input set, the vector $\{x_j : j \in \mathbb{N}_n\}$ appearing in Section II is identified with the function $\{K_{t_j} : j \in \mathbb{N}_n\}$ where $K_{t_j}(t) = K(t_j, t), j \in \mathbb{N}_n, t \in \mathcal{T}$. The Gram matrix of the function $\{K_{t_j} : j \in \mathbb{N}_n\}$ is given as $G = (K(t_i, t_j))_{i,j\in\mathbb{N}_n}$.

Next, we choose the exact function $g \in H$ and then compute the vector $D_g := (g(t_j) : j \in \mathbb{N}_n)$. Then, we corrupt the data by additive noise. Thus, we define $d = D_g + e$. Indeed, our problem becomes as follows. Given $t_0 \in \mathcal{T}$, we want to estimate $f(t_0)$ knowing that $||f||_K \leq \delta$ and $|d - Qf|_2^2 \leq \varepsilon$ where $Qf := (f(t_j) = \langle f, K_{t_j}\rangle : j \in \mathbb{N}_n)$ and $|\cdot|_2$ is a Euclidean norm on $\mathbb{R}^n$. As we briefly described the regularization method in Section 1, we give $\rho > 0$ and we choose the function which minimizes this functional over H on the following

$$|d - Qf|_2^2 + \rho||f||_K^2.$$

Then, we obtain the minimizer function

$$f_\rho(t) = \sum_{j\in\mathbb{N}_n} c(\rho)_j K(t, t_j), \quad t \in \mathcal{T}$$

where $(G + \rho I)c(\rho) = d$. We choose $f_\rho(t_0)$ as our estimation and define

$$\varepsilon_\rho^2 = |d - Qf|_2^2 = \sum_{j\in\mathbb{N}_n} (1 - \frac{\lambda_j}{\rho+\lambda_j})^2 \gamma_j^2$$

and

$$\delta_\rho^2 = \|f_\rho\|_K^2 = \sum_{j\in\mathbb{N}_n} \frac{\lambda_j \gamma_j^2}{(\rho+\lambda_j)^2}$$

where $0 \leq \lambda_1 \leq \ldots \leq \lambda_n$ are the eigenvalues of the Gram matrix G corresponding to the orthonormal eigenvectors $u^j : j \in \mathbb{N}_n$ and $d = \sum_{j\in\mathbb{N}_n} \gamma_j u^j$.

As we want to compare this method to the midpoint algorithm, we then define the interval of uncertainty

$$I(t_0, \varepsilon_\rho, \delta_\rho) = \{f(t_0) : |d - Qf|_2 \leq \varepsilon_\rho, ||f||_K \leq \delta_\rho\}.$$

Clearly, $f_\rho(t_0)$ in $I(t_0, \varepsilon_\rho, \delta_\rho)$, However, the hyperellipse $\mathcal{H}_2(d|E(\delta_\rho))$ consists of only *one point*, namely f_ρ. To prove this, choose any $h \in \mathcal{H}_2(d|E(\delta_\rho))$. This mean that

$$||h||_K^2 \leq ||f_\rho||_K^2 = \delta_\rho$$

and

$$|d - Qh|_2^2 \leq |d - Qf_\rho|_2^2 = \varepsilon_\rho.$$

Consequently, we have that

$$|d - Qh|_2^2 + \rho||h||_K^2 \quad \leq \quad |d - Qf_\rho|_2^2 + ||f_\rho||_K^2$$

Since f_ρ is unique minimizer of R_ρ, $h = f_\rho$.

Therefore, our strategy in comparing the regularization and midpoint estimator, is to consider a bigger value of ε_ρ and δ_ρ. We choose $\varepsilon = \varepsilon_\rho$ and $\delta = \alpha\delta_\rho$ where α is in $A = \{1.5, 3, 6, 12, 24\}$. Consequently, the function $V_{\alpha\delta_\rho}$ in (11) becomes

$$V_{\alpha\delta_\rho}(c) := \alpha\delta_\rho||x_0 - Q^T c|| + \varepsilon|c|_2 + (c, d) \qquad (15)$$

and the corresponding hyperellipse is given by

$$\mathcal{H}_2(d|E(\alpha\delta_\rho)) = \{x : ||x|| \leq \alpha\delta_\rho, |d - Qx|_2 \leq \varepsilon_\rho\}.$$

For the computation $m_+(x_0, \pm d|\delta, \varepsilon)$, we need to find the minimum of the function $V_{\alpha\delta_\rho}$ defined for $c \in \mathbb{R}^n$ as

$$V_{\alpha\delta_\rho}(c) \quad = \quad \alpha\delta_\rho\sqrt{f_{\alpha\delta_\rho}(c)} + \varepsilon\sqrt{\sum_{j\in\mathbb{N}_n} c_j^2} \pm \sum_{j\in\mathbb{N}_n} c_j d_j.$$

where $f_{\alpha\delta_\rho} : \mathbb{R}^n \longrightarrow \mathbb{R}$ and $f_{\alpha\delta_\rho}(c)$

$$= \alpha\delta_\rho\sqrt{K(t_0, t_0) - 2\sum_{j\in\mathbb{N}_n} c_j K(t_0, t_j) + \sum_{i,j\in\mathbb{N}_n} c_i c_j K(t_i, t_j)}.$$

Moreover, we desire here not only to estimate the value of function f at one t_0 but also we estimate the value of function f at $t_j^* \in T_0$ where $T_0 = \{t_j^* : j \in \mathbb{N}_k\}$ and $T_0 \subseteq \mathcal{T}\backslash T$. To compare both methods for any point $t_j^* \in T_0$, we then compute a sum square error between exact function g at the point t_j^* and the function learned by using regularization estimator $f_\rho(t_j^*)$ and midpoint estimator $m(t_j^*, d|E(\delta))$ with different values of δ. That is, we define the sum square error of the regularization estimator by

$$E_\rho(T_0) = \sum_{j\in\mathbb{N}_k} |g(t_j^*) - f_\rho(t_j^*)|^2$$

and the sum square error of the midpoint estimator by

$$E_m(T_0) = \max_{\alpha \in A} e_m(T_0, d|E(\alpha\delta_\rho))$$

where

$$e_m(T_0, d|E(\alpha\delta_\rho)) = \sum_{j \in \mathbb{N}_k} |g(t_j^*) - m(t_j^*, d|E(\alpha\delta_\rho))|^2.$$

The results of sum square error are shown in Tables I, II and III for three learning approaches.

The algorithm for finding the value of a function by using the regularization estimator and the midpoint estimator has shown below.

The Algorithm

Step 0 : Given $\rho > 0$.

Step 1 : Calculate $c(\rho)$ by the formula:
$c(\rho) = (G + \rho I)^{-1} d.$

Step 2 : Let $f_\rho(t) = \sum_{j \in \mathbb{N}_n} c(\rho)_j K(t, t_j)$
and find $f_\rho(t_j^*) = \sum_{j \in \mathbb{N}_n} c(\rho)_j K(t_j^*, t_j)$
for all $t_j^* \in T_0$.

Step 3 : Calculate ε_ρ^2 and δ_ρ^2 by the formula:
$\varepsilon_\rho^2 = \sum_{j \in \mathbb{N}_n} (1 - \frac{\lambda_j}{\rho + \lambda_j})^2 \gamma_j^2.$
$\delta_\rho^2 = \sum_{j \in \mathbb{N}_n} \frac{\lambda_j \gamma_j^2}{(\rho + \lambda_j)^2}.$

Step 4 : Set $\varepsilon = \varepsilon_\rho$ and $\delta = \alpha\delta_\rho$ where $\alpha \in A$.
Find $m_+(t_j^*, \pm d|E(\alpha\delta_\rho))$, we use the program fminunc in the optimization toolbox of MATLAB 7.3.0.

$$m_+(t_j^*, \pm d|E(\alpha\delta_\rho)) = \min_{c \in \mathbb{R}^n} \alpha\delta||K_{t_j^*} - Q^T c|| + \varepsilon|c|_2 \pm (c, d).$$

Step 5 : Calculate $m(t_j^*, d|E(\alpha\delta_\rho))$ by the formula

$$m(t_j^*, d|E(\alpha\delta)) = \frac{1}{2}\Big(m_+(t_j^*, d \mid E(\alpha\delta)) - m_+(t_j^*, -d \mid E(\alpha\delta)\Big)$$

Step 6 : Calculate $E_\rho(T_0) = \sum_{j \in \mathbb{N}_k} |g(t_j^*) - f_\rho(t_j^*)|^2$
and $E_m(T_0) = \max_{\alpha \in A} e_m(T_0, d|E(\alpha\delta_\rho))$.

A. *Experiment 1*

For the first experiment, we use the gaussian kernel on $\mathbb{R}$. Specifically, we choose the exact function g as following

$$g(t) = K_0(t) + 15K_{2.7}(t) - K_{4.7}(t). \quad (16)$$

where

$$K(t, s) = K_s(t) = exp(-\frac{(t-s)^2}{50}) \quad t, s \in \mathbb{R}. \quad (17)$$

Graph of the exact function as Eq. (16) is shown in Fig.III-A

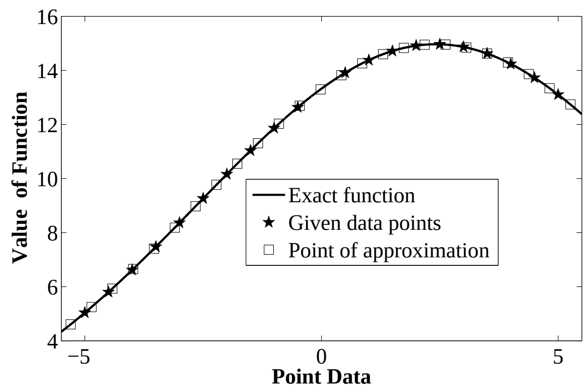

Fig. 1. Graph of the exact function for Gaussian kernel on $\mathbb{R}$.

The set T consists of twenty equally spaced points given by the formulae $t_1 = -5.0, t_{j+1} = t_j + 0.5$ and $t_{11} = 0.5, t_{j+11} = t_{10+j} + 0.5$, for all $j \in \mathbb{N}_9$. We then generate the data vector $d = (d_j : j \in \mathbb{N}_{20})$ by setting $d_j = g(t_j) + e_j$, $j \in \mathbb{N}_{20}$, where the error vector e is generated randomly from a uniform distribution and given by the formulae $e_{1+j} = (-1)^j 0.00207, e_{2+j} = (-1)^j 0.00607, e_{3+j} = (-1)^j 0.0063, e_{4+j} = (-1)^j 0.0037, e_{5+j} = (-1)^j 0.00575, j = 0, 5, 10, 15.$

Next, we choose the set T_0 which consists of twenty five equally spaced points given by the formula $t_1^* = -5.3, t_{j+1}^* = t_j^* + 0.44$ for all $j \in \mathbb{N}_{24}$. We shall estimate the value of the function $f(t_j^*)$ when $f \in \mathcal{H}_2(d|E(\delta))$ and for any $t_j^* \in T_0$.

TABLE I
THE SUM SQUARE ERROR OBTAINED FROM GAUSSIAN KERNEL ON $\mathbb{R}$ FOR BOTH METHODS FOR DIFFERENT VALUES OF THE REGULARIZATION PARAMETER ρ.

ρ	Sum Square Error	
	$E_\rho(T_0)$	$E_m(T_0, d\|E(\delta))$
10^{-5}	0.0310	0.0278
10^{-4}	0.0585	0.4343
10^{-3}	0.1577	0.0397
10^{-2}	0.6382	0.0579
10^{-1}	5.4352	0.2309
1	146.4015	8.0028
5	1.0456e+003	735.4432
10	1.7518e+003	1.5720e+003

Our computation in Tables I, II and III show each of these quantities as the values of ρ in the first column and the sum square errors of regularization estimator in the second column and those of the midpoint estimator in the third column. Table I presents the sum square errors between the exact function and the function learned from the regularization estimator and the midpoint estimator.

Our computation indicates that the midpoint estimator for almost all the range of the regularization parameter is better than the regularization estimator although we pick up E_m, which is the largest sum square error of the midpoint estimator with the value of $\delta = \alpha\delta_\rho$ for all $\alpha \in A = \{1.5, 3, 6, 12, 24\}$.

B. Experiment 2

In our second experiment, we choose the exact function

$$g(t) = K_0(t) - \frac{1}{2}K_{\frac{1}{2}}(t) - K_{-\frac{1}{3}}(t) \quad (18)$$

where

$$K(t,s) = K_s(t) = \frac{1}{1-ts} \quad t, s \in (-1,1) \quad (19)$$

is the rational kernel on $(-1,1)$.

Graph of the exact function as Eq. (16) is shown in Fig.III-B.

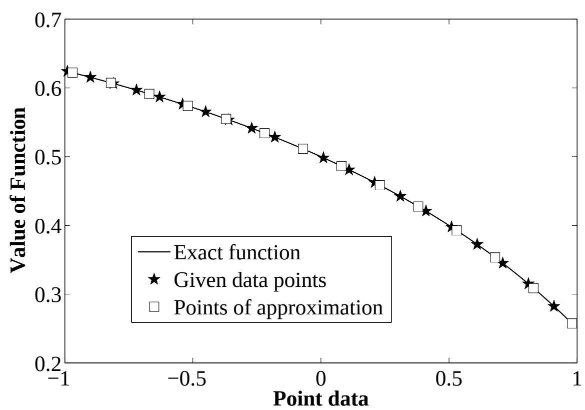

Fig. 2. Graph of the exact function for Rational kernel on $\mathbb{R}$.

The set up is similar to that in Experiment 1. Data d_j are set as $d_j = g(t_j) + e_j$, $j \in \mathbb{N}_{20}$ with e_j are similar to that in Experiment 1. Points t_j are the point of exact values in $T = \{t_j : j \in \mathbb{N}_{20}\}$. The set of T consists of twenty equally spaced points given by the formulae $t_1 = -0.99, t_{j+1} = t_j + 0.99$ and $t_{11} = 0.01, t_{j+11} = t_{10+j} + 0.1$, for all $j \in \mathbb{N}_9$. In this experiment, we choose the set of T_0 which consists of fourteen equally spaced points given by the formula $t_1^* = -0.97, t_{j+1}^* = t_j^* + 0.15$ for all $j \in \mathbb{N}_{13}$.

From Table II, comparing the results of the sum square error, we see that the midpoint estimator seems to perform better for almost all the range of the regularization parameter.

TABLE II
THE SUM SQUARE ERROR OBTAINED FROM RATIONAL KERNEL ON $\mathbb{R}$ FOR BOTH METHODS FOR DIFFERENT VALUES OF THE REGULARIZATION PARAMETER ρ.

ρ	Sum Square Error	
	$E_\rho(T_0)$	$E_m(T_0, d\|E(\delta))$
10^{-5}	0.0192	0.0112
10^{-4}	0.0030	0.0050
10^{-3}	0.0087	0.0076
10^{-2}	3.8727e-004	0.0055
10^{-1}	2.6605e-004	2.4927e-004
1	0.0117	0.0082
5	0.1466	0.0030
10	0.4198	0.1306

C. Experiment 3

For the third computational experiment, we choose the gaussian kernel on $\mathbb{R}^2$ and choose the exact function

$$g(t) = \tfrac{1}{2}K_{(1,1)}(t) + \tfrac{1}{6}K_{(1,-1)}(t) + \tfrac{1}{6}K_{(-1,-1)}(t) + \tfrac{1}{3}K_{(-1,1)}(t) \quad (20)$$

where

$$K(t,s) = K_s(t) = e^{|t-s|_2^2}, \ t, s \in \mathbb{R}^2. \quad (21)$$

Graph of the exact function as Eq. (16) is shown in Fig.III-C

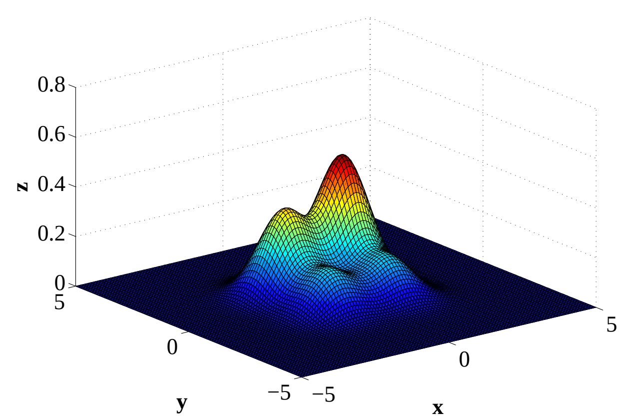

Fig. 3. Graph of the exact function for Rational kernel on $\mathbb{R}^2$.

The set up is similar to Experiments 1 and 2. Data d_j are set as $d_j = g(t_j)+e_j$, $j \in \mathbb{N}_{20}$ with e_j. Points t_j are the point of exact values in $T = \{t_j : j \in \mathbb{N}_{20}\}$. The set of T consists of twenty data points given by the formula

$$t_j = \left(x_j, \frac{(-1)^j\sqrt{36-4x_j^2}}{3}\right) \quad (22)$$

where $x_1 = -2.8, x_{1+j} = x_j + 0.3, x_{11} = 0.1, x_{11+j} = x_{10+j} + 0.3$ and $j \in \mathbb{N}_9$.

In this experiment, we choose the value of $T_0 = \{t_j^* : j \in \mathbb{N}_{14}\}$ which consists of ten data points given by the formula in

Eq. (22)where $x_1^* = -3.0, x_{1+j}^* = x_j^* + 0.6, x_6^* = 0.5, x_{6+j}^* = x_{5+j}^* + 0.6$ and $j \in \mathbb{N}_4$.

TABLE III
THE SUM SQUARE ERROR OBTAINED FROM GAUSSIAN KERNEL ON $\mathbb{R}^2$ FOR BOTH METHODS FOR DIFFERENT VALUES OF THE REGULARIZATION PARAMETER ρ.

ρ	Sum Square Error	
	$E_\rho(T_0)$	$E_m(T_0, d\|E(\delta))$
10^{-5}	2.4986e-004	2.4989e-004
10^{-4}	2.4964e-004	2.4989e-004
10^{-3}	2.4752e-004	2.4987e-004
10^{-2}	2.3487e-004	2.4783e-004
10^{-1}	4.4568e-004	2.4904e-004
1	0.0110	0.0019
5	0.0503	0.0437
10	0.0705	0.0676

Table III depicts the sum square error evaluated on ten data points for the regularization estimator and the midpoint estimator. Our computation again indicates that the midpoint algorithm provides, at least in this numerical experiment, better result than the regularization method.

IV. CONCLUSION

In this paper, we have provided some basic facts about the Hypercircle inequality for data error We provided it in the case that the unit ball B is replaced by δB where δ is any positive number. Moreover, we also discussed some important facts of $Hide$ for practical computation. In Section III, we discussed some results of our numerical experiments of learning the value of a function in RKHS. All our computation indicated that the midpoint algorithm on the learning tasks provided, at least in our computational numerical experiments, better results than the regularization approach.

ACKNOWLEDGMENT

The authors are grateful to Prof. Charles A. Micchelli for his helpful suggestions and comments on this paper. The authors would also like to thank the referee for valuable comments on an earlier version of this paper.

REFERENCES

[1] N. Aronszajn, Theory of reproducing kernels. *Trans. Amer. Math. Soc.* 68(1950), 337-404.

[2] Philip J. Davis, Interpolation and Approximation. *Dover Publications, New York*. 1975.

[3] T. Evgeniou, M. Pontil, T. Poggio. Regularization networks and support vector machines, *Adv. Comput. Math.* 13(2000), 1-50.

[4] G. E. Forsythe, G. H. Golub. On the stationary values of a second-degree polynomial on the unit sphere, *J. Soc. Indust. Appl. Math*. 13(1965), 1050-1068.

[5] G. Gnecco, M. Sanguineti. Weight-decay regularization in Reproducing Kernel Hilbert Spaces by variable-basis schemes. *WSEAS Transactions on Mathematics*. 8(2009), 1109-2769.

[6] U. Hamarik, T. Raus. Choice of the regularization parameter in ill-posed problems with rough estimate of the noise level of data, *WSEAS Transactions on Mathematics*. 4(2005), 7681.

[7] K. Khompurngson. The Hypercircle Inequality for Inaccurate Data, *Ph.D. Thesis, Mahidol University, Bangkok, Thailand.* 2010.

[8] K. Khompurngson, C. A. Micchelli. Hide, *Jaen Journal on Approximation*. (Accepted).

[9] C. A. Micchelli, T. J. Rivlin. A survey of optimal recovery, *Optimal Estimation in Approximation Theory, C. A. Micchelli and T. J. Rivlin, Eds., Plenum Publishing Co., New York*. 1977, 1-53.

[10] B. Scholkopf, A. J. Smola. Learning with kernels, *The MIT Press*. 2002.

[11] J. Shawe-Taylor, N. Cristianini. Kernel methods for pattern analysis, *Cambridge University Press*. 2004.

[12] G. Wahba. Spline Models for Observational Data, *Philadelphia: Society of Industrial and Applied Mathematics*. 59(1990).

[13] L. Wu. A parameter choice method for Tikhonov regularization, *Electron. Trans. Numer. Anal.* 16(2003), 107-128.