Function Representation Using Hypercircle Inequality for Data Error

D. Poltem, K. Khompurngson, B. Novaprateep

***Abstract*—**In this paper, the study of Hypercircle inequality for data error ($Hide$) is briefly reviewed. Within the framework of $Hide$ and to find a function representation from inaccurate data, the midpoint algorithm is provided. We give a new result for a function representation that has the form of the representer theorem. We illustrate some important facts for a practical computation and study the problem in the learning value of a function for a learning kernel. We demonstrate the potential of this framework by comparing our result to the regularization method, which is the standard method in the learning value of a function. The present example compares the performance of the methods when the optimal values of regularization parameters are used.

***Keywords*—**Hypercircle Inequality, Reproducing kernel Hilbert space, Regularization, Convex optimizations, Noise data

I. Introduction

ONE of the most important issue of the scientific problem is to find a function representation from data when only finite samples are known exactly [2], [3], [8], [10], [19]. Tikhonov regularization in reproducing kernel Hilbert spaces (RKHS), known as the representer theorem, is one of the most important role in a learning problem. It plays an important role in approximation as it allows to write the solution in learning problem easily.

In the regularization theory, for the target function from domain $\mathcal{T}$ to the range $Y \subseteq \mathbb{R}$ we call $\mathcal{T}$ the input set and Y the output set. Suppose that a finite set $\{(t_j, d_j) : j \in \mathbb{N}_n\} \subseteq X \times Y$ of samples of target function is available. For the simplicity of enumerating with the finite sets, we set $\mathbb{N}_n = \{1, 2, ..., n\}$ for $n \in \mathbb{N}$. Following the framework of Tikhonov regularization in the machine learning, we let the hypothesis space H be a reproducing kernel Hilbert space (RKHS) of real value function on a set $\mathcal{T}$. For each $t \in \mathcal{T}$, there exists a function $K_t \in H$ (called representer of t) with the reproducing property

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

for all $f \in H$. Since K_t is a function in H, by the reproducing property, for each of $s \in \mathcal{T}$ we can write

$$K_t(s) = K(t, s)$$

D. Poltem is with the Department of Mathematics, Faculty of Science, Burapha University, and Centre of Excellence in Mathematics, PERDO, CHE, THAILAND e-mail: duangkamolp@buu.ac.th.

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

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.

Aronszajn's theory of reproducing kernel Hilbert spaces [1] states that a function $K : \mathcal{T} \times \mathcal{T} \longrightarrow \mathbb{R}$ is a reproducing kernel if it is a symmetric, that is $K(s,t) = K(t,s)$, and positive definite:

$$\sum_{i,j=1}^{n} a_j a_i K(t_j, t_i) \geq 0$$

for any $n \in \mathbb{N}$ and the choice of inputs $T = \{t_j : j \in \mathbb{N}_n\} \subseteq \mathcal{T}$ and $a = (a_1, ..., a_n) \in \mathbb{R}^n$. Moreover, we know that 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 [17].

Given $t_0 \in \mathcal{T}$, we are able to determine a meaningful approximation of 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 basic idea in learning a problem is to determine a functional representation from the data. A possible way to efficiently solve the learning problem is provided by regularization networks, which amounts to minimization of the following R_ρ functional defined for $f \in H$

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

where ρ is a positive number. According to representer theorem [4], [12], [13], [15], [16], [19], [20], the form of the solution of equation (1) is, under general condition:

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

for some real vector $c(\rho) = (G + \rho I)^{-1} d$ where I is $n \times n$ identity matrix and Gram matrix $G = (K(t_i, t_j) : i, j \in \mathbb{N}_n)$.

However, the regularization theory is only applied to circumstances for which data is know exactly [6], [7]. Then our previous work [11], [14] extends it to circumstances for which there is known data error. We have discussed the Hypercircle Inequality for data error ($Hide$) in a learning problem. The midpoint algorithm for finding the value of function at the given points was proposed. As the midpoint algorithm, we 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, the best choice for this number is a function whose values at t_0 is a midpoint of the interval $I(t_0, \varepsilon_\rho, \delta_\rho)$. The most important issue in a learning problem is the choice of the data representation. Then the purpose of this paper is to present the approximation function that we obtain from the midpoint algorithm. We then compare our result to the regularization method. To compare these two methods, the regularization estimator $f_\rho(t_0)$ can be viewed as an element in the interval $I(t_0, \varepsilon_\rho, \delta_\rho)$. In the previous work [11], [14], we found that the learned function has the form of representer theorem (2), but the choice of the coefficients in equation (2) are generally different from those obtained from a regularization procedure. Therefore, a learned function that has the form of representer theorem for a practical computation is described and analyzed in this paper. In addition, this present paper proposes a new choice to choose value of δ_ρ. We also report some results from numerical experiment in the learning value of a function by using midpoint algorithm with different values of δ_ρ.

This paper is organized as follows: In section 2, we recall the Hypercircle Inequality for data error measured with square loss and the way to get a function representation to the midpoint algorithm. Section 3 deals with the midpoint algorithm. After a short introduction about the framework of $Hide$ and the midpoint algorithm, we present some numerical simulations based on our analysis and discuss some extensions of our framework.

II. Hypercircle inequality for data error

In this section, we describe the necessary background. Most of the results below are based on results from our previous work [11], [14]. Some improvements and simplifications of our previous work on $Hide$ are provided.

Let Hilbert space H over the real number with inner product $\langle \cdot, \cdot \rangle$ and choose a finite set of linearly independent elements $\mathcal{X} = \{x_j : j \in \mathbb{N}_n\}$ in H. Let M be the $n-$dimensional linear subspace of H spanned by the vectors in $\mathcal{X}$. That is, we denote

$$M := \{\sum_{j \in \mathbb{N}_n} a_i x_i : a \in \mathbb{R}^n\}$$

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 follows

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

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 follows

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

Consequently, the Gram's 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, [3], [12], and discuss what we need for $Hide$.

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

$$Qx(d) = d.$$

We have $x(d) = Q^T G^{-1} d$.

Moreover, from this formula we obtain the useful equation

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

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 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 [5].

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.

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

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 [11] for more information about the proof of $Hide$ measured with any norm on $\mathbb{R}^n$.

Let d be a given vector in $\mathbb{R}^n$ and

$$E_2 = \{e : e \in \mathbb{R}^n, |e|_2 \leq \varepsilon\}$$

where $|\cdot|_2 : \mathbb{R}^n \to \mathbb{R}_+$ is a Euclidean norm on $\mathbb{R}^n$ and ε is some prescribed positive number. The *hyperellipse*, $\mathcal{H}(d|\delta E_2)$, is a subset of H which is defined by

$$\mathcal{H}(d|\delta E_2) = \{x : x \in \delta B, Qx - d \in E_2\}$$

where $\delta B := \{x : x \in H, ||x|| \leq 1\}$ is the *unit ball* in H.

In the special case that $\varepsilon = 0$, we denote that the hyperellipse becomes to hypercircle as follows

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

Moreover, we point out for any $d \in \mathbb{R}^n$ there is $x(d) = Q^T G^{-1} d$ such that

$$Qx(d) = d \text{ and } ||x(d)||^2 = (d, G^{-1}d)$$

where $(\cdot, \cdot)$ is Euclidean inner product on $\mathbb{R}^n$.

Moreover, we point out that if $\mathcal{H}(d, \delta) \neq \phi$ then the vector $x(d) = L^T G^{-1} d$ is the best estimator to estimate the value of $\langle x, x_0 \rangle$ when $x \in \mathcal{H}(d, \delta)$. That means, the best estimator has the form of linear combination of the vector in $\mathcal{X}$ and the coefficient is given by $G^{-1}d$. In additional, the best estimator $x(d)$ is independent of vector x_0.

Indeed, let us add the relation between them as shown below

$$\mathcal{H}(d|\delta E_2) = \bigcup_{e \in E_2} \mathcal{H}(d + e, \delta).$$

So far, we obtain that for each $e \in E_2$ there is the vector

$$x(d + e) = Q^T G^{-1}(d + e) \in M$$

such that

$$Qx(d + e) = d + e \text{ and } ||x(d + e)||^2 = (d + e, G^{-1}(d + e)).$$

Now we ready to disscus when $\mathcal{H}(d|\delta\mathbb{E}_2) \neq \phi$. Let us recall the following facts.

Definition 1: 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

$$A(x - d) = \lambda x. \tag{6}$$

Since G is positive definite matrix, we then assume that $0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$ are eigenvalue of G^{-1} and

$$\{u^j : j \in \mathbb{N}_n\}$$

is a corresponding orthonormal set of eigenvector. Next, we 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 $I := \{j : \lambda_j d_j = 0\}$.

Theorem 1: The spectrum of the pair $(\varepsilon^2 G^{-1}, \frac{d}{\varepsilon})$ consists of all real λ such that

$$g(\lambda) = \sum_{i \in I} \frac{\lambda_i^2 \gamma_i^2}{(\varepsilon^2 \lambda_i - \lambda)^2} = 1 \tag{7}$$

togerther with each eigenvalue λ_k of $\varepsilon^2 D_{_I}$ for which

$$g(\lambda_k) < 1$$

where $k \in I := \{j : \lambda_j d_j = 0\}$.

Proof. We refer the reader to [5] for the proof.

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

$$\min_{\substack{e \in \mathbb{R}^n \\ |e| \leq \varepsilon}} (d + e, G^{-1}(d + e)) =$$
$$\begin{cases} \Lambda_{_I} + \Lambda_{_I} \sum_{j \notin I} \frac{\lambda_j |\gamma_j|^2}{\Lambda_{_I} - \varepsilon^2 \lambda_j}, & \text{if } |d| > \varepsilon \\ 0, & \text{if } |d| \leq \varepsilon \end{cases}$$

and

$$\max_{\substack{e \in \mathbb{R}^n \\ |e| \leq \varepsilon}} (d + e, G^{-1}(d + e)) =$$
$$\begin{cases} \Lambda'_{_I} + \Lambda'_{_I} \sum_{j \notin I} \frac{\lambda'_j |\gamma_j|^2}{\Lambda'_{_I} - \varepsilon^2 \lambda'_j}, & \text{if } |d| > \varepsilon \\ 0, & \text{if } |d| \leq \varepsilon \end{cases}$$

Proof. We refer the reader to the paper for the proof [5].

Theorem 2: Let Λ be the least value in the spectrum of the pair $(\varepsilon^2 G^{-1}, \frac{d}{\varepsilon})$. If $|d| > \varepsilon$ then $\mathcal{H}(d|\delta\mathbb{E}_2) \neq \phi$ if and only if

$$\Lambda + \Lambda \sum_{j \notin I} \frac{\lambda_j |\gamma_j|^2}{\Lambda - \varepsilon^2 \lambda_j} + \leq \delta^2$$

Proof. We refer the reader to [5] for the proof.

Before we state the next theorem, let us define

$$\mathbb{M} := \{x(d + e) = L^T G^{-1}(d + e) : e \in E\}.$$

Next, we have the following

Theorem 3: Let Λ' be the greatest value in the spectrum of the pair $(\varepsilon^2 G^{-1}, \frac{d}{\varepsilon})$.
If $|d| > \varepsilon$ then

$$\mathbb{M} \subseteq \mathcal{H}(d|\delta\mathbb{E}_2)$$

if and only if

$$\Lambda' + \Lambda' \sum_{j \notin I} \frac{\lambda'_j |\gamma_j|^2}{\Lambda' - \varepsilon^2 \lambda_j} \leq \delta^2 \tag{8}$$

Therefore, we point out that if we choose

$$\Lambda' + \Lambda' \sum_{j \notin I} \frac{\lambda'_j |\gamma_j|^2}{\Lambda' - \varepsilon^2 \lambda_j} \leq \delta^2$$

then the best estimator, $x(d + e_0)$, is in the interval

$$I(x_0, x(d + e)) = \{\langle x, x_0 \rangle : x \in \mathcal{H}(d|\delta E_2)\}.$$

The original idea of $Hide$ to optimally estimate one feature of an $x \in \mathcal{H}(d|\delta E_2)$ was presented in the previous work [11]. We define a feature of $x \in H$ as the value of a prescribed linear functional F_{x_0} defined at x as

$$F_{x_0}(x) := \langle x, x_0 \rangle.$$

The interval of uncertainty for the feature F_{x_0} is defined as

$$I(x_0, d|\delta E_2) = \{F_{x_0}(x) : x \in \mathcal{H}(d|\delta E_2)\}$$

which is closed and bounded. We also define the interval in $\mathbb{R}$ by

$$I(x_0, d|\delta E_2) = [m_-(x_0, d|\delta E_2), m_+(x_0, d|\delta E_2)] \quad (9)$$

where

$$m_+(x_0, d|\delta E_2) := \max\{F_{x_0}(x) : x \in \mathcal{H}(d|\delta E_2)\} \quad (10)$$

and

$$m_-(x_0, d|\delta E_2) := \min\{F_{x_0}(x) : x \in \mathcal{H}(d|\delta E_2)\}.$$

We observe that

$$m_-(x_0, d|\delta E_2) = -m_+(x_0, -d|\delta E_2).$$

Consequently, a *center* of the above interval is

$$m(x_0, d|\delta E_2) = \frac{m_+(x_0, d|\delta E_2) - m_+(x_0, -d|\delta E_2)}{2}. \quad (11)$$

To state the main result of this section, a function representation from midpoint algorithm, we introduce the following terminology.

The proposition below is important and useful to determine the best estimator in the hyperellipse $\mathcal{H}_2(d, \varepsilon)$ for the feature F_{x_0}. To this end, let us define $P : H \to M$ which is the orthogonal projection of H onto M as for any $x \in H$

$$Px = Q^T G^{-1}(Qx).$$

Proposition 1: If $\mathcal{H}(d|E_2(\delta)) \neq \emptyset$ and $x_0 \notin M$ then there exist $e_\pm \in E_2$ such that

$$F_{x_0}(x_\pm(d + e_\pm)) = m_\pm(x_0, d|E_2(\delta)). \quad (12)$$

Moreover, the vector

$$x_\pm(d + e_\pm) = x(d + e_\pm) \pm \frac{\sqrt{\delta^2 - ||x(d + e_\pm)||^2}}{\text{dist}(x_0, M)}(x_0 - Px_0)$$

Proof. We refer the reader to [11] for the proof.

Theorem 4: If $\mathcal{H}(d|E_2(\delta)) \neq \emptyset$ then there is a $e_0 \in E$ such that $x(d + e_0)$ is the best estimator for the feature F_{x_0}. Moreover, the vector $e_0 \in \mathbb{R}^n$ can be chosen on the line segment joining e_- and e_+.

Proof. We refer the reader to [11] for the proof.

That is, we have that

$$F_{x_0}(x(d + e_0)) = m(x_0, d|\delta E_2)$$

and $x(d+e_0) = Q^T G^{-1}(d+e_0) \in M$. Therefore, we can see that the best estimator is still the form of linear combination of the vector in $\mathcal{X}$ and the coefficient is given by

$$G^{-1}(d + e_0)$$

for some $e_0 \in E_2$. However, we observe that the best estimator is depend on the vector of x_0.

Specifically, to find the best estimator we only need to evaluate the two number $m_+(x_0, \pm d|E_2(\delta))$ and then compute

$$\frac{1}{2}(m_+(x_0, d|E_2(\delta)) - m_+(x_0, -d|E_2(\delta))).$$

These points are arguments that may suggest the main results of this section. For the purpose of establishing a function representation from midpoint algorithm, we know that the learned function has the form of representer theorem (2), but the choice of the coefficients in equation (2) are generally different from those obtained from a regularization procedure. Therefore, the following proposition formally establishes the way to get the coefficients in equation (2) that obtains from the midpoint algorithm.

Next, we will describe a duality formula for the right-hand side of the interval of uncertainty and then show how to find the vector e_0.

Proposition 2: If $\mathcal{H}(d|E_2(\delta))$ contains more than one point, $x_0 \notin M$, and $\frac{x_0}{||x_0||} \notin \mathcal{H}(d|E_2(\delta))$, then

$$m_+(x_0, d|E) = \min\{\delta||x_0 - Q^T c|| + \varepsilon|c|_2 + (c, d) : c \in \mathbb{R}^n\} \quad (13)$$

where $(\cdot, \cdot)$ is a Euclidean inner product on $\mathbb{R}^n$. Moreover, the minimum $c^* \in \mathbb{R}^n$ is the unique solution of the nonlinear equation

$$-\delta Q\left(\frac{x_0 - Q^T c^*}{||x_0 - Q^T c^*||}\right) + \varepsilon \frac{c^*}{|c^*|_2} + d = 0 \quad (14)$$

and

$$x_+(d) := \delta \frac{x_0 - Q^T c^*}{||x_0 - Q^T c^*||} \quad (15)$$

satisfies

$$x_+(d) = \arg\ \max\{F_{x_0}(x) : x \in \mathcal{H}(d|E_2(\delta))\}. \quad (16)$$

Proof. We refer the reader to [11] for the proof.

Before we state the main result of this paper, let us define the real-valued function $h : [0, 1] \to \mathbb{R}^n$ as following

$$h(\lambda) = \langle x(d + \lambda e_+ + (1 - \lambda)e_-, x_0\rangle.$$

Proposition 3: If $\mathcal{H}(d|E_2(\delta)) \neq \emptyset$ then there is a $\lambda_0 \in [0, 1]$ such that $h(\lambda_0) = m(x_0, d|E_2(\delta))$ where

$$\lambda_0 = \frac{m(x_0, d|E_2) - (G^{-1}(d + e_-), Qx_0)}{(G^{-1}(e_+ - e_-), Qx_0)}.$$

Proof. By Proposition 2, there is $x_+(d) \in \mathcal{H}(d|E_2(\delta))$ such that

$$F_{x_0}(x_+(d)) = m_+(x_0, d|E).$$

That is, $Qx_+(d) = d + e_+$ for some $e_+ \in E_2$. Since $m_-(x_0, d|E) = -m_+(x_0, -d|E)$ and by Proposition 2 again, we have that there is $x_+(-d) \in \mathcal{H}(-d|E_2(\delta))$ such that

$$F_{x_0}(x_+(-d)) = m_+(x_0, -d|E).$$

That is, $Qx_+(-d) = -d + e'$ for some $e' \in E_2$. Therefore,

$$-x_+(-d) \in \mathcal{H}(d|E_2(\delta)) \text{ and } Q(-x_+(-d)) = d - e'.$$

We set $e_- = -e'$. Since $h(0) \leq m(x_0, d|E_2)$ and $h(1) \geq m(x_0, d|E_2)$, we can solve the linear equation above to find λ_0 such that $h(\lambda_0) = m(x_0, d|E_2(\delta))$. Consequently, that value is given by

$$\lambda_0 = \frac{m(x_0, d|E_2) - (G^{-1}(d + e_-), Qx_0)}{(G^{-1}(e_+ - e_-), Qx_0)}.$$

The above proposition can be easily used for a practical computation of a function representation from midpoint algorithm. In the next section, we will describe some numerical experiments in learning the value of a function and also present a function representation of a learned function. We also compare our results to the regularization method which is the standard method in learning value of a function.

III. Numerical Experiments

As discussed in the introduction and review of $Hide$, we conducted an experiment to compare the regularization method and our midpoint algorithm. We will show a function representation of a learned function by using our main goal results that we obtained from section 2.

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 2 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 $Qf := (f(t_j) =< K(t_j, \cdot), f >: j \in \mathbb{N}_n)$ 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

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

where $(G + \rho I)c(\rho) = d$. We 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 $0 \leq \lambda_1 \leq ... \leq \lambda_n$ 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 the regularization 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

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

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

Since, the interval of uncertainty for the feature F_{t_0} is defined as

$$I(t_0, d|\delta E_2) = \{F_{t_0}(x) : x \in \mathcal{H}(d|\delta E_2)\}$$

which is closed and bounded. We also define the interval in $\mathbb{R}$ by

$$I(t_0, d|\delta E_2) = [m_-(t_0, d|\delta E_2), m_+(t_0, d|\delta E_2)]$$

where

$$m_+(t_0, d|\delta E_2) := \max\{F_{t_0}(x) : x \in \mathcal{H}(d|\delta E_2)\}$$

and

$$m_-(t_0, d|\delta E_2) := \min\{F_{t_0}(x) : x \in \mathcal{H}(d|\delta E_2)\}.$$

We obtain

$$m_-(t_0, d|\delta E_2) = -m_+(t_0, -d|\delta E_2).$$

Consequently, a $center$ of the above interval is

$$m(t_0, d|\delta E_2) = \frac{m_+(t_0, d|\delta E_2) - m_+(t_0, -d|\delta E_2)}{2}.$$

For the computation $m_+((t_0, d|\delta E_2))$, we use the program $fminunc$ in the optimization toolbox of MATLAB 7.3.0 [8]. 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 $\delta > \delta_\rho$ as we report an important result for the choice of δ in section II. In order to establish a function representation from midpoint algorithm, we know that the learned function has the form of representer theorem (2). We then compute the coefficients in equation (2) that we obtain from the midpoint algorithm by using the following expression:

$$h(\lambda) = \langle t(d + \lambda e_+ + (1 - \lambda)e_-, t_0 \rangle$$

where

$$\lambda_0 = \frac{m(t_0, d|E_2) - (G^{-1}(d+e_-), Qt_0)}{(G^{-1}(e_+ - e_-), Qt_0)}.$$

Therefore, a function representation can be written in the the form of the representer theorem as

$$t(d+e_0) = Q^T G^{-1}(d + \lambda_0 e_+ + (1-\lambda_0)e_-).$$

A. Experiment

For the experiment, we use the Gaussian kernel on $\mathbb{R}$. Specifically, we choose

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

and the function g is chosen to be

$$g(t) = 2K_{1.8}(t) + K_{3.3}(t) - K_{1.2}(t). \tag{18}$$

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

TABLE I
THE MAXIMUM OF VALUE OF δ_ρ FOR THE RANGE OF THE REGULARIZATION PARAMETER ρ

ρ	δ_ρ	Λ'	maximum	δ_{max}
10^{-5}	1.97393	0.2506	4.5999	2.15
10^{-4}	1.94015	0.7045	5.2487	2.34
10^{-3}	1.93651	1.2334	5.9426	2.44
10^{-2}	1.92595	19.3955	26.1935	5.12
10^{-1}	1.85888	1016.1	1040.5	32.23
1	1.48779	46251	4.6392×10^4	215.39
2	1.23694	120772	1.2099×10^5	347.84

To find a function representation from our inaccurate data, we choose δ that satisfy equation (8) then $\mathbb{M}$ in Theoem 3 is smallest subset of M which is contained the best estimator of $< x, x_q >$ when $x \in \mathcal{H}_2(d|\delta\mathbb{E}_2)$. Table I shows the choices of δ in the different values of regularization parameters ρ. This table shows the minimum and maximum values of δ For midpoint algorithm, for example, we consider the value of a function in the case that $\rho = 10^{-5}$. From section II, we then obtain that $\mathcal{H}_2(d|\delta E_2) \neq \emptyset$ if and only if $\delta > 1.97393$. We also show that $\mathbb{M} \subseteq \mathcal{H}_2(d|\delta\mathbb{E}_2) \neq \emptyset$ if and only if $\delta > 2.15$ as shown in table I.

We shall estimate the value of a function at $f(2.7)$ where f is unknown function in the Hilbert space corresponding to the Gaussian kernel. The result of the computation is indicated by using $\varepsilon = \varepsilon(\rho)$ and $\delta > \delta(\rho)$ with differerent values of the regularization parameter ρ. Our computation indicates that the midpoint algorithm provides a better result than the regularization approach ($f_\rho(2.7)$) to the exact value of $g(2.7) = 1.8466$ (see table II).

TABLE II
THE VALUE OF A FUNCTION OBTAINED FROM GAUSSIAN KERNEL ON $\mathbb{R}$ FOR BOTH METHODS WITH THE REGULARIZATION PARAMETER ρ AND DIFFERENT VALUES OF δ CHOOSING FROM INTERVAL OF THE DELTA

ρ	$f_\rho(2.7)$	$m(t_0\|\delta = 2)$	$m(t_0\|\delta = 5)$	$m(t_0\|\delta = 50)$
10^{-5}	1.6114	1.8493	1.8497	1.8497
10^{-4}	1.6114	1.8487	1.8497	1.8496
10^{-3}	1.6114	1.8484	1.8498	1.8496
10^{-2}	1.6114	1.8481	1.8498	1.8496
10^{-1}	1.6114	1.8476	1.8498	1.8496
ρ	$f_\rho(2.7)$	$m(t_0\|\delta = 2)$	$m(t_0\|\delta = 100)$	$m(t_0\|\delta = 370)$
1	1.6114	1.7478	1.8497	1.8497
2	1.6114	1.5851	1.8497	1.8497

TABLE III
THE COEFFICIENTS OF FUNCTION APPROXIMATION OBTAINED FROM GAUSSIAN KERNEL ON $\mathbb{R}$ FOR BOTH METHODS WITH THE REGULARIZATION PARAMETER $\rho = 10^{-5}$ AND DIFFERENT VALUES OF δ CHOOSING FROM INTERVAL OF THE DELTA .

Coefficients	Parameters				
	$\rho = 10^{-5}$	$\delta = 2$	$\delta = 5$	$\delta = 10$	$\delta = 50$
c_1	-0.132	-0.010	-0.012	-0.011	-0.009
c_2	0.240	0.059	0.065	0.061	0.057
c_3	0.492	-0.227	-0.235	-0.230	-0.224
c_4	0.510	1.254	1.261	1.257	1.252
c_5	0.405	1.028	1.026	1.027	1.029
c_6	0.263	-0.383	-0.386	-0.384	-0.382
c_7	0.134	0.189	0.195	0.191	0.186
c_8	0.032	-0.085	-0.093	-0.088	-0.082
c_9	-0.013	0.029	0.035	0.031	0.027
c_{10}	-0.018	-0.006	-0.008	-0.006	-0.005

For the purpose of establishing a function representation from the regularization method and the midpoint algorithm, the learned function has the form of representer theorem:

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

Our computation shows that the choice of the coefficients in the above equation are generally different from those obtained from both methods as shown in table III. For midpoint estimator, after we increase value of δ, the coefficients are close to the same value.

The function representations from the regularization method and the midpoint algorithm compared to the exact function are shown in figures 1 and 2. Our computation indicates that the midpoint estimators for the regularization parameters $\rho = 10^{-5}$ and $\rho = 2$ give a better choice of

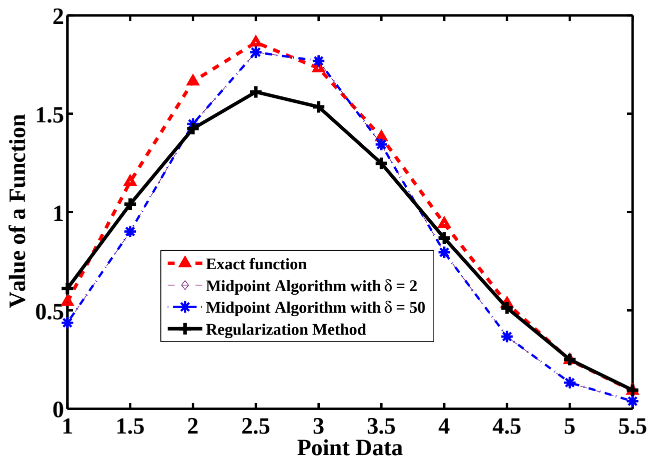

Fig. 1. Graph of the exact function for Gaussian kernel on $\mathbb{R}$ and function approximation using coefficients obtained from regularization method and midpoint algorithm with $\rho = 10^{-5}$.

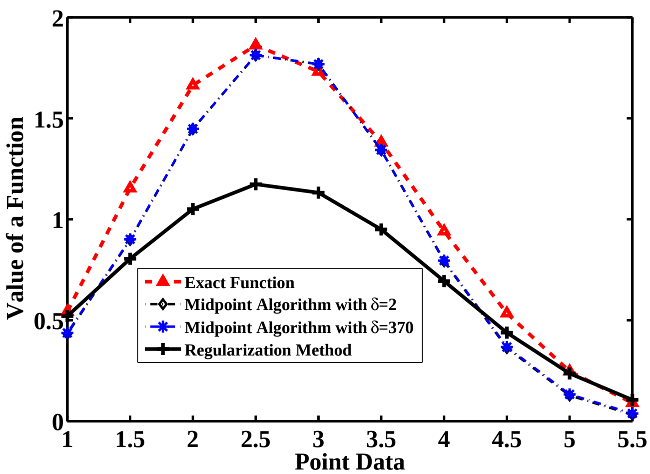

Fig. 2. Graph of the exact function for Gaussian kernel on $\mathbb{R}$ and function approximation using coefficients obtained from regularization method and midpoint algorithm with $\rho = 2$.

function approximation than the regularization estimator. As we see from figures 1 and 2, the function representation that we obtain from the midpoint algorthim with different values of δ is close to the same function.

Figures 3 and 4 show the value of a function obatianed from midpoint algorithm with different values of regularization parameter. We simulate the the result in the range of delta as shown in table 1. From these two figures indicate that the value of a function tend to the same value when the value of δ increased.

IV. Conclusion

In this paper, some basic facts of the Hypercircle Inequality were provided. The function representation from midpoint algorithm was presented in section 2. In section 3, we presented a numerical experiment to present a function representation from midpoint algorithm. Our computation indicated that the

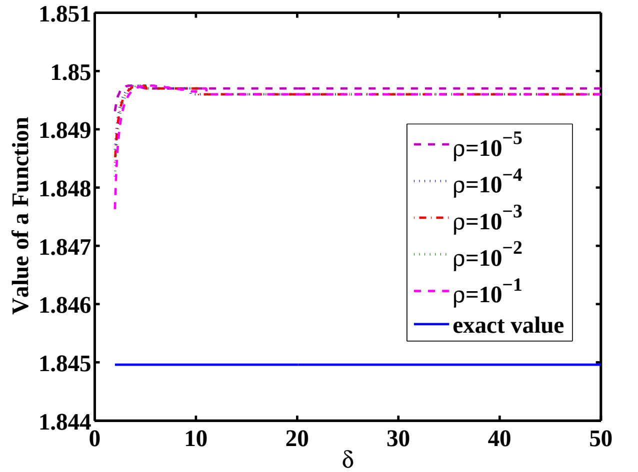

Fig. 3. The value of a function obtained from exact function and the midpoint algorithm from the range of δ at $t_0 = 2.7$.

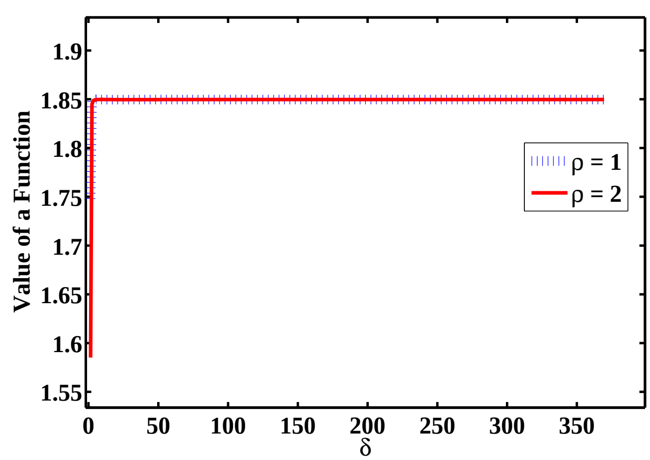

Fig. 4. The value of a function obtained from midpoint algorithm from the range of δ at $t_0 = 2.7$.

midpoint algorithm on the learning tasks provided, at least in our computational numerical experiments, better results than the regularization approach.

Acknowledgment

This research is supported by the Centre of Excellence in Mathematics, the Commission on Higher Education, Thailand. 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.*, vol. 68, 1950, pp. 337-404.

[2] O.H. Choon, L.C. Hoong, T.S. Huey “A functional approximation comparison between neural networks and polynomial regression”, *WSEAS Transactions on Mathematics*, vol. 6, 2008, pp. 353-363.

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

[4] T. Evgeniou, M. Pontil, T. Poggio, “Regularization networks and support vector machines”, *Adv. Comput. Math.*, vol. 13, 2000, pp. 1-50.

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

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

[7] Groetsch, W. Charles,"The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind", *Pitman: Boston*, 1984.

[8] 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*, vol. 4, 2005, pp. 7681.

[9] P.C. Hansen, "Regularization tools version 4.0 for Matlab 7.3", *SIAM Numerical Algorithms*, vol. 46, 2007, pp. 189-194.

[10] M. Iqbal, "Technique for a priori choice of regularizing parameters in Tikhonov regularization", *International Mathematical Forum*, vol. 1, 2006, pp. 337-255.

[11] K. Khompurngson, C. A. Micchelli, "Hide", *Jaen Journal on Approximation*, vol.3(1), 2011, pp.87-115.

[12] 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, pp. 1-53.

[13] V.A. Morozov, "A. On the solution of functional equations by the method of regularization", *Soviet Math. Dokl.*, vol. 7, 1996, pp. 414-417.

[14] B. Novaprateep, K. Khompurngson, D. Poltem, Learning the value of a function by using hypercircle inequality for data error, *International Journal of Mathematics and Computers in Simulation*. vol. 5, 2011, pp. 1-8.

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

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

[17] G. Taylor, R. Parr, "Kernelized value function approximation for reinforcement learning", *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada 2009.

[18] G. Wahba, "Practical approximate solutions to linear operator equations, when data are noisy", *SIAM. J. Numer. Anal*, vol. 14, 1997, pp. 651-677.

[19] G. Wahba, "Spline Models for observational data", *Philadelphia: Society of Industrial and Applied Mathematics*, vol. 59, 1990.

[20] L. Wu, "A parameter choice method for Tikhonov regularization", *Electron. Trans. Numer. Anal.*, vol 16, 2003, pp. 107-128.