The Use of Hide in Learning the Value of a Function

K. KHOMPURNGSON[1,4], D. POLTEM[2,4], A. YAMARAT[3], B. NOVAPRATEEP[3,4*]

[1]School of Science and Technology, Phayao University, THAILAND
[2]Kanchanaburi Campus, Mahidol University, THAILAND
[3]Dept of Mathematics, Faculty of Science, Mahidol University, THAILAND
[4]Centre of Excellence in Mathematics, PERDO, CHE, THAILAND
[*]Corresponding author: scbnv@mahidol.ac.th

Abstract: In this paper, we briefly review some recent work on 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. We 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. Moreover, we compare our numerical experiment to the method of regularization, which is the standard method for learning problem. We also discuss the effect of the values of δ on the learning task under consideration.

Key–Words: Hypercircle inequality, Reproducing kernel Hilbert space, Regularization, Convex optimization and noise data

1 Introduction

In this paper, we briefly review some recent work on 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. We 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. Moreover, we compare our numerical experiment to the method of regularization, which is the standard method for learning problem. We also discuss the effect of the values of δ on the learning task under consideration.

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 $K_t(s) = K(t, s)$. 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 semidefinite. 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 $Qf := (f(t_i) = \langle f, K_{t_i} \rangle : i \in \mathbb{N}_n)$ and $|\cdot|_2$ is a euclidean norm on $\mathbb{R}^n$. The standard method for learning $f(t_0)$ is the method of regularization. Given $\rho > 0$, we choose the function which minimizes the R_ρ functional defined for $f \in H$ as

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

According to the Representer Theorem [4, 10, 11, 14], 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 the 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 viewed 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 is to consider a bigger value of $\varepsilon(\rho)$ and $\delta(\rho)$. For this reason, we shall discuss and continue to report some results from numerical experiments of learning the value of a function in RKHS by midpoint algorithm with different values of δ in section 3. In section 2, we briefly review Hypercircle inequality for data error measured with square loss and discuss what we need for section 3.

2 Hypercircle inequality for data error

In this section we begin with a Hilbert space H over the real numbers 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 from H onto $\mathbb{R}^n$, which is defined for any $x \in H$ as

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

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

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

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

$$G = QQ^T = (\langle x_j, x_l \rangle : j, l \in \mathbb{N}_n) \tag{5}$$

which is symmetric and positive definite. Now, let us describe Hypercircle inequality for data error $(Hide)$. We provide it in the case that the data error is measured with the euclidean norm. We refer the readers to the paper [7] for more information about the proof of Hide measured with any norm on $\mathbb{R}^n$.

Definition 1 *Let H be the Hilbert space over the real numbers 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 the subset of H, which is defined by*

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

Thus, we begin this section by discussing when $\mathcal{H}_2(d|E(\delta)) \neq \emptyset$.

Lemma 2 *$\mathcal{H}_2(d|E(\delta)) \neq \emptyset$ if and only if*

$$\min_{|e|_2 \leq 1} (d + \varepsilon e, G^{-1}(d + \varepsilon e)) \leq \delta \tag{6}$$

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

Next, we want to find the best estimator to optimally estimate one feature of an $x \in \mathcal{H}_2(d|E(\delta))$ when we define a feature of $x \in H$ as the value of 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 a convex subset of H which is sequentially compact in the weak topology on H, we obtain the uncertainty set that 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. Next, we observe that

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

Then, we only need to evaluate the two numbers $m_+(x_0, \pm d \mid E((\delta)))$ and then compute the midpoint $m_\delta(x_0, d|E) = \frac{1}{2}(m_+(x_0, d \mid E(\delta)) - m_+(x_0, -d \mid E(\delta))$.

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 $c \in \mathbb{R}^n$

$$V_\delta(c) := \delta||x_0 - Q^T c|| + \varepsilon|c|_2 + (c, d). \tag{7}$$

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

Theorem 3 *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) $\dfrac{\delta x_0}{||x_0||} = \arg\ \max\{\langle x, x_0 \rangle : x \in \mathcal{H}_2(d|E(\delta))\}$.

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

Theorem 4 *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*

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

3 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 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 this method to the midpoint algorithm, we then define the interval of uncertainty

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

Since there is only one $f_\rho(t_0)$ in $I(t_0, \varepsilon_\rho, \delta_\rho)$, our strategy in comparing the regularization and midpoint estimator, is to consider a bigger value of $\varepsilon(\rho)$ and $\delta(\rho)$. We choose $\varepsilon = \varepsilon(\rho)$ and $\delta = \alpha\delta(\rho)$ where α is in $A = \{1.5, 3, 6, 12, 24\}$. 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 $T_0 = \{t_{-j} : j \in \mathbb{N}_k\}$ for some $k \in \mathbb{N}$ 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 method $f_\rho(t_{-j})$ and midpoint algorithm $m_\delta(t_{-j})$ 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

$$e_m(T_0, d|E(\delta)) = \sum_{j\in\mathbb{N}_k} |g(t_{-j}) - m_\delta(t_{-j})|^2$$

and

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

For the computation $m_+(x_0, \pm d|E(\delta))$, we use the program fminunc in the optimization tool box of MATLAB 7.3.0. The results of sum square error are shown in Tables 1 and 2 for both of the learning approaches.

3.1 Experiment 1

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

$$K(t, s) = K_s(t) = exp(-\frac{(t - s)^2}{50}) \quad t, s \in \mathbb{R} \quad (8)$$

and the function g is chosen to be

$$g(t) = K_0(t) + 15K_{2.7}(t) - K_{4.7}(t). \quad (9)$$

The set T consists of 20 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 25 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 $t_{-j} \in T_0$.

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

Table 1: The sum square error obtained from Gaussian kernel for two methods for different values of the regularization parameter ρ.

Our computation above shows 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 1 presents the sum square errors between the exact function and the function learned from the regularization method and the midpoint algorithm.

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 $\alpha \in A = \{1.5, 3, 6, 12, 24\}$.

3.2 Experiment 2

In our second experiment, we choose the exact function

$$g(t) = K_0(t) - \frac{1}{2}K_{\frac{1}{2}}(t) - K_{-\frac{1}{3}}(t) \qquad (10)$$

where

$$K(t,s) = K_s(t) = \frac{1}{1-ts} \quad t, s \in (-1, 1) \qquad (11)$$

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

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 20 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 14 equally spaced points given by the formula $t_{-1} = -0.97, t_{-j-1} = t_{-j} + 0.15$ for all $j \in \mathbb{N}_{13}$.

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

Table 2: The sum square error obtained from rational kernel for two methods for different values of the regularization parameter ρ.

Table 2 depicts the sum square error evaluated on 14 data points for the regularization method and the midpoint algorithm. Our computation again indicates that the midpoint algorithm provides, at least in this numerical experiment, better result than the regularization method.

4 Conclusion

In this paper, we have provided some basic facts about the Hypercircle inequality and discussed what we need for section 3, which is the major theme of this paper. In section 3, 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.

Acknowledgements:

The authors are grateful to Prof. Charles A. Michelli 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*. 686, 1950, pp. 337–404.

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

[3] Peter Duren, Theory of H^p spaces, *Second editon, Dover Publications, New York*. 2000.

[4] T. Evgeniou and M. Pontil and T. Poggio, Regularization networks and support vector machines, *Adances in Computational Mathematics*. 13, 2000, pp. 1-50.

[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*. 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] K. Khompurngson and C. A. Micchelli, Hide, *Journal of Mahcine Learing Research* (Accepted) .

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

[9] 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.

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

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

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

[13] 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.

[14] G. Wahba, Spline Models for Observational Data, *SIAM, Philadelphia, Series in Applied Mahtematics*. 59, 1990.

[15] L. Wu, A parameter choice method for Tikhonov regularization, *Electronic Transactions on Numerical Analysis*. 16, 2003, pp. 107-128.