IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol.E84-A, no.9, pp.2319–2329, Sep. 2001.

Active Learning for Optimal Generalization in Trigonometric Polynomial Models

Masashi Sugiyama Hidemitsu Ogawa

Department of Computer Science,
Graduate School of Information Science and Engineering,
Tokyo Institute of Technology.
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan.

sugi@og.cs.titech.ac.jp
http://ogawa-www.cs.titech.ac.jp/~sugi/

Abstract

In this paper, we consider the problem of active learning, and give a necessary and sufficient condition of sample points for the optimal generalization capability. By utilizing the properties of pseudo orthogonal bases, we clarify the mechanism of achieving the optimal generalization capability. We also show that the condition does not only provide the optimal generalization capability but also reduces the computational complexity and memory required for calculating learning result functions. Based on the optimality condition, we give design methods of optimal sample points for trigonometric polynomial models. Finally, the effectiveness of the proposed active learning method is demonstrated through computer simulations.

Keywords

machine learning, supervised learning, active learning, generalization capability, trigonometric polynomial space, pseudo orthogonal bases.

1 Introduction

Supervised learning is obtaining an underlying rule from training examples made up of input points and corresponding output values. If the input-output rule is successfully acquired, then we can estimate appropriate output values corresponding to unknown input points. This ability is called the *generalization capability*. It is known that higher levels of the generalization capability can be acquired if we actively design sample points [10]. In this paper, we discuss the problem of designing sample points, called *active learning* [4][25][9], for the optimal generalization capability. Active learning is also referred to as *optimal experiments* [12][7][3] or *query construction* [22].

Active learning has been studied from two stand points depending on the optimality. One is the *global optimality*, where a set of all sample points is optimal [7][10][26]. The other is the *greedy optimality*, where the next sample point to add is optimal in each step [14][3][9][24]. In this paper, we focus on the former global optimal case and give an active learning method especially in trigonometric polynomial models. The present paper is an extended version of the reference [23] with some new results.

This paper is organized as follows. In Section 2, the supervised learning problem is formulated as an inverse problem from the functional analytic point of view. Within this framework, the generalization measure and learning method are described. In Section 3, our main result, a necessary and sufficient condition for the optimal generalization capability, is derived. Since no approximation is employed in its derivation, the condition gives exactly the optimal generalization capability. By utilizing the properties of pseudo orthogonal bases, we clarify the mechanism of achieving the optimal generalization capability. Also, an efficient calculation method of learning result functions is provided. In Section 4, design methods of optimal sample points for trigonometric polynomial models are given. We also show that one of the designs of sample points further reduces the computational complexity and memory. Finally, Section 5 is devoted to computer simulations demonstrating the effectiveness of the proposed active learning method.

2 Formulation of supervised learning

In this section, the supervised learning problem is formulated from the functional analytic point of view [17].

2.1 Supervised learning as an inverse problem

Let us consider the supervised learning problem of obtaining an approximation to a target function from a set of *training examples*. Let the learning target function $f(\boldsymbol{x})$ be a complex function of L variables defined on a subset $\mathcal{D}$ of the L-dimensional Euclidean space $\mathbf{R}^L$. The training examples are made up of *sample points* $\boldsymbol{x}_m$ in $\mathcal{D}$ and corresponding *sample values* y_m in $\mathbf{C}$:

$$\{(\boldsymbol{x}_m, y_m) \mid y_m = f(\boldsymbol{x}_m) + \epsilon_m,\ m = 1, 2, \ldots, M\}, \tag{1}$$

where y_m is degraded by additive noise ϵ_m with mean zero. Let $\boldsymbol{y}$ and $\boldsymbol{\epsilon}$ be M-dimensional vectors defined as

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_M)^\top, \tag{2}$$
$$\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_M)^\top, \tag{3}$$

where $\top$ denotes the transpose of a vector. $\boldsymbol{y}$ is called a *sample value vector*, and the space to which $\boldsymbol{y}$ belongs is called the *sample value space.*

In this paper, the learning target function $f(\boldsymbol{x})$ is assumed to belong to a *reproducing kernel Hilbert space* H [2]. The reproducing kernel $K(\boldsymbol{x}, \boldsymbol{x}')$ is a bivariate function defined on $\mathcal{D} \times \mathcal{D}$ that satisfies the following conditions:

- For any fixed $\boldsymbol{x}'$ in $\mathcal{D}$, $K(\boldsymbol{x}, \boldsymbol{x}')$ belongs to H as a function of $\boldsymbol{x}$.
- For any function f in H and for any $\boldsymbol{x}'$ in $\mathcal{D}$, it holds that

$$\langle f(\cdot), K(\cdot, \boldsymbol{x}') \rangle = f(\boldsymbol{x}'), \tag{4}$$

where $\langle \cdot, \cdot \rangle$ stands for the inner product in H.

Note that the reproducing kernel is unique if it exists. If a function $\psi_m(\boldsymbol{x})$ is defined as

$$\psi_m(\boldsymbol{x}) = K(\boldsymbol{x}, \boldsymbol{x}_m), \tag{5}$$

then the value of a function f at a sample point $\boldsymbol{x}_m$ is expressed as

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

For this reason, ψ_m is called a *sampling function.*

Let A be a linear operator from H to $\mathbf{C}^M$ defined as

$$A = \sum_{m=1}^{M} \left(\boldsymbol{e}_m \otimes \overline{\psi_m} \right), \tag{7}$$

where $\boldsymbol{e}_m$ is the m-th vector of the so-called standard basis in $\mathbf{C}^M$ and $(\cdot \otimes \bar{\cdot})$ stands for the *Neumann-Schatten product*[1]. The operator A is a mapping from a function f to the M-dimensional vector with the m-th element being $f(\boldsymbol{x}_m)$:

$$Af = (f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots, f(\boldsymbol{x}_M))^\top. \tag{8}$$

For this reason, A is called a *sampling operator.*

[1] For any fixed g in a Hilbert space H_1 and any fixed f in a Hilbert space H_2, the *Neumann-Schatten product* $(f \otimes \overline{g})$ is an operator from H_1 to H_2 defined by using any $h \in H_1$ as [21]

$$(f \otimes \overline{g})h = \langle h, g \rangle f.$$

Then the relationship between f and $\boldsymbol{y}$ can be expressed as

$$\boldsymbol{y} = Af + \boldsymbol{\epsilon}. \tag{9}$$

Let us denote a mapping from $\boldsymbol{y}$ to a learning result function $\hat{f}$ by X:

$$\hat{f} = X\boldsymbol{y}. \tag{10}$$

X is called a *learning operator*. Consequently, the supervised learning problem is reformulated as an inverse problem of obtaining X that minimizes a certain generalization error J_G.

2.2 Generalization measure and learning method

We adopt the following J_G as the generalization error of the learning result function $\hat{f}$:

$$J_G = \mathrm{E}_{\boldsymbol{\epsilon}} \|\hat{f} - f\|^2, \tag{11}$$

where $\mathrm{E}_{\boldsymbol{\epsilon}}$ denotes the expectation over the noise $\boldsymbol{\epsilon}$. It is known that Eq.(11) can be decomposed into the *bias* and *variance* [6]:

$$J_G = \|\mathrm{E}_{\boldsymbol{\epsilon}}\hat{f} - f\|^2 + \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f} - \mathrm{E}_{\boldsymbol{\epsilon}}\hat{f}\|^2. \tag{12}$$

Substituting Eqs.(10) and (9) into Eq.(12), we have

$$J_G = \|XAf - f\|^2 + \mathrm{E}_{\boldsymbol{\epsilon}}\|X\boldsymbol{\epsilon}\|^2. \tag{13}$$

Here, we shall minimize the variance under the constraint of the bias being zero. For this purpose, we let

$$X = A^\dagger, \tag{14}$$

where $A^\dagger$ denotes the *Moore-Penrose generalized inverse*[2] of A. Note that the learning result function $\hat{f}$ given by Eqs.(10) and (14) minimizes the *training error* [17]:

$$\frac{1}{M}\sum_{m=1}^{M}\left|\hat{f}(\boldsymbol{x}_m) - y_m\right|^2. \tag{15}$$

Then J_G is reduced to

$$J_G = \|P_{\mathcal{R}(A^*)}f - f\|^2 + \mathrm{E}_{\boldsymbol{\epsilon}}\|A^\dagger\boldsymbol{\epsilon}\|^2, \tag{16}$$

[2] An operator X is called the *Moore-Penrose generalized inverse* of an operator A if X satisfies the following four conditions [1].

$$AXA = A, \;\; XAX = X, \;\; (AX)^* = AX, \;\; (XA)^* = XA.$$

The Moore-Penrose generalized inverse is unique and denoted by $A^\dagger$.

where A^* denotes the adjoint operator of A, $\mathcal{R}(\cdot)$ denotes the range of an operator, and P_S denotes the orthogonal projection operator onto a subspace S. For the bias being zero, we assume

$$\mathcal{R}(A^*) = H. \tag{17}$$

Assumption (17) holds only if the dimension μ of H is finite and the number M of training examples is larger than or equal to μ. Then the generalization measure J_G yields

$$J_G = \mathrm{E}_{\boldsymbol{\epsilon}} \|A^\dagger \boldsymbol{\epsilon}\|^2. \tag{18}$$

3 Active learning for optimal generalization

In this section, we discuss the problem of active learning, i.e., designing a set $\{\boldsymbol{x}_m\}_{m=1}^M$ of sample points for the optimal generalization capability.

3.1 Necessary and sufficient condition for optimal generalization capability

We shall derive a necessary and sufficient condition for minimizing the generalization error J_G in terms of the sampling operator A.

Theorem 1 *Let H be a finite dimensional reproducing kernel Hilbert space such that the reproducing kernel of H satisfies*

$$K(\boldsymbol{x}, \boldsymbol{x}) = r \ \textit{ for any } \boldsymbol{x} \in \mathcal{D}, \tag{19}$$

where r is a non-negative constant. Let the noise covariance matrix Q be given as

$$Q = \mathrm{E}_{\boldsymbol{\epsilon}} \left(\boldsymbol{\epsilon} \otimes \overline{\boldsymbol{\epsilon}} \right) = \sigma^2 I_M, \tag{20}$$

where the noise variance σ^2 is a (generally unknown) positive scalar and I_M denotes the M-dimensional identity matrix. Then the generalization error J_G defined by Eq.(11) is minimized with respect to the sampling operator A under the constraint (17) if and only if

$$\frac{\mu}{rM} A^* A = I, \tag{21}$$

where μ is the dimension of H and I denotes the identity operator on H. In this case, the minimum value of J_G is given as

$$\frac{\sigma^2 \mu^2}{rM}. \tag{22}$$

Proofs of all theorems and lemmas are provided in B.

3.2 Mechanism of achieving optimal generalization capability

Eq.(21) is equivalent to that a set $\{\sqrt{\frac{\mu}{rM}}\psi_m\}_{m=1}^{M}$ of sampling functions forms a *pseudo orthonormal basis* (PONB) [19][18] in H. The concept of PONBs is an extension of orthonormal bases to linearly dependent over-complete systems. The rigorous definition and properties of PONBs are described in A. Using the properties of PONBs, we have the following lemma.

Lemma 1 *When the sampling operator A satisfies Condition (21), it holds that*

$$\|Af\| = \sqrt{\tfrac{rM}{\mu}}\|f\| \quad \text{for any } f \in H, \tag{23}$$

$$\|A^\dagger \boldsymbol{u}\| = \begin{cases} \sqrt{\frac{\mu}{rM}}\|\boldsymbol{u}\| & \text{for any } \boldsymbol{u} \in \mathcal{R}(A), \\ 0 & \text{for any } \boldsymbol{u} \in \mathcal{R}(A)^\perp, \end{cases} \tag{24}$$

where $\mathcal{R}(A)^\perp$ denotes the orthogonal complement of $\mathcal{R}(A)$.

Eqs.(23) and (24) imply that $\sqrt{\frac{\mu}{rM}}A$ becomes an *isometry* and $\sqrt{\frac{rM}{\mu}}A^\dagger$ becomes a *partial isometry* with the initial space $\mathcal{R}(A)$.

Lemma 1 gives interpretation of Theorem 1. Let us decompose the noise $\boldsymbol{\epsilon}$ into $\tilde{\boldsymbol{\epsilon}}$ in $\mathcal{R}(A)$ and $\tilde{\boldsymbol{\epsilon}}^\perp$ in $\mathcal{R}(A)^\perp$:

$$\boldsymbol{\epsilon} = \tilde{\boldsymbol{\epsilon}} + \tilde{\boldsymbol{\epsilon}}^\perp. \tag{25}$$

Then the sample value vector $\boldsymbol{y}$ is rewritten as

$$\boldsymbol{y} = Af + \tilde{\boldsymbol{\epsilon}} + \tilde{\boldsymbol{\epsilon}}^\perp. \tag{26}$$

From Eq.(17), it holds for any f in H that

$$A^\dagger A f = P_{\mathcal{R}(A^*)} f = f, \tag{27}$$

which implies that the signal component Af is transformed to the original function f by $A^\dagger$. From Eq.(24), $A^\dagger$ suppresses the magnitude of the noise $\tilde{\boldsymbol{\epsilon}}$ in $\mathcal{R}(A)$ by $\sqrt{\frac{\mu}{rM}}$ and completely removes the noise $\tilde{\boldsymbol{\epsilon}}^\perp$ in $\mathcal{R}(A)^\perp$:

$$\|A^\dagger \tilde{\boldsymbol{\epsilon}}\| = \sqrt{\tfrac{\mu}{rM}}\|\tilde{\boldsymbol{\epsilon}}\|, \tag{28}$$

$$A^\dagger \tilde{\boldsymbol{\epsilon}}^\perp = 0. \tag{29}$$

The above analysis is summarized in Fig. 1.

In general, it is difficult to suppress the effect of the noise $\tilde{\boldsymbol{\epsilon}}$ in $\mathcal{R}(A)$ since it can not be distinguished from the signal component Af. However, the above analysis suggests that the effect of the noise $\tilde{\boldsymbol{\epsilon}}$ is minimized if the mean magnification of $A^\dagger$ is minimized. Since minimizing the mean magnification of $A^\dagger$ is equivalent to maximizing the mean magnification of A, the effect of the noise $\tilde{\boldsymbol{\epsilon}}$ is minimized if the norm of Af is maximized in the average sense. This principle well agrees with our intuition that the sampling with the highest signal-to-noise ratio in the sample value vector $\boldsymbol{y}$ provides the optimal generalization capability.

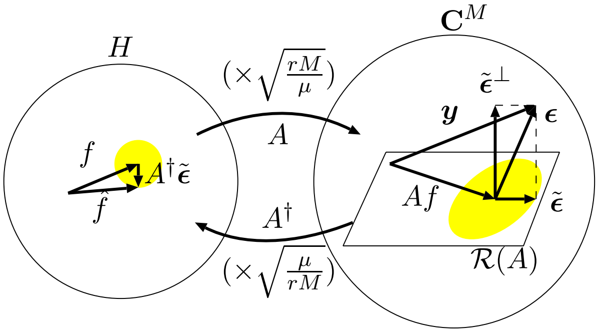

H
$\mathbf{C}^M$
$(\times\sqrt{\frac{rM}{\mu}})$
A
$A^\dagger$
$(\times\sqrt{\frac{\mu}{rM}})$
f
$\hat{f}$
$A^\dagger\tilde{\boldsymbol{\epsilon}}$
$\boldsymbol{y}$
$\tilde{\boldsymbol{\epsilon}}^\perp$
$\boldsymbol{\epsilon}$
Af
$\tilde{\boldsymbol{\epsilon}}$
$\mathcal{R}(A)$

Figure 1: Mechanism of achieving optimal generalization capability by Theorem 1. If sampling operator A satisfies $\frac{\mu}{rM}A^*A = I$, then $A^\dagger Af = f$, $\|A^\dagger\tilde{\boldsymbol{\epsilon}}\| = \sqrt{\frac{\mu}{rM}}\|\tilde{\boldsymbol{\epsilon}}\|$, and $A^\dagger\tilde{\boldsymbol{\epsilon}}^\perp = 0$.

3.3 Calculation of learning result functions

Now we discuss the calculation method of the learning result function $\hat{f}(\boldsymbol{x})$ given by Eqs.(10) and (14).

Let $\{\varphi_p(\boldsymbol{x})\}_{p=1}^{\mu}$ be an orthonormal basis in H. Then the following proposition holds.

Proposition 1 *[6] For general sample points, the learning result function $\hat{f}(\boldsymbol{x})$ can be calculated as*

$$\hat{f}(\boldsymbol{x}) = \sum_{p=1}^{\mu}\left(\sum_{p'=1}^{\mu}[C^{-1}]_{p,p'}\sum_{m=1}^{M}\overline{\varphi_{p'}(\boldsymbol{x}_m)}y_m\right)\varphi_p(\boldsymbol{x}), \tag{30}$$

where $[\,\cdot\,]_{p,p'}$ denotes the (p,p')-th element of a matrix and $\overline{\,\cdot\,}$ denotes the complex conjugate of a scalar. C is a μ-dimensional matrix defined as

$$[C]_{p,p'} = \sum_{m=1}^{M}\overline{\varphi_p(\boldsymbol{x}_m)}\varphi_{p'}(\boldsymbol{x}_m). \tag{31}$$

When the sample points satisfy Condition (21), the following theorem holds.

Theorem 2 *When Condition (21) holds, the learning result function $\hat{f}(\boldsymbol{x})$ can be calculated as*

$$\hat{f}(\boldsymbol{x}) = \sum_{p=1}^{\mu}\left(\frac{\mu}{rM}\sum_{m=1}^{M}\overline{\varphi_p(\boldsymbol{x}_m)}y_m\right)\varphi_p(\boldsymbol{x}). \tag{32}$$

Let us measure the computational complexity by the number of scalar multiplications. For general sample points, the computational complexity and memory required for calculating $\hat{f}(\boldsymbol{x})$ by Eq.(30) are $\mathcal{O}(\mu^2(M+\mu))$ and $\mathcal{O}(M+\mu^2)$, respectively. In contrast, Theorem 2 states that if the sample points satisfy Condition (21), then the computational complexity and memory can be reduced to $\mathcal{O}(\mu M)$ and $\mathcal{O}(M+\mu)$, respectively.

Table 1: Computational complexity and memory required for calculating the learning result function $\hat{f}(\boldsymbol{x})$. M is the number of training examples and μ is the dimension of H. In Corollary 3, H is a trigonometric polynomial space and $M = T\mu$ where T is a positive integer.

Sample Points	Calculation Method	Computational Complexity	Memory
General	Proposition 1	$\mathcal{O}(\mu^2(M+\mu))$	$\mathcal{O}(M+\mu^2)$
Condition (21)	Theorem 2	$\mathcal{O}(\mu M)$	$\mathcal{O}(M+\mu)$
Theorem 4 with Eq.(54)	Corollary 3	$\mathcal{O}(\mu^2)$	$\mathcal{O}(\mu)$

This shows that Theorems 1 and 2 do not only provide the optimal generalization capability but also reduce the computational complexity and memory. These results are summarized in Table 1.

4 Optimal design of sample points in trigonometric polynomial space

We have given the optimality condition of sample points for a finite dimensional reproducing kernel Hilbert space such that Eq.(19) holds. In this section, we introduce the *trigonometric polynomial space* that meets the above requirements, and give design methods of optimal sample points.

4.1 Trigonometric polynomial space

Let us denote the L-dimensional input vector $\boldsymbol{x}$ by

$$\boldsymbol{x} = (\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(L)})^\top. \tag{33}$$

Then the trigonometric polynomial space is defined as follows.

Definition 1 *For $l = 1, 2, \ldots, L$, let N_l be a non-negative integer and $\mathcal{D}_l = [-\pi, \pi]$. Then a function space H is called a trigonometric polynomial space of order $(N_1, N_2, \ldots, N_L)$ if H is spanned by the functions*

$$\left\{ \prod_{l=1}^{L} \exp(in_l\xi^{(l)}) \,\middle|\, n_l = -N_l, -N_l+1, \ldots, N_l \ \text{ for } l = 1, 2, \ldots, L \right\} \tag{34}$$

defined on $\mathcal{D} = \mathcal{D}_1 \times \mathcal{D}_2 \times \cdots \times \mathcal{D}_L$, and the inner product in H is defined as

$$\langle f, g \rangle = \frac{1}{(2\pi)^L} \int_{\mathcal{D}} f(\boldsymbol{x})\overline{g(\boldsymbol{x})}d\boldsymbol{x}. \tag{35}$$

The dimension μ of a trigonometric polynomial space of order $(N_1, N_2, \ldots, N_L)$ is

$$\mu = \prod_{l=1}^{L} (2N_l + 1), \tag{36}$$

and the reproducing kernel of this space is expressed as

$$K(\boldsymbol{x}, \boldsymbol{x}') = \prod_{l=1}^{L} K_l(\xi^{(l)}, \xi^{(l)\prime}), \tag{37}$$

where

$$K_l(\xi^{(l)}, \xi^{(l)\prime}) = \begin{cases} \dfrac{\sin\Big((2N_l+1)(\xi^{(l)} - \xi^{(l)\prime})/2\Big)}{\sin\Big((\xi^{(l)} - \xi^{(l)\prime})/2\Big)} & \text{if } \xi^{(l)} \neq \xi^{(l)\prime}, \\ 2N_l + 1 & \text{if } \xi^{(l)} = \xi^{(l)\prime}. \end{cases} \tag{38}$$

When the dimension L of the input vector $\boldsymbol{x}$ is 1, a trigonometric polynomial space of order N is spanned by

$$\Big\{ \exp(inx) \;\Big|\; n = -N, -N+1, \ldots, N \Big\} \tag{39}$$

defined on $\mathcal{D} = [-\pi, \pi]$, and the inner product is defined as

$$\langle f, g \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\overline{g(x)}dx. \tag{40}$$

The dimension μ of a trigonometric polynomial space of order N is

$$\mu = 2N + 1, \tag{41}$$

and the reproducing kernel of this space is expressed as

$$K(x, x') = \begin{cases} \dfrac{\sin\left((2N+1)(x - x')/2\right)}{\sin\left((x - x')/2\right)} & \text{if } x \neq x', \\ 2N + 1 & \text{if } x = x'. \end{cases} \tag{42}$$

In the case of the trigonometric polynomial space, the generalization error J_G defined by Eq.(11) is expressed as

$$J_G = \mathrm{E}_{\boldsymbol{\epsilon}} \frac{1}{(2\pi)^L} \int_{\mathcal{D}} \left| \hat{f}(\boldsymbol{x}) - f(\boldsymbol{x}) \right|^2 d\boldsymbol{x}. \tag{43}$$

In many statistical learning theories [3][4][10][9], the generalization measure is defined as

$$\mathrm{E}_{\boldsymbol{\epsilon}} \int_{\mathcal{D}} \left| \hat{f}(\boldsymbol{x}) - f(\boldsymbol{x}) \right|^2 p(\boldsymbol{x}) d\boldsymbol{x}, \tag{44}$$

where $p(\boldsymbol{x})$ is the (generally unknown) probability density function of test input points $\boldsymbol{x}$. In our setting, we assume that $p(\boldsymbol{x})$ in Eq.(44) is the uniform distribution on the domain $\mathcal{D}$. Note that, in many active learning methods, the assumption that $p(\boldsymbol{x})$ is known in advance is made [3][4][26][25][24].

When H is a trigonometric polynomial space, the constant r in Eq.(19) is μ:

$$r = \mu, \tag{45}$$

where μ is the dimension of H. Therefore, Eq.(22) is reduce to

$$\frac{\sigma^2 \mu}{M}, \tag{46}$$

which is equivalent to the asymptotic generalization error by passive learning [10]. This means that Eq.(46) can be attained with a finite number of training examples if Condition (21) holds.

4.2 Optimal design of sample points

For the trigonometric polynomial space, we shall give design methods of sample points $\{\boldsymbol{x}_m\}_{m=1}^{M}$ that satisfy the optimality condition (21).

Theorem 3 *For $l = 1, 2, \ldots, L$, let M_l be a positive integer such that $M_l \geq 2N_l + 1$ and c_l be an arbitrary constant such that $-\pi \leq c_l \leq -\pi + \frac{2\pi}{M_l}$. Let the number M of training examples be*

$$M = \prod_{l=1}^{L} M_l. \tag{47}$$

If a set

$$\left\{ \boldsymbol{x}_m \,\middle|\, m = \sum_{l=2}^{L} \left((m_l - 1) \prod_{l'=1}^{l-1} M_{l'} \right) + m_1, \;\; m_l = 1, 2, \ldots, M_l \;\text{ for } l = 1, 2, \ldots, L \right\} \tag{48}$$

of M sample points is fixed to

$$\boldsymbol{x}_m = (\xi_m^{(1)}, \xi_m^{(2)}, \ldots, \xi_m^{(L)})^\top, \tag{49}$$

where

$$\xi_m^{(l)} = c_l + \frac{2\pi}{M_l}(m_l - 1) \;\text{ for } l = 1, 2, \ldots, L, \tag{50}$$

then Condition (21) holds.

Theorem 4 *Let $M = T\mu$ where T is a positive integer and μ is the dimension of H. For $t = 1, 2, \ldots, T$ and $l = 1, 2, \ldots, L$, let $c_{t,l}$ be an arbitrary constant such that $-\pi \leq c_{t,l} \leq -\pi + \frac{2\pi}{2N_l+1}$. If a set*

$$\Bigg\{ \boldsymbol{x}_m \;\Bigg|\; m = (t-1)\mu + \sum_{l=2}^{L} \Bigg((n_l - 1) \prod_{l'=1}^{l-1} (2N_{l'} + 1) \Bigg) + n_1, \quad t = 1, 2, \ldots, T,\ n_l = 1, 2, \ldots, 2N_l + 1 \ \text{for } l = 1, 2, \ldots, L \Bigg\} \tag{51}$$

of M sample points is fixed to

$$\boldsymbol{x}_m = (\xi_m^{(1)}, \xi_m^{(2)}, \ldots, \xi_m^{(L)})^\top, \tag{52}$$

where

$$\xi_m^{(l)} = c_{t,l} + \frac{2\pi}{2N_l + 1}(n_l - 1) \ \text{for } l = 1, 2, \ldots, L, \tag{53}$$

then Condition (21) holds.

Theorem 3 means that M sample points are fixed to regular intervals in the domain $\mathcal{D}$ (Fig. 2). In contrast, Theorem 4 means that for each t, μ sample points are fixed to regular intervals in the domain $\mathcal{D}$ (Fig. 3). Especially when

$$c_{1,l} = c_{2,l} = \cdots = c_{T,l} = c_l \ \text{for } l = 1, 2, \ldots, L, \tag{54}$$

μ sample points are fixed to regular intervals in the domain $\mathcal{D}$ and sample values are gathered T times at each point (Fig. 4). Note that the design of sample points shown in Theorem 3 is also *D-optimal* [7].

When the dimension L of the input vector $\boldsymbol{x}$ is 1, the above theorems are reduced to simpler forms.

Corollary 1 *Let $M \geq \mu$ and c be an arbitrary constant such that $-\pi \leq c \leq -\pi + \frac{2\pi}{M}$. If a set $\{x_m\}_{m=1}^{M}$ of M sample points as fixed to*

$$x_m = c + \frac{2\pi}{M}(m - 1), \tag{55}$$

then Condition (21) holds.

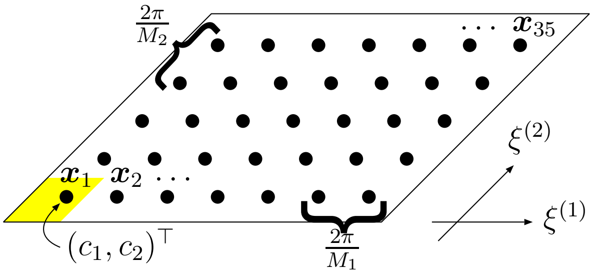
Corollary 2 *Let $M = T\mu$ where T is a positive integer. For $t = 1, 2, \ldots, T$, let c_t be an arbitrary constant such that $-\pi \leq c_t \leq -\pi + \frac{2\pi}{\mu}$. If a set*

$$\{x_m \mid m = (t-1)\mu + p, t = 1, 2, \ldots, T,\ p = 1, 2, \ldots, \mu\} \tag{56}$$

of M sample points is fixed to

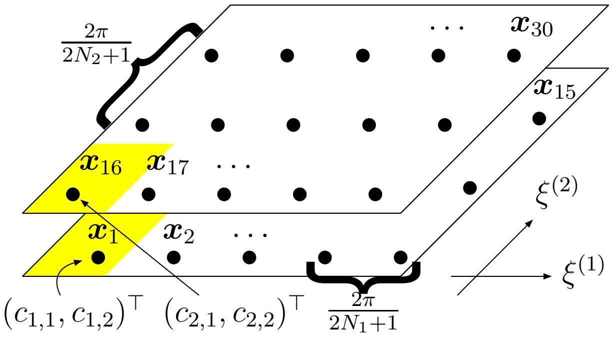
$$x_m = c_t + \frac{2\pi}{\mu}(p - 1), \tag{57}$$

then Condition (21) holds.



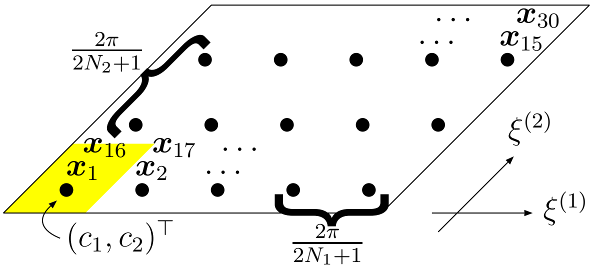
$\frac{2\pi}{M_2}$
$\cdots$ $\boldsymbol{x}_{35}$
$\xi^{(2)}$
$\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\cdots$
$\xi^{(1)}$
$(c_1, c_2)^\top$
$\frac{2\pi}{M_1}$

Figure 2: Example of optimal sample points (Theorem 3). H is a trigonometric polynomial space of order $(2, 1)$. The number M of training examples is $M = M_1 \times M_2 = 7 \times 5 = 35$.



$\frac{2\pi}{2N_2+1}$
$\cdots$ $\boldsymbol{x}_{30}$
$\boldsymbol{x}_{15}$
$\boldsymbol{x}_{16}$ $\boldsymbol{x}_{17}$ $\cdots$
$\xi^{(2)}$
$\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\cdots$
$\xi^{(1)}$
$(c_{1,1}, c_{1,2})^\top$ $(c_{2,1}, c_{2,2})^\top$
$\frac{2\pi}{2N_1+1}$

Figure 3: Example of optimal sample points (Theorem 4). H is a trigonometric polynomial space of order $(2, 1)$. The number M of training examples is $M = T \times \mu = 2 \times 15 = 30$.



$\cdots$ $\boldsymbol{x}_{30}$
$\cdots$ $\boldsymbol{x}_{15}$
$\frac{2\pi}{2N_2+1}$
$\xi^{(2)}$
$\boldsymbol{x}_{16}$ $\boldsymbol{x}_{17}$ $\cdots$
$\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\cdots$
$\xi^{(1)}$
$(c_1, c_2)^\top$
$\frac{2\pi}{2N_1+1}$

Figure 4: Example of optimal sample points (Theorem 4 with Eq.(54)). H is a trigonometric polynomial space of order $(2, 1)$. The number M of training examples is $M = T \times \mu = 2 \times 15 = 30$.

4.3 Calculation of learning result functions

As shown in Theorem 2, the learning result function can be efficiently calculated if Condition (21) holds. In the case of the trigonometric polynomial space, the efficiency can be further improved. Let $\{\varphi_p(\boldsymbol{x})\}_{p=1}^{\mu}$ be an orthonormal basis in the trigonometric polynomial space H, e.g., it is given by Eq.(34). Then we have the following corollary.

Corollary 3 *When sample points are designed following Theorem 4 with Eq.(54), the learning result function $\hat{f}(\boldsymbol{x})$ can be calculated as*

$$\hat{f}(\boldsymbol{x}) = \sum_{p=1}^{\mu} \left(\frac{1}{\mu} \sum_{p'=1}^{\mu} \overline{\varphi_p(\boldsymbol{x}_{p'})} \tilde{y}_{p'} \right) \varphi_p(\boldsymbol{x}), \tag{58}$$

where $\tilde{y}_{p'}$ is the mean sample value at $\boldsymbol{x}_{p'}$:

$$\tilde{y}_{p'} = \frac{1}{T} \sum_{t=1}^{T} y_{p'+(t-1)\mu}. \tag{59}$$

Corollary 3 is clear from Theorem 2, so the proof is omitted.

If sample points are designed following Theorem 4 with Eq.(54) and the learning result function $\hat{f}(\boldsymbol{x})$ is calculated following Corollary 3, then the computational complexity and memory can be further reduced to $\mathcal{O}(\mu^2)$ and $\mathcal{O}(\mu)$, respectively (Table 1). This is extremely efficient since the dimension μ of H does not depend on the number M of training examples.

5 Simulations

In this section, the effectiveness of the proposed active learning method is demonstrated through computer simulations.

Let the dimension L of the input vector $\boldsymbol{x}$ be 1 and H be a trigonometric polynomial space of order 100. Let the noise covariance matrix Q be $Q = I_M$, i.e., the noise variance σ^2 be 1. Let us consider the following sampling schemes.

(A) Optimal sampling: Sample points are determined following Theorem 3.

(B) Two-stage active learning: Eqs.(5.6) and (5.7) in the reference [24] are adopted as the active learning criteria. Sample points are determined by multi-point search with 3 randomly created candidates.

(C) Experiment design: Eq.(10) in the reference [3] is adopted as the active learning criterion. Sample points are also determined by multi-point search with 3 randomly created candidates.

(D) Passive learning: Sample points are randomly created.

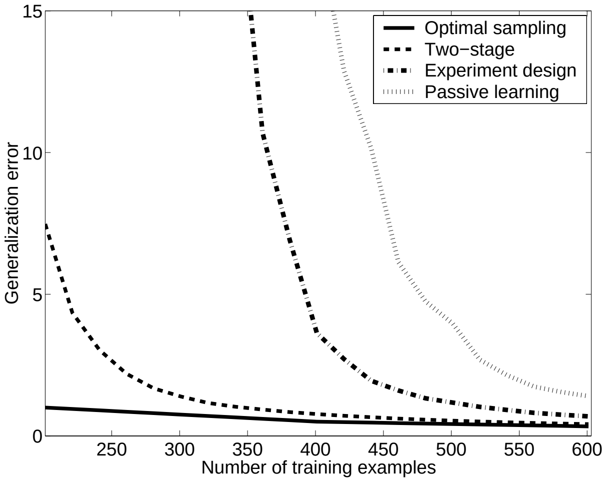


Figure 5: Relation between the number of training examples and the generalization error.

Note that the sampling scheme (A) is a global optimal method while the sampling schemes (B) and (C) are greedy optimal methods. The information that $p(x)$ is the uniform distribution is utilized in the sampling schemes (A), (B), and (C) (see Eq.(44)).

Fig. 5 displays the relation between the number of training examples and the generalization error. The horizontal and vertical axes denote the number of training examples and the generalization error J_G measured by Eq.(43) with $L = 1$, respectively. The solid curve shows the generalization error by the sampling scheme (A). The dashed, dash-dotted, and dotted curves denote the mean generalization errors of 10 trials by the sampling schemes (B), (C), and (D), respectively. When the number of training examples is 201 ($= \dim H$), the generalization errors of the sampling schemes (A), (B), (C), and (D) are 1.00, 7.48, 3.18×10^4, and 8.75×10^4, respectively.

The four curves show that the proposed sampling scheme gives much better generalization capability than other sampling schemes with a small number of training examples.

The mean computation times of the sampling schemes (A), (B), (C), and (D) are 2.64×10^{-2}, 85.2, 84.6, and 80.1 seconds, respectively. Therefore, learning with the sampling scheme (A) is much faster than learning with other sampling schemes.

The sampling schemes (B) and (C) can be applied to any Hilbert spaces while the proposed method (Theorems 3 and 4) is restricted to the trigonometric polynomial space. This simulation suggests that when H is the trigonometric polynomial space, the proposed method is applicable and the optimal generalization capability can be acquired. Otherwise, the incremental active learning method shown in the reference [24] seems to work well.

6 Conclusion

We gave a necessary and sufficient condition of sample points for the optimal generalization capability. By utilizing the properties of pseudo orthogonal bases, we clarified the mechanism of achieving the optimal generalization capability. We showed that the condition provides exactly the optimal generalization capability and at the same time, it reduces the computational complexity and memory required for calculating learning result functions. Based on the optimality condition, we gave design methods of optimal sample points for trigonometric polynomial models.

A Pseudo orthonormal bases

In Section 3.2, the necessary and sufficient condition for the optimal generalization capability was characterized by using the properties of pseudo orthonormal bases (PONBs). PONBs are a special type of pseudo orthogonal bases (POBs). In this section, we briefly review the concepts of POBs and PONBs, and show their fundamental properties.

Let H be a finite μ-dimensional Hilbert space and M be a finite integer larger than or equal to μ:

$$M \geq \mu. \tag{60}$$

Then POBs are defined as follows.

Definition 2 *[19] A set $\{\phi_m\}_{m=1}^{M}$ of elements in H is called a POB if any f in H is expressed as*

$$f = \sum_{m=1}^{M} \langle f, \phi_m \rangle \phi_m. \tag{61}$$

The concept of POBs is an extension of orthonormal bases (ONBs) to linearly dependent over-complete systems. It is clear that a POB is reduced to an ONB in H if M is equal to the dimension of H. POBs and their extension, pseudo biorthogonal bases [15][18], have been successfully applied to various real world problems including signal restoration [16][18], computerized tomography [20], neural network learning [17], and robust construction of neural networks [13][11].

The following proposition shows basic characteristics of POBs.

Proposition 2 *[19] The following conditions are mutually equivalent.*

1. *A set $\{\phi_m\}_{m=1}^{M}$ is a POB in H.*
2. $\|f\|^2 = \sum_{m=1}^{M} |\langle f, \phi_m \rangle|^2$ for any $f \in H$.
3. $\langle f, g \rangle = \sum_{m=1}^{M} \langle f, \phi_m \rangle \overline{\langle g, \phi_m \rangle}$ for any $f, g \in H$.

Condition 2 implies that a POB is a *tight frame with frame bound one* [5] or a *normalized tight frame* [8] in the *frame* terminology. When M is equal to the dimension of H, Conditions 2 and 3 are reduced to *Parseval's equalities.*

Now let us consider a finite M-dimensional Hilbert space H'. Let a set $\{\varphi'_m\}_{m=1}^M$ be an ONB in H' and U be an operator defined as

$$U = \sum_{m=1}^{M} \left(\varphi'_m \otimes \overline{\phi_m}\right). \tag{62}$$

Then the following proposition holds.

Proposition 3 *[19] The following conditions are mutually equivalent.*

1. *A set $\{\phi_m\}_{m=1}^M$ is a POB in H.*
2. *$U^*U = I$, where I is the identity operator on H.*
3. *$\|Uf\| = \|f\|$ for any $f \in H$.*
4. *$\langle Uf, Ug\rangle = \langle f, g\rangle$ for any $f, g \in H$.*

It follows from Condition 2 that

$$\begin{aligned}\sum_{m=1}^{M} \|\phi_m\|^2 &= \mathrm{tr}\left(\sum_{m=1}^{M}\left(\phi_m \otimes \overline{\phi_m}\right)\right) = \mathrm{tr}\left(U^*U\right) \\ &= \mathrm{tr}\left(I\right) = \dim(H) = \mu,\end{aligned} \tag{63}$$

where $\mathrm{tr}\left(\cdot\right)$ denotes the trace of an operator. Condition 3 means that the operator U is an *isometry.* From these properties, we have the following construction method of POBs.

Proposition 4 *[19] Let U be an isometry from H to H' and a set $\{\varphi'_m\}_{m=1}^M$ be an ONB in H'. If we let*

$$\phi_m = U^*\varphi'_m \text{ for } m = 1, 2, \ldots, M, \tag{64}$$

then a set $\{\phi_m\}_{m=1}^M$ becomes a POB in H.

Note that all POBs can be constructed by changing U with a fixed ONB $\{\varphi'_m\}_{m=1}^M$ or by changing $\{\varphi'_m\}_{m=1}^M$ with a fixed U.

If a set $\{\phi_m\}_{m=1}^M$ is a POB and

$$\|\phi_1\| = \|\phi_2\| = \cdots = \|\phi_M\|, \tag{65}$$

then the set $\{\phi_m\}_{m=1}^M$ is called a *pseudo orthonormal basis* (PONB). In this case, it follows from Eq.(63) that

$$\|\phi_m\| = \sqrt{\frac{\mu}{M}} \text{ for } m = 1, 2, \ldots, M. \tag{66}$$

Finally, we show a construction method of PONBs that plays an important role in the proof of Theorem 4.

Theorem 5 *Let $M = T\mu$ where T is a positive integer and μ is the dimension of H. Then a set $\{\phi_m\}_{m=1}^M$ becomes a PONB in H if a set $\{\sqrt{T}\phi_m\}_{m=1}^M$ consists of T sets of ONBs in H.*

B Proofs of theorems and lemmas

B.1 Theorem 1

It follows from Eq.(20) that Eq.(18) is reduced to

$$\begin{aligned} J_G &= \mathrm{E}_{\boldsymbol{\epsilon}}\|A^{\dagger}\boldsymbol{\epsilon}\|^2 = \mathrm{tr}\left(A^{\dagger}\mathrm{E}_{\boldsymbol{\epsilon}}\left(\boldsymbol{\epsilon}\otimes\overline{\boldsymbol{\epsilon}}\right)(A^{\dagger})^*\right) \\ &= \sigma^2\mathrm{tr}\left(A^{\dagger}(A^{\dagger})^*\right) = \sigma^2\mathrm{tr}\left((A^*A)^{\dagger}\right). \end{aligned} \tag{67}$$

Because of Eq.(17), A^*A is *positive definite*. Therefore, it has μ positive eigenvalues $\{\lambda_p\}_{p=1}^{\mu}$ considering the *geometric multiplicity*. Then it holds that

$$\mathrm{tr}\left(A^*A\right) = \sum_{p=1}^{\mu}\lambda_p, \tag{68}$$

$$\mathrm{tr}\left((A^*A)^{-1}\right) = \sum_{p=1}^{\mu}\frac{1}{\lambda_p}. \tag{69}$$

It is well-known that the arithmetic and harmonic means have the following relation:

$$\frac{\sum_{p=1}^{\mu}\lambda_p}{\mu} \geq \frac{\mu}{\sum_{p=1}^{\mu}\frac{1}{\lambda_p}}, \tag{70}$$

where equality holds if and only if

$$\lambda_1 = \lambda_2 = \cdots = \lambda_{\mu}. \tag{71}$$

From Eqs.(67), (69), (70), and (68), we have

$$J_G \geq \frac{\sigma^2\mu^2}{\mathrm{tr}\left(A^*A\right)}. \tag{72}$$

Since it follows from Eqs.(7), (6), (5), and (19) that

$$\begin{aligned} \mathrm{tr}\left(A^*A\right) &= \mathrm{tr}\left(\sum_{m=1}^{M}\left(\psi_m\otimes\overline{\psi_m}\right)\right) = \sum_{m=1}^{M}\|\psi_m\|^2 \\ &= \sum_{m=1}^{M}\langle\psi_m,\psi_m\rangle = \sum_{m=1}^{M}\psi_m(\boldsymbol{x}_m) \\ &= \sum_{m=1}^{M}K(\boldsymbol{x}_m,\boldsymbol{x}_m) = \sum_{m=1}^{M}r = rM, \end{aligned} \tag{73}$$

Eq.(72) yields

$$J_G \geq \frac{\sigma^2\mu^2}{rM}. \tag{74}$$

From Eqs.(68), (73), and (71), equality in Eq.(74) holds if and only if

$$\lambda_1 = \lambda_2 = \cdots = \lambda_\mu = \frac{rM}{\mu}. \tag{75}$$

Because of Eq.(17), Eq.(75) is equivalent to

$$A^*A = \frac{rM}{\mu}I, \tag{76}$$

which implies Eq.(21). Eq.(22) is clear from Eq.(74) with equality. ■

B.2 Lemma 1

If we let $\varphi'_m = \boldsymbol{e}_m$ and $\phi_m = \sqrt{\frac{\mu}{rM}}\psi_m$ in U defined by Eq.(62), then Eq.(23) is clear from Items 2 and 3 in Proposition 3 in A. It follows from Eq.(21) that

$$\begin{aligned}
\|A^\dagger \boldsymbol{u}\| &= \sqrt{\|A^\dagger \boldsymbol{u}\|^2} = \sqrt{\langle (A^\dagger)^* A^\dagger \boldsymbol{u}, \boldsymbol{u}\rangle} \\
&= \sqrt{\langle (A^\dagger)^* (A^*A)^{-1} A^* \boldsymbol{u}, \boldsymbol{u}\rangle} \\
&= \sqrt{\langle (A^\dagger)^* (\tfrac{rM}{\mu} I)^{-1} A^* \boldsymbol{u}, \boldsymbol{u}\rangle} \\
&= \sqrt{\tfrac{\mu}{rM}\langle (A^*)^\dagger A^* \boldsymbol{u}, \boldsymbol{u}\rangle} = \sqrt{\tfrac{\mu}{rM}\langle P_{\mathcal{R}(A)} \boldsymbol{u}, \boldsymbol{u}\rangle} \\
&= \sqrt{\tfrac{\mu}{rM}\|P_{\mathcal{R}(A)} \boldsymbol{u}\|^2} = \sqrt{\tfrac{\mu}{rM}}\|P_{\mathcal{R}(A)} \boldsymbol{u}\|,
\end{aligned} \tag{77}$$

which implies Eq.(24). ■

B.3 Theorem 2

Let W be an operator from $\mathbf{C}^\mu$ to H defined as

$$W = \sum_{p=1}^{\mu} \left(\varphi_p \otimes \overline{\boldsymbol{e}_p}\right), \tag{78}$$

where $\boldsymbol{e}_p$ is the p-th vector of the so-called standard basis in $\mathbf{C}^\mu$. Note that the operator W is *unitary*, i.e., it holds that

$$W^* = W^{-1}. \tag{79}$$

Then it follows from Eqs.(7) and (78) that

$$[AW]_{m,p} = \varphi_p(\boldsymbol{x}_m). \tag{80}$$

Hence, it follows from Eq.(31) that

$$C = W^*A^*AW. \tag{81}$$

When the sample points satisfy Condition (21), it follows from Eqs.(81) and (79) that

$$C = W^*(\frac{rM}{\mu}I)W = \frac{rM}{\mu}W^{-1}W = \frac{rM}{\mu}I_\mu, \tag{82}$$

where I_μ is the μ-dimensional identity matrix. Substituting Eq.(82) into Eq.(30), we have Eq.(32). ■

B.4 Theorem 3

Any function $f(\boldsymbol{x})$ in a trigonometric polynomial space of order $(N_1, N_2, \ldots, N_L)$ can be expressed as

$$
\begin{aligned}
f(\boldsymbol{x}) \;=\; & \sum_{n_1=-N_1}^{N_1} \sum_{n_2=-N_2}^{N_2} \cdots \sum_{n_L=-N_L}^{N_L} \\
& a_{n_1,n_2,\ldots,n_L} \prod_{l=1}^{L} \exp\left(i n_l \xi^{(l)}\right), \qquad (83)
\end{aligned}
$$

where $a_{n_1,n_2,\ldots,n_L}$ is a coefficient. It follows from Eqs.(6), (83), (49), and (50) that

$$
\begin{aligned}
& \sum_{m_1=1}^{M_1} \sum_{m_2=1}^{M_2} \cdots \sum_{m_L=1}^{M_L} \left| \langle f, \frac{1}{\sqrt{M}} \psi_m \rangle \right|^2 \\
&= \frac{1}{M} \sum_{m_1=1}^{M_1} \sum_{m_2=1}^{M_2} \cdots \sum_{m_L=1}^{M_L} |f(\boldsymbol{x}_m)|^2 \\
&= \frac{1}{M} \sum_{m_1=1}^{M_1} \sum_{m_2=1}^{M_2} \cdots \sum_{m_L=1}^{M_L} \left| \sum_{n_1=-N_1}^{N_1} \sum_{n_2=-N_2}^{N_2} \cdots \sum_{n_L=-N_L}^{N_L} a_{n_1,n_2,\ldots,n_L} \prod_{l=1}^{L} \exp(i n_l \xi_m^{(l)}) \right|^2 \\
&= \frac{1}{M} \sum_{m_1=1}^{M_1} \sum_{m_2=1}^{M_2} \cdots \sum_{m_L=1}^{M_L} \sum_{n_1=-N_1}^{N_1} \sum_{n_2=-N_2}^{N_2} \cdots \sum_{n_L=-N_L}^{N_L} \\
& \qquad \sum_{n'_1=-N_1}^{N_1} \sum_{n'_2=-N_2}^{N_2} \cdots \sum_{n'_L=-N_L}^{N_L} a_{n_1,n_2,\ldots,n_L} \overline{a_{n'_1,n'_2,\ldots,n'_L}} \prod_{l=1}^{L} \exp\left(i(n_l - n'_l)\xi_m^{(l)}\right) \\
&= \frac{1}{M} \sum_{n_1=-N_1}^{N_1} \sum_{n_2=-N_2}^{N_2} \cdots \sum_{n_L=-N_L}^{N_L} \sum_{n'_1=-N_1}^{N_1} \sum_{n'_2=-N_2}^{N_2} \cdots \sum_{n'_L=-N_L}^{N_L} a_{n_1,n_2,\ldots,n_L} \overline{a_{n'_1,n'_2,\ldots,n'_L}} \\
& \qquad \times \prod_{l=1}^{L} \left[\sum_{m_l=1}^{M_l} \exp\left(i(n_l - n'_l) \frac{2\pi m_l}{M_l} \right) \right] \prod_{l=1}^{L} \exp\left(i(n_l - n'_l)(c_l - \frac{2\pi}{M_l}) \right). \qquad (84)
\end{aligned}
$$

For any integers n_l and n'_l, it generally holds that

$$
\sum_{m_l=1}^{M_l} \exp\left(i(n_l - n'_l) \frac{2\pi m_l}{M_l} \right) = \begin{cases} M_l & \text{if } n_l = n'_l, \\ 0 & \text{if } n_l \neq n'_l. \end{cases} \qquad (85)
$$

Therefore, it follows from Eqs.(84), (85), (47), and (83) that

$$
\sum_{m_1=1}^{M_1} \sum_{m_2=1}^{M_2} \cdots \sum_{m_L=1}^{M_L} \left| \langle f, \frac{1}{\sqrt{M}} \psi_m \rangle \right|^2
$$

$$
\begin{aligned}
&= \frac{1}{M}\sum_{n_1=-N_1}^{N_1}\sum_{n_2=-N_2}^{N_2}\cdots\sum_{n_L=-N_L}^{N_L}|a_{n_1,n_2,\ldots,n_L}|^2\prod_{l=1}^{L}M_l\prod_{l=1}^{L}\exp(0)\\
&= \sum_{n_1=-N_1}^{N_1}\sum_{n_2=-N_2}^{N_2}\cdots\sum_{n_L=-N_L}^{N_L}|a_{n_1,n_2,\ldots,n_L}|^2\\
&= \|f\|^2. \qquad (86)
\end{aligned}
$$

According to Items 1 and 2 in Proposition 2 in A with $\phi_m = \frac{1}{\sqrt{M}}\psi_m$, Eq.(86) is equivalent to that a set $\{\frac{1}{\sqrt{M}}\psi_m\}_{m=1}^{M}$ forms a POB in H. Therefore, Items 1 and 2 in Proposition 3 with $\varphi'_m = \boldsymbol{e}_m$ and $\phi_m = \frac{1}{\sqrt{M}}\psi_m$ yield Eq.(21) with r given by Eq.(45). ■

B.5 Theorem 4

For a set $\{\boldsymbol{x}_m\}_{m=(t-1)\mu+1}^{t\mu}$ of μ sample points with a fixed t, it follows from Eqs.(6), (5), (52), (53), (37), and (38) that

$$
\begin{aligned}
\langle\sqrt{\frac{T}{M}}\psi_{m'},\sqrt{\frac{T}{M}}\psi_m\rangle &= \frac{1}{\mu}\psi_{m'}(\boldsymbol{x}_m)\\
&= \frac{1}{\mu}K(\boldsymbol{x}_m,\boldsymbol{x}_{m'}) = \begin{cases} 1 & \text{if } m=m', \\ 0 & \text{if } m\neq m'. \end{cases} \qquad (87)
\end{aligned}
$$

Eq.(87) implies that for each t, a set

$$\{\sqrt{\frac{T}{M}}\psi_m\}_{m=(t-1)\mu+1}^{t\mu} \qquad (88)$$

of μ elements in H forms an orthonormal in H. Therefore, a set $\{\frac{1}{\sqrt{M}}\psi_m\}_{m=1}^{M}$ forms a PONB in H from Theorem 5 in A. This is equivalent to Eq.(21) with r given by Eq.(45) according to Items 1 and 2 in Proposition 3 in A with $\varphi'_m = \boldsymbol{e}_m$ and $\phi_m = \frac{1}{\sqrt{M}}\psi_m$. ■

B.6 Theorem 5

For any ONB $\{\varphi_p\}_{p=1}^{\mu}$ in H, it holds that

$$\sum_{p=1}^{\mu}(\varphi_p\otimes\overline{\varphi_p}) = I. \qquad (89)$$

Hence, if a set $\{\sqrt{T}\phi_m\}_{m=1}^{M}$ of elements in H consists of T sets of ONBs, it follows from Eq.(62) that

$$
\begin{aligned}
U^*U &= \sum_{m=1}^{M}\left(\phi_m\otimes\overline{\phi_m}\right)\\
&= \frac{1}{T}\sum_{m=1}^{M}\left(\sqrt{T}\phi_m\otimes\overline{\sqrt{T}\phi_m}\right) = I. \qquad (90)
\end{aligned}
$$

According to Items 1 and 2 in Proposition 3, Eq.(90) is equivalent to that a set $\{\phi_m\}_{m=1}^{M}$ forms a POB in H. In this case, the set $\{\phi_m\}_{m=1}^{M}$ is a PONB in H since $\|\phi_m\| = \frac{1}{\sqrt{T}}$ for $m = 1, 2, \ldots, M$. ■

Acknowledgement

The authors would like to thank anonymous reviewers for their valuable comments.

References

[1] A. Albert, Regression and the Moore-Penrose Pseudoinverse, Academic Press, New York and London, 1972.

[2] N. Aronszajn, "Theory of reproducing kernels," Trans. American Math. Soc., vol. 68, pp. 337–404, 1950.

[3] D. A. Cohn, "Neural network exploration using optimal experiment design," Neural Networks, vol. 9, no. 6, pp. 1071–1083, 1996.

[4] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," J. Artificial Intelligence Research, vol. 4, pp. 129–145, 1996.

[5] I. Daubechies, Ten Lectures on Wavelets, Soc. for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1992.

[6] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993.

[7] V. V. Fedorov, Theory of Optimal Experiments, Academic Press, New York, 1972.

[8] M. Frank and D. R. Larson, "A module frame concept for Hilbert C^*-modules," Functional and Harmonic Analysis of Wavelets, Contemporary Mathematics, vol. 247, American Mathematical Soc., San Antonio, TX, 1999.

[9] K. Fukumizu, "Statistical active learning in multilayer perceptrons," IEEE Trans. Neural Networks, vol. 11, no. 1, pp. 17–21, 2000.

[10] K. Fukumizu and S. Watanabe, "Optimal Training Data and Predictive Error of Polynomial Approximation," IEICE Trans., vol. J79-A, no. 5, pp. 1100–1108, 1996. (In Japanese)

[11] H. Iwaki, H. Ogawa, and A. Hirabayashi, "Optimally generalizing neural networks with ability to recover from stuck-at r faults," IEICE Trans., Vol. J83-D-II, no. 2, pp. 805–813, 2000. (In Japanese)

[12] J. Kiefer, "Optimal experimental designs," J. R. Stat. Soc., series B, vol. 21, pp. 272–304, 1959.

[13] S. Nakazawa and H. Ogawa, "Optimal realization of optimally generalizing neural networks," IEICE Technical Report, NC96-60, pp. 17–24, 1996. (In Japanese)

[14] D. J. C. MacKay, "Information-based objective functions for active data selection," Neural Computation, vol. 4, no. 4, pp. 590–604, 1992.

[15] H. Ogawa, "A theory of pseudo biorthogonal bases," IEICE Trans., vol. J64-D, no. 7, pp. 555–562, 1981. (In Japanese)

[16] H. Ogawa, "A unified approach to generalized sampling theorems," Proc. ICASSP'86, Intl. Conf. Acoustics, Speech, and Signal Processing, pp. 1657–1660, Tokyo, Japan, 1986.

[17] H. Ogawa, "Neural network learning, generalization and over-learning," Proc. ICIIPS'92, Intl. Conf. Intelligent Information Processing & System, vol. 2, pp. 1–6, Beijing, China, 1992.

[18] H. Ogawa, "Theory of pseudo biorthogonal bases and its application," Research Institute for Mathematical Science, RIMS Kokyuroku, vol. 1067, Reproducing Kernels and their Applications, pp. 24–38, 1998.

[19] H. Ogawa and T. Iijima, "A theory of pseudo orthogonal bases," IECE Trans., vol. J58-D, no. 5, pp. 271–278, 1975. (In Japanese)

[20] H. Ogawa and I. Kumazawa, "Radon transform and analog coding." Mathematical Methods in Tomography, Lecture Notes in Mathematics, vol. 1497, pp. 229–241, Springer-Verlag, 1991.

[21] R. Schatten, Norm Ideals of Completely Continuous Operators, Springer-Verlag, Berlin, 1970.

[22] P. Sollich, "Query construction, entropy and generalization in neural network models," Phys. Rev. E, vol. 49, pp. 4637–4651, 1994.

[23] M. Sugiyama and H. Ogawa, "Training data selection for optimal generalization in trigonometric polynomial networks," Advances in Neural Information Processing Systems, vol. 12, pp. 624–630, The MIT Press, Cambridge, 2000.

[24] M. Sugiyama and H. Ogawa, "Incremental active learning for optimal generalization," Neural Computation, vol. 12, no. 12, pp. 2909–2940.

[25] S. Vijayakumar and H. Ogawa, "Improving generalization ability through active learning," IEICE Trans. Inf. & Syst., vol. E82-D, no. 2, pp. 480–487, 1999.

[26] R. X. Yue and F. J. Hickernell, "Robust designs for fitting linear models with misspecification," Statistica Sinica, vol. 9, pp. 1053–1069, 1999.