Incremental Active Learning for Optimal Generalization

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

The problem of designing input signals for optimal generalization is called active learning. In this paper, we give a two-stage sampling scheme for reducing both the bias and variance, and based on this scheme, we propose two active learning methods. One is the multi-point-search method applicable to arbitrary models. The effectiveness of this method is shown through computer simulations. The other is the optimal sampling method in trigonometric polynomial models. This method precisely specifies the optimal sampling locations.

Keywords

Active learning, generalization capability, projection learning, incremental projection learning, trigonometric polynomial space.

1 Introduction

Supervised learning is obtaining an underlying rule from sampled information. Depending on the type of sampling, supervised learning can be classified into two different categories. One is the case where information is given unilaterally from the environment. For example, in *time series prediction*, sample points are fixed to regular intervals and learners can not change the interval. The other is the case where learners can design input signals by themselves and sample corresponding output signals. For example, it is possible to design input signals in many scientific experiments or learning of *sensorimotor maps* of multi-joint robot arms. Learning can be performed more efficiently if we can actively design input signals.

The problem of designing input signals for optimal generalization is called *active learning* (Cohn, Ghahramani, & Jordan, 1996; Fukumizu, 1996; Vijayakumar & Ogawa, 1999). It is also referred to as *optimal experiments* (Kiefer, 1959; Fedorov, 1972; Cohn, 1994) or *query construction* (Sollich, 1994). *Reinforcement learning* (Kaelbling, 1996), which has been extensively studied recently in the field of machine learning, can be regarded as another form of active learning.

In mathematical statistics, an active learning criterion called the *D-optimal design* has been thoroughly studied (Kiefer, 1959; Kiefer & Wolfowitz, 1960; Fedorov, 1972). The D-optimal design minimizes the determinant of the dispersion matrix of the estimator. One of the advantages of the D-optimal design is that it is invariant under all affine transformations in the input space (Kiefer, 1959). Kiefer and Wolfowitz (1960) showed that the D-optimal design agrees with the *minimax design* when the noise variance is the same magnitude all over the domain. The minimax design is aimed at finding the sample points $\{x_j\}$ minimizing the maximum of the noise variance:

$$\underset{\{x_j\}}{\operatorname{argmin}} \max_{x} E_n|f_0(x) - f(x)|^2, \tag{1}$$

where E_n, f_0, and f are the ensemble average over the noise, a learning result, and the learning target function, respectively. In order to find the optimal design of sample points, the learning criterion prescribing the mapping from training examples to a learning result has to be determined. In a general approach to the D-optimal design, *best linear unbiased estimation* is adopted as a learning criterion, which is aimed at minimizing the mean noise variance over the domain under the constraint of *unbiasedness.* This implies that the criterion for the D-optimal design is inconsistent with the criterion for best linear unbiased estimation, causing a crucial problem for acquiring the optimal generalization capability.

Within the framework of Bayesian statistics, MacKay (1992) derived a criterion for selecting the most informative training data for specifying the parameters of neural networks. Cohn (1994, 1996) and Cohn, Ghahramani, and Jordan (1996) gave an active learning criterion for minimizing the variance of the estimator. Fukumizu (1996) proposed an active learning method in multi-layer perceptrons using asymptotic approximation for estimating the generalization error. Essentially, the criteria derived in these papers are equivalent to the A-optimal design shown in Fedorov (1972). However, it is generally intractable to calculate the value of the criteria and difficult to find the optimal solution. MacKay (1992) proposed to use a fixed set of reference points for estimating the value of the criterion. Cohn, Ghahramani, and Jordan (1996) recommended to use the Monte Carlo sampling for estimating the value of the criterion, and Cohn (1996) used the gradient method for finding a local optimum of the criterion. Fukumizu (1996) introduced a parametric family of the density function for generating sampling locations.

In the above approaches, the active learning criteria are aimed at minimizing the variance of the estimator. However, as shown in Geman, Bienenstock, and Doursat (1992), the generalization error consists of the bias and variance. This implies that the above methods assume that the bias is zero or small enough to be neglected. Yue and Hickernell (1999) showed an upper bound of the generalization error and derived an active learning criterion for minimizing the upper bound. However, this criterion includes an unknown controlling parameter of the trade-off between the bias and variance, so the optimal solution can not be obtained. Cohn (1997) used resampling methods such as the *bootstrapping* (Efron & Tibshirani, 1993) and the *cross-validation* (Stone, 1974) for estimating the bias, and proposed an active learning method for reducing the bias. His experiments showed that the bias-only approach outperforms the variance-only approach. From the functional analytic point of view, Vijayakumar and Ogawa (1999) gave a necessary and sufficient condition of the sampling locations to provide the optimal generalization capability in the absence of noise. In this condition, the bias is explicitly evaluated by utilizing the knowledge of the distribution of the learning target functions. Vijayakumar, Sugiyama, and Ogawa (1998) extended the condition to the noisy case by dividing the sampling scheme into two stages. The first stage is for reducing the bias and the second stage is for reducing the variance with the small bias attained in the first stage maintained.

In this paper, we propose two active learning methods in the presence of noise. One is the multi-point-search method applicable to arbitrary models. The effectiveness of this method is shown through computer simulations. The other is the optimal sampling method in trigonometric polynomial models. This method precisely specifies the optimal sampling locations. Both methods are based on the idea of the two-stage sampling scheme proposed in Vijayakumar, Sugiyama, and Ogawa (1998). The difference is that a priori knowledge of the distribution of the target functions is not required in the present paper. This paper is organized as follows. In Section 2, the supervised learning problem is formulated. Section 3 describes a general learning process and requirements for acquiring the optimal generalization capability. Section 4 is devoted to giving a basic sampling strategy. Based on this strategy, Section 5 gives the multi-point-search method and Section 6 gives the optimal sampling method in trigonometric polynomial models. Finally, computer simulations are performed in Section 7, demonstrating the effectiveness of the proposed methods.

2 Formulation of supervised learning problem

In this section, the supervised learning problem is formulated from the functional analytic point of view (see Ogawa, 1989, 1992).

Let us consider a supervised learning problem of obtaining the optimal approximation to a target function $f(x)$ of L variables from a set of m training examples. The training examples are made up of input signals x_j in $\mathcal{D}$, where $\mathcal{D}$ is a subset of the L-dimensional Euclidean space $\mathbf{R}^L$, and corresponding output signals y_j in the unitary space $\mathbf{C}$:

$$\{(x_j, y_j) \mid y_j = f(x_j) + n_j\}_{j=1}^{m}, \tag{2}$$

where y_j is degraded by zero-mean additive noise n_j. Let $n^{(m)}$ and $y^{(m)}$ be m-dimensional vectors whose j-th elements are n_j and y_j, respectively. $y^{(m)}$ is called a *sample value vector*, and a space to which $y^{(m)}$ belongs is called a *sample value space*. In this paper, the target function $f(x)$ is assumed to belong to a reproducing kernel Hilbert space H (Aronszajn, 1950; Bergman, 1970; Saitoh, 1988, 1997; Wahba, 1990). If H is unknown, then it can be estimated by model selection methods (e.g. Akaike, 1974; Sugiyama & Ogawa, 1999d). The reproducing kernel $K(x, x')$ is a bivariate function defined on $\mathcal{D} \times \mathcal{D}$ which satisfies the following conditions.

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

  $$\langle f(\cdot), K(\cdot, x') \rangle = f(x'), \tag{3}$$

  where $\langle \cdot, \cdot \rangle$ denotes the inner product.

Note that the reproducing kernel is unique if it exists. In the theory of the Hilbert space, arguments are developed by regarding a function as a point in that space. Thus, the value of a function at a point can not be discussed within the general framework of the Hilbert space. However, if the Hilbert space has the reproducing kernel, then it is possible to deal with the value of a function at a point. Indeed, if a function $\psi_j(x)$ is defined as

$$\psi_j(x) = K(x, x_j), \tag{4}$$

then the value of f at a sample point x_j is expressed as

$$f(x_j) = \langle f, \psi_j \rangle. \tag{5}$$

For this reason, ψ_j is called a *sampling function*. Let A_m be an operator mapping f to an m-dimensional vector whose j-th element is $f(x_j)$:

$$A_m f = (f(x_1), f(x_2), \cdots, f(x_m))^\top, \tag{6}$$

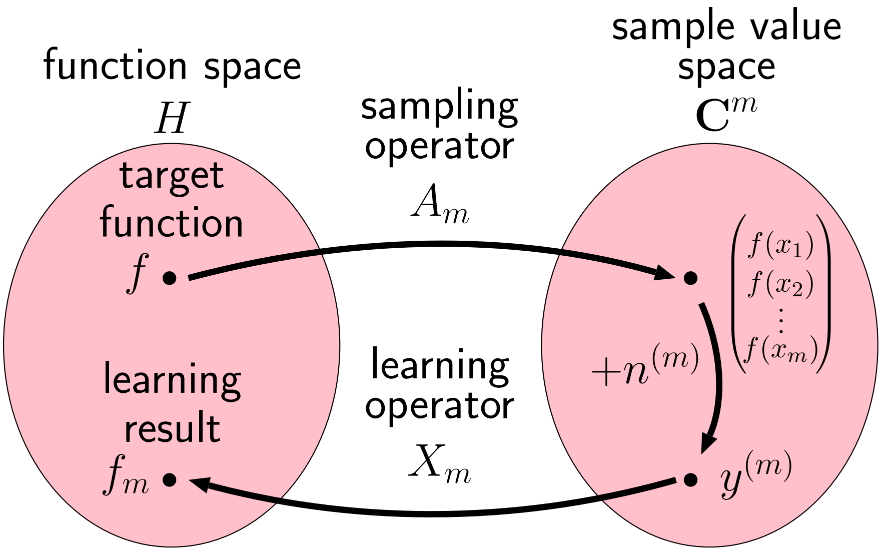


Figure 1: Supervised learning as an inverse problem.

where $\top$ denotes the transpose of a vector. We call A_m a *sampling operator*. Note that A_m is always a linear operator even when we are concerned with a non-linear function $f(x)$. Indeed, A_m can be expressed as

$$A_m = \sum_{j=1}^{m} \left(e_j^{(m)} \otimes \overline{\psi_j} \right), \tag{7}$$

where $e_j^{(m)}$ is the j-th vector of the so-called standard basis in $\mathbf{C}^m$ and $(\cdot \otimes \bar{\cdot})$ stands for the *Neumann-Schatten product*[1]. Then, the relationship between f and $y^{(m)}$ can be expressed as

$$y^{(m)} = A_m f + n^{(m)}. \tag{8}$$

Let f_m be a learning result obtained from m training examples and let us denote a mapping from $y^{(m)}$ to f_m by X_m:

$$f_m = X_m y^{(m)}, \tag{9}$$

where X_m is called a *learning operator*. Then, the supervised learning problem is reformulated as an inverse problem of obtaining X_m providing the best approximation f_m to f under a certain learning criterion (Fig.1).

[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 (Schatten, 1970)

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

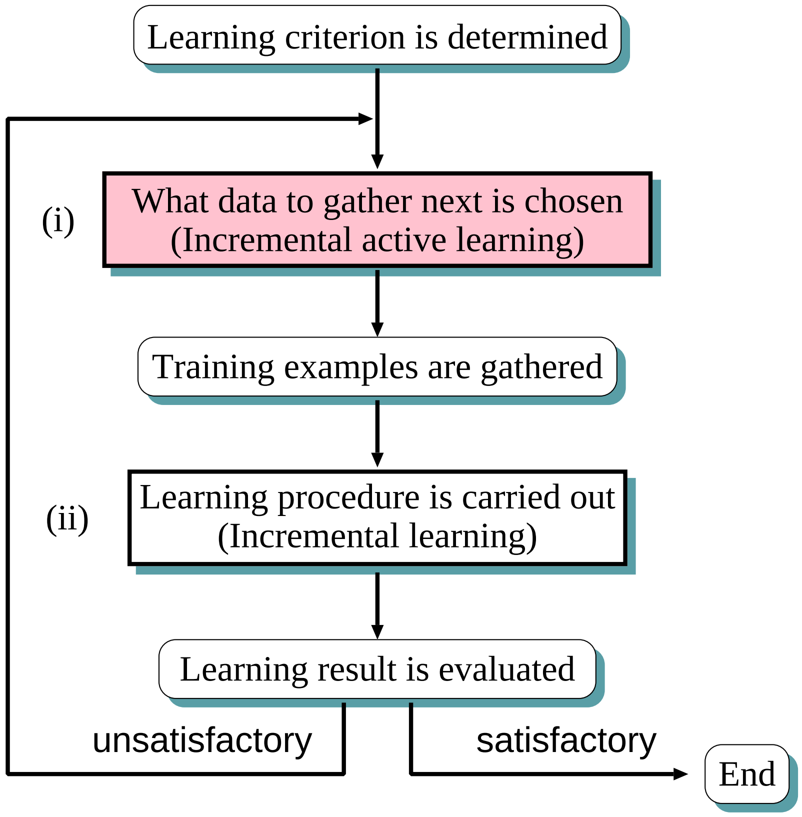

Learning criterion is determined
(i)
What data to gather next is chosen
(Incremental active learning)
Training examples are gathered
(ii)
Learning procedure is carried out
(Incremental learning)
Learning result is evaluated
unsatisfactory
satisfactory
End

Figure 2: General process for supervised learning.

3 Learning process

In this section, we show a general process for supervised learning and describe requirements for optimal generalization.

3.1 Requirements for optimal generalization

Supervised learning is generally processed as illustrated in Fig.2. First of all, the learning criterion is determined in accordance with the purpose of learning. Then, (i) what data to gather is decided and sample values are gathered at the decided locations. By using the gathered training examples, (ii) a learning procedure is carried out and the obtained learning result is evaluated. If the learning result is satisfactory, then the learning process is completed. Otherwise, training examples are added to improve the learning result until it becomes satisfactory. In this paper, training examples are sampled and added one by one along with the process.

In practical situations, the number of training examples is always finite. Hence, for acquiring the optimal generalization capability through this learning process, the following requirements should be met in the non-asymptotic sense.

(a) The criterion for active learning is consistent with the purpose of learning.

(b) The active learning method precisely specifies the optimal sample points.

(c) The incremental learning method provides exactly the same generalization capability as that obtained by batch learning with all training examples.

Strictly speaking, *optimal* in the requirement (b) has two meanings. One is the *globally* optimal, where a set of all training examples is optimal (e.g. Sugiyama & Ogawa, 2000). The other is the *greedy* optimal, where the next training example to sample is optimal in each step (e.g. MacKay, 1992; Cohn, 1994, 1996; Fukumizu, 1996). In this paper, we focus on the latter greedy case and devise an incremental active learning method meeting the above requirements. In the rest of this section, we review *projection learning* and a method of *incremental projection learning* which meets the requirement (c).

3.2 Projection learning

As mentioned in Section 2, function approximation is performed on the basis of a learning criterion. Our purpose of learning in this paper is to minimize the generalization error of the learning result f_m measured by

$$J_G = E_n \|f_m - f\|^2. \tag{10}$$

Then, the following proposition holds.

Proposition 1 *(Takemura, 1991) It holds that*

$$J_G = \|E_n f_m - f\|^2 + E_n \|f_m - E_n f_m\|^2. \tag{11}$$

The first and second terms of Eq.(11) is called the *bias* and *variance* of f_m, respectively. Let us restrict our discussion within the case where the learning operator X_m in Eq.(9) is linear. Then, it follows from Eqs.(9) and (8) that the learning result f_m can be decomposed as

$$f_m = X_m A_m f + X_m n^{(m)}. \tag{12}$$

In this case, it follows from Eq.(12) that

$$E_n f_m = X_m A_m f, \tag{13}$$

and hence the mean of f_m over the noise belongs to $\mathcal{R}(X_m A_m)$, where $\mathcal{R}(\cdot)$ denotes the range of an operator. Let P_S be the orthogonal projection operator onto a subspace S. In order to minimize the bias of f_m, $X_m A_m f$ should agree with the orthogonal projection of f onto $\mathcal{R}(X_m A_m)$:

$$X_m A_m f = P_{\mathcal{R}(X_m A_m)} f. \tag{14}$$

From Albert (1972), the operator equation

$$X_m A_m = P_S \tag{15}$$

has a solution if and only if $S \subset \mathcal{R}(A_m^*)$, where A_m^* denotes the adjoint operator of A_m. Since bigger $\mathcal{R}(X_m A_m)$ provides better approximation, we adopt the largest one:

$$\mathcal{R}(X_m A_m) = \mathcal{R}(A_m^*). \tag{16}$$

For this reason, $\mathcal{R}(A_m^*)$ is called the *approximation space*. In order to reduce the generalization error, the variance of f_m should be minimized. This learning method is called *projection learning*:

Definition 1 (Projection learning) *(Ogawa, 1987) An operator X_m is called the projection learning operator if X_m minimizes the functional*

$$J_P[X_m] = E_n \|X_m n^{(m)}\|^2 \tag{17}$$

under the constraint

$$X_m A_m = P_{\mathcal{R}(A_m^*)}. \tag{18}$$

Let $A_m^\dagger$ be the *Moore-Penrose generalized inverse*[2] of A_m. Then, the following proposition holds.

Proposition 2 *(Ogawa, 1987) A general form of the projection learning operator is expressed as*

$$X_m = V_m^\dagger A_m^* U_m^\dagger + Y_m(I_m - U_m U_m^\dagger), \tag{19}$$

where Y_m is an arbitrary operator from $\mathbf{C}^m$ to H and

$$\begin{aligned} Q_m &= E_n\left(n^{(m)} \otimes \overline{n^{(m)}}\right), && (20)\\ U_m &= A_m A_m^* + Q_m, && (21)\\ V_m &= A_m^* U_m^\dagger A_m. && (22) \end{aligned}$$

Substituting Eqs.(12), (13), and (18) into Eq.(11), we have

$$J_G = \|P_{\mathcal{R}(A_m^*)} f - f\|^2 + E_n \|X_m n^{(m)}\|^2. \tag{23}$$

Eq.(23) implies that projection learning reduces the bias of f_m to a certain level and minimizes the variance of f_m.

There are various methods for calculating the projection learning operator X_m and the projection learning result f_m by matrix operation. Here, we show one of the simplest methods valid for all finite dimensional Hilbert spaces H.

[2] An operator X is called the *Moore-Penrose generalized inverse* of an operator A if X satisfies the following four conditions (Albert, 1972; Ben-Israel & Greville, 1974).

$$AXA = A, \quad XAX = X, \quad (AX)^* = AX, \quad \text{and} \quad (XA)^* = XA.$$

Note that the Moore-Penrose generalized inverse is unique and denoted as $A^\dagger$.

When the dimension of H, denoted by μ, is finite, functions in H can be expressed in the form of

$$f(x) = \sum_{k=1}^{\mu} a_k \varphi_k(x), \tag{24}$$

where $\{\varphi_k\}_{k=1}^{\mu}$ is an orthonormal basis in H and $\{a_k\}_{k=1}^{\mu}$ is its coefficients. Let us consider a μ-dimensional parameter space in which functions in H are expressed as

$$f = (a_1, a_2, \cdots, a_\mu)^\top. \tag{25}$$

If we regard this parameter space as H, then the sampling function ψ_j is expressed as

$$\psi_j = (\varphi_1(x_j), \varphi_2(x_j), \cdots, \varphi_\mu(x_j))^*, \tag{26}$$

where $(a_1, a_2, \cdots, a_\mu)^*$ denotes the complex conjugate of the transpose of $(a_1, a_2, \cdots, a_\mu)$. Hence, the sampling operator A_m becomes an $m \times \mu$ matrix whose (j, k)-element is

$$[A_m]_{jk} = \varphi_k(x_j). \tag{27}$$

This A_m is sometimes called the *design matrix* (Efron & Tibshirani, 1993). Then, the projection learning operator obtained by Eq.(19) becomes an $\mu \times m$ matrix, and the projection learning result f_m obtained by Eq.(9) becomes a μ-dimensional vector:

$$f_m = (b_1, b_2, \cdots, b_\mu)^\top. \tag{28}$$

From this, we have the learning result function $f_m(x)$ as

$$f_m(x) = \sum_{k=1}^{\mu} b_k \varphi_k(x). \tag{29}$$

In practice, the calculation of the Moore-Penrose generalized inverse is sometimes unstable. To overcome the unstableness, we recommend to use *Tikhonov's regularization* (Tikhonov & Arsenin, 1997):

$$A_m^\dagger \longleftarrow A_m^*(A_m A_m^* + \epsilon I)^{-1}, \tag{30}$$

where ϵ is a small constant, say $\epsilon = 10^{-4}$.

It has been shown that learning results obtained by projection learning are invariant under the inner product in the sample value space (Yamashita & Ogawa, 1992). Hence, without loss of generality, the Euclidean inner product is adopted in the sample value space.

When the noise covariance matrix Q_m is in the form of

$$Q_m = \sigma^2 I_m \tag{31}$$

with $\sigma^2 > 0$, the projection learning operator is given as (Ogawa, 1987)

$$X_m = A_m^\dagger. \tag{32}$$

This implies that projection learning agrees with usual least mean squares learning aimed at minimizing the *training error*

$$\sum_{j=1}^{m} |f_m(x_j) - y_j|^2. \tag{33}$$

3.3 Incremental projection learning

Let us consider the case where a new training example (x_{m+1}, y_{m+1}) is added after a learning result f_m has been obtained from $\{(x_j, y_j)\}_{j=1}^{m}$. It follows from Eq.(9) that a learning result f_{m+1} obtained from $\{(x_j, y_j)\}_{j=1}^{m+1}$ in a batch manner can be expressed as

$$f_{m+1} = X_{m+1} y^{(m+1)}. \tag{34}$$

In order to devise an exact incremental learning method meeting the requirement (c) shown in Section 3.1, let us calculate f_{m+1} in Eq.(34) by using f_m and (x_{m+1}, y_{m+1}). Let the noise characteristics of (x_{m+1}, y_{m+1}) be

$$\begin{aligned} q_{m+1} &= E_n(\overline{n_{m+1}} n^{(m)}), && (35)\\ \sigma_{m+1} &= E_n |n_{m+1}|^2, && (36) \end{aligned}$$

where $\overline{n_{m+1}}$ denotes the complex conjugate of n_{m+1}. Note that q_{m+1} is an m-dimensional vector while σ_{m+1} is a scalar. Let $\mathcal{N}(A_m)$ be the null space of A_m and the following notation is defined.

Vectors:

$$\begin{aligned} s_{m+1} &= A_m \psi_{m+1} + q_{m+1}, && (37)\\ t_{m+1} &= U_m^{\dagger} s_{m+1}. && (38) \end{aligned}$$

Scalars:

$$\begin{aligned} \alpha_{m+1} &= \psi_{m+1}(x_{m+1}) + \sigma_{m+1} - \langle t_{m+1}, s_{m+1} \rangle, && (39)\\ \beta_{m+1} &= y_{m+1} - f_m(x_{m+1}) - \langle y^{(m)} - A_m f_m, t_{m+1} \rangle. && (40) \end{aligned}$$

Functions:

$$\begin{aligned} \tilde{\psi}_{m+1} &= P_{\mathcal{N}(A_m)} \psi_{m+1} \quad (= \psi_{m+1} - A_m^{\dagger} A_m \psi_{m+1}), && (41)\\ \xi_{m+1} &= \psi_{m+1} - A_m^{*} t_{m+1}, && (42)\\ \tilde{\xi}_{m+1} &= V_m^{\dagger} \xi_{m+1}. && (43) \end{aligned}$$

From Eq.(5), $\psi_{m+1}(x_{m+1})$ in Eq.(39) agrees with $\|\psi_{m+1}\|^2$.

As shown in Sugiyama and Ogawa (1999a, 1999c), the additional training examples such that $\alpha_{m+1} = 0$ can be rejected since they have no effect on learning results. Hence, from here on, we focus on the training examples such that $\alpha_{m+1} \neq 0$. Then, the exact incremental learning method called *incremental projection learning* (IPL) is given as follows.

Proposition 3 (Incremental projection learning) *(Sugiyama & Ogawa, 1999a, 1999b) When α_{m+1} defined by Eq.(39) is not zero, a posterior projection learning result f_{m+1} can be obtained by using prior results f_m, A_m, $U_m^{\dagger}$, $V_m^{\dagger}$, and $y^{(m)}$ as follows.*

(a) When $\psi_{m+1} \notin \mathcal{R}(A_m^*)$,

$$f_{m+1} = f_m + \frac{\beta_{m+1}}{\tilde{\psi}_{m+1}(x_{m+1})} \tilde{\psi}_{m+1}. \tag{44}$$

(b) When $\psi_{m+1} \in \mathcal{R}(A_m^*)$,

$$f_{m+1} = f_m + \frac{\beta_{m+1}}{\alpha_{m+1} + \langle \tilde{\xi}_{m+1}, \xi_{m+1} \rangle} \tilde{\xi}_{m+1}. \tag{45}$$

Note that f_{m+1} obtained by Proposition 3 exactly agrees with the learning result obtained by batch projection learning (Eq.(34)) with $\{(x_j, y_j)\}_{j=1}^{m+1}$. $\tilde{\psi}_{m+1}(x_{m+1})$ in Eq.(44) is equivalent to $\|\tilde{\psi}_{m+1}\|^2$ (see Eqs.(5) and (41)). The condition $\psi_{m+1} \notin \mathcal{R}(A_m^*)$ means that ψ_{m+1} is linearly independent of $\{\psi_j\}_{j=1}^m$, i.e., the approximation space $\mathcal{R}(A_{m+1}^*)$ becomes wider than $\mathcal{R}(A_m^*)$. In contrast, $\psi_{m+1} \in \mathcal{R}(A_m^*)$ means that ψ_{m+1} is linearly dependent of $\{\psi_j\}_{j=1}^m$, and hence the approximation space $\mathcal{R}(A_{m+1}^*)$ is equal to $\mathcal{R}(A_m^*)$.

Let us consider the case where the noise covariance matrix Q_{m+1} is positive definite[3] and diagonal, i.e.,

$$Q_{m+1} = \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_{m+1}), \tag{46}$$

where $\sigma_j > 0$ for all j. Let β'_{m+1} and V'_m be

$$\begin{aligned} \beta'_{m+1} &= y_{m+1} - f_m(x_{m+1}), & (47)\\ V'_m &= A_m^* Q_m^{-1} A_m. & (48) \end{aligned}$$

In β'_{m+1}, the third term in β_{m+1} vanishes. In V'_m, $U_m^\dagger$ in V_m is replaced with Q_m^{-1}. Then, IPL is reduced to the following simpler form.

Proposition 4 *(Sugiyama & Ogawa, 1999a, 1999c) If Q_{m+1} is given by Eq.(46) with $\sigma_j > 0$ for all j, then a posterior projection learning result f_{m+1} can be obtained by using prior results f_m and $V_m'^\dagger$ as follows.*

(a) When $\psi_{m+1} \notin \mathcal{R}(A_m^*)$,

$$f_{m+1} = f_m + \frac{\beta'_{m+1}}{\tilde{\psi}_{m+1}(x_{m+1})} \tilde{\psi}_{m+1}. \tag{49}$$

(b) When $\psi_{m+1} \in \mathcal{R}(A_m^*)$,

$$f_{m+1} = f_m + \frac{\beta'_{m+1}}{\sigma_{m+1} + \langle V_m'^\dagger \psi_{m+1}, \psi_{m+1} \rangle} V_m'^\dagger \psi_{m+1}. \tag{50}$$

Compared with Proposition 3, β_{m+1} is replaced with β'_{m+1}, and α_{m+1}, ξ_{m+1}, and $\tilde{\xi}_{m+1}$ are not required in Proposition 4. Again, f_{m+1} obtained by Proposition 4 exactly agrees with the learning result obtained by batch projection learning with $\{(x_j, y_j)\}_{j=1}^{m+1}$. If $\sigma_1 = \sigma_2 = \cdots = \sigma_{m+1} = \sigma^2$ (> 0), i.e.,

$$Q_{m+1} = \sigma^2 I_{m+1}, \tag{51}$$

[3] An operator T is said to be *positive definite* if $\langle Tf, f \rangle > 0$ for any $f \neq 0$.

then Eq.(50) yields

$$f_{m+1} = f_m + \frac{\beta'_{m+1}}{1 + \langle (A_m^* A_m)^\dagger \psi_{m+1}, \psi_{m+1} \rangle} (A_m^* A_m)^\dagger \psi_{m+1}. \tag{52}$$

Eqs.(49) and (52) imply that the value of σ^2 is not required for IPL when Q_{m+1} is in the form of Eq.(51).

4 Basic sampling strategy

This section is devoted to giving a sampling strategy which is the basis for devising active learning methods in the following sections.

Let J_b and J_v be the changes in the bias and variance of f_m through the addition of a training example (x_{m+1}, y_{m+1}), respectively, i.e.,

$$\begin{aligned} J_b &= \|P_{\mathcal{R}(A_{m+1}^*)} f - f\|^2 - \|P_{\mathcal{R}(A_m^*)} f - f\|^2, && (53) \\ J_v &= E_n \|X_{m+1} n^{(m+1)}\|^2 - E_n \|X_m n^{(m)}\|^2. && (54) \end{aligned}$$

Then, the following proposition holds.

Proposition 5 *(Sugiyama & Ogawa, 1999a, 1999c) For any additional training example (x_{m+1}, y_{m+1}) such that $\alpha_{m+1} \neq 0$, the following relations hold.*

(a) When $\psi_{m+1} \notin \mathcal{R}(A_m^*)$,

$$J_b \leq 0 \text{ and } J_v \geq 0. \tag{55}$$

(b) When $\psi_{m+1} \in \mathcal{R}(A_m^*)$,

$$J_b = 0 \text{ and } J_v \leq 0. \tag{56}$$

Proposition 5 states that an additional training example such that $\psi_{m+1} \notin \mathcal{R}(A_m^*)$ reduces or maintains the bias while it increases or maintains the variance. In contrast, an additional training example such that $\psi_{m+1} \in \mathcal{R}(A_m^*)$ maintains the bias while it reduces or maintains the variance.

Let us consider the case where the dimension of the Hilbert space H is finite, and the total number M of training examples to sample is larger than or equal to the dimension of H. In this case, it follows from Eq.(23) that the bias of learning results is zero for any f in H if and only if $\mathcal{N}(A_m) = \{0\}$. For this reason, we comply with the *two-stage sampling scheme* shown in Fig.3.

We start from $m = 0$. In Stage 1, training examples such that $\psi_{m+1} \notin \mathcal{R}(A_m^*)$ are added to reduce the bias until it reaches zero. Let μ be the dimension of H. Stage 1 ends if a training example such that $\psi_{m+1} \notin \mathcal{R}(A_m^*)$ is added μ times by which $\mathcal{N}(A_\mu) = \{0\}$ can be attained. Then, in Stage 2, training examples such that $\psi_{m+1} \in \mathcal{R}(A_m^*)$ are added to reduce the variance until the number of added training examples becomes M. Note

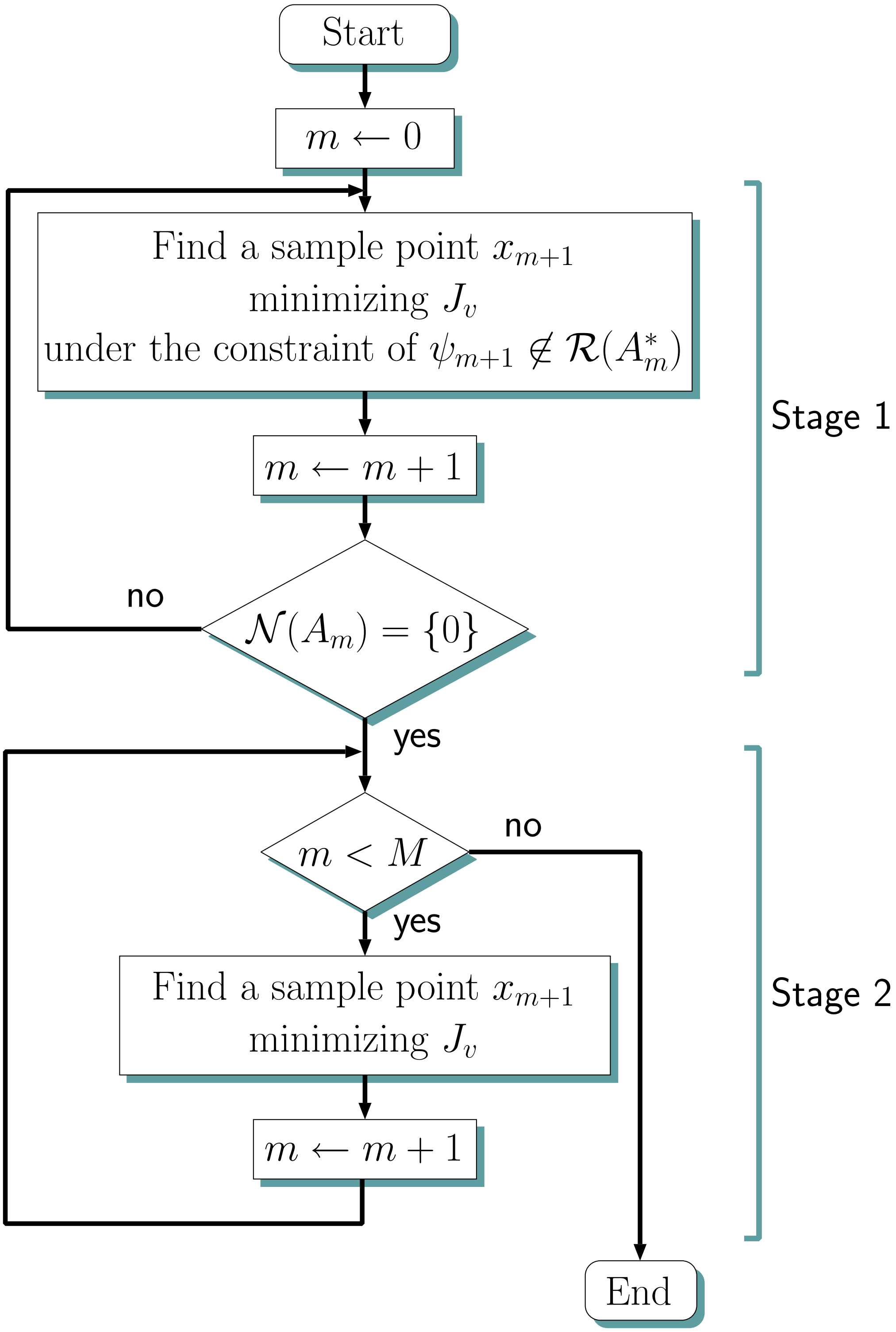

Start
$m \leftarrow 0$
Find a sample point x_{m+1}
minimizing J_v
under the constraint of $\psi_{m+1} \notin \mathcal{R}(A_m^*)$
$m \leftarrow m+1$
no
$\mathcal{N}(A_m) = \{0\}$
yes
Stage 1
no
$m < M$
yes
Find a sample point x_{m+1}
minimizing J_v
$m \leftarrow m+1$
Stage 2
End

Figure 3: Two-stage sampling scheme.

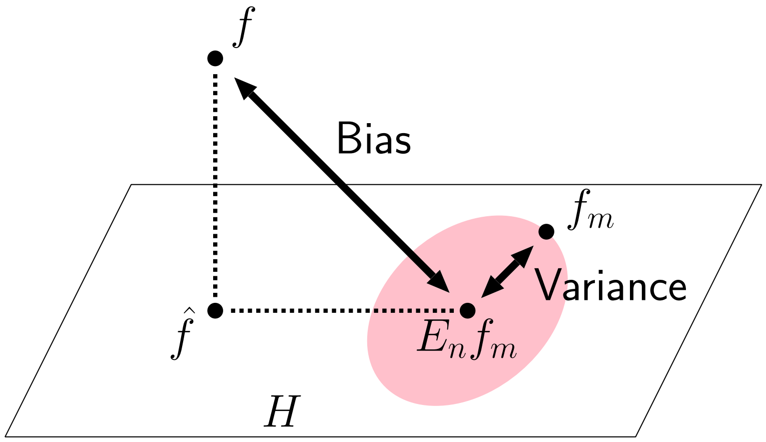


Figure 4: The interpretation of the assumptions in statistical active learning methods and our method. Let $\hat{f}$ be a function which is the best approximation to f in H. Statistical active learning methods assume that $f = \hat{f}$ and $\hat{f} = E_n f_m$. Namely, f belongs to H and the mean of f_m over the noise agrees with $\hat{f}$. In contrast, our method only assumes that $f \in H$. The difference between $\hat{f}$ and $E_n f_m$ is explicitly evaluated in Stage 1.

that the additional training examples such that $\psi_{m+1} \in \mathcal{R}(A_m^*)$ maintain the bias (see Proposition 5 (b)). Hence, the bias remains zero throughout Stage 2.

As mentioned in Section 3.1, the criterion for active learning should be consistent with the purpose of learning, i.e., the active learning criterion should be aimed at minimizing the generalization error. Therefore, our active learning problems in both stages become as follows.

Stage 1: Find a sample point minimizing J_v under the constraint of $\psi_{m+1} \notin \mathcal{R}(A_m^*)$.

Stage 2: Find a sample point minimizing J_v under the constraint of $\psi_{m+1} \in \mathcal{R}(A_m^*)$.

Note that all additional training examples in Stage 2 satisfy $\psi_{m+1} \in \mathcal{R}(A_m^*)$. This means that, in Stage 2, the constraint $\psi_{m+1} \in \mathcal{R}(A_m^*)$ does not have to be taken into account. The condition $\psi_{m+1} \notin \mathcal{R}(A_m^*)$ in Stage 1 can be easily verified since $\psi_{m+1} \notin \mathcal{R}(A_m^*)$ if and only if

$$P_{\mathcal{N}(A_m)}\psi_{m+1} = \tilde{\psi}_{m+1} \neq 0. \tag{57}$$

In practice, we recommend to use the following criterion.

$$\texttt{if } \|\tilde{\psi}_{m+1}\|^2 > \epsilon \texttt{ then } \psi_{m+1} \notin \mathcal{R}(A_m^*),$$

where ϵ is a small constant, say $\epsilon = 10^{-4}$.

In the statistical active learning methods devised so far, the bias of the estimator is assumed to be zero (MacKay, 1992; Cohn, 1994; Fukumizu, 1996). The interpretation of this assumption is illustrated in Fig.4. Let $\hat{f}$ be a function which is the best approximation to f in H. Then, the assumption of zero-bias is equivalent to $f = \hat{f}$ and $\hat{f} = E_n f_m$. Namely, f belongs to H and the mean of f_m over the noise agrees with $\hat{f}$. In contrast, the condition assumed in our framework is only $f \in H$. The difference between $\hat{f}$ and $E_n f_m$ is explicitly evaluated in Stage 1.

Based on the two-stage sampling scheme shown in Fig.3, we propose two active learning methods in the following sections.

5 Multi-point-search active learning

In this section, we propose an active learning method based on the multi-point-search.

In the derivation of the multi-point-search method, the following theorem plays a central role.

Theorem 1 *J_v defined by Eq.(54) can be expressed as follows.*

(a) When $\psi_{m+1} \notin \mathcal{R}(A_m^*)$,

$$J_v = \frac{\alpha_{m+1} + \langle \tilde{\xi}_{m+1}, \xi_{m+1} \rangle}{\tilde{\psi}_{m+1}(x_{m+1})} - 1. \tag{58}$$

(b) When $\psi_{m+1} \in \mathcal{R}(A_m^*)$,

$$J_v = -\frac{\|\tilde{\xi}_{m+1}\|^2}{\alpha_{m+1} + \langle \tilde{\xi}_{m+1}, \xi_{m+1} \rangle}. \tag{59}$$

Proofs of all theorems are given in Appendix A. Theorem 1 implies that J_v can be calculated without y_{m+1}. Namely, we can evaluate the quality of additional training examples only by using their sampling locations. It should be noted that Eq.(59) is conceptually similar to the criteria shown in Fedorov (1972), MacKay (1992), and Cohn (1994). The difference is that the noise is assumed to be *i.i.d.* in their methods while correlated noise can be treated in our method if the noise covariance matrix is available. When the noise is uncorrelated, Theorem 1 can be reduced to the following simpler form.

Theorem 2 *If Q_{m+1} is given by Eq.(46) with $\sigma_j > 0$ for all j, then J_v can be expressed as follows.*

(a) When $\psi_{m+1} \notin \mathcal{R}(A_m^*)$,

$$J_v = \frac{\sigma_{m+1} + \langle V_m'^{\dagger} \psi_{m+1}, \psi_{m+1} \rangle}{\tilde{\psi}_{m+1}(x_{m+1})}. \tag{60}$$

(b) When $\psi_{m+1} \in \mathcal{R}(A_m^*)$,

$$J_v = -\frac{\|V_m'^{\dagger} \psi_{m+1}\|^2}{\sigma_{m+1} + \langle V_m'^{\dagger} \psi_{m+1}, \psi_{m+1} \rangle}. \tag{61}$$

Compared with Theorem 1, α_{m+1}, ξ_{m+1}, and $\tilde{\xi}_{m+1}$ are not required in Theorem 2. If $\sigma_1 = \sigma_2 = \cdots = \sigma_{m+1} = \sigma^2$ (> 0), i.e.,

$$Q_{m+1} = \sigma^2 I_{m+1}, \tag{62}$$

then Theorem 2 becomes as follows.

Corollary 1 *If Q_{m+1} is given by Eq.(62), then J_v can be expressed as follows.*

(a) When $\psi_{m+1} \notin \mathcal{R}(A_m^*)$,

$$J_v = \sigma^2 \frac{1 + \langle (A_m^* A_m)^\dagger \psi_{m+1}, \psi_{m+1} \rangle}{\tilde{\psi}_{m+1}(x_{m+1})}. \tag{63}$$

(b) When $\psi_{m+1} \in \mathcal{R}(A_m^*)$,

$$J_v = -\sigma^2 \frac{\|(A_m^* A_m)^\dagger \psi_{m+1}\|^2}{1 + \langle (A_m^* A_m)^\dagger \psi_{m+1}, \psi_{m+1} \rangle}. \tag{64}$$

Corollary 1 implies that when Q_{m+1} is given by Eq.(62) with $\sigma^2 > 0$, the value of σ^2 is not required for the minimization of J_v. Based on the above theorems and corollary, the algorithm of the multi-point search active learning method is described in Fig.5.

Strictly, the algorithm shown in Fig.5 does not meet the requirement (b) mentioned in Section 3.1. However, it will be experimentally shown through computer simulations in Section 7 that the multi-point-search method specifies a better sampling location.

If the dimension of the input x is very large, many candidates may be required for finding a better sampling location. One of the measures is to employ the gradient method for finding a local maximum, i.e., with some initial value, x_{m+1} is updated until a certain convergence criterion holds as

$$x_{m+1} \longleftarrow x_{m+1} - \epsilon \Delta J_v(x_{m+1}), \tag{65}$$

where ϵ is a small positive constant and $\Delta J_v(x_{m+1})$ is the gradient of J_v at x_{m+1}.

6 Optimal sampling in the trigonometric polynomial space

In the previous section, we gave an active learning method for general finite dimensional Hilbert spaces. In this section, we focus on the trigonometric polynomial space and devise a more effective active learning method. This method strictly meets the requirement (b) described in Section 3.1.

First, our model, the trigonometric polynomial space is defined as follows.

```
$m \leftarrow 0$;
while $\mathcal{N}(A_m) \neq \{0\}$ {
    Generate $c$ locations $\{x_{m+1}^{(j)}\}_{j=1}^{c}$ such that $\psi_{m+1} \notin \mathcal{R}(A_m^*)$ as candidates;
    $j_0 \leftarrow \mathop{\mathrm{argmin}}_j J_v(x_{m+1}^{(j)})$;
    Sample $y_{m+1}$ at $x_{m+1}^{(j_0)}$;
    Carry out IPL with $(x_{m+1}^{(j_0)}, y_{m+1})$;
    $m \leftarrow m + 1$;
}
while $m < M$ {
    Generate $c$ locations $\{x_{m+1}^{(j)}\}_{j=1}^{c}$ as candidates;
    $j_0 \leftarrow \mathop{\mathrm{argmin}}_j J_v(x_{m+1}^{(j)})$;
    Sample $y_{m+1}$ at $x_{m+1}^{(j_0)}$;
    Carry out IPL with $(x_{m+1}^{(j_0)}, y_{m+1})$;
    $m \leftarrow m + 1$;
}
```

Figure 5: Algorithm of the multi-point-search active learning method.

Definition 2 (Trigonometric polynomial space) *Let $x = (\xi^{(1)}, \xi^{(2)}, \cdots, \xi^{(L)})^\top$. For $1 \leq l \leq L$, 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 $(N_1, N_2, \cdots, N_L)$ if H is spanned by*

$$\left\{ \prod_{l=1}^{L} \exp(in_l \xi^{(l)}) \right\}_{n_1=-N_1, n_2=-N_2, \cdots, n_L=-N_L}^{N_1, N_2, \cdots, N_L} \tag{66}$$

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

$$\langle f, g \rangle = \frac{1}{(2\pi)^L} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} f(x) \overline{g(x)} d\xi^{(1)} d\xi^{(2)} \cdots d\xi^{(L)}. \tag{67}$$

Note that the function space spanned by $\{\exp(in\xi)\}_{n=-N}^{N}$ is equal to the function space spanned by $\{1, \cos n\xi, \sin n\xi\}_{n=1}^{N}$. The dimension μ of a trigonometric polynomial space of order $(N_1, N_2, \cdots, N_L)$ is

$$\mu = \prod_{l=1}^{L} (2N_l + 1). \tag{68}$$

The reproducing kernel of a trigonometric polynomial space of order $(N_1, N_2, \cdots, N_L)$ is expressed as

$$K(x, x') = \prod_{l=1}^{L} K_l(\xi^{(l)}, \xi^{(l)\prime}), \tag{69}$$

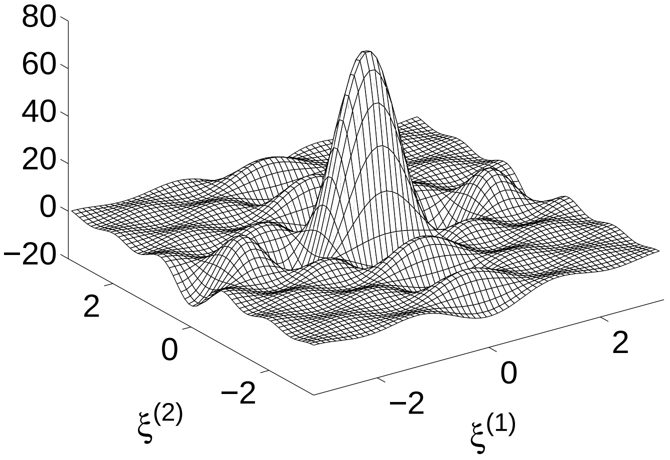


Figure 6: Profile of the reproducing kernel of a trigonometric polynomial space of order (5,3) with $x' = (0,0)^\top$.

where

$$K_l(\xi^{(l)}, \xi^{(l)\prime}) = \begin{cases} \sin \dfrac{(2N_l+1)(\xi^{(l)} - \xi^{(l)\prime})}{2} \Big/ \sin \dfrac{\xi^{(l)} - \xi^{(l)\prime}}{2} & \text{if } \xi^{(l)} \neq \xi^{(l)\prime}, \\ 2N_l + 1 & \text{if } \xi^{(l)} = \xi^{(l)\prime}. \end{cases} \tag{70}$$

The profile of Eq.(69) is illustrated in Fig.6. Then, we have the following theorem.

Theorem 3 *Suppose that the noise covariance matrix Q_M is*

$$Q_M = \sigma^2 I_M, \tag{71}$$

with $\sigma^2 > 0$. For $0 \leq k \leq \lfloor \frac{M-1}{\mu} \rfloor$ and $1 \leq l \leq L$, let $c_k^{(l)}$ be an arbitrary constant such that $-\pi \leq c_k^{(l)} \leq -\pi + \frac{2\pi}{2N_l+1}$, where $\lfloor c \rfloor$ denotes the maximum integer less than or equal to c. If we put

$$x_{m+1} = (\xi_{m+1}^{(1)}, \xi_{m+1}^{(2)}, \cdots, \xi_{m+1}^{(L)})^\top \ : \ \xi_{m+1}^{(l)} = c_p^{(l)} + \frac{2\pi}{2N_l+1} q_l \tag{72}$$

where

$$p = \left\lfloor \frac{m}{\mu} \right\rfloor, \tag{73}$$

$$q_l = \begin{cases} m \bmod (2N_1+1) & \text{if } l = 1, \\ \left\lfloor \frac{m}{\prod_{r=1}^{l-1}(2N_r+1)} \right\rfloor \bmod (2N_l+1) & \text{if } 2 \leq l \leq L, \end{cases} \tag{74}$$

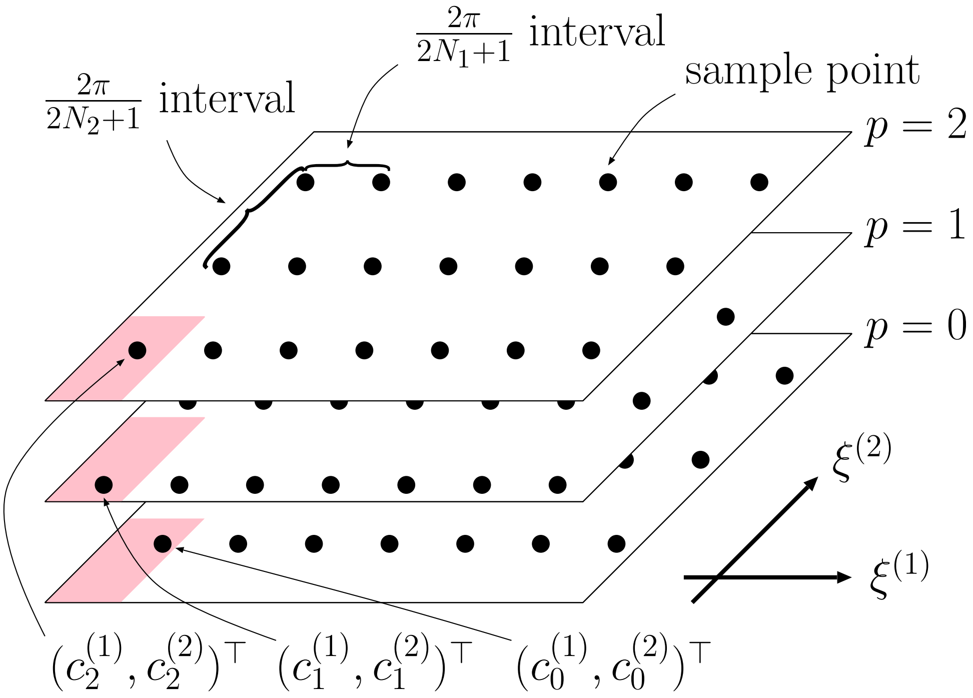

$\frac{2\pi}{2N_1+1}$ interval
$\frac{2\pi}{2N_2+1}$ interval
sample point
$p=2$
$p=1$
$p=0$
$\xi^{(2)}$
$\xi^{(1)}$
$(c_2^{(1)}, c_2^{(2)})^\top$
$(c_1^{(1)}, c_1^{(2)})^\top$
$(c_0^{(1)}, c_0^{(2)})^\top$

Figure 7: Optimal sample points in a trigonometric polynomial space of order $(3, 1)$. The number M of training examples is $7 \times 3 \times 3 = 63$.

then x_{m+1} minimizes J_v under the constraint that $\{x_j\}_{j=1}^m$ are successively determined by Eq.(72).

Theorem 3 states that for each p, μ locations are fixed at regular intervals. Fig.7 illustrates a set of 63 sample points determined by Theorem 3 when $L = 2$, $N_1 = 3$, and $N_2 = 1$. For each p, the base point $(c_p^{(1)}, c_p^{(2)})^\top$ is fixed in the dark region, and 21 sample points are fixed at regular intervals. From Eq.(69), a set $\{\frac{1}{\sqrt{\mu}}\psi_{p\mu+q}\}_{q=1}^{\mu}$ of sampling functions forms an orthonormal basis in H for each p. Although this sampling scheme is derived as the greedy optimal scheme, this scheme is in fact globally optimal at the same time when $(M \bmod \mu) = 0$ (see Sugiyama & Ogawa, 2000).

7 Computer simulations

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

7.1 Illustrative simulation

Let us consider learning in a trigonometric polynomial space of order 3. Let the target function $f(x)$ be

$$f(x) = \sqrt{2}\sin x + \sqrt{2}\cos x + \frac{1}{2\sqrt{2}}\sin 2x + \frac{1}{\sqrt{2}}\cos 2x - \sqrt{2}\sin 3x + \sqrt{2}\cos 3x. \tag{75}$$

Let the noise covariance matrix be

$$Q_M = I_M. \tag{76}$$

We shall compare the performance of following sampling schemes.

(a) Optimal sampling: Training examples are sampled following Theorem 3 with $c_k^{(1)} = -\pi + \frac{\pi}{2\times 3+1}$ for all k.

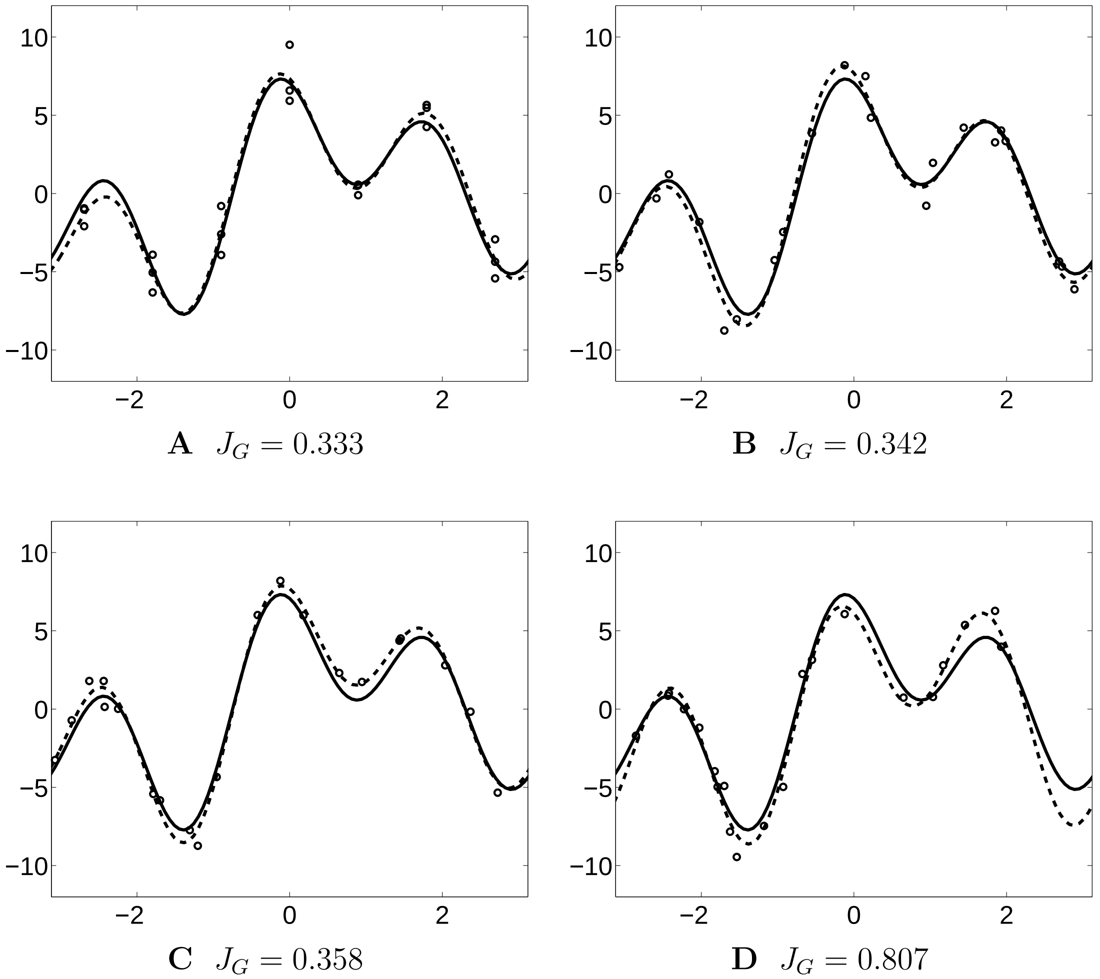
(b) Multi-point-search: Training examples are sampled following the multi-point-search method shown in Fig.5. Let the number c of candidates be 3 and they are randomly generated in the domain $[-\pi, \pi]$.

(c) Experiment design: Eq.(2) in Cohn (1994) is adopted as the active learning criterion. The value of this criterion is evaluated by the Monte Carlo sampling with 30 reference points. The next sampling location is determined by the multi-point-search with 3 candidates. Namely, 3 locations are randomly created in the domain and the one minimizing the criterion is selected.

(d) Passive learning: Training examples are randomly supplied from the domain.

Learning results obtained by the above sampling schemes with 21 training examples are shown in Fig.8. The solid and dashed lines show the target function $f(x)$ and learning results, respectively. $\circ$ denotes a training example. Fig.8 **A**–**D** show the learning results obtained by the sampling schemes (a)–(d), where the generalization errors measured by Eq.(10) are 0.333, 0.342, 0.358, and 0.807, respectively. These results show that the sampling schemes (a), (b), and (c) give 58.7, 57.6, and 55.7 percent reductions in the generalization error, respectively, compared with the sampling scheme (d). The generalization capability acquired by the sampling schemes (b) and (c) are close to that obtained by the sampling scheme (a). This means that the sampling schemes (b) and (c) works quite well with a small number of candidates since the sampling schemes (a) gives the optimal generalization capability (see Section 6).

The changes in the variance through the addition of training examples are shown in Fig.9. The horizontal axis denotes the number of training examples while the vertical axis denotes the variance measured by Eq.(17). The solid line shows the sampling scheme (a). The dashed, dash-dotted, and dotted lines denote the means of 100 trials by the sampling schemes (b), (c), and (d), respectively. In the sampling schemes (a) and (b), it always hold that $\mathcal{N}(A_7) = \{0\}$ because the dimension of H is 7 (see Section 4). In the sampling schemes (c) and (d), $\mathcal{N}(A_7) = \{0\}$ was attained in all 100 trials in this simulation. Hence,



A $J_G = 0.333$

B $J_G = 0.342$

C $J_G = 0.358$

D $J_G = 0.807$

Figure 8: Results of learning simulation in a trigonometric polynomial space of order 3 with the noise covariance matrix $Q_{21} = I_{21}$. The solid line shows the target function $f(x)$. Dashed lines in **A**–**D** show the learning results obtained by the optimal sampling method (Theorem 3), the multi-point-search method (Fig.5), the experimental design method, and passive learning, respectively. $\circ$ denotes a training example. The generalization error J_G is measured by Eq.(10).

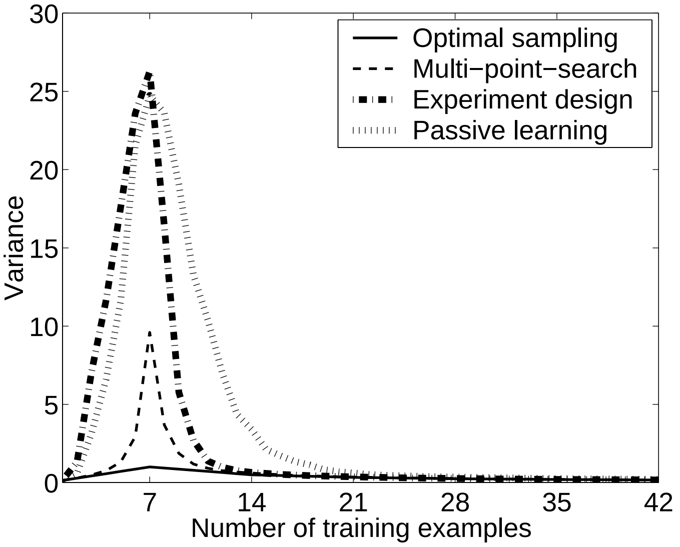


Figure 9: Relation between the number of training examples and the noise variance in a trigonometric polynomial space of order 3 with the noise covariance matrix $Q_M = I_M$. The horizontal axis denotes the number of training examples while the vertical axis denotes the noise variance measured by Eq.(17). The vertical axis can be regarded as the generalization error when the number of training examples is larger than or equal to 7.

it follows from Eq.(23) that the vertical axis in Fig.9 can be regarded as the generalization error when $m \geq 7$.

This graph shows that, when $m \leq 7$, the variance of all sampling schemes increases, this phenomenon is in good agreement with Proposition 5. In this case, the sampling schemes (a) and (b) give much lower variance than the sampling schemes (c) and (d). It should be noted that the sampling scheme (c) gives higher variance than the sampling scheme (d), passive learning. This may be caused by the fact that the bias is not zero, so the criterion for the optimal experiment design is no longer valid. When $m > 7$, the variances of all sampling schemes decrease as shown in Proposition 5. The sampling schemes (a), (b), and (c) suppress the variance more efficiently than the sampling scheme (d).

This simulation shows that the sampling schemes (a) and (b) outperform the sampling scheme (c) especially when the number of training examples is small, thanks to the two-stage sampling scheme. Also, the multi-point-search method with a small number of candidates is shown to work well.

7.2 Learning of the sensorimotor map of a two-joint robot arm

In this subsection, we apply the multi-point-search active learning method proposed in Section 5 to a real world problem. Let us consider learning of sensorimotor maps of a two-joint robot arm shown in Fig.10. A sensorimotor map of the k-th joint ($k = 1, 2$) is

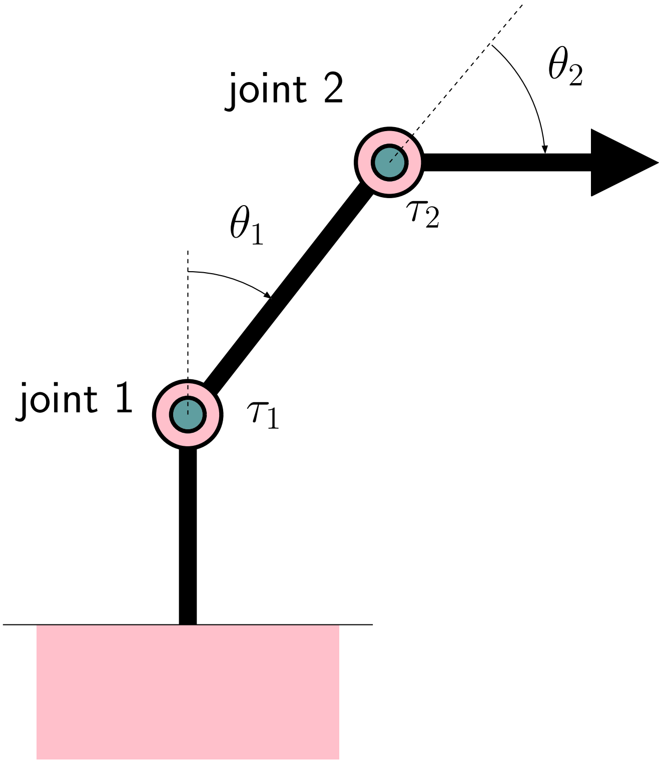


Figure 10: Two-joint robot arm.

a mapping from joint angle θ_k, angular velocity $\dot{\theta}_k$, and angular acceleration $\ddot{\theta}_k$ to torque τ_k which should be applied to the k-th joint:

$$\tau_k = f^{(k)}(\theta_1, \theta_2, \dot{\theta}_1, \dot{\theta}_2, \ddot{\theta}_1, \ddot{\theta}_2) \quad \text{for } k = 1, 2, \tag{77}$$

where $-\pi \leq \theta_1, \theta_2 \leq \pi$, $-a_1 \leq \dot{\theta}_1 \leq a_1$, $-a_2 \leq \dot{\theta}_2 \leq a_2$, $-b_1 \leq \ddot{\theta}_1 \leq b_1$, and $-b_2 \leq \ddot{\theta}_2 \leq b_2$.

Function spaces H_k to which $f^{(k)}$ belong are given as follows (Vijayakumar, 1998).

$$\begin{aligned} H_1 &= \mathcal{L}(\ddot{\theta}_1,\ \ddot{\theta}_2,\ \ddot{\theta}_1 \cos\theta_2,\ \ddot{\theta}_2 \cos\theta_2, \\ &\qquad \dot{\theta}_2^{\,2} \sin\theta_2,\ \dot{\theta}_1\dot{\theta}_2 \sin\theta_2,\ \sin\theta_1,\ \sin\theta_1\cos\theta_2,\ \sin\theta_2\cos\theta_1), \quad (78) \\ H_2 &= \mathcal{L}(\ddot{\theta}_1,\ \ddot{\theta}_2,\ \ddot{\theta}_1 \cos\theta_2,\ \dot{\theta}_1^{\,2} \sin\theta_2,\ \sin\theta_1,\ \sin\theta_1\cos\theta_2,\ \sin\theta_2\cos\theta_1), \quad (79) \end{aligned}$$

where $H = \mathcal{L}(\varphi_1, \varphi_2, \cdots, \varphi_k)$ means that H is spanned by $\varphi_1, \varphi_2, \cdots, \varphi_k$. The inner product in H_k is defined as

$$\langle f, g \rangle = \frac{1}{64\pi^2 a_1 a_2 b_1 b_2} \int_{-b_2}^{b_2} \int_{-b_1}^{b_1} \int_{-a_2}^{a_2} \int_{-a_1}^{a_1} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} f(x)\overline{g(x)}\, d\theta_1 d\theta_2 d\dot{\theta}_1 d\dot{\theta}_2 d\ddot{\theta}_1 d\ddot{\theta}_2. \tag{80}$$

We shall perform a learning simulation of the sensorimotor map $f^{(1)}$. From Eqs.(78)

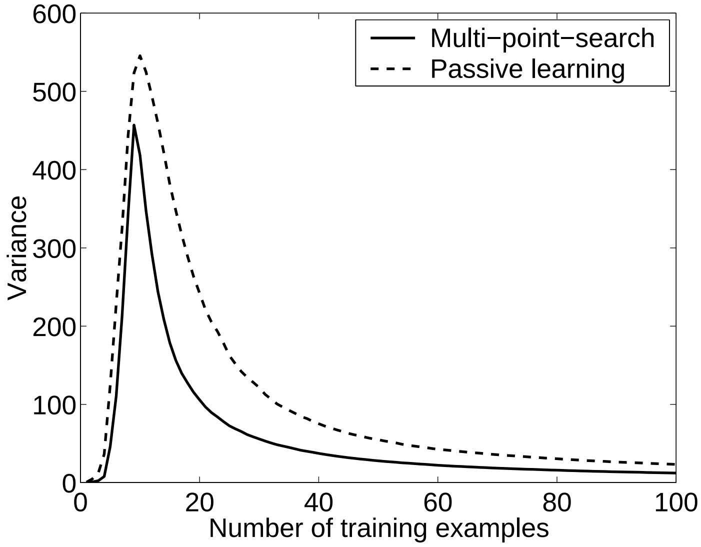


Figure 11: Results of learning of the sensorimotor map $f^{(1)}$. The horizontal axis denotes the number of training examples while the vertical axis denotes the noise variance measured by Eq.(17). The solid and dotted lines show the means of 100 trials by the multi-point-search method and passive learning, respectively. The vertical axis can be regarded as the generalization error when the number of training examples is larger than or equal to 10.

and (80), an orthonormal basis $\{\varphi_j^{(1)}\}_{j=1}^{9}$ in H_1 is given as follows.

$$\begin{array}{rclrclrcl}
\varphi_1^{(1)} &=& \frac{\sqrt{3}}{b_1}\ddot{\theta}_1, & \varphi_2^{(1)} &=& \frac{\sqrt{3}}{b_2}\ddot{\theta}_2, & \varphi_3^{(1)} &=& \frac{\sqrt{6}}{b_1}\ddot{\theta}_1\cos\theta_2, \\
\varphi_4^{(1)} &=& \frac{\sqrt{6}}{b_2}\ddot{\theta}_2\cos\theta_2, & \varphi_5^{(1)} &=& \frac{\sqrt{10}}{a_2^2}\dot{\theta}_2^{\,2}\sin\theta_2, & \varphi_6^{(1)} &=& \frac{\sqrt{18}}{a_1a_2}\dot{\theta}_1\dot{\theta}_2\sin\theta_2, \\
\varphi_7^{(1)} &=& \sqrt{2}\sin\theta_1, & \varphi_8^{(1)} &=& 2\sin\theta_1\cos\theta_2, & \varphi_9^{(1)} &=& 2\sin\theta_2\cos\theta_1.
\end{array}$$

Hence, the reproducing kernel of H_1 is given as

$$K_1(x,x') = \sum_{j=1}^{9} \varphi_j^{(1)}(x)\overline{\varphi_j^{(1)}(x')}. \tag{81}$$

Suppose that training examples are degraded by additive noise. Let us consider the following two sampling schemes.

(a) Multi-point-search: Training examples are sampled following the multi-point-search method shown in Fig.5. Let the number c of candidates be 3 and they are randomly generated in the domain.

(b) Passive learning: Training examples are randomly supplied from the domain.

Fig.11 shows simulation results. The horizontal axis denotes the number of training examples while the vertical axis denotes the variance measured by Eq.(17). The solid and dashed lines show the mean variances of 100 trials by the sampling schemes (a) and (b), respectively. In the sampling scheme (a), it holds that $\mathcal{N}(A_9) = \{0\}$ because the dimension of H is equal to 9. In the sampling scheme (b), $\mathcal{N}(A_{10}) = \{0\}$ was attained in all 100 trials in this simulation. Hence, it follows from Eq.(23) that the vertical axis in Fig.11 can be regarded as the generalization error when $m \geq 10$. This graph shows that the multi-point-search method with three candidates provides better generalization capability than passive learning. However, the performance of the multi-point-search method in this simulation is not so excellent as that in the previous one. The reason may be that the number c of candidates is small in contrast to the size of the domain.

8 Conclusion

In this paper, we gave a basic sampling strategy called the two-stage sampling scheme for reducing both the bias and variance, and based on this scheme, we proposed two active learning methods. One is the multi-point-search method applicable to arbitrary models, and the other is the optimal sampling method in the trigonometric polynomial space. The effectiveness of the proposed methods was demonstrated through computer simulations. As well as usual active learning methods devised so far, our methods assume that a model to which the learning target belongs is available. If the model is unknown, it should be estimated by model selection methods. To evaluate the robustness of the presented methods when they are combined with model selection is important future work.

Appendix

A Proof of Theorem 1

Let tr$(\cdot)$ stand for the trace of an operator. Then, for $f, g \in H$, it hold that

$$\mathrm{tr}\left(f \otimes \overline{g}\right) = \langle f, g \rangle. \tag{82}$$

It follows from Eqs.(20), (21), (19), (18), and (22) that

$$\begin{aligned}
E_n\|X_m n^{(m)}\|^2 &= E_n \mathrm{tr}\left(X_m \left(n^{(m)} \otimes \overline{n^{(m)}}\right) X_m^*\right) \\
&= \mathrm{tr}\left(X_m Q_m X_m^*\right) \qquad (83) \\
&= \mathrm{tr}\left(X_m U_m X_m^*\right) - \mathrm{tr}\left(X_m A_m A_m^* X_m^*\right) \\
&= \mathrm{tr}\left(V_m^\dagger A_m^* U_m^\dagger U_m U_m^\dagger A_m V_m^\dagger\right) - \mathrm{tr}\left(P_{\mathcal{R}(A_m^*)} P_{\mathcal{R}(A_m^*)}^*\right) \\
&= \mathrm{tr}\left(V_m^\dagger\right) - \mathrm{tr}\left(P_{\mathcal{R}(A_m^*)}\right). \qquad (84)
\end{aligned}$$

Eqs.(54) and (84) yield

$$J_v = \mathrm{tr}\left(V_{m+1}^{\dagger}\right) - \mathrm{tr}\left(P_{\mathcal{R}(A_{m+1}^*)}\right) - \mathrm{tr}\left(V_m^{\dagger}\right) + \mathrm{tr}\left(P_{\mathcal{R}(A_m^*)}\right). \tag{85}$$

First, we shall prove the case where $\psi_{m+1} \notin \mathcal{R}(A_m^*)$. It follows from Eq.(7) that

$$\mathrm{tr}\left(P_{\mathcal{R}(A_{m+1}^*)}\right) = \mathrm{tr}\left(P_{\mathcal{R}(A_m^*)}\right) + 1. \tag{86}$$

From Sugiyama and Ogawa (1999b), $V_{m+1}^{\dagger}$ is expressed by using $V_m^{\dagger}$ as

$$V_{m+1}^{\dagger} = V_m^{\dagger} + \frac{\alpha_{m+1} + \langle \tilde{\xi}_{m+1}, \xi_{m+1}\rangle}{\tilde{\psi}_{m+1}(x_{m+1})^2} \tilde{\psi}_{m+1} \otimes \overline{\tilde{\psi}_{m+1}} - \frac{\tilde{\xi}_{m+1} \otimes \overline{\tilde{\psi}_{m+1}} + \tilde{\psi}_{m+1} \otimes \overline{\tilde{\xi}_{m+1}}}{\tilde{\psi}_{m+1}(x_{m+1})}. \tag{87}$$

Since it holds from Ogawa (1987) that

$$\mathcal{R}(V_m^{\dagger}) = \mathcal{R}(A_m^*), \tag{88}$$

Eqs.(41) and (43) yield

$$\langle \tilde{\psi}_{m+1}, \tilde{\xi}_{m+1}\rangle = \langle P_{\mathcal{N}(A_m)}\psi_{m+1}, V_m^{\dagger}\xi_{m+1}\rangle = 0. \tag{89}$$

It follows from Eqs.(87), (89), (41), and (5) that

$$\begin{aligned} \mathrm{tr}\left(V_{m+1}^{\dagger}\right) &= \mathrm{tr}\left(V_m^{\dagger}\right) + \frac{\alpha_{m+1} + \langle \tilde{\xi}_{m+1}, \xi_{m+1}\rangle}{\tilde{\psi}_{m+1}(x_{m+1})^2} \|\tilde{\psi}_{m+1}\|^2 \\ &= \mathrm{tr}\left(V_m^{\dagger}\right) + \frac{\alpha_{m+1} + \langle \tilde{\xi}_{m+1}, \xi_{m+1}\rangle}{\tilde{\psi}_{m+1}(x_{m+1})}. \end{aligned} \tag{90}$$

Substituting Eqs.(86) and (90) into Eq.(85), we have Eq.(58).

We shall prove the case where $\psi_{m+1} \in \mathcal{R}(A_m^*)$. It follows from Eq.(7) that

$$\mathcal{R}(A_{m+1}^*) = \mathcal{R}(A_m^*). \tag{91}$$

From Sugiyama and Ogawa (1999b), it holds that

$$V_{m+1}^{\dagger} = V_m^{\dagger} - \frac{\tilde{\xi}_{m+1} \otimes \overline{\tilde{\xi}_{m+1}}}{\alpha_{m+1} + \langle \tilde{\xi}_{m+1}, \xi_{m+1}\rangle}. \tag{92}$$

Substituting Eqs.(91) and (92) into Eq.(85), we have Eq.(59). ■

B Proof of Theorem 2

When Q_{m+1} is given by Eq.(46) with $\sigma_j > 0$ for all j, it holds from Ogawa (1987) that the projection learning operator is expressed as

$$X_m = V_m'^{\dagger} A_m^* Q_m^{-1}. \tag{93}$$

Eqs.(83), (93), and (48) yield

$$\begin{aligned} E_n \|X_m n^{(m)}\|^2 &= \mathrm{tr}\,(X_m Q_m X_m^*) \\ &= \mathrm{tr}\left(V_m'^{\dagger} A_m^* Q_m^{-1} Q_m Q_m^{-1} A_m V_m'^{\dagger}\right) \\ &= \mathrm{tr}\left(V_m'^{\dagger}\right). \end{aligned} \tag{94}$$

It follows from Eqs.(54) and (94) that

$$J_v = \mathrm{tr}\left(V_{m+1}'^{\dagger}\right) - \mathrm{tr}\left(V_m'^{\dagger}\right). \tag{95}$$

First, we shall prove the case where $\psi_{m+1} \notin \mathcal{R}(A_m^*)$. It follows from Sugiyama and Ogawa (1999c) that

$$\begin{aligned} V_{m+1}'^{\dagger} &= V_m'^{\dagger} + \frac{\sigma_{m+1} + \langle V_m'^{\dagger}\psi_{m+1}, \psi_{m+1}\rangle}{\tilde{\psi}_{m+1}(x_{m+1})^2} \tilde{\psi}_{m+1} \otimes \overline{\tilde{\psi}_{m+1}} \\ &\quad - \frac{V_m'^{\dagger}\psi_{m+1} \otimes \overline{\tilde{\psi}_{m+1}} + \tilde{\psi}_{m+1} \otimes \overline{V_m'^{\dagger}\psi_{m+1}}}{\tilde{\psi}_{m+1}(x_{m+1})}. \end{aligned} \tag{96}$$

Since $\mathcal{R}(V_m'^{\dagger}) = \mathcal{R}(A_m^*)$, it holds from Eq.(41) that

$$\langle V_m'^{\dagger}\psi_{m+1}, \tilde{\psi}_{m+1}\rangle = \langle V_m'^{\dagger}\psi_{m+1}, P_{\mathcal{N}(A_m)}\psi_{m+1}\rangle = 0. \tag{97}$$

It follows from Eqs.(96), (97), (41), and (5) that

$$\begin{aligned} \mathrm{tr}\left(V_{m+1}'^{\dagger}\right) &= \mathrm{tr}\left(V_m'^{\dagger}\right) + \frac{\sigma_{m+1} + \langle V_m'^{\dagger}\psi_{m+1}, \psi_{m+1}\rangle}{\tilde{\psi}_{m+1}(x_{m+1})^2} \|\tilde{\psi}_{m+1}\|^2 \\ &= \mathrm{tr}\left(V_m'^{\dagger}\right) + \frac{\sigma_{m+1} + \langle V_m'^{\dagger}\psi_{m+1}, \psi_{m+1}\rangle}{\tilde{\psi}_{m+1}(x_{m+1})}. \end{aligned} \tag{98}$$

Substituting Eq.(98) into Eq.(95), we have Eq.(60).

We shall prove the case where $\psi_{m+1} \in \mathcal{R}(A_m^*)$. It follows from Sugiyama and Ogawa (1999c) that

$$V_{m+1}'^{\dagger} = V_m'^{\dagger} - \frac{V_m'^{\dagger}\psi_{m+1} \otimes \overline{V_m'^{\dagger}\psi_{m+1}}}{\sigma_{m+1} + \langle V_m'^{\dagger}\psi_{m+1}, \psi_{m+1}\rangle}. \tag{99}$$

Eqs.(95) and (99) yield Eq.(61). ■

C Proof of Theorem 3

If $m = 0$, then it follows from Eq.(63) that

$$J_v = \frac{\sigma^2}{\|\tilde{\psi}_{m+1}\|^2} = \frac{\sigma^2}{\|\psi_{m+1}\|^2} = \frac{\sigma^2}{\mu}. \tag{100}$$

Hence, J_v is minimized by any x_1.

Now we focus on the case where $m \geq 1$. Suppose that sample points $\{x_j\}_{j=1}^m$ are successively determined by Eq.(72). For a fixed integer s such that $0 \leq s \leq \lfloor \frac{m-1}{\mu} \rfloor$, let us consider the sample points x_j and $x_{j'}$ such that

$$s\mu + 1 \leq j, j' \leq \begin{cases} (s+1)\mu & \text{if } (s+1)\mu \leq m, \\ m & \text{if } s\mu + 1 \leq m < (s+1)\mu. \end{cases} \tag{101}$$

Let $t = (j \bmod \mu)$ and $t' = (j' \bmod \mu)$. Then, it follows from Eqs.(69) and (70) that

$$\langle \psi_{j'}, \psi_j \rangle = \psi_{j'}(x_j) = K(x_j, x_{j'}) = \mu\delta_{tt'}, \tag{102}$$

where $\delta_{tt'}$ denotes *Kronecker's delta.*

First, we shall prove the case where $1 \leq m \leq \mu - 1$. It follows from Eqs.(7) and (102) that

$$A_m^* A_m = \sum_{j=1}^m \left(\psi_j \otimes \overline{\psi_j} \right) = \mu \sum_{j=1}^m \left(\frac{\psi_j}{\sqrt{\mu}} \otimes \overline{\frac{\psi_j}{\sqrt{\mu}}} \right) = \mu P_{\mathcal{R}(A_m^*)}, \tag{103}$$

which yields

$$(A_m^* A_m)^\dagger \psi_{m+1} = \frac{1}{\mu} P_{\mathcal{R}(A_m^*)} \psi_{m+1} = \frac{1}{\mu}(\psi_{m+1} - \tilde{\psi}_{m+1}). \tag{104}$$

By using the fact that $\|\tilde{\psi}_{m+1}\|^2 \leq \|\psi_{m+1}\|^2 = \mu$, it follows from Eqs.(63), (104), and (102) that

$$\begin{aligned} J_v &= \sigma^2 \frac{1 + \frac{1}{\mu}\langle \psi_{m+1} - \tilde{\psi}_{m+1}, \psi_{m+1} \rangle}{\|\tilde{\psi}_{m+1}\|^2} \\ &= \sigma^2 \frac{1 + \frac{1}{\mu}(\|\psi_{m+1}\|^2 - \|\tilde{\psi}_{m+1}\|^2)}{\|\tilde{\psi}_{m+1}\|^2} \\ &\geq \sigma^2 \frac{1 + \frac{1}{\mu}(\|\psi_{m+1}\|^2 - \|\psi_{m+1}\|^2)}{\|\psi_{m+1}\|^2} \\ &= \frac{\sigma^2}{\mu}, \end{aligned} \tag{105}$$

where equality holds if and only if

$$\psi_{m+1} = \tilde{\psi}_{m+1} \tag{106}$$

because of Eq.(41). Since Eq.(102) implies Eq.(106), J_v is minimized.

We shall prove the case where $\mu \leq m \leq M-1$. Let $p = \lfloor \frac{m}{\mu} \rfloor$. It follows from Eqs.(7) and (102) that

$$\begin{aligned} A_m^* A_m &= \mu \left[\sum_{j=1}^{p\mu} \left(\frac{\psi_j}{\sqrt{\mu}} \otimes \overline{\frac{\psi_j}{\sqrt{\mu}}} \right) + \sum_{j=p\mu+1}^{m} \left(\frac{\psi_j}{\sqrt{\mu}} \otimes \overline{\frac{\psi_j}{\sqrt{\mu}}} \right) \right] \\ &= p\mu (I_H + \frac{1}{p} P), \end{aligned} \tag{107}$$

where I_H and P denote the identity operator on H and the orthogonal projection operator onto $\mathcal{L}\left(\{\psi_j\}_{j=p\mu+1}^m\right)$, respectively. Then, it holds that

$$(A_m^* A_m)^{-1} = \frac{1}{p\mu}(I_H - \frac{1}{p+1}P). \tag{108}$$

It follows from Eqs.(64), (108), and (102) that

$$\begin{aligned} J_v &= -\sigma^2 \frac{\|\frac{1}{p\mu}\psi_{m+1} - \frac{1}{p(p+1)\mu}P\psi_{m+1}\|^2}{1 + \langle \frac{1}{p\mu}\psi_{m+1} - \frac{1}{p(p+1)\mu}P\psi_{m+1}, \psi_{m+1}\rangle} \\ &= -\frac{\sigma^2 \left(1 - \frac{(2p+1)\|P\psi_{m+1}\|^2}{(p+1)^2\mu}\right)}{p(p+1)\mu - \frac{p}{p+1}\|P\psi_{m+1}\|^2}. \end{aligned} \tag{109}$$

Let a function $g(t)$ be

$$g(t) = -\frac{\sigma^2 \left(1 - \frac{(2p+1)t}{(p+1)^2\mu}\right)}{p(p+1)\mu - \frac{p}{p+1}t}, \tag{110}$$

where $\|P\psi_{m+1}\|^2$ in Eq.(109) is replaced with t. It follows from Eq.(102) that

$$0 \leq \|P\psi_{m+1}\|^2 \leq \|\psi_{m+1}\|^2 = \mu. \tag{111}$$

Hence, we focus on $[0, \mu]$ as the domain of $g(t)$. Since the derivative of g with respect to t is

$$\frac{dg}{dt} = \frac{\frac{2\sigma^2 p^2}{p+1}}{\left(p(p+1)\mu - \frac{p}{p+1}t\right)^2} > 0 \tag{112}$$

and $g(t)$ is continuous, it is minimized if and only if $t = 0$. This implies that J_v is minimized if and only if $\|P\psi_{m+1}\|^2 = 0$, which is attained by Eq.(102). ■

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, AC-19(6)*, 716–723.

Albert, A. (1972). *Regression and the Moore-Penrose pseudoinverse.* New York and London: Academic Press.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions on American Mathematical Society, 68*, 337–404.

Ben-Israel, A., & Greville, T. N. E. (1974). *Generalized inverses: Theory and applications.* New York: John Wiley & Sons.

Bergman, S. (1970). *The kernel function and conformal mapping.* Providence, Rhode Island: American Mathematical Society.

Cohn, D. A. (1994). Neural network exploration using optimal experiment design. In J. D. Cowan, G. Tesauro, & J. Alspector (Eds.), *Advances in Neural Information Processing Systems 6* (pp. 679–686). Morgan-Kaufmann.

Cohn, D. A. (1996). Neural network exploration using optimal experiment design. *Neural Networks, 9(6)*, 1071–1083.

Cohn, D. A. (1997). Minimizing statistical bias with queries. In M. C. Mozer, M. I. Jordan, & T. Petsche (Eds.), *Advances in Neural Information Processing Systems 9* (pp. 417–423). The MIT Press.

Cohn, D. A., Ghahramani, Z., & Jordan, M. I. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research, 4*, 129–145.

Efron, B., & Tibshirani, B. (1993). *An introduction to bootstrap.* New York: Chapman and Hall.

Fedorov, V. V. (1972). *Theory of optimal experiments.* New York: Academic Press.

Fukumizu, K. (1996). Active learning in multilayer perceptrons. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in Neural Information Processing Systems 8* (pp. 295–301). Cambridge: The MIT Press.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation, 4(1)*, 1–58.

Kaelbling, L. P. (Ed.) (1996). A special issue on reinforcement learning. *Machine Learning, 22(1/2/3).*

Kiefer, J. (1959). Optimal experimental designs. *Journal of the Royal Statistics Society, Series B, 21*, 272–304.

Kiefer, J., & Wolfowitz, J. (1960). The equivalence of two extremum problems. *Annals of Mathematical Statistics, 32*, 298.

MacKay, D. J. C. (1992). Information-based objective functions for active data selection. *Neural Computation, 4(4)*, 590–604.

Ogawa, H. (1987). Projection filter regularization of ill-conditioned problem. *Proceedings of SPIE, Inverse Problems in Optics, 808*, 189–196.

Ogawa, H. (1989). Inverse problem and neural networks. *Proceedings of IEICE 2nd Karuizawa Workshop on Circuits and Systems* (pp. 262–268). Karuizawa, Japan. (in Japanese)

Ogawa, H. (1992). Neural network learning, generalization and over-learning. *Proceedings of the ICIIPS'92, International Conference on Intelligent Information Processing & System, 2* (pp. 1–6). Beijing, China.

Saitoh, S. (1988). *Theory of reproducing kernels and its applications.* Pitsman Research Notes in Mathematics Series, 189. UK: Longman Scientific & Technical.

Saitoh, S. (1997). *Integral transform, reproducing kernels and their applications.* Pitsman Research Notes in Mathematics Series, 369. UK: Longman.

Schatten, R. (1970). *Norm ideals of completely continuous operators.* Berlin: Springer-Verlag.

Sollich, P. (1994). Query construction, entropy and generalization in neural network models. *Physical Review E, 49*, 4637–4651.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B, 36*, 111–133.

Sugiyama, M., & Ogawa, H. (1999a). Exact incremental projection learning in the presence of noise. *Proceedings of the 11th Scandinavian Conference on Image Analysis* (pp. 747–754). Kangerlussuaq, Greenland.

Sugiyama, M., & Ogawa, H. (1999b). Incremental projection learning for optimal generalization. *Technical Report TR99-0007*, Department of Computer Science, Tokyo Institute of Technology, Japan. (available at http://www.cs.titech.ac.jp/TR/tr99.html)

Sugiyama, M., & Ogawa, H. (1999c). Properties of incremental projection learning. *Technical Report TR99-0008*, Department of Computer Science, Tokyo Institute of Technology, Japan. (available at http://www.cs.titech.ac.jp/TR/tr99.html)

Sugiyama, M., & Ogawa, H. (1999d). Functional analytic approach to model selection — subspace information criterion. *Proceedings of 1999 Workshop on Information-Based Induction Sciences (IBIS'99)* (pp. 93–98). Izu, Japan. (Its complete version is available at http://www.cs.titech.ac.jp/TR/tr99.html, Subspace information criterion for model selection. *Technical Report TR99-0009*, Department of Computer Science, Tokyo Institute of Technology, Japan.)

Sugiyama, M., & Ogawa, H. (2000). Training data selection for optimal generalization in trigonometric polynomial networks. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in Neural Information Processing Systems 12* (pp. 624–630). The MIT Press.

Takemura, A. (1991). *Modern mathematical statistics.* Tokyo: Sobunsya. (in Japanese)

Tikhonov, A. N., & Arsenin, V. Y. (1977). *Solutions of Ill-Posed Problems.* Washington DC: V. H. Winston.

Vijayakumar, S. (1998). Computational theory of incremental and active learning for optimal generalization. *Ph.D Thesis*, Department of Computer Science, Tokyo Institute of Technology, Japan.

Vijayakumar, S., & Ogawa, H. (1999). Improving generalization ability through active learning. *IEICE Transactions on Information and Systems, E82-D(2)*, 480–487.

Vijayakumar, S., Sugiyama, M., & Ogawa, H. (1998). Training data selection for optimal generalization with noise variance reduction in neural networks. In M. Marinaro & R. Tagliaferri (Eds.), *Neural Nets WIRN Vietri-98* (pp. 153–166). Springer-Verlag.

Wahba, H. (1990). *Spline model for observational data.* Philadelphia and Pennsylvania: Society for Industrial and Applied Mathematics.

Yamashita, Y., & Ogawa, H. (1992). Optimum image restoration and topological invariance. *The Transactions of the IEICE D-II, J75-D-II(2)*, 306–313. (in Japanese)

Yue, R. X., & Hickernell, F. J. (1999). Robust designs for fitting linear models with misspecification. *Statistica Sinica, 9*, 1053–1069.