Learning the Value of a Function from Inaccurate Data with Different Error Tolerance of Data Error

K. Khompurngson, B. Novaprateep, D. Poltem

***Abstract*—The intend of learning problem is to identify the best predictor from given data. Specifically, the well−known hypercircle inequality was applied to kernel−based machine learning when data is know exactly. In our previous work, this lead us to extend it to circumstance for which data is known within error. In this paper, we continues the study of this subject by improving the hypothesis of nonlinear optimization problem which is used to obtain the best predictor. In additional, we apply our results to special problem of learning the value of a function from inaccurate data with different error tolerance of data error.**

***Index Terms*—Hypercircle inequality, Reproducing Kernel Hilbert space, Convex Optimization and Noise Data.**

I. Introduction

THE principal goal of learning problem is to identify the best predictor from the available data. An example of this is the hypercircle inequality which has a long history in applied mathematics. Specifically, it was applied to kernel−based machine learning when there is known data exactly [4], [10]. Motivated by this fact, we have extended it to circumstance for which there is known data error [10]. The main propose of this paper is to improve the hypothesis of nonlinear optimization problem which is used to obtain the best predictor. From a variety of reason, our data error may have different error tolerance in real situation. We then apply our results to the special problem of learning the value of a function from inaccurate data with different error tolerance of data error in reproducing kernel Hilbert space.

We assume that H is a Hilbert space over the real numbers with inner product $\langle\cdot,\cdot\rangle$. We denote $\mathbb{N}_n = \{1, 2, ..., n\}$ and choose

$$\mathcal{X} = \{x_j : j \in \mathbb{N}_n\} \tag{1}$$

which is the set of *linearly independent* in H. Consequently, let M be the $n-$dimensional subspace of H spanned by the vectors in $\mathcal{X}$. That is, we have that

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

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

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

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.

Next, we define the linear operator $L : H \to \mathbb{R}^n$ as $x \in H$

$$Lx := \big(\langle x, x_j\rangle : j \in \mathbb{N}_n\big).$$

Consequently, the adjoint operator $L^T : \mathbb{R}^n \to H$ is given by

$$L^T a := \sum_{j\in\mathbb{N}_n} a_j x_j$$

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

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

We point out that G is positive definite which is useful tools. We choose $|\cdot| : \mathbb{R}^n \to \mathbb{R}$ is some norm on $\mathbb{R}^n$ and ε is some positive number. Consequently, we define

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

which contains information about data error. For any $d \in \mathbb{R}^n$, the $hyperellipse$, [10], is defined by

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

where $\delta B = \{x : x \in H, \parallel x \parallel \leq \delta\}$. Clearly, data error is determined by some norm on $\mathbb{R}^n$ and error tolerance. That is, we have that

$$|Lx - d| \leq \varepsilon$$

In the special case that $\varepsilon = 0$, we know that the hyperellipse becomes to $hypercircle$, [4], as shown below

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

That is, we have that hyprecircle is the intersection of a hyperplan of fintie codimension $, Lx = d$, and the closed ball of ratio δ, δB. Therefore, the hypercircle inequality, [4], becomes as the following.

Let $x(d)$ be the element in hyperplane which is nearest the origin. That is, we define

$$x(d) := arg\min\{||x|| : Lx = d\}$$

Moreover, we observe that $x(d) = L^T G^{-1} d$ and

$$||x(d)||^2 = (d, G^{-1}d)$$

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

For any $x \in \mathcal{H}(d,\delta)$ and $x_0 \in H$ then

$$|\langle x(d), x_0\rangle - \langle x, x_0\rangle| \le \text{dist }(x_0, M)\sqrt{\delta^2 - ||x(d)||^2}. \quad (2)$$

Moreover, there is $x_\pm(d) \in \mathcal{H}(d+e)$ for which equality above holds

$$x_\pm(d) = \pm\delta\frac{x_0 - L^T a_\pm}{||x_0 - L^T a_\pm||} \quad (3)$$

where the vector $a_\pm \in \mathbb{R}^n$ is given by the formula

$$a_\pm := G^{-1}(Lx_0 \mp \frac{\text{dist}(x_0, M)}{\sqrt{\delta^2 - ||x(d)||^2}}d). \quad (4)$$

Specifically, we define

$$\text{dist }(x_0, M) = \min\{||x_0 - w|| : w \in M\}.$$

Alternatively, we have the relation between them as the following

$$\mathcal{H}(d|\delta E) = \bigcup_{e\in E}\mathcal{H}(d+e,\delta) \quad (5)$$

Given $x_0 \in H$, our main gold here is to estimate $\langle x, x_0\rangle$ when $x \in \mathcal{H}(d|\delta E)$. According to midpoint algorithm, we then define

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

We point out that $I(x_0, d|\delta E)$ is a closed and bounded subset in $\mathbb{R}$. Therefore, we obtain that

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

where $m_-(x_0, d|\delta E) = \min\{\langle x, x_0\rangle : x \in \mathcal{H}(d|\delta E)\}$ and $m_+(x_0, d|\delta E) = \max\{\langle x, x_0\rangle : x \in \mathcal{H}(d|\delta E)\}$ respectively. Hence, the best estimator is the midpoint of this interval. Moreover, the solution of this problem has the form of linear combination of the vector in $\mathcal{X}$ but the choice of the coefficient are dependent on the vector x_0. That means, there is $e \in E$ such that

$$\langle x(d+e), x_0\rangle = m(x_0, d|\delta E)$$

where $x(d+e) := arg\min\{||x|| : Lx = d+e\}$ and $m(x_0, d|\delta E)$ is the midpoint of the interval $I(x_0, d|\delta E)$.

To this end, we need to evaluate the right and left hand endpoint of the interval $I(x_0, d|\delta E)$. According to our previous work, we give the formula for the right hand endpoint. We recall the conjugate norm of $|\cdot|$ which is defined for all $c \in \mathbb{R}^n$ as

$$|c|_* = \max_{\substack{w\in\mathbb{R}^n \\ |w|\le 1}}(c, w).$$

Moreover, if $c \ne 0$ then there is a $\hat{c} \in \mathbb{R}^n$ such that $|\hat{c}| = 1$ and $|c|_* = (c, \hat{c})$.

If $\mathcal{H}(d,\delta) \ne \phi$ and $|\cdot|_*$ is conjugate norm of $|\cdot|$ then

$$m_+(x_0, d|\delta E) := \min_{c\in\mathbb{R}^n} V_\delta(c) \quad (6)$$

where $V_\delta(c) = \delta||x_0 - L^T c|| + \varepsilon|c|_* + (c, d)$ and $(\cdot,\cdot)$ is Euclidean inner product on $\mathbb{R}^n$.

Therefore, to find the best predictor, we only need evaluate the two numbers $m_+(x_0, \pm d|\delta E)$ and then compute

$$m(x_0, d|\delta E) = \frac{1}{2}\big(m_+(x_0, d|\delta E) - m_+(x_0, -d|\delta E)\big)$$

In the special case that our data is measured with square loss, we assume that

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

and the hyperellipse is given by

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

We found that the minimum of the function V_δ exists with the different hypothesis as following below.

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

$$m_+(x_0, d|\delta E_2) := \min_{c\in\mathbb{R}^n} V_{2,\delta}(c)$$

where $V_{2,\delta}(c) = \delta||x_0 - L^T c|| + \varepsilon|c|_2 + (c, d)$.

For more detail, we refer the reader to the paper [10]. Furthermore, we observe that $\mathcal{H}(d,\delta) \ne \phi$ implies for any data error set E , $\mathcal{H}(d|\delta E)$ contains more than one point and we also provide the example that the infimum is not archived if $\mathcal{H}(d|\delta E)$ contain only one point in paper [10] .

The article is organized as follows: Section II, we improve our result by proving the minimum of the function V_δ in (6) achieve under the same hypothesis in the case of square loss. In section III, we specializes the results of Section II to circumstance for which data error is measured with different norms on $\mathbb{R}^n$ and different error tolerance. Section IV, we will report on further computational experiment of learning the value of a function from inaccurate data with different error tolerance in a reproducing kernel Hilbert space.

II. HYPERCIRCLE INEQUALITY FOR DATA ERROR (HIDE)

The main purpose of this section is to improve the hypothesis how to identify the minimum of the function V_δ in (6). For this propose, we introduce the following terminology.

Form the equation (5), we point out that for each $e \in E$ and $|e| \le \varepsilon$. There is a unique vector

$$x(d+e) \in M$$

which is defined as

$$L^T G^{-1}(d+e) = x(d+e) := arg\min\{||x|| : Lx = d+e\}$$

and

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

Now, let us point out when $\mathcal{H}(d|\delta E) \ne \emptyset$.

Lemma 1: $\mathcal{H}(d|\delta E) \neq \emptyset$ if and only if

$$\min_{|e|\leq\varepsilon}(d+e, G^{-1}(d+e)) \leq \delta^2 \quad (7)$$

Proof. See [10]

In that case that

$$\min_{|e|\leq\varepsilon}(d+e, G^{-1}(d+e)) < \delta^2.$$

if only if $\mathcal{H}(d|\delta E) \neq \emptyset$ contain more than on point. Next, let us provide this lemma before we state the main result of this section.

Lemma 2: If $\mathcal{H}(d|\delta E)$ contains more than one point then there exits $\hat{e} \in E$ and $|\hat{e}| < \varepsilon$ such that $x(d+\hat{e}) \in \mathcal{H}(d|\delta E)$ where $x(d+\hat{e}) = L^T G^{-1}(d+\hat{e})$.

Proof. By our assumption, there exits $e \in E$ such that

$$||x(d+e)|| < \delta \text{ and } x(d+e) = L^T G^{-1}(d+e).$$

That is, we have that

$$||x(d+e)||^2 = (d+e, G^{-1}(d+e)) < \delta^2.$$

Let $\alpha_n \in (0,1)$ and $\alpha_n \to 1$ as $n \to \infty$. We define $e_n = \alpha_n e$ and get that $e_n = \alpha_n e \to e$ as $n \to \infty$. Since E is compact subset of $\mathbb{R}^n$ and the map $e \to (d+e, G^{-1}(d+e))$ is continuous function on E. Hence we obtaint that

$$(d+e_n, G^{-1}(d+e_n)) \to (d+e, G^{-1}(d+e))$$

as $n \to \infty$. Since $(d+e, G^{-1}(d+e)) < \delta^2$, there is

$$\hat{e} = e_n = \alpha_n e \quad \text{and} \quad |\hat{e}| < \varepsilon$$

for some $n \in \mathbb{N}_n$ such that

$$(d+\hat{e}, G^{-1}(d+\hat{e})) < \delta^2 \text{ and } |\hat{e}| < \varepsilon.$$

Hence, we conclude that there is the vector

$$x(d+\hat{e}) = L^T G^{-1}(d+\hat{e}) \in \mathcal{H}(d|\delta E)$$

. □

Theorem 3: Let $|\cdot|$ be a norm on $\mathbb{R}^n$ and $|\cdot|_*$ be conjugate norm of $|\cdot|$. Then

$$(c, w) \leq |c||w|_*$$

and

$$(c, w) \leq |c|_*|w|$$

for all $c, w \in \mathbb{R}^n$.

These facts can be found in [9]. Finally, we recall a useful version of the Von Neumann Minimax Theorem which appears in [2].

Theorem 4: Let $f : \mathcal{C} \times \mathcal{U} \to \mathbb{R}$ where $\mathcal{C}$ is a closed convex subset of a Hausdorff topological vector space U and $\mathcal{U}$ is a convex subset of a vector space Y. If for every $x \in \mathcal{U}$ the function $c \to f(c,x)$ is convex and lower semi-continuous on $\mathcal{C}$ and for every $c \in \mathcal{C}$ the function $x \to f(c,x)$ is concave on $\mathcal{U}$ and there is an $\hat{x} \in \mathcal{U}$ such that for all $\lambda \in \mathbb{R}$ the set

$$\{c : c \in \mathcal{C}, f(c, \hat{x}) \leq \lambda\}$$

is a compact subset of X then there is a $c_0 \in \mathcal{C}$ such that

$$\sup_{x\in\mathcal{U}} f(c_0, x) := \sup_{x\in\mathcal{U}} \inf_{c\in\mathcal{C}} f(c,x).$$

In particular, we have that

$$\min_{c\in\mathcal{C}} \sup_{x\in\mathcal{U}} f(c,x) = \sup_{x\in\mathcal{U}} \inf_{c\in\mathcal{C}} f(c,x)$$

From the hypothesis above, we recall the lower semi-continuity means the set $\{c : c \in \mathcal{C}, f(c,x) \leq \lambda\}$ is a closed subset of $\mathcal{C}$, for all $\lambda \in \mathbb{R}$ and $x \in \mathcal{U}$.

Theorem 5: If $|\cdot| : \mathbb{R}^n \longrightarrow \mathbb{R}_+$ is a norm and $|\cdot|_* : \mathbb{R}^n \longrightarrow \mathbb{R}_+$ its conjugate norm. For any $d \in \mathbb{R}^n$, if $\mathcal{H}(d|\delta E)$ contains more than one point then

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

where the function $V_\delta(c) := \delta||x_0 - L^T c|| + \varepsilon|c|_* + (d,c)$ for all $c \in \mathbb{R}^n$.

Proof. For any $x \in \mathcal{H}(d|\delta E)$, $c \in \mathbb{R}^n$ and $x_0 \in H$ we have that

$$\begin{aligned} \langle x, x_0 \rangle &= \langle x_0 - L^T c, x \rangle + (c, Lx - d) + (c, d). \\ &\leq \delta||x_0 - L^T c|| + \varepsilon|c|_* + (c, d). \end{aligned}$$

Thus, we obtain that

$$\begin{aligned} m_+(x_0, d|\delta E) &= \max_{x\in\mathcal{H}(d|\delta E)} \langle x, x_0 \rangle \\ &\leq \inf_{c\in\mathbb{R}^n} \delta||x_0 - L^T c|| + \varepsilon|c|_* + (c, d). \end{aligned}$$

Moreover, we observer that

$$\inf_{c\in\mathbb{R}^n} \delta||x_0 - L^T c|| + \varepsilon|c|_* + (c,d) = \inf_{c\in\mathbb{R}^n} \max_{x\in\delta B} f(c,x)$$

where the function $f : \delta B \times \mathbb{R}^n \to \mathbb{R}$ is defined at $c \in \mathbb{R}^n$ and $x \in \delta B$ as

$$f(x,c) := \langle x, x_0 - L^T c \rangle + \varepsilon|c|_* + (c,d).$$

According to Theorem VN, we identify $\mathcal{C} = \mathbb{R}^n$ and $\mathcal{U} = \delta B$. For each $x \in \delta B$, we see that the function $c \to f(c,x)$ is convex and lower seimi−continuous on $\mathcal{C}$. For any $x \in \mathcal{U}$, $x \to f(c,x)$ is concave on $\mathcal{U}$. Since $\mathcal{H}(d|\delta E)$ contains more than one point and lemma 2, there is an $\hat{x} \in \delta B$ and $\hat{e} \in \mathbb{R}^n$ with $L\hat{x} = d + \hat{e}$ and $|\hat{e}| < \varepsilon$. For any $c \in \mathcal{C}$, we claim that the set

$$\{c : c \in \mathbb{R}^n, f(c, \hat{x}) \leq \lambda\}$$

is a compact subset of $\mathbb{R}^n$. Clearly, $\{c : c \in \mathbb{R}^n, f(c, \hat{x}) \leq \lambda\}$ is closed subset of $\mathbb{R}^n$. Next, we observe that for all $c \in \mathcal{C}$

$$\begin{aligned} f(c, \hat{x}) &= \langle \hat{x}, x_0 - L^T c \rangle + \varepsilon|c|_* + (c, d) \\ &= \langle \hat{x}, x_0 \rangle + \varepsilon|c|_* - (\hat{e}, c) \leq \lambda \end{aligned}$$

and we obtain that

$$|c|_* \leq \frac{\lambda - \langle \hat{x}, x_0 \rangle}{\varepsilon - |\hat{e}|}$$

Therefore, $\{c : c \in \mathbb{R}^n, f(c, \hat{x}) \leq \lambda\}$ is bounded subset of $\mathbb{R}^n$. Hence, we have the claim. From Theorem VN, we applies and have that

$$\begin{aligned} &\min_{c\in\mathbb{R}^n} \delta||x_0 - L^T c|| + \varepsilon|c|_* + (c, d) \\ &= \sup_{x\in\delta B} \inf_{c\in\mathbb{R}^n} \langle x_0 - L^T c, x\rangle + \varepsilon|c|_* + (c, d). \\ &= \sup_{x\in\delta B} \inf_{c\in\mathbb{R}^n} \{\varepsilon|c|_* + (c, d - Lx). \end{aligned}$$

Next, we claim that

$$\inf_{c\in\mathbb{R}^n} \varepsilon|c|_* + (c, d - Lx) = \left\{ \begin{array}{ll} -\infty, & |Lx - d| > \varepsilon \\ 0, & |Lx - d| \leq \varepsilon. \end{array}\right. \quad (8)$$

First, we consider $|Lx - d| \leq \varepsilon$ then we have

$$\begin{array}{rcl} \varepsilon|c|_* + (c, d - Lx) & \geq & \varepsilon|c|_* - |c|_*|Lx - d| \\ & = & |c|_*(\varepsilon - |Lx - d|) \\ & \geq & 0 \end{array}$$

Therefore, the infimum on the left hand side is achieved for $c = 0$. If $|Lx - d| > \varepsilon$ then we choose $\hat{c} \in \mathbb{R}^n \setminus \{0\}$ so that

$$(\hat{c}, d - Lx) = |Lx - d|\ |\hat{c}|_*$$

Hence, for any $t > 0$ we have that

$$\begin{array}{rcl} \inf\limits_{c\in\mathbb{R}^n} \varepsilon|c|_* + (c, d - Lx) & \leq & \varepsilon| - t\hat{c}|_* + (-t\hat{c}, d - Lx) \\ & = & \varepsilon t|\hat{c}|_* - t|Lx - d||\hat{c}|_* \\ & = & t(\varepsilon - |Lx - d|)|\hat{c}|_* \end{array}$$

Therefore,

$$\inf_{c\in\mathbb{R}^n} \varepsilon|c|_* + (c, d - Lx) = -\infty$$

when $t \to \infty$. □

Let us add one remark which ensures that V_δ has a unique minimum. If $\mathcal{H}(d|\delta E)$ contains more than one point and either $x_0 \notin M$ or $|\cdot|_*$ is strictly convex then V_δ has a unique minimum.

III. Hide with Different Error Tolerance of Data Error

In this section, we specializes the results of section II to circumstance for which there is known different error tolerance of data error. We denote I which is the subset of $\mathbb{N}_n$ and we let m be the number of element in the set I. Consequently, we denote $J = \mathbb{N}_n \backslash I$.

For each $e = (e_1, ..., e_n) \in \mathbb{R}^n$, we define the notation

$$e_I = (e_i : i \in I) \text{ and } e_J = (e_i : i \in J).$$

We defined $|\cdot| : \mathbb{R}^m \longrightarrow \mathbb{R}_+$ and $|||\cdot||| : \mathbb{R}^{n-m} \longrightarrow \mathbb{R}_+$ be norms on $\mathbb{R}^m$ and $\mathbb{R}^{n-m}$ respectively. We assume that

$$\mathbb{E} = \{e : e \in \mathbb{R}^n, |e|_\infty \leq 1\}$$

where we define $|\cdot|_\infty$ as following

$$|e|_\infty = \max\{\frac{1}{\varepsilon}|e_I|, \frac{1}{\varepsilon'}|||e_J|||\}.$$

where $\varepsilon, \varepsilon' > 0$. In this case, we define the *hyperellipse* in the following ways

$$\mathcal{H}(d|\delta\mathbb{E}) = \{x : x \in \delta B, |Lx - d|_\infty \leq 1\}$$

That is, for each $x \in \mathcal{H}(d|\delta\mathbb{E})$

$$|(Lx - d)_I| \leq \varepsilon$$

and

$$|||(Lx - d)_J||| \leq \varepsilon'.$$

First, let us begin by discussing when $\mathcal{H}(d|\delta\mathbb{E}) \neq \phi$. Since G is positive definite, we then assume that

$$0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$$

are eigenvalues of G^{-1}. Consequently. we let $\{u^j : j \in \mathbb{N}_n\}$ is a corresponding orthonormal set of eigenvector and write the vector d in the form

$$d = \sum_{j\in\mathbb{N}_n} \gamma_j u^j$$

for some constants $\gamma_j \in \mathbb{R}$. We denote the vector

$$d^*_I = (\gamma_i : i \in I) \quad \text{and} \quad d^*_J = (\gamma_i : i \in J) \quad (9)$$

respectively. For any $e = \sum\limits_{j\in\mathbb{N}_n} e_j u^j \in \mathbb{E}$, we obtain that

$$\begin{aligned} &(d + e, G^{-1}(d + e)) = \sum_{i\in\mathbb{N}_n} (\gamma_i + e_i)^2\lambda_i \\ &= \sum_{i\in I}(\gamma_i + e_i)^2\lambda_i + \sum_{i\in J}(\gamma_i + e_i)^2\lambda_i \\ &= (d^*_I + e_I, D_I(d^*_I + e_I)) \quad + (d^*_J + e_J, D_J(d^*_J + e_J)) \end{aligned}$$

where D_I and D_J are m and $n - m$ dimensional diagonal matrices when the element on diagonal of D_I and D_J are λ_i for $i \in I$ and $i \in J$ respectively.

As we show above, we then obtain the following fact.

Theorem 6: $\mathcal{H}(d|\delta\mathbb{E}) \neq \phi$ if and only if

$$(d^*_I + \hat{e}_I, D_I(d^*_I + \hat{e}_I)) + (d^*_J + \hat{e}_J, D_J(d^*_J + \hat{e}_J)) \leq \delta^2$$

where we define

$$\min_{\substack{e\in\mathbb{R}^m \\ |e|\leq\varepsilon}} (d^*_I + e, D_I(d^*_I + e)) = (d^*_I + \hat{e}_I, D_I(d^*_I + \hat{e}_I))$$

$$\min_{\substack{e\in\mathbb{R}^{n-m} \\ |||e|||\leq\varepsilon'}} (d^*_J + e, D_I(d^*_J + e)) = (d^*_J + \hat{e}_J, D_J(d^*_J + \hat{e}_J))$$

As we said above, our main gold here is to estimate $\langle x, x_0\rangle$ when $x \in \mathcal{H}(d|\delta\mathbb{E})$. That is, our data is generally measured with different norm and error tolerance. Using midpoint algorithm, we then define

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

where $m_-(x_0, d|\delta\mathbb{E}) = \min\{\langle x, x_0\rangle : x \in \mathcal{H}(d|\delta\mathbb{E})\}$ and $m_+(x_0, d|\delta\mathbb{E}) = \min\{\langle x, x_0\rangle : x \in \mathcal{H}(d|\delta\mathbb{E}\}$ respectively. We need the following theorem to obtain the best predictor.

According to theorem 5, we found that the conjugate norm of $|\cdot|_\infty$ is given by for each $e \in \mathbb{R}^n$

$$|e|_1 = \varepsilon|e_{_I}|_* + \varepsilon'|||e_{_J}|||_*$$

where $|\cdot|_*$ and $|||\cdot|||_*$ are conjugate norm of $|\cdot|$ and $|||\cdot|||$ respectively.

Theorem 7: Let $|\cdot|_*$ and $|||\cdot|||_*$ be the conjugate norms of $|\cdot|$ and $|||\cdot|||$ respectively. If $\mathcal{H}(d|\delta\mathbb{E})$ contains more than one point then

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

where the function

$$\mathbb{V}_\delta(c) := \delta||x_0 - L^T c|| + \varepsilon|c_{_I}|_* + \varepsilon'|||c_{_J}|||_* + (d, c)$$

for all $c \in \mathbb{R}^n$. Moreover, either $x_0 \notin M$ or $|\cdot|_*$ or $|||\cdot|||_*$ is strictly convex then it has a unique minimum.

Next, we discuss the special case that our data error is measured with square loss. We define

$$\mathbb{E}_2 = \{e : e \in \mathbb{R}^n, |e|_\infty \leq 1\}$$

where $|\cdot|_\infty$ is a norm on $\mathbb{R}^n$ which is defined by

$$|e|_\infty = \max\{\frac{1}{\varepsilon}|e_{_I}|_2, \frac{1}{\varepsilon'}|||e_{_J}|||_2\}.$$

We use this notation for hyperelipse

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

Consequently, we have that

$$|(Lx-d)_{_I}|_2 \leq \varepsilon \text{ and } |||(Lx-d)_{_J}|||_2 \leq \varepsilon'.$$

In this case, we have the formula for checking when $\mathcal{H}(d|\delta\mathbb{E}_2) \neq \phi$. Form the equation (5), we will show the sufficient values of δ such that

$$\mathcal{H}(d+e, \delta) \neq \phi$$

for each $e \in \mathbb{E}_2$. That is, we obtain that

$$x(d+e) = L^T G^{-1}(d+e) \in \mathcal{H}(d|\delta\mathbb{E})$$

for each $e \in \mathbb{E}_2$. We begin with the following terminology.

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

$$A(x - d) = \lambda x. \tag{10}$$

Theorem 9: The spectrum of the pair $(\varepsilon^2 D_{_I}, \frac{d^*_{_I}}{\varepsilon})$ consists of all real λ such that

$$g(\lambda) = \sum_{i\in\mathbb{I}} \frac{\lambda_i^2\gamma_i^2}{(\varepsilon^2\lambda_i - \lambda)^2} = 1 \tag{11}$$

togerther with each eigenvalue λ_k of $\varepsilon^2 D_{_I}$ for which $g(\lambda_k) < 1$ where $k \in \mathbb{I} = \{i : i \in I, \lambda_i\gamma_i = 0\}$.
Proof. See [5]

Accoring to theorem 6, we obtain the following theorem.
Theorem 10: If $\Lambda_{_I}$ and $\Lambda'_{_I}$ are the least and greates value in the spectrum of the pair $(\varepsilon^2 D_{_I}, \frac{d^*_{_I}}{\varepsilon})$ respectively. Then we have the following

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

and

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

Proof. See [5]

Theorem 11: Let $\Lambda_{_I}$ and $\Lambda_{_J}$ be the least value in the spectrum of the pair $(\varepsilon^2 D_{_I}, \frac{d^*_{_I}}{\varepsilon})$ and $(\varepsilon'^2 D_{_J}, \frac{d^*_{_J}}{\varepsilon'})$ respectively. If $|d^*_{_I}| > \varepsilon$ and $|||d^*_{_J}||| > \varepsilon'$ then $\mathcal{H}(d|\delta\mathbb{E}_2) \neq \phi$ if and only if

$$\Lambda_{_I} + \Lambda_{_J} + \Lambda_{_I}\sum_{j\notin\mathbb{I}} \frac{\lambda_j|\gamma_j|^2}{\Lambda_{_I} - \varepsilon^2\lambda_j} + \Lambda_{_J}\sum_{j\notin\mathbb{J}} \frac{\lambda_j|\gamma_j|^2}{\Lambda_{_J} - \varepsilon^2\lambda_j} \leq \delta^2$$

where we denote $\mathbb{I} = \{j : j \in I, \lambda_j\gamma_j = 0\}$ and
$\mathbb{J} = \{j : j \in J, \lambda_j\gamma_j = 0\}$.

Proof. See [5]

Before we state the next theorem, let us define

$$\mathbb{M} := \{x(d+e) = L^T G^{-1}(d+e) : e \in \mathbb{E}_2\}.$$

Theorem 12: Let $\Lambda'_{_I}$ and $\Lambda'_{_J}$ be the greatest value in the spectrum of the pair $(\varepsilon^2 D_{_I}, \frac{d^*_{_I}}{\varepsilon})$ and $(\varepsilon'^2 D_{_J}, \frac{d^*_{_J}}{\varepsilon'})$ respectively. If $|d^*_{_I}| > \varepsilon$ and $|||d^*_{_J}||| > \varepsilon'$ then

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

if and only if

$$\Lambda'_{_I} + \Lambda'_{_J} + \Lambda'_{_I}\sum_{j\notin\mathbb{I}} \frac{\lambda_j|\gamma_j|^2}{\Lambda'_{_I} - \varepsilon^2\lambda_j} + \Lambda'_{_J}\sum_{j\notin\mathbb{J}} \frac{\lambda_j|\gamma_j|^2}{\Lambda'_{_J} - \varepsilon^2\lambda_j} \leq \delta^2 \tag{12}$$

Proof. See [5]

Therfore, if we choose δ which is satisfied equation (12) then $\mathbb{M}$ is smallest subset of H which contain the best predictor to estimate the value of $\langle x, x_0\rangle$ when $x \in \mathcal{H}(d|\delta\mathbb{E}_2)$.

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

$$m_+(x_0, d|\delta\mathbb{E}_2) = \min_{c\in\mathbb{R}^n} \mathbb{V}_{2,\delta}(c)$$

where $\mathbb{V}_{2,\delta}(c) := \delta||x_0 - L^T c|| + \varepsilon|c_I|_2 + \varepsilon'|||c_J|||_2 + (d, c)$. Moreover, the function $\mathbb{V}_{2,\delta}$ has a unique mimimum.

We specialize this results to the case that $\mathcal{X}$ is an othonormal set of vector. First, we point out that the Gram matrix is the identity matrix. Consequently, all eigenvalue of this case are 1 and $\{u^1, u^2, ..., u^n\}$ is standart basis of $\mathbb{R}^n$ which is the eigenvector corresponding to eigenvalue. In our case, we write the vector d in the folloing way

$$d = \sum_{j\in\mathbb{N}_n} d_j u^j$$

The vector d^* in equation (9) are identified with $d^*_I = d_I$ and $d^*_J = d_J$ respectively. Consequently, we obtain the spectrum of the order pair $(\varepsilon^2 D_I, \frac{d^*_I}{\varepsilon})$ and $(\varepsilon'^2 D_J, \frac{d^*_J}{\varepsilon'})$ when we obtain form the euation (11) as the following

$$\Big\{\varepsilon \pm \sqrt{\sum_{i\notin\mathbb{I}} d_i^2}\Big\} \quad \text{and} \quad \Big\{\varepsilon' \pm \sqrt{\sum_{i\notin\mathbb{J}} d_i^2}\Big\}.$$

If $\displaystyle\sum_{i\notin\mathbb{I}} \frac{d_i^2}{(\varepsilon-1)^2} < 1$ and $\displaystyle\sum_{i\notin\mathbb{J}} \frac{d_i^2}{(\varepsilon-1)^2} < 1$ then $\lambda_k = 1$ is also the spectrum of $(\varepsilon^2 D_I, \frac{d^*_I}{\varepsilon})$ and $(\varepsilon'^2 D_J, \frac{d^*_J}{\varepsilon'})$. According to Theorem 4, we obtain

$$\Lambda_I = \min\{\varepsilon \pm \sqrt{\sum_{i\notin\mathbb{I}} d_i^2}, 1\} \text{ and } \Lambda'_I = \max\{\varepsilon \pm \sqrt{\sum_{i\notin\mathbb{I}} d_i^2}, 1\}$$

$$\Lambda_J = \min\{\varepsilon \pm \sqrt{\sum_{i\notin\mathbb{J}} d_i^2}, 1\} \text{ and } \Lambda'_J = \max\{\varepsilon \pm \sqrt{\sum_{i\notin\mathbb{J}} d_i^2}, 1\}$$

where we denote

$$\mathbb{I} = \{j : j \in I, d_j = 0\} \text{ and } \mathbb{J} = \{j : j \in J, d_j = 0\}.$$

To this end, we add final remark to the case $x_0 \in \mathcal{X}$. That is, we assume that $x_0 = x_i$ for some $i \in \mathbb{N}_n$. We obtain the uncertainty interval as the following

$$\begin{aligned} I(x_0, d|\mathbb{E}_2) &:= \{\langle x(d+e), x_0\rangle : x(d+e) \in \mathcal{H}(d|\delta\mathbb{E}_2)\} \\ &= \{\langle L^T(d+e), x_0\rangle : x(d+e) \in \mathcal{H}(d|\delta\mathbb{E}_2)\} \\ &= \{((d+e), Lx_0) : (d+e, d+e) \le \delta, e \in \mathbb{E}_2\} \\ &= \{d_i + e_i : (d+e, d+e) \le \delta, e \in \mathbb{E}_2\} \end{aligned}$$

If $\mathbb{M} \subseteq \mathcal{H}(d|\delta\mathbb{E}_2)$ then the midpoint is given by d_i.

IV. Numerical Experiments

In this section, we shall report some results of numerical experiment in learning the value of a function in reproducing kernel Hilbert spapce $(RKHS)$ by the midpoint algorithm when data error measured with square loss and its have different error tolerance.

Let H_K be the reproducing kernel Hilbert space of real valued function on a set $\mathcal{T}$. That is, for any element $f \in H_K$ we have that $f : \mathcal{T} \to \mathbb{R}$. The real function $K(t, s)$ 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 defined for any $s \in \mathcal{T}$ as $K_t(s) := K(t, s)$. The Aronszajn-Moore theorem,[1], gives an *intrinsic* characterization of reproducing kernels K for a RKHS. This result states that K is a reproducing kernel for some RKHS if and only if for any inputs $T = \{t_j : j \in \mathbb{N}_n\}$ the $n \times n$ matrix $(K(t_i, t_j) : i, j \in \mathbb{N}_n)$ is positive semi-definite. Moreover, for any kernel K there is a unique RKHS with K as its reproducing kernel.

In the frist computaional experiment, we choose the gaussian kernel on $\mathbb{R}$. Our example below is organzied in the flowing way. Frist, we choose $\{t_j : j \in \mathbb{N}_{20}\} \subseteq \mathbb{R}$ and $t_0 = 0$. Consequently, we get the gram matrix is given as $G(K(t_i, t_j))_{i,j\in\mathbb{N}_{20}}$. So far, $\mathcal{X}$ in equation (1) is identified with the function $\{K_{t_j} : j \in \mathbb{N}_{20}\} \subseteq H_K$ and the vector x_0 with the function K_{t_0}. Next, we choose $g \in H_K$ and define $d = (g(t_j) + e_j : j \in \mathbb{N}_{20})$ where we choose the vector e representing "noise" which is separated into two groups with different error tolerance. Consequently, the hyperellipse becomes

$$\begin{aligned} \mathcal{H}(d|\delta\mathbb{E}_2) \quad &= \Big\{f : ||f|| \le \delta, (\sum_{j\in I}(f(t_j) - d_j)^2)^{\frac{1}{2}} \le \varepsilon_I, \\ &\quad (\sum_{j\in J}(f(t_j) - d_j)^2)^{\frac{1}{2}} \le \varepsilon_J\Big\} \end{aligned}$$

We start by compute the sufficient the value of δ such that $\mathcal{H}(d|\delta\mathbb{E}_2) \neq \phi$ which is obtained by theorem 6 and 11. As midpoint algorithm and Theorem 5, we need to find numerically the minimum of the function $\mathbb{V}_{\delta\pm}$ which is defined for $c \in \mathbb{R}$ as $\mathbb{V}_{\delta\pm}(c) = \delta\sqrt{K(t_0, t_0) - 2\sum_{j\in\mathbb{N}_n} c_j K(t_0, t_j) + \sum_{i,j\in\mathbb{N}_n} c_i c_j K(t_i, t_j)} + \varepsilon_I\sqrt{\sum_{j\in I} c_j^2} + \varepsilon_J\sqrt{\sum_{j\in J} c_j^2} \pm \sum_{j\in\mathbb{N}_n} c_j d_j$.

Threfore, we obtain the midpoint

$$m(x_0, d|\delta E) = \frac{1}{2}\big(\mathbb{V}_{\delta+}(c^*_+) - \mathbb{V}_{\delta-}(c^*_-)\big)$$

where we denote

$$m_+(x_0, d|\delta E) = \mathbb{V}_{\delta+}(c^*_+) := \min\{\mathbb{V}_{\delta+}(c) : c \in \mathbb{R}^{20}\}$$

and

$$m_+(x_0, -d|\delta E) = \mathbb{V}_{\delta-}(c^*_-) := \min\{\mathbb{V}_{\delta-}(c) : c \in \mathbb{R}^{20}\}.$$

We use the program fiminunc in the optimization toolbox of Matlab 7.3.0 to obtain the minimum of the function $\mathbb{V}_{\delta\pm}$

Example 1 We choose the gaussian kernel on $\mathbb{R}$ which is defined by

$$K(t,s) := e^{-\frac{(t-s)^2}{10}} \quad t,s \in \mathbb{R}.$$

We choose the exact function

$$\begin{aligned} g(t) \;\; &:= \;\; 4e^{-\frac{(t-7.5)^2}{10}} + 2e^{-\frac{(t-2.5)^2}{10}} \\ & \quad\; -0.5e^{-\frac{(t+2.5)^2}{10}} + 5e^{-\frac{(t+7.5)^2}{10}} \end{aligned}$$

and choose $t_0 = 0$. We generated a training set of twenty points $T = \{(t_j, d_j) : j \in \mathbb{N}_{20}\} \subseteq \mathbb{R} \times \mathbb{R}$ obtained by the function g. Specifically, we choose $t_1 = -20, t_{j+1} = t_j + 2$ and $t_{11} = 2, t_{j+11} = t_{10+j} + 2$, for all $j \in \mathbb{N}_9$. We set $d_j = g(t_j) + e_j$, $j \in \mathbb{N}_{20}$ where

$$e_i \in \begin{cases} (0.1, 0.2) & \text{,if } i \in I = \{1,2,...,5,16,...,20\} \\ (0, 0.01) & \text{,if } i \in J = \{6,7,...,15\} \end{cases}$$

Consequently, we choose $\varepsilon_{_I} = 0.75$ and $\varepsilon_{_J} = 0.09$ respectively.

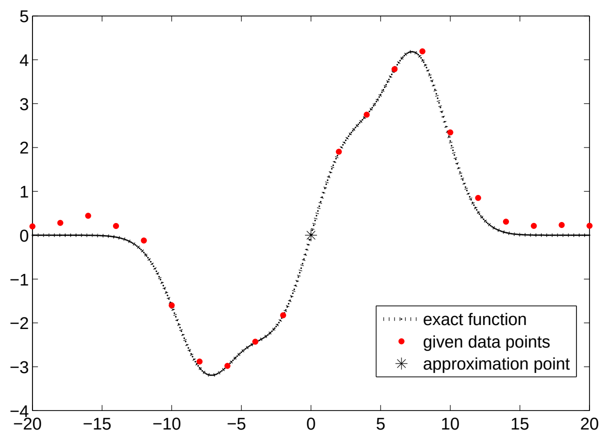

Fig. 1. Exact function obtained from guassian kernel on $\mathbb{R}$

By Theorem 6, 11 and 12, we compute and obtain that

$$\min_{\substack{e\in\mathbb{R}^{10} \\ |e|_2 \le \varepsilon_1}} (d^*_{_I} + e, D_{_I}(d^*_{_I} + e)) = 31.4079$$

$$\min_{\substack{e\in\mathbb{R}^{10} \\ 2|e|_2 \le \varepsilon_2}} (d^*_{_J} + e, D_{_I}(d^*_{_J} + e)) = 8.6115$$

and

$$\max_{\substack{e\in\mathbb{R}^{10} \\ |e|_2 \le \varepsilon_1}} (d^*_{_I} + e, D_{_I}(d^*_{_I} + e)) = 71.2428$$

$$\max_{\substack{e\in\mathbb{R}^{10} \\ 2|e|_2 \le \varepsilon_2}} (d^*_{_J} + e, D_{_I}(d^*_{_J} + e)) = 10.0067$$

Consequently, we have that

$$\min_{e\in\mathbb{E}_2} (d + e, G^{-1}(d + e)) = 40.0194$$

and

$$\max_{e\in\mathbb{E}_2} (d + e, G^{-1}(d + e)) = 81.2495$$

Therefore, we obtain that $\mathcal{H}_2(d|\delta E_\infty) \neq \phi$ if and only if we choose $\delta > 6.3261$. According to theorem 6, we obtain that $\mathbb{M} \subseteq \mathcal{H}(d|\delta\mathbb{E}_2)$ if and only if $\delta > 9.0139$. Therefore, we have the uncertainty interval as the following

$$\begin{aligned} I(t_0, d|\delta E) \;\; &= \;\; \{\langle K_{t_0}, f\rangle : f \in \mathcal{H}(d|\delta\mathbb{E}_2)\} \\ &= \;\; \{f(0) : f \in \mathcal{H}(d|\delta\mathbb{E}_2)\} \end{aligned}$$

The result of the computation is indicate in Figure 2. while the exact value $g(0) = 0.7993$.

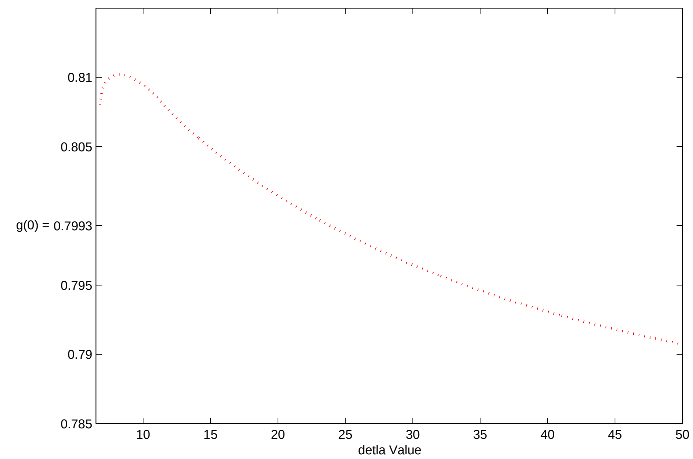

Fig. 2. Midpoint estimator from Gaussian kernel on $\mathbb{R}$

Example 2 We choose the gaussian kernel on $\mathbb{R}^2$. That is, we define

$$K(x,y) := e^{-|x-y|_2^2} \quad x, y \in \mathbb{R}^2.$$

We choose the exact function

$$g := 0.5K_{(1,1)} + 2.75K_{(1,-1)} - 3.25K_{(-1,-1)} + 1.5K_{(-1,1)}$$

where we define $K_x(y) = K(x,y)$. We choose the values on $T = \{t_i : i \in \mathbb{N}_{20}\}$ on a spiral curve surrounding the origin as show in Figure 3 .

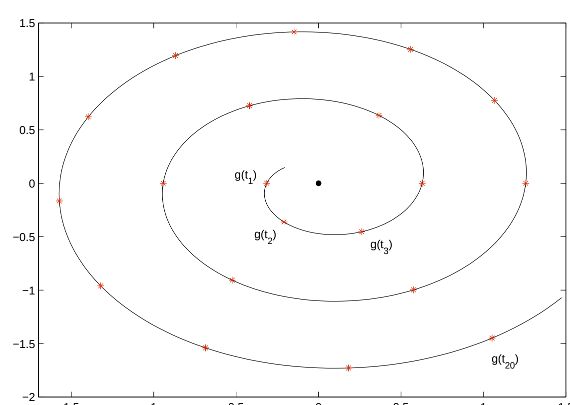

Fig. 3. Point on spiral

We set $d_j = g(t_j) + e_j$, $j \in \mathbb{N}_{20}$ where

$$e_i \in \begin{cases} (0.1, 0.2) & \text{,if } i \in I_1 = \{1, 2, ..., 5, \} \\ (0.09, 0.1) & \text{,if } i \in I_2 = \{6, 7, ..., 10\} \\ (0.01, 0.02) & \text{,if } i \in I_3 = \{11, 12..., 15\} \\ (0.009, 0.01) & \text{,if } i \in I_4 = \{16, 17, ..., 20\} \end{cases}$$

Consequently, we choose $\varepsilon_i = \begin{cases} 0.0361 & \text{,if } i = 1 \\ 0.2135 & \text{,if } i = 2 \\ 0.3489 & \text{,if } i = 3 \\ 0.0214 & \text{,if } i = 4 \end{cases}$

and we define hyperellipse as the following

$$\mathcal{H}(d|\delta\mathbb{E}_2) = \{f : ||f|| \leq \delta, (\sum_{j \in I_i} (f(t_j) - d_j)^2) \leq \varepsilon_i^2, i \in \mathbb{N}_4\}$$

From Theorem 13 and choose $t_0 = 0$, we obtain the function $\mathbb{V}_{2,\delta}$ as follows

$$\mathbb{V}_{2,\delta}(c) = \delta \sqrt{1 - 2 \sum_{j \in \mathbb{N}_{20}} c_j K(t_0, t_j) + \sum_{i,j \in \mathbb{N}_{20}} c_i c_j K(t_i, t_j)}$$

$$+\varepsilon_1 \sqrt{\sum_{j \in I_1} c_j^2} + \varepsilon_2 \sqrt{\sum_{j \in I_2} c_j^2} + \varepsilon_3 \sqrt{\sum_{j \in I_3} c_j^2} + \varepsilon_4 \sqrt{\sum_{j \in I_4} c_j^2} + \sum_{j \in \mathbb{N}_{20}} c_j d_j.$$

I_i	Λ_{I_i}	Λ'_{I_i}	minimun	maximum
I_1	-0.0779	0.0789	14.8174	15.131
I_2	-0.4838	0.6237	3.9087	6.1205
I_3	-0.3810	1.2769	0.4273	3.6079
I_4	-0.0537	0.1380	0.2635	0.6102

TABLE I
THE MAXIMUM AND MINIMUM OF A POSITVE DEFINITE QUADRATIC POLYNOMIAL ON A SHERE

Consequently, we have that

$$\min_{e \in \mathbb{E}_2} (d + c, G^{-1}(d + c)) = 19.4169$$

and

$$\max_{e \in \mathbb{E}_2} (d + c, G^{-1}(d + c)) = 25.4696$$

Therefore, we obtain that $\mathcal{H}_2(d|\delta E_\infty) \neq \phi$ if and only if we choose $\delta > 4.4065$. Moreover, we also obtain that for each $e \in \mathbb{E}_2$ $\mathcal{H}(d + e) \neq \phi$ if and only if $\delta > 5.0467$. Again, the result of the computation is indicate in Figure 4 while the exact value $g(0) = 1.0827$.

V. Conclusion

In this paper, we have improved the recent work on hypercircle inequality for data error. Indeed, we have improved the hypothesis of nonlinear optimization problem which is used to obtain the best predictor. Morover, we have provided material about the Hypercircle inequality for data error which have different error tolerance and reported further numerical experiment in Section 4.

Acknowledgements:

This research is supported by the Centre of Excellence in Mathematics, the Commission on Higher Education, Thailand.

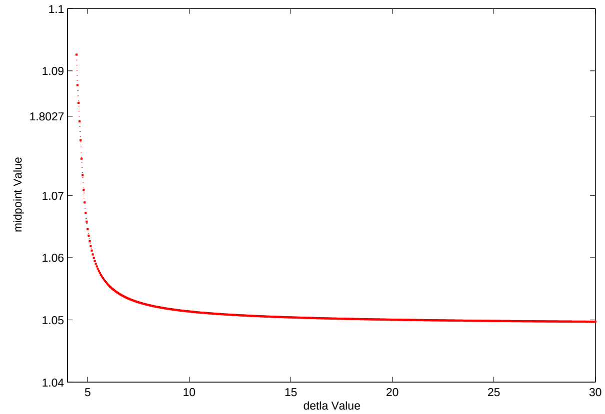

Fig. 4. Midpoint estimator from Gaussian on $\mathbb{R}^2$

References

[1] N. Aronszajn,"Theory of reproducing kernels", *Trans. Amer, Math. Soc.* 686, 1950, pp. 337–404.

[2] J. P. Aubin, "Mathematical Methods of Game and Economic Theory, Studies in Mathematics and its Applications", *Vol 7, North Holland* 1982

[3] 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.

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

[5] George E. Forsythe and Gene H. "Golub,On the stationary values of a second-degree polynomial on the unit sphere" , *SIAM J. Appl. Math. V.13* 1965, pp. 1050-1068.

[6] M. Golomb and H. F. Weinberger, " Optimal approximation and error bounds", *In R. E. Langer, editor, The University of Wisconsin Press*. 1959, pp. 117-190.

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

[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*. 4(2005), 7681.

[9] R. A. Horn and C. R. Johnson, " Matrix Analysis", *Cambridge University Press*. 1985.

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

[11] C.A. Micchelli and A. Pinkus, "Variational problem arising from balancing several error criteria", *Rendiconti di Mathematica*. Serie VII,14Roma, 1994, pp. 37-86.

[12] C.A Micchelli and T. J Rivlin," A survey of optimal recovery", *Optimal Estimation in Approximation Theory*. C.A Micchelli, and T. J Rivlin, eds., Plenum Press, 1977, pp. 1-53.

[13] 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.

[14] B. Scholkopf and A. J. Smola," Learning with Kernels", *The MIT Press, Cambridge, MA, USA*. 2002.

[15] J. Shawe-Taylor and N. Cristianini, "Kernel Methods for Pattern Analysis", *Cambridge University Press*. 2004.

[16] E. Spjotvoll,"A note on a theorem of Forsythe and Golub", *SIAM J. Appl. Math*. 3rd edition. 23, 1972, pp. 307-311.

[17] I. Steinwart and D. Hush and C. Scovel, " An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels", IEEE, *Transactions on Information Theory*. 52, 2006, pp. 4635-4643.