数理解析研究所講究録 1352 巻 2004 年 114-126

Active Learning for Maximal Generalization Capability

東京工業大学・大学院情報理工学研究科 小川 英光 (Hidemitsu Ogawa)
杉山 将 (Masashi Sugiyama)
Graduate School of Information Science and Engineering
Tokyo Institute of Technology

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 input points. In this paper, we discuss the problem of designing input points for the maximal generalization capability. This problem is referred to as *active learning* [4, 25, 8] or *experimental design* [11, 6, 3].

Active learning has been studied from two stand points depending on the optimality. One is the *global optimality*, where a set of all input points is optimal [6, 9, 26]. The other is the *greedy optimality*, where the next input point to add is optimal in each step [13, 3, 8, 24]. In this paper, we focus on the former global optimal case and give an active learning method especially in the trigonometric polynomial model.

2 Formulation of the problem

Let $f(\boldsymbol{x})$ be a learning target function. It is a complex valued function of L variables defined on a subset D of the L-dimensional Euclidean space $\mathbb{R}^L$. Assume that f belongs to a reproducing kernel Hilbert space (RKHS) H [2, 22].

The training examples are made up of finite M number of sample points $\boldsymbol{x}_m$ in D and corresponding M sample values y_m in $\mathbb{C}$:

$$y_m = f(\boldsymbol{x}_m) + \epsilon_m \quad : 1 \leq m \leq M,$$

where y_m is degraded by additive noise ϵ_m. Let $\boldsymbol{y}$ and $\boldsymbol{\epsilon}$ be M-dimensional vectors consist of (y_m) and (ϵ_m), respectively. Let A be an operator which transforms f to the M-dimensional vector with the m-th element being $f(\boldsymbol{x}_m)$. The operator A is called the *sampling operator*. Then, we have

$$\boldsymbol{y} = Af + \boldsymbol{\epsilon}. \tag{1}$$

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

$$\hat{f} = X\boldsymbol{y}. \tag{2}$$

The *supervised learning problem* is an inverse problem of obtaining X that minimizes a *generalization error*. We adopt the following J_G as the generalization error of $\hat{f}$:

$$J_G = \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f} - f\|^2, \tag{3}$$

where $\mathrm{E}_{\boldsymbol{\epsilon}}$ is the expectation with respect to noise ensemble $\{\boldsymbol{\epsilon}\}$.

Assume that X is linear. Then, from Eqs.(1) and (2), we have

$$\hat{f} = XAf + X\epsilon. \tag{4}$$

The first and second terms on the right-hand side of this equation are called the *signal component* and the *noise component* of $\hat{f}$, respectively. We require that for any given f in H, the signal component agrees with the target function f. The requirement can be satisfied if and only if

$$XA = I, \tag{5}$$

where I is the identity operator on H. Let $T^\dagger$ be the Moore-Penrose generalized inverse of an operator T. In connection with the requirement (5), we have

Lemma 1 *The following four statements are mutually equivalent.*
(i) The operator equation $XA = I$ has a solution.
(ii) $A^\dagger A = I$.
(iii) $N(A) = \{0\}$.
(iv) $R(A^*) = H$.
In this case, A^*A becomes non-singular.

Proof. First of all, we notice that $R(A)$ and $R(A^*)$ are closed because $\mathbb{C}^M$ is of finite dimension. Then, $A^\dagger$ in (ii) is well-defined and (iv) is equivalent to (iii). The mutual equivalence among (i)-(iii) is a well-known result [1]. (Q.E.D.)

Since $R(A^*) = H$ means that H is of finite dimension, we shall concentrate our attention on the finite dimensional H. Hence, all subspaces appeared in this paper are closed. Let N be the dimension of H. Since $R(A^*) = H$, N is less than or equal to the number of training data, i.e. $N \leq M$.

Definition 2 (Optimal learning operator) *For a fixed A, if an operator X minimizes J_G in Eq.(3) subject to $XA = I$, then X is called an* optimal learning operator *and denoted by X_0.*

Active learning problem is a problem to design sample points $\{\boldsymbol{x}_m : 1 \leq m \leq M\}$ so that $\hat{f}$ minimizes the generalization error J_G. It is equivalent to design the sampling operator A so that $\hat{f}$ minimizes J_G. In order to solve this problem, we first provide the optimal learning operator X_0 for each A. Then we shall devise the optimal A which minimizes $J_G[X_0]$.

Note that $\hat{f}$ is an unbiased estimate of the original target function f if the mean of the noise ensemble is zero because of Eqs.(4) and (5).

3 Optimal learning operator

In this section, we shall devise a closed form of the optimal learning operator X_0 for each A. Let us denote the trace of an operator T by $\mathrm{tr}\,(T)$. Let Q be the noise operator defined by

$$Q = \mathrm{E}_\epsilon\,(\epsilon \otimes \overline{\epsilon}), \tag{6}$$

where $(\cdot \otimes \overline{\cdot})$ is the Neumann-Schatten product. The rigorous definitions and properties of the trace and the Neumann-Schatten product are described in Appendix.

Theorem 3 (Optimal learning operator) *Assume that $R(A^*) = H$. For each A, the optimal learning operator always exists. Its general form is given by*

$$X_0 = V^{-1}A^*U^\dagger + Y(I_M - UU^\dagger), \tag{7}$$

where I_M is the identity operator on $\mathbb{C}^M$, Y is an arbitrary operator from $\mathbb{C}^M$ to H, and

$$U = AA^* + Q, \tag{8}$$
$$V = A^*U^\dagger A. \tag{9}$$

Furthermore, the minimum value, say J_0, of J_G with respect to X is given by

$$J_0 = \mathrm{tr}\left(V^{-1}\right) - N. \tag{10}$$

In order to prove this theorem, we shall prepare several lemmas.

Lemma 4 *[1] For any fixed operators T_1 and T_2, the following statements are mutually equivalent.*
(i) The equation $XT_1 = T_2$ has a solution.
(ii) $N(T_1) \subseteq N(T_2)$.
(iii) $T_2T_1^\dagger T_1 = T_2$.
In this case, a general solution of $XT_1 = T_2$ is given by

$$X = T_2T_1^\dagger + Y(I - T_1T_1^\dagger),$$

where I is the identity operator and Y is an arbitrary operator.

Lemma 5 (Properties of U) *The operator U in Eq.(8) is positive semidefinite, and it holds that*

$$N(U) = N(A^*) \cap N(Q), \tag{11}$$
$$R(U) = R(A) + R(Q), \tag{12}$$
$$UU^\dagger A = A, \tag{13}$$
$$A^*U^\dagger U = A^*. \tag{14}$$

Proof. Since Q is positive semidefinite, for any $\boldsymbol{u} \in \mathbb{C}^M$, Eq.(8) yields

$$\langle U\boldsymbol{u}, \boldsymbol{u}\rangle = \langle (AA^* + Q)\boldsymbol{u}, \boldsymbol{u}\rangle = \|A^*\boldsymbol{u}\|^2 + \langle Q\boldsymbol{u}, \boldsymbol{u}\rangle \geq 0,$$

which implies $U \geq 0$. Furthermore, if $\boldsymbol{u} \in N(U)$, then $\|A^*\boldsymbol{u}\|^2 = 0$ and $\langle Q\boldsymbol{u}, \boldsymbol{u}\rangle = 0$. Hence, $A^*\boldsymbol{u} = 0$ and $Q\boldsymbol{u} = 0$. That is, $N(U) \subseteq N(A^*) \cap N(Q)$. The converse is clear from Eq.(8). Hence, Eq.(11) holds. Taking the orthogonal complement of Eq.(11) yields Eq.(12). Eq.(12) yields $R(U) \supseteq R(A)$, which implies Eq.(13). Eq.(11) yields $N(U) \subseteq N(A^*)$, which implies Eq.(14) because of Lemma 4. (Q.E.D.)

Lemma 6 (Properties of V) *Assume that $R(A^*) = H$. The operator V in Eq.(9) is self-adjoint and non-singular.*

Proof. Since U is self-adjoint, V is also self-adjoint because of Eq.(9). Let $u \in N(V)$. Since $Au \in R(A) \subseteq R(U)$ because of Eq.(12), there exists $\boldsymbol{v}$ such that $Au = U\boldsymbol{v}$. Hence,

$$\begin{aligned}\langle U\boldsymbol{v}, \boldsymbol{v}\rangle &= \langle UU^\dagger U\boldsymbol{v}, \boldsymbol{v}\rangle = \langle U^\dagger U\boldsymbol{v}, U\boldsymbol{v}\rangle = \langle U^\dagger Au, Au\rangle \\ &= \langle A^*U^\dagger Au, u\rangle = \langle Vu, u\rangle = 0.\end{aligned}$$

Then, $U\boldsymbol{v} = 0$ and $Au = 0$. That is, $N(V) \subseteq N(A) = \{0\}$ because of Lemma 1. Hence, $N(V) = \{0\}$. Taking the orthogonal complement of $N(V) = \{0\}$ yields $R(V) = H$ because $V^* = V$. That is, V is a bijection on H, which implies V is non-singular. (Q.E.D.)

The following lemma characterizes the optimal learning operator.

Lemma 7 *Assume that $R(A^*) = H$. An operator X is an optimal learning operator if and only if X together with an operator C satisfies*

$$XA = I, \tag{15}$$
$$XQ = CA^*. \tag{16}$$

In this case, the minimum value J_0 of J_G is given by

$$J_0 = \mathrm{tr}\,(C). \tag{17}$$

Proof. Assume that $XA = I$. It follows from Eq.(4) that

$$\|\hat{f} - f\|^2 = \|X\epsilon\|^2 = \mathrm{tr}\left((X\epsilon) \otimes \overline{(X\epsilon)}\right) = \mathrm{tr}\,(X\,(\epsilon \otimes \bar{\epsilon})\,X^*)\,.$$

Then, Eqs.(3) and (6) yield

$$J_G = \mathrm{E}_\epsilon\|X\epsilon\|^2 = \mathrm{tr}\,(XQX^*) = \langle XQ, X\rangle, \tag{18}$$

where $\langle XQ, X\rangle$ is the Schmidt inner product of operators. The rigorous definition and properties of the Schmidt inner product are described in Appendix. Let C be the Lagrange multiplier operator. The conditional problem of variation for the optimal learning operator is reduced to the following unconditional problem of variation with respect to X and C:

$$J_G[X, C] = \langle XQ, X\rangle - 2\mathrm{Re}\langle XA - I, C\rangle,$$

where Re stands for the real part of a complex number. Equating the partial derivative of $J_G[X, C]$ with respect to C to zero yields Eq.(5), which is equal to Eq.(15). Equating the partial derivative of $J_G[X, C]$ with respect to X to zero yields

$$\begin{aligned}\delta J_G &= \langle \delta XQ, X\rangle + \langle XQ, \delta X\rangle - 2\mathrm{Re}\langle \delta XA, C\rangle \\ &= 2\mathrm{Re}\langle \delta X, XQ - CA^*\rangle = 0,\end{aligned}$$

where δX is an arbitrary operator. Then we have Eq.(16).

We shall show that Eqs.(15) and (16) have solutions. The assumption $R(A^*) = H$ guarantees that V is non-singular because of Lemma 6. If we let

$$X = V^{-1}A^*U^\dagger \quad \text{and} \quad C = V^{-1} - I. \tag{19}$$

then it follows from Eq.(9) that

$$XA = V^{-1}A^*U^\dagger A = V^{-1}V = I.$$

Hence, X in Eq.(19) satisfies Eq.(15). It follows from Eqs.(8),(19), and (14) that

$$\begin{aligned}XQ &= X(U - AA^*) = XU - XAA^* = V^{-1}A^*U^\dagger U - A^* \\ &= V^{-1}A^* - A^* = CA^*.\end{aligned}$$

Then X and C in Eq.(19) satisfy Eq.(16). That is, Eqs.(15) and (16) have solutions.

Let X_0 be a solution of Eqs.(15) and (16). We shall show that for any X which satisfies Eq.(15), it holds that $J_G[X] \geq J_G[X_0]$. From the definition of X_0

$$X_0A = I \quad \text{and} \quad X_0Q = CA^*. \tag{20}$$

Then, Eq.(15) yields

$$\begin{aligned} XQX_0^* &= X(X_0Q)^* = X(CA^*)^* = X(AC^*) \\ &= (XA)C^* = C^* = (X_0A)C^* = X_0(AC^*) \\ &= X_0(CA^*)^* = X_0(X_0Q)^* = X_0QX_0^*, \end{aligned}$$

and hence $XQX_0^* = X_0QX_0^*$. Since $X_0QX_0^*$ is self-adjoint, XQX_0^* is also self-adjoint and it holds that

$$X_0QX_0^* = XQX_0^* = X_0QX^*.$$

Therefore, Eq.(18) and $Q \geq 0$ yield

$$\begin{aligned} J_G[X] - J_G[X_0] &= \mathrm{tr}\,(XQX^*) - \mathrm{tr}\,(X_0QX_0^*) \\ &= \mathrm{tr}\,(XQX^* - XQX_0^* - X_0QX^* + X_0QX_0^*) \\ &= \mathrm{tr}\,((X - X_0)Q(X - X_0)^*) \geq 0. \end{aligned} \tag{21}$$

Then we have $J_G[X] \geq J_G[X_0]$. That is, X_0 is an optimal learning operator.

Conversely, assume that $J_G[X] = J_G[X_0]$ for X which satisfies Eq.(15). Eq.(21) yields $(X - X_0)Q = 0$, and hence $XQ = X_0Q = CA^*$. That is, X satisfies Eq.(16). It means that all optimal learning operators are given as solutions of Eqs.(15) and (16).

Finally we shall show Eq.(17). It follows from Eqs.(18) and (20) that

$$J_G[X_0] = \mathrm{tr}\,(X_0QX_0^*) = \mathrm{tr}\,(CA^*X_0^*) = \mathrm{tr}\,(C(X_0A)^*) = \mathrm{tr}\,(C),$$

which implies Eq.(17). (Q.E.D.)

This proof states that the operator C in Eq.(16) is the Lagrange multiplier and the minimal value of J_G is given by the trace of C.

Lemma 7 has transformed a variational problem for the optimal learning operator to an algebraic problem. It characterizes the optimal learning operator by using the system of two equations. The following lemma characterizes it by using only one equation.

Lemma 8 *Assume that $R(A^*) = H$. An operator X is an optimal learning operator if and only if X satisfies*

$$XU = V^{-1}A^*. \tag{22}$$

Proof. The assumption $R(A^*) = H$ guarantees that V is non-singular. From Lemma 7, it is enough to show that the system of Eqs.(15) and (16) is equivalent to Eq.(22). Let X and C be solutions of Eqs.(15) and (16). It follows from Eqs.(8), (15), (16), (9), and (13) that

$$\begin{aligned} XU &= X(AA^* + Q) = XAA^* + XQ = A^* + CA^* \\ &= (I + C)A^* = (I + C)VV^{-1}A^* \\ &= (I + C)(A^*U^\dagger A)V^{-1}A^* = ((I + C)A^*)U^\dagger AV^{-1}A^* \\ &= XUU^\dagger AV^{-1}A^* = XAV^{-1}A^* = V^{-1}A^*, \end{aligned}$$

which implies Eq.(22). This proof also guarantees the existence of a solution of Eq.(22).

Conversely, let X be a solution of Eq.(22) and $C = V^{-1} - I$. It follows from Eqs.(13), (22), and (9) that

$$XA = XUU^\dagger A = V^{-1}A^*U^\dagger A = V^{-1}V = I,$$

which implies Eq.(15). It follows from Eqs.(8), (22), and (15) that

$$\begin{aligned} XQ &= X(U - AA^*) = XU - XAA^* \\ &= V^{-1}A^* - A^* = (V^{-1} - I)A^* = CA^*, \end{aligned}$$

which implies Eq.(16). (Q.E.D.)

In the light of these lemmas, we shall prove Theorem 3

(Proof of Theorem 3)

As is shown in the proof of Lemma 8, Eq.(22) has a solution. Its general form is given by Eq.(7) because of Lemma 4. We shall show Eq.(10). As is shown in the proof of Lemma 7, we can use $C = V^{-1} - I$ as C in Eq.(17), which implies Eq.(10). (Q.E.D.)

4 Simpler expressions of the optimal learning operator

According to the noise characteristics Q, the expression of the optimal learning operator in Eq.(7) becomes much simpler. In order to show that, the following lemma is useful.

Lemma 9 (Operator pseudo-inversion lemma) *[16] Let T_1 be an operator from a Hilbert space H_2 to a Hilbert space H_1. Let T_2 be a positive semidefinite operator on H_1. If and only if $R(T_1) \subseteq R(T_2)$, it holds*

$$(T_1T_1^* + T_2)^\dagger = T_2^\dagger - T_2^\dagger T_1(I_2 + T_1^*T_2^\dagger T_1)^{-1}T_1^*T_2^\dagger,$$

where I_2 is the identity operator on H_2.

This lemma leads us to the following theorem.

Theorem 10 *Assume that $R(A^*) = H$. If $R(Q) \supseteq R(A)$, then Eqs.(7) and (10) reduce to*

$$X_0 = (A^*Q^\dagger A)^{-1}A^*Q^\dagger + Y(I - QQ^\dagger). \tag{23}$$

and

$$J_0 = \mathrm{tr}\left((A^*Q^\dagger A)^{-1}\right). \tag{24}$$

Proof. Since $R(Q) \supseteq R(A)$, Eq.(12) yields $R(U) = R(Q)$. Then, $UU^\dagger = P_{R(U)} = P_{R(Q)} = QQ^\dagger$ and the second terms of the right-hand sides of Eqs.(7) and (23) agree with each other.

In order to prove the first terms of the right-hand sides of Eqs.(7) and (23) agree with each other, let us temporarily denote $A^*Q^\dagger A$ by B. Since $Q^* = Q$, when $R(A) \subseteq R(Q)$, it follows from the operator pseudo-inversion lemma that

$$(AA^* + Q)^\dagger = Q^\dagger - Q^\dagger A(I + A^*Q^\dagger A)^{-1}A^*Q^\dagger. \tag{25}$$

Eqs.(8) and (25) yield

$$\begin{aligned} A^*U^\dagger &= A^*(AA^* + Q)^\dagger \\ &= A^*[Q^\dagger - Q^\dagger A(I + B)^{-1}A^*Q^\dagger] \\ &= A^*Q^\dagger - (A^*Q^\dagger A)(I + B)^{-1}A^*Q^\dagger \\ &= [I - B(I + B)^{-1}]A^*Q^\dagger \\ &= (I + B)^{-1}A^*Q^\dagger, \end{aligned}$$

and hence

$$A^*U^\dagger = (I + B)^{-1}A^*Q^\dagger. \tag{26}$$

Eqs.(26) and (9) yield $V = A^*U^\dagger A = (I + B)^{-1}B$. Then we have $B = (I + B)V$. Since $I + B$ and V are non-singular, B is also non-singular. Hence,

$$V^{-1} = B^{-1}(I + B). \tag{27}$$

From Eqs.(27) and (26), the first term of the right-hand side of Eq.(7) becomes

$$V^{-1}A^*U^\dagger = [B^{-1}(I+B)][(I+B)^{-1}A^*Q^\dagger] = B^{-1}A^*Q^\dagger.$$

This is the first term of the right-hand side of Eq.(23).

Finally, we shall prove Eq.(24). It follows from Eq.(27) that $V^{-1} - I = B^{-1}$. Then, Eq.(10) yields Eq.(24). (Q.E.D.)

This theorem states that if $R(Q) \supseteq R(A)$, then we can replace U in Eq.(7) with Q. Since $Q = \sigma^2 I_M$ implies $R(Q) \supseteq R(A)$, the following is a direct consequence of the theorem.

Corollary 11 *Assume that $R(A^*) = H$. If $Q = \sigma^2 I_M$ ($\sigma > 0$), then the optimal learning operator is uniquely determined and $X_0 = A^\dagger$. Furthermore, it holds that*

$$J_0 = \sigma^2 \mathrm{tr}\left((A^*A)^{-1}\right). \tag{28}$$

5 Optimal sampling operator

Active learning is a problem to design sample points $\{\boldsymbol{x}_m : 1 \leq m \leq M\}$ so that $\hat{f}$ minimizes the generalization error J_G. It is equivalent to design the sampling operator A so that A minimizes the minimum value J_0 in Eq.(10). Such an operator A is called an *optimal sampling operator*. In this section, we shall provide a necessary and sufficient condition for A to be an optimal sampling operator under some assumptions.

5.1 Optimal sampling operator

Let $K(\boldsymbol{x}, \boldsymbol{x}')$ be a reproducing kernel of H. Let us define functions $\{\psi_m : 1 \leq m \leq M\}$ by

$$\psi_m(\boldsymbol{x}) = K(\boldsymbol{x}, \boldsymbol{x}_m) : 1 \leq m \leq M. \tag{29}$$

Then, the sampling operator A is expressed by

$$A = \sum_{m=1}^{M} \left(\boldsymbol{e}_m \otimes \overline{\psi_m}\right). \tag{30}$$

Furthermore, it holds that

$$A^* = \sum_{m=1}^{M} \left(\psi_m \otimes \overline{\boldsymbol{e}_m}\right), \tag{31}$$

$$A^*A = \sum_{m=1}^{M} \left(\psi_m \otimes \overline{\psi_m}\right). \tag{32}$$

Equation (31) means that $R(A^*)$ is the subspace spanned by the set $\{\psi_m : 1 \leq m \leq M\}$. Hence, $R(A^*) = H$ holds if and only if the set $\{\psi_m : 1 \leq m \leq M\}$ spans the whole space H.

Theorem 12 *Assume that*

(i) H is a finite N-dimensional RKHS whose reproducing kernel satisfies

$$K(\boldsymbol{x}, \boldsymbol{x}) = \kappa \quad \textit{for any } \boldsymbol{x} \textit{ in } D, \tag{33}$$

where κ is a positive constant.

(ii) $R(A^*) = H$.

(iii) $Q = \sigma^2 I_M$: $\sigma > 0$. (34)

Then, J_0 in Eq.(10) is minimized if and only if

$$A^*A = \frac{\kappa M}{N} I. \tag{35}$$

In this case, the minimum value, say J^*, of J_0 is given by

$$J^* = \frac{\sigma^2 N^2}{\kappa M}. \tag{36}$$

Proof. It follows from Eqs.(29) and (33) that

$$\mathrm{tr}\,(\psi_m \otimes \overline{\psi_m}) = \|\psi_m\|^2 = K(\boldsymbol{x}_m, \boldsymbol{x}_m) = \kappa. \tag{37}$$

Then, Eq.(32) yields

$$\mathrm{tr}\,(A^*A) = \kappa M. \tag{38}$$

Let $\|T\|_2$ be the Schmidt norm of an operator T. Its rigorous definition and properties are described in Appendix. Let us temporarily denote $(A^*A)^{1/2}$ by B. The operator B is self-adjoint and non-singular because of Lemma 1. From the Schwarz inequality and Eq.(38), we have

$$\begin{aligned} N^2 &= \mathrm{tr}\,(I)^2 = \mathrm{tr}\left(BB^{-1}\right)^2 = \langle B, B^{-1}\rangle^2 \\ &\le \|B\|_2^2 \|B^{-1}\|_2^2 = \mathrm{tr}\left(B^2\right) \mathrm{tr}\left((B^2)^{-1}\right) \\ &= \mathrm{tr}\,(A^*A)\,\mathrm{tr}\left((A^*A)^{-1}\right) = \kappa M \mathrm{tr}\left((A^*A)^{-1}\right). \end{aligned}$$

Hence, Eq.(28) yields

$$J_0 = \sigma^2 \mathrm{tr}\left((A^*A)^{-1}\right) \ge \frac{\sigma^2 N^2}{\kappa M}. \tag{39}$$

Since $B \neq 0$, the equality in Eq.(39) holds if and only if $B = \lambda B^{-1}$ with λ a positive constant. That is, the equality holds if and only if

$$A^*A = \lambda I. \tag{40}$$

In this case, $\mathrm{tr}\,(A^*A) = \lambda N$. Hence, Eq.(38) yields $\lambda = \kappa M/N$. Then, Eq.(40) is equivalent to Eq.(35). Eq.(36) is clear from Eqs.(39). (Q.E.D.)

Based on this theorem, we can obtain an optimal set of sample points $\{\boldsymbol{x}_m : 1 \le m \le M\}$. It is a subject in Section 6.

5.2 Mechanism of achieving maximal generalization capability

In this subsection, we investigate how the generalization capability is maximized by Theorem 12. For this purpose, the following corollary is useful.

Corollary 13 *Under the assumptions in Theorem 12, it holds that*

$$\|Af\| = \sqrt{\tfrac{\kappa M}{N}}\|f\| \quad : f \in H. \tag{41}$$

$$\|A^\dagger \boldsymbol{u}\| = \begin{cases} \sqrt{\tfrac{N}{\kappa M}}\|\boldsymbol{u}\| & : \boldsymbol{u} \in R(A), \\ 0 & : \boldsymbol{u} \in R(A)^\perp. \end{cases} \tag{42}$$

Proof. It follows from Eq.(35) that for $f \in H$,

$$\|Af\|^2 = \langle A^*Af, f\rangle = \frac{\kappa M}{N}\|f\|^2,$$

which implies Eq.(41). For $\boldsymbol{u} \in \mathbb{C}^M$, it holds that

$$\begin{aligned}\|A^\dagger \boldsymbol{u}\|^2 &= \langle (A^\dagger)^* A^\dagger \boldsymbol{u}, \boldsymbol{u}\rangle = \langle (A^\dagger)^* (A^* A)^{-1} A^* \boldsymbol{u}, \boldsymbol{u}\rangle \\ &= \frac{N}{\kappa M}\langle (A^\dagger)^* A^* \boldsymbol{u}, \boldsymbol{u}\rangle = \frac{N}{\kappa M}\langle (AA^\dagger)^* \boldsymbol{u}, \boldsymbol{u}\rangle \\ &= \frac{N}{\kappa M}\langle P_{R(A)}\boldsymbol{u}, \boldsymbol{u}\rangle = \frac{N}{\kappa M}\|P_{R(A)}\boldsymbol{u}\|^2.\end{aligned}$$

This implies Eq.(42). (Q.E.D.)

Corollary 13 implies that $\sqrt{\frac{N}{\kappa M}}A$ becomes an *isometry* and $\sqrt{\frac{\kappa M}{N}}A^\dagger$ becomes a *partial isometry* with the initial space $R(A)$.

Using Corollary 13, we show how Theorem 12 maximizes the generalization capability. In the following, we assume that $\kappa M/N > 1$. Let us decompose the noise ϵ into $\bar{\epsilon}$ in $\mathcal{R}(A)$ and $\bar{\epsilon}^\perp$ in $\mathcal{R}(A)^\perp$:

$$\epsilon = \bar{\epsilon} + \bar{\epsilon}^\perp.$$

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

$$\boldsymbol{y} = Af + \bar{\epsilon} + \bar{\epsilon}^\perp.$$

Since $A^\dagger$ is the optimal learning operator, the signal component Af is transformed to the original function f by $A^\dagger$. Indeed, using $R(A^*) = H$, we have

$$A^\dagger A f = P_{\mathcal{R}(A^*)} f = f.$$

From Eq.(42), $A^\dagger$ suppresses the magnitude of the noise $\bar{\epsilon}$ in $\mathcal{R}(A)$ by $\sqrt{\frac{N}{\kappa M}}$ and completely removes the noise $\bar{\epsilon}^\perp$ in $\mathcal{R}(A)^\perp$:

$$\begin{aligned}\|A^\dagger \bar{\epsilon}\| &= \sqrt{\tfrac{N}{\kappa M}}\|\bar{\epsilon}\|, \\ A^\dagger \bar{\epsilon}^\perp &= 0.\end{aligned}$$

In general, it is difficult to suppress the effect of the noise $\bar{\epsilon}$ 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 $\bar{\epsilon}$ is minimized if the magnification of $A^\dagger$ for each sample value vector $\boldsymbol{y}$ is minimized. Since minimizing the magnification of $A^\dagger$ is equivalent to maximizing the magnification of A, the effect of the noise $\bar{\epsilon}$ is minimized if the norm of Af is maximized for each f in H. 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 maximal generalization capability.

6 Optimal design of sample points

The condition (35) in Theorem 12 can be characterized by using the concept of the pseudo orthogonal basis. It leads us to a method for designing the optimal set of sample points $\{\boldsymbol{x}_m : 1 \leq m \leq M\}$.

6.1 Pseudo orthogonal bases

Definition 14 *[19, 18] A set $\{u_m : 1 \leq m \leq M\}$ in an N-dimensional Hilbert space H is called a* pseudo orthogonal basis *(POB) if any f in H is expressed as*

$$f = \sum_{m=1}^{M} \langle f, u_m \rangle u_m. \tag{43}$$

Eq.(43) means $M \geq N$. That is, the concept of POB is an extension of the orthonormal basis (ONB) to linearly dependent over-complete systems. Clearly, a POB reduces to an ONB if $M = N$. POBs and their extensions, pseudo biorthogonal bases (PBOB) [14, 18], have been successfully applied to various real world problems including signal restoration [15, 18], computerized tomography [20], neural network learning [17], and robust construction of neural networks [10, 12].

Lemma 15 *[19] A set $\{u_m : 1 \leq m \leq M\}$ is a POB in H if and only if*

$$\|f\|^2 = \sum_{m=1}^{M} |\langle f, u_m \rangle|^2 \quad \text{for any } f \text{ in } H.$$

This equation is an extension of the Parseval equality. It implies that a POB is a tight frame with frame bound one [5] or a normalized tight frame [7] in the frame terminology. Eq.(43) is equivalent to

$$\sum_{m=1}^{M} (u_m \otimes \overline{u_m}) = I. \tag{44}$$

Taking the trace of Eq.(44) gives the following invariant for the POB:

$$\sum_{m=1}^{M} \|u_m\|^2 = N, \tag{45}$$

where N is the dimension of H. Note that the left-hand side of this equation is independent of not only the number of elements M but also the chosen elements $\{u_m : 1 \leq m \leq M\}$.

The following two lemmas give construction methods of POBs.

Lemma 16 *[19] Let T be an isometry from H to an M-dimensional Hilbert space H' and $\{v_m : 1 \leq m \leq M\}$ be an ONB in H'. If we let*

$$u_m = T^* v_m \quad \text{for } m = 1, 2, \ldots, M, \tag{46}$$

then the set $\{u_m : 1 \leq m \leq M\}$ becomes a POB in H.

Note that all POBs can be constructed by changing T with a fixed ONB $\{v_m : 1 \leq m \leq M\}$ or by changing $\{v_m : 1 \leq m \leq M\}$ with a fixed T.

If a set $\{u_m : 1 \leq m \leq M\}$ is a POB and $\|u_1\| = \|u_2\| = \cdots = \|u_M\|$, then the set is called a *pseudo orthonormal basis* (PONB). In this case, it follows from Eq.(45) that

$$\|u_m\| = \sqrt{\frac{N}{M}} \quad \text{for } m = 1, 2, \ldots, M. \tag{47}$$

Lemma 17 *[21] Let $M = \mu N$, where μ is a positive integer and N is the dimension of H. Then, a set $\{u_m : 1 \leq m \leq M\}$ becomes a PONB in H if a set $\{\sqrt{\mu} u_m : 1 \leq m \leq M\}$ consists of μ sets of ONBs in H.*

6.2 Optimal design of sample points

In this subsection, we shall provide a method for designing the optimal set of sample points $\{\boldsymbol{x}_m : 1 \leq m \leq M\}$ by using the concept of POB.

Theorem 18 *Let*

$$\varphi_m = \sqrt{\frac{N}{\kappa M}} \psi_m \quad : 1 \leq m \leq M. \tag{48}$$

Under the assumptions in Theorem 12, J_G in Eq.(3) is minimized if and only if $\{\varphi_m : 1 \leq m \leq M\}$ is a POB in H.

Proof. It follows from Eqs.(32) and (48) that

$$A^*A = \sum_{m=1}^{M} (\psi_m \otimes \overline{\psi_m}) = \frac{\kappa M}{N} \sum_{m=1}^{M} (\varphi_m \otimes \overline{\varphi_m}).$$

Then Eq.(35) holds if and only if $\{\varphi_m : 1 \leq m \leq M\}$ is a POB in H because of Eq.(44). (Q.E.D.)

Note that if $\{\varphi_m : 1 \leq m \leq M\}$ in Eq.(48) is a POB, then it is a PONB, because of Eq.(37). In the following subsection, Theorem 18 is applied to the trigonometric polynomial model.

6.3 Trigonometric polynomial space

In this subsection, we show optimal sets of sample points $\{\boldsymbol{x}_m : 1 \leq m \leq M\}$ for the trigonometric polynomial model based on Theorem 18.

Let us discuss functions defined on $[-\pi, \pi]$. It is easily extended to a general L-variable functions. Let H be a trigonometric polynomial space of order N_1, which is denoted by $T_{N_1}[-\pi, \pi]$. That is, $T_{N_1}[-\pi, \pi]$ is a space spanned by the functions $\{\exp(inx) : 0 \leq |n| \leq N_1\}$ with the inner product defined by

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

The dimension of $T_{N_1}[-\pi, \pi]$ is $N = 2N_1 + 1$. The reproducing kernel of $T_{N_1}[-\pi, \pi]$ is given by

$$K(x, x') = \begin{cases} \dfrac{\sin\left((2N_1+1)(x-x')/2\right)}{\sin\left((x-x')/2\right)} & \text{if } x \neq x', \\ 2N_1 + 1 & \text{if } x = x'. \end{cases} \tag{49}$$

It follows from Eq.(49) that $T_{N_1}[-\pi, \pi]$ is a RKHS that satisfies the condition in Eq.(33) with $\kappa = 2N_1 + 1 = N$. Hence, the condition (35) reduces to $A^*A = MI$. Therefore, related to Theorem 18 and Lemma 17, we have the following two sets of optimal sample points.

Theorem 19 *Let $M \geq N = 2N_1+1$ and c be an arbitrary constant such that $-\pi \leq c \leq -\pi+2\pi/M$. If we let*

$$x_m = c + \frac{2\pi}{M}(m-1) \quad : 1 \leq m \leq M, \tag{50}$$

then $\{x_m : 1 \leq m \leq M\}$ is the optimal set of sample points.

Theorem 20 *Let $M = \mu N = \mu(2N_1 + 1)$ with μ a positive integer. For $t = 1, 2, .., \mu$, let c_t be an arbitrary constant such that $-\pi \leq c_t \leq -\pi + 2\pi/N$. If we let*

$$x_m = c_t + \frac{2\pi}{N}(p-1) \quad : m = (t-1)N + p, \;\; t = 1, 2, \ldots, \mu, \;\; p = 1, 2, \ldots, N, \tag{51}$$

then $\{x_m : 1 \leq m \leq M\}$ is the optimal set of sample points.

7 Conclusion

We derived a general form of the optimal learning operator for a given sampling operator. Using the optimal learning operator, we gave a necessary and sufficient condition of sample points for maximizing the generalization capability. By utilizing the properties of pseudo orthogonal bases, we clarified the mechanism of achieving the maximal generalization capability. Based on the optimality condition, we gave design methods of optimal sample points for the trigonometric polynomial model.

Acknowledgement

M. S. acknowledges the partial financial support from the Alexander von Humboldt Foundation.

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

[7] 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.

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

[9] 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)

[10] 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)

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

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

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

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

[15] 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.

[16] H. Ogawa: "An operator pseudo-inversion lemma," SIAM J. Appl. Math., vol.48, no.6, pp.1527–1531, Dec. 1988.

[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] M. Sato and T. Iijima, "A suppression method for proportional noise in transmission system of spatial information," IECE Trans., vol. J58-D, no. 5, pp. 279–286, 1975. (In Japanese)

[22] S. Saitoh, Theory of Reproducing Kernels and Its Applications, Longman Scientific and Technical, Harlow, England, 1988.

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

[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.

Appendix: Mathematical preliminaries

For readers' convenience, we first briefly review the Schmidt inner product, the Schmidt norm, and the trace of operators. After that, the Neumann-Schatten product is introduced [23]. Let T_1 and T_2 be linear operators from an N-dimensional Hilbert space H_1 to a Hilbert space H_2. Let $\{u_n : 1 \leq n \leq N\}$ be an orthonormal basis in H_1. The following sum is independent of the chosen $\{u_n : 1 \leq n \leq N\}$.

$$\langle T_1, T_2 \rangle = \sum_{n=1}^{N} \langle T_1 u_n, T_2 u_n \rangle.$$

$\langle T_1, T_2 \rangle$ is called the Schmidt inner product. Furthermore, $\|T_1\|_2 = \sqrt{\langle T_1, T_1 \rangle}$ is called the Schmidt norm of T_1 and $\mathrm{tr}\,(T) = \langle T, I \rangle$ is called the trace of T, where T is a linear operator from H_1 to H_1 and I is the identity operator on H_1. The following formulas are used in this paper.

$$\begin{aligned} \langle T_1 X, T_2 \rangle &= \langle T_1, T_2 X^* \rangle, \\ \langle X T_1, T_2 \rangle &= \langle T_1, X^* T_2 \rangle, \\ \mathrm{tr}\,(T^2) &= \|T\|_2^2 \quad \text{if } T \text{ is self-adjoint.} \end{aligned}$$

Let u and v be given elements in Hilbert spaces H_1 and H_2, respectively. Let $u \otimes \overline{v}$ be an operator from H_2 to H_1 defined by

$$(u \otimes \overline{v})\, w = \langle w, v \rangle u,$$

where w is any element in H_2. The operator is called the Neumann-Schatten product. The following formulas are used in this paper.

$$\begin{aligned} (u \otimes \overline{v})^* &= (v \otimes \overline{u}), \\ (T_1 u) \otimes \overline{(T_2 v)} &= T_1 (u \otimes \overline{v}) T_2^*, \\ \mathrm{tr}\,(u \otimes \overline{u}) &= \|u\|^2. \end{aligned}$$