IEICE Transactions on Information and Systems,
vol.E86-D, no.12, pp.2753–2763, Dec. 2003.

Active Learning with Model Selection — Simultaneous Optimization of Sample Points and Models for Trigonometric Polynomial Models

Masashi Sugiyama
Hidemitsu Ogawa

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

`sugi@cs.titech.ac.jp`
`http://sugiyama-www.cs.titech.ac.jp/~sugi/`

Abstract

In supervised learning, the selection of sample points and models is crucial for acquiring a higher level of the generalization capability. So far, the problems of active learning and model selection have been independently studied. If sample points and models are simultaneously optimized, then a higher level of the generalization capability is expected. We call this problem active learning with model selection. However, active learning with model selection can not be generally solved by simply combining existing active learning and model selection techniques because of the active learning / model selection dilemma: the model should be fixed for selecting sample points and conversely the sample points should be fixed for selecting models. In this paper, we show that the dilemma can be dissolved if there is a set of sample points that is optimal for all models in consideration. Based on this idea, we give a practical procedure for active learning with model selection in trigonometric polynomial models. The effectiveness of the proposed procedure is demonstrated through computer simulations.

Keywords

supervised learning, generalization capability, active learning, model selection, trigonometric polynomial space.

1 Introduction: supervised learning and the active learning / model selection dilemma

Supervised learning is obtaining an underlying rule from training examples made up of sample points and corresponding sample values. If the rule is successfully acquired, then appropriate output values corresponding to unknown input points can be estimated. This ability is called the *generalization capability*.

In supervised learning, there are two factors we can control for optimal generalization: *sample points* and *models*. A sample point corresponds to a query to the oracle, and a model refers to, for example, the type and number of basis functions used for learning. The problem of designing sample points is called *active learning* (or experimental design), and the problem of determining the model is called *model selection*.

So far, extensive and profound studies have been conducted to solve the problems of active learning [8, 11, 6, 7, 10, 9, 20, 21] and model selection [12, 1, 18, 19, 15, 5, 22]. However, it seems that these two methods have been studied independently. One would naturally think that if sample points and models are simultaneously optimized, then a higher level of the generalization capability is expected. This is the problem that we would like to tackle in this paper. We refer to the problem as *active learning with model selection*[1].

In general, the model should be fixed for active learning[2]. On the other hand, the training examples gathered at fixed sample points are generally required for model selection. These facts imply that the problem of active learning with model selection can not be generally solved by simply combining existing active learning and model selection techniques. We call this the *active learning / model selection dilemma*.

One of the possible approaches to dissolving this dilemma is to perform active learning and model selection alternately in an incremental manner. That is, first decide an initial model and design a set of sample points for the initial model. Then, select a new model using the samples, and design a set of additional sample points for the new model to further improve the generalization capability, likewise. This incremental approach seems reasonable in practice. However, only the greedy optimality can be achieved, i.e., the additional set of sample points is optimal for the current model.

In this paper, we therefore propose a basic strategy for simultaneously optimizing sample points and models for the optimal generalization capability in a batch manner. Specifically, our proposal is to select sample points that are optimal for all models in consideration. Although this strategy seems rather idealistic, we show that this strategy can be practically realized for trigonometric polynomial models. The usefulness of the proposed method will be experimentally shown for both realizable and unrealizable

[1] There is also an interesting way of relating active learning to model selection [4, 11], which is aimed at selecting sample points so that they maximally discriminate between models. Although this method is extremely useful when we would like to specify the true model, this is different from what we are heading for in the current paper.

[2] Some active learning methods are incremental so it is possible to change the model through the incremental learning process. Even so, those methods essentially work for a fixed model.

learning target functions.

The rest of this paper is organized as follows. Section 2 mathematically formulates the problem of active learning with model selection. Section 3 gives a basic strategy for active learning with model selection. In Sections 4 and 5, a practical procedure for active learning with model selection under a certain setting is given. The effectiveness of the proposed procedure is experimentally investigated through computer simulations in Section 6. Finally, Section 7 gives concluding remarks and future prospects.

2 Problem formulation

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

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

where y_m is degraded by additive noise ϵ_m. The purpose of supervised learning is to find a learning result function $\hat{f}(\boldsymbol{x})$ that minimizes a certain generalization error J_G.

Let $\mathcal{X}$ be a set of M sample points $\{\boldsymbol{x}_m\}_{m=1}^{M}$ and let S be a *model*, which refers to, for example, the type and number of basis functions used for learning. Let $\mathcal{M}$ be a set of models from which the model is selected. In this paper, we discuss the problem of simultaneously optimizing the sample points $\mathcal{X}$ and model S, called the *active learning with model selection.*

Definition 1 (Active learning with model selection) *Determine sample points $\mathcal{X}$ and select a model from a set $\mathcal{M}$ so that the generalization error J_G is minimized:*

$$\min_{\mathcal{X},\ S\in\mathcal{M}} J_G[\mathcal{X}, S]. \tag{2}$$

3 Basic strategy

As we pointed out in Section 1, the problem of active learning with model selection can not be generally solved, in a batch manner, by simply combining existing active learning and model selection techniques because of the *active learning / model selection dilemma*: the model should be fixed for active learning and conversely sample points should be fixed for model selection.

Some readers may think the problem of active learning with model selection can be naively solved by

$$\min_{\mathcal{X}_S\in\mathrm{argmin}_{\mathcal{X}} J_G[\mathcal{X},S],\ S\in\mathcal{M}} J_G[\mathcal{X}_S, S], \tag{3}$$

i.e., first deciding the sample points $\mathcal{X}_S$ for every S in $\mathcal{M}$, and then selecting the model S using $\mathcal{X}_S$. However, this naive approach does not work in practice because when selecting

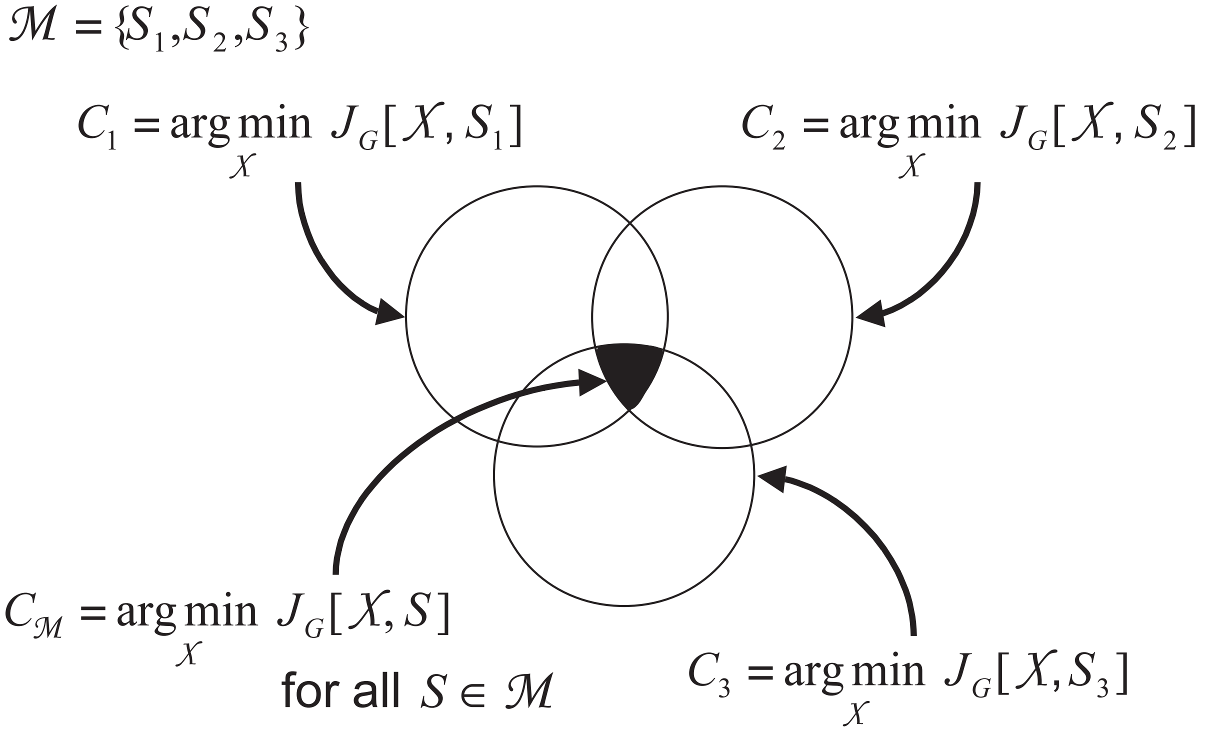

$\mathcal{M} = \{S_1, S_2, S_3\}$
$C_1 = \arg\min_{\mathcal{X}} J_G[\mathcal{X}, S_1]$
$C_2 = \arg\min_{\mathcal{X}} J_G[\mathcal{X}, S_2]$
$C_{\mathcal{M}} = \arg\min_{\mathcal{X}} J_G[\mathcal{X}, S]$
for all $S \in \mathcal{M}$
$C_3 = \arg\min_{\mathcal{X}} J_G[\mathcal{X}, S_3]$

Figure 1: Basic strategy for active learning with model selection. Let the set $\mathcal{M}$ of models be $\{S_1, S_2, S_3\}$. The top-left circle denotes a set C_1 of optimal $\mathcal{X}$ for the model S_1, i.e., an element in C_1 is a set $\mathcal{X}$ of sample points $\{\boldsymbol{x}_m\}_{m=1}^M$ that minimizes $J_G[\mathcal{X}, S_1]$. Similarly, the top-right and bottom circles denote sets of optimal $\mathcal{X}$ for S_2 and S_3, respectively. If there exists $\mathcal{X}$ that is commonly optimal for all models in $\mathcal{M}$, i.e., $C_{\mathcal{M}}$ is not empty, then the problem of active learning with model selection can be straightforwardly solved by using the commonly optimal sample points.

a model with existing model selection methods, we need sample values $\{y_m^{(S)}\}_{m=1}^M$ at the sample points $\{\boldsymbol{x}_m^{(S)}\}_{m=1}^M$. Namely, sample values $\{y_m^{(S)}\}_{m=1}^M$ for all S in $\mathcal{M}$ (totally $M \times |\mathcal{M}|$ sample values) should be available, which is just waste of sample values and sampling cost.

In order to solve the problem of active learning with model selection, we propose selecting sample points for a set of models in consideration, not for a fixed model. Specifically, if there is a set $\mathcal{X}$ of sample points that is optimal for all models in the set $\mathcal{M}$, the problem of active learning with model selection can be straightforwardly solved as follows. First, $\mathcal{X}$ is determined so that it is optimal for all models in the set $\mathcal{M}$, and sample values $\{y_m\}_{m=1}^M$ are gathered at the optimal points $\{\boldsymbol{x}_m\}_{m=1}^M$. Then model selection is carried out with the optimal training examples $\{(\boldsymbol{x}_m, y_m)\}_{m=1}^M$. Consequently, we obtain the optimal model with optimal sample points because the sample points are optimal for any selected model. This basic strategy is summarized in Figure 1.

Although the above strategy seems idealistic, we show in the following sections that the strategy can be practically realized under certain conditions. Possible generalization of this strategy will be finally discussed in Section 7.

4 Setting

In this section, the setting is described.

4.1 Trigonometric polynomial space

We assume that the learning target function $f(\boldsymbol{x})$ belongs to S_N, which is a *trigonometric polynomial space* of order $N = (N_1, N_2, \ldots, N_L)$. The trigonometric polynomial space is defined as follows.

Definition 2 (Trigonometric polynomial space) *Let us denote the L-dimensional input vector $\boldsymbol{x}$ by*

$$\boldsymbol{x} = (\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(L)})^\top. \tag{4}$$

For $l = 1, 2, \ldots, L$, let n_l be a non-negative integer and $\mathcal{D}_l = [-\pi, \pi]$. Let

$$\mathcal{D} = \mathcal{D}_1 \times \mathcal{D}_2 \times \cdots \times \mathcal{D}_L. \tag{5}$$

Then, a function space S_n is called a trigonometric polynomial space of order $n = (n_1, n_2, \ldots, n_L)$ if S_n is spanned by the functions

$$\left\{ \prod_{l=1}^{L} \exp(ip_l \xi^{(l)}) \;\middle|\; p_l = -n_l, -n_l + 1, \ldots, n_l \right. \\ \left. \text{for } l = 1, 2, \ldots, L \right\} \tag{6}$$

defined on $\mathcal{D}$, and the inner product, denoted by $\langle \cdot, \cdot \rangle$, is defined by

$$\begin{aligned} \langle f, g \rangle \;\; = \;\; & \frac{1}{(2\pi)^L} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} \\ & f(\boldsymbol{x}) \overline{g(\boldsymbol{x})} d\xi^{(1)} d\xi^{(2)} \cdots d\xi^{(L)}, \end{aligned} \tag{7}$$

where $\overline{\cdot}$ denotes the complex conjugate of a complex number.

The dimension of S_n is

$$\dim S_n = \prod_{l=1}^{L} (2n_l + 1), \tag{8}$$

and the reproducing kernel[3] of S_n, denoted by $K_n(\boldsymbol{x},\boldsymbol{x}')$, is expressed as

$$K_n(\boldsymbol{x},\boldsymbol{x}') = \prod_{l=1}^{L} K_n^{(l)}(\xi^{(l)},\xi^{(l)\prime}), \tag{9}$$

where

$$K_n^{(l)}(\xi^{(l)},\xi^{(l)\prime}) = \begin{cases} \dfrac{\sin\dfrac{(2n_l+1)(\xi^{(l)}-\xi^{(l)\prime})}{2}}{\sin\dfrac{\xi^{(l)}-\xi^{(l)\prime}}{2}} \\ \qquad\qquad\qquad \text{if } \xi^{(l)} \neq \xi^{(l)\prime}, \\ 2n_l+1 \qquad\quad \text{if } \xi^{(l)} = \xi^{(l)\prime}. \end{cases} \tag{10}$$

When the dimension L of the input x is 1, a trigonometric polynomial space of order n is spanned by

$$\Big\{\exp(ipx) \;\Big|\; p=-n,-n+1,\ldots,n\Big\} \tag{11}$$

defined on $[-\pi,\pi]$, and the inner product is defined by

$$\langle f,g\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\overline{g(x)}dx. \tag{12}$$

The dimension of a trigonometric polynomial space of order n is

$$\dim S_n = 2n+1, \tag{13}$$

and the reproducing kernel of this space is expressed as

$$K_n(x,x') = \begin{cases} \dfrac{\sin\dfrac{(2n+1)(x-x')}{2}}{\sin\dfrac{x-x'}{2}} & \text{if } x \neq x', \\ 2n+1 & \text{if } x = x'. \end{cases} \tag{14}$$

[3] The reproducing kernel $K(\boldsymbol{x},\boldsymbol{x}')$ is a bivariate function defined on $\mathcal{D}\times\mathcal{D}$ that satisfies the following conditions [3]:

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

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

Note that the reproducing kernel is unique if it exists.

4.2 Least squares learning

We adopt the usual *least squares (LS) learning* as the learning criterion. LS learning is aimed at finding a learning result function $\hat{f}(\boldsymbol{x})$ in a subspace S of S_N that minimizes the *training error* J_{TE}:

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

In the LS learning case, a subspace S is the model. Since S_N has the reproducing kernel (see Section 4.1), a subspace S also has the reproducing kernel. Let $K(\boldsymbol{x}, \boldsymbol{x}')$ be the reproducing kernel of S and A be a linear operator defined by

$$A = \sum_{m=1}^{M}\left(\boldsymbol{e}_m \otimes \overline{K(\cdot, \boldsymbol{x}_m)}\right), \tag{16}$$

where $(\cdot \otimes \bar{\cdot})$ denotes the *Neumann-Schatten product*[4], and $\boldsymbol{e}_m$ is the m-th vector of the so-called standard basis in $\mathbb{C}^M$. Note that the operator A is called the sampling operator since it holds for any function f in S that $Af = (f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots, f(\boldsymbol{x}_M))^\top$, where $\top$ denotes the transpose of a vector. Let $A^\dagger$ be the *Moore-Penrose generalized inverse* [2] of A and $\boldsymbol{y}$ be an M-dimensional vector whose m-th element is the sample value y_m:

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_M)^\top. \tag{17}$$

Then, the LS learning result function $\hat{f}(\boldsymbol{x})$ is given by

$$\hat{f} = A^\dagger \boldsymbol{y}. \tag{18}$$

4.3 Generalization measure

We measure the generalization error, denoted by J_G, by the expected squared norm in S_N:

$$\begin{aligned} J_G &= \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f} - f\|^2 \\ &= \mathrm{E}_{\boldsymbol{\epsilon}}\frac{1}{(2\pi)^L}\int_{\mathcal{D}}|\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})|^2 d\boldsymbol{x}, \end{aligned} \tag{19}$$

where $\|\cdot\|$ denotes the norm and $\mathrm{E}_{\boldsymbol{\epsilon}}$ denotes the expectation over the noise.

In many statistical active learning learning methods [6, 7, 20], the generalization measure is defined by

$$\mathrm{E}_{\boldsymbol{\epsilon}}\int_{\mathcal{D}}\left|\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})\right|^2 p(\boldsymbol{x})d\boldsymbol{x}, \tag{20}$$

[4] For any fixed g in a Hilbert space S and any fixed f in a Hilbert space S', the *Neumann-Schatten product* $(f \otimes \overline{g})$ is an operator from S to S' defined by using any h in S as (see [17])

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

where $p(\boldsymbol{x})$ is the probability density function of test input points $\boldsymbol{x}$. In our setting, we assume that $p(\boldsymbol{x})$ in Eq.(20) is the uniform distribution on the domain $\mathcal{D}$.

4.4 The number of training examples

We assume that the number M of training examples is

$$M = \prod_{l=1}^{L} M_l, \tag{21}$$

where M_l is a positive integer such that

$$M_l \geq 2N_l + 1 \ \text{ for } l = 1, 2, \ldots, L. \tag{22}$$

We should admit that this assumption is rather restrictive when L is large. Assuming Eq.(21) virtually means that we focus on a small L, say at the largest 3 or 4.

4.5 Noise characteristics

We assume that the noise is independently drawn from a distribution with mean zero and variance σ^2. σ^2 does not have to be known.

4.6 Model candidates

In the LS learning case, a subset of basis functions is the model (see Section 4.2). Let $\mathcal{M}$, the set of models from which the model is selected, be a set of all trigonometric polynomial spaces included in S_N:

$$\begin{aligned}\mathcal{M} = \{S_n \ \mid \ & n = (n_1, n_2, \ldots, n_L), \ n_l = 0, 1, \ldots, N_l \\ & \text{for } l = 1, 2, \ldots, L\}.\end{aligned} \tag{23}$$

5 Active learning with model selection for trigonometric polynomial models

In this section, we give a practical procedure for active learning with model selection under the setting described in Section 4.

Let $\hat{f}_n(\boldsymbol{x})$ be a learning result function obtained with the model S_n. $\hat{f}_n$ is given by

$$\hat{f}_n = A_n^{\dagger}\boldsymbol{y}, \tag{24}$$

where A_n is defined with the reproducing kernel $K_n(\boldsymbol{x}, \boldsymbol{x}')$ of S_n by

$$A_n = \sum_{m=1}^{M} \left(\boldsymbol{e}_m \otimes \overline{K_n(\cdot, \boldsymbol{x}_m)}\right). \tag{25}$$

It is known that the generalization error of $\hat{f}_n$ defined by Eq.(19) is decomposed into the *bias* and *variance* terms:

$$\begin{aligned} J_G &= \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f}_n - f\|^2 \\ &= \|\mathrm{E}_{\boldsymbol{\epsilon}}\hat{f}_n - f\|^2 + \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f}_n - \mathrm{E}_{\boldsymbol{\epsilon}}\hat{f}_n\|^2. \end{aligned} \tag{26}$$

Note that the bias term can not be zero unless the learning target function f belongs to S_n. First, we review a necessary and sufficient condition for a set $\mathcal{X}$ of sample points $\{\boldsymbol{x}_m\}_{m=1}^{M}$ so that the generalization error J_G is minimized for a *fixed* model S_n.

Proposition 1 *[21] For a model S_n to which the learning target function f belongs, the generalization error of $\hat{f}_n$ is minimized with respect to sample points $\mathcal{X}$ under the constraint of bias-free if and only if sample points $\mathcal{X}$ satisfy*

$$\frac{1}{M}A_n^* A_n = I_{S_n}, \tag{27}$$

where A_n^* is the adjoint operator of A_n and I_{S_n} denotes the identity operator on S_n.

There are infinitely many sets of sample points such that Condition (27) holds for a fixed model S_n [21]. Here, we give a design method of sample points that satisfy Condition (27) for all models in the set $\mathcal{M}$.

Theorem 1 *Let c_l be an arbitrary constant such that*

$$-\pi \le c_l \le -\pi + \frac{2\pi}{M_l} \quad \text{for } l = 1, 2, \ldots, L. \tag{28}$$

If a set

$$\Bigg\{\boldsymbol{x}_m \;\Bigg|\; m = \sum_{l=2}^{L}\left((m_l - 1)\prod_{l'=1}^{l-1} M_{l'}\right) + m_1, \\ m_l = 1, 2, \ldots, M_l \;\text{ for } l = 1, 2, \ldots, L\Bigg\} \tag{29}$$

of M sample points is let

$$\boldsymbol{x}_m = (\xi_m^{(1)}, \xi_m^{(2)}, \ldots, \xi_m^{(L)})^\top, \tag{30}$$

where

$$\xi_m^{(l)} = c_l + \frac{2\pi}{M_l}(m_l - 1) \quad \text{for } l = 1, 2, \ldots, L, \tag{31}$$

then

$$\frac{1}{M}A_n^* A_n = I_{S_n} \quad \text{for all } S_n \in \mathcal{M}. \tag{32}$$

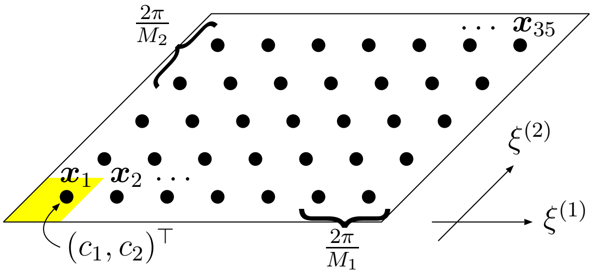


Figure 2: Example of sample points such that Condition (32) holds. The number M of training examples is $M = M_1 \times M_2 = 7 \times 5 = 35$.

A proof of Theorem 1 is provided in Appendix[5].

Eq.(31) means that M sample points are fixed to regular intervals in the domain $\mathcal{D}$. An example of sample points designed by Eq.(31) is illustrated in Figure 2.

Theorem 1 and Proposition 1 assert that the sample points designed by Eq.(31) are optimal for all models to which the learning target function f belongs. For a model S_n to which the learning target function f does not belong, the sample points designed by Eq.(31) has the property that they minimize the variance under the constraint that the range of A_n^* agrees with S_n [21]. Computer simulations in Section 6 experimentally show that the sample points designed by Eq.(31) do not only give the optimal generalization capability for models to which the learning target function f *belongs*, but also give a higher level of the generalization capability for models to which the learning target function f does *not* belong. Therefore, the sample points designed by Eq.(31) can be practically regarded as a good design for all models in the set $\mathcal{M}$.

With training examples $\{(\boldsymbol{x}_m, y_m)\}_{m=1}^M$ gathered at the optimal sample points $\{\boldsymbol{x}_m\}_{m=1}^M$ designed by Eq.(31), model selection is carried out. Then we may obtain a learning result function that has a higher level of the generalization capability.

Another advantage of using Eq.(31) is that LS learning with sample points designed by Eq.(31) is computationally very efficient since $A_n^\dagger$ is given by $\frac{1}{M}A_n^*$ [21].

When the dimension L of the input x is 1, Theorem 1 is reduced to a simpler form.

Corollary 1 *Let $M \geq 2N+1$ and c be an arbitrary constant such that*

$$-\pi \leq c \leq -\pi + \frac{2\pi}{M}. \tag{33}$$

If a set $\{x_m\}_{m=1}^M$ of sample points is let

$$x_m = c + \frac{2\pi}{M}(m-1), \tag{34}$$

then Eq.(32) holds.

[5] Note that we can also prove this theorem from the fact that the equidistance design is D-optimal for all trigonometric polynomial models and D- and A-optimalities are equivalent for trigonometric polynomial models [8].

6 Computer simulations

In this section, the effectiveness of active learning with model selection is demonstrated through computer simulations.

6.1 Realizable case: illustrative example

Let the dimension L of the input x is 1 and let the order N of the largest trigonometric polynomial space be 100. Let the learning target function $f(x)$ be

$$f(x) = \frac{1}{10}\sum_{n=1}^{50}(\sin nx + \cos nx). \tag{35}$$

Note that f belongs to S_n for $n \geq 50$. The noise ϵ_m is drawn from the normal distribution with mean 0 and variance σ^2. Let the set $\mathcal{M}$ of model candidates be

$$\mathcal{M} = \{S_0, S_1, S_2, \ldots, S_{100}\}. \tag{36}$$

We measure the error of a learning result function $\hat{f}(x)$ by

$$\text{Error} = \|\hat{f} - f\|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\hat{f}(x) - f(x)|^2 dx. \tag{37}$$

We compare the performance of the following two sampling schemes.

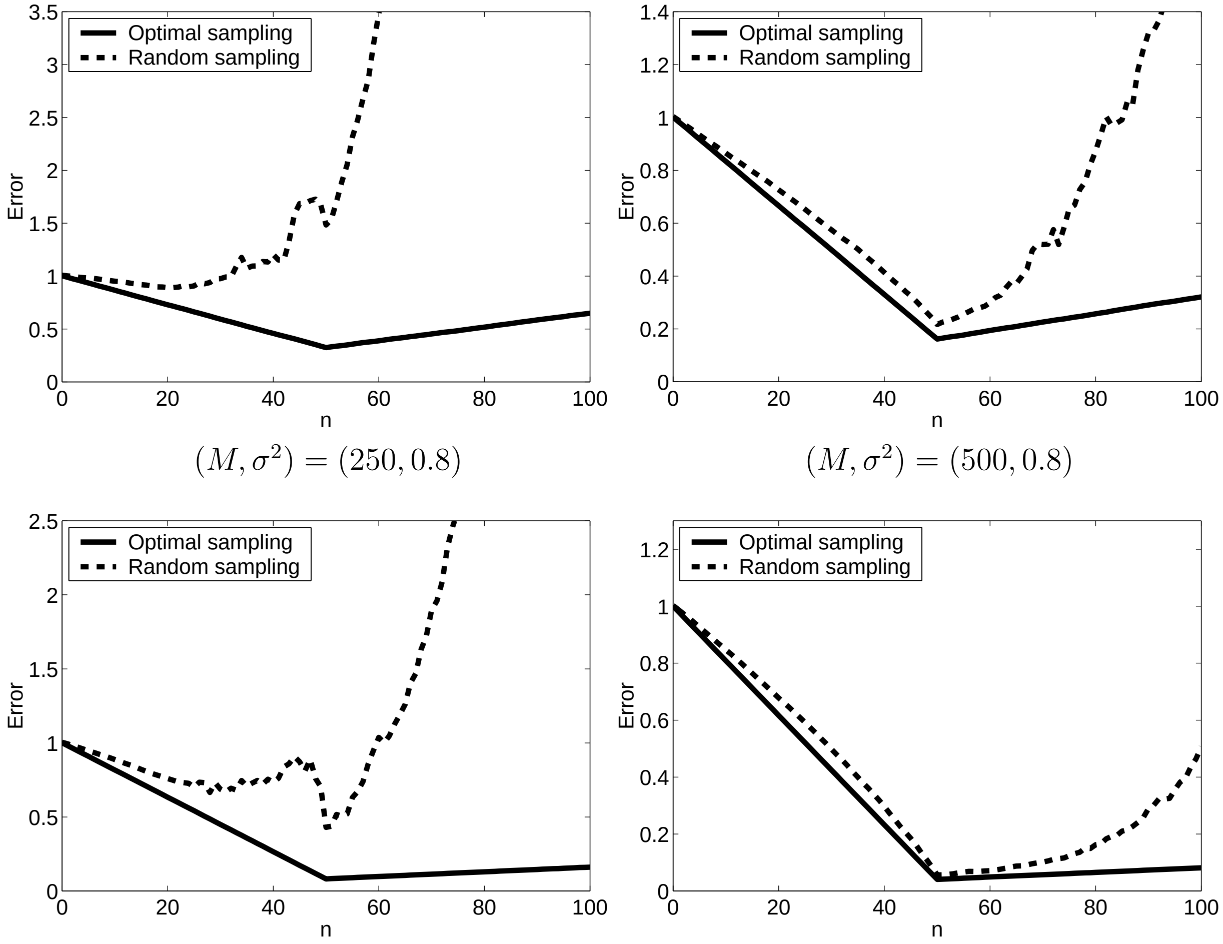
(i) Optimal sampling: Sample points $\{x_m\}_{m=1}^{M}$ are designed by Eq.(34).

(ii) Random sampling: Sample points are randomly created in the domain $[-\pi, \pi]$.

Figure 3 shows the simulation results for

$$\begin{aligned}(M, \sigma^2) &= (250, 0.8), (500, 0.8), \\ &\quad (250, 0.2), (500, 0.2).\end{aligned} \tag{38}$$

The horizontal axis denotes the order n of the model while the vertical axis denotes the error measured by Eq.(37). The solid and dashed lines show the mean errors over 100 trials by (i) Optimal sampling and (ii) Random sampling, respectively. These graphs show that the proposed sampling method provides better generalization capability than random sampling irrespective of the number M of training examples, noise variance σ^2, and order n of the model. Especially, when M is small and σ^2 is large, its effectiveness is remarkable.



$(M, \sigma^2) = (250, 0.8)$ $(M, \sigma^2) = (500, 0.8)$

$(M, \sigma^2) = (250, 0.2)$ $(M, \sigma^2) = (500, 0.2)$

Figure 3: Results of illustrative simulation.

6.2 Unrealizable case

In the previous experiment, the learning target function f belongs to S_N. Here, we perform a simulation for a practical unrealizable case that f does not belong to S_N.

Let us consider the chaotic series created by the *Mackey-Glass delay-difference equation* (e.g.[16]):

$$g(t+1) = \begin{cases} (1-b)g(t) + \dfrac{a\, g(t-\tau)}{1+g(t-\tau)^{10}} & \\ \qquad\qquad\qquad\qquad \text{for } t \geq \tau + 1, & \\ 0.3 \qquad\qquad\quad\;\; \text{for } 0 \leq t \leq \tau, & \end{cases} \tag{39}$$

where $a = 0.2$, $b = 0.1$, and $\tau = 17$. Let $\{h_t\}_{t=1}^{600}$ be

$$h_t = g(t + \tau + 1). \tag{40}$$

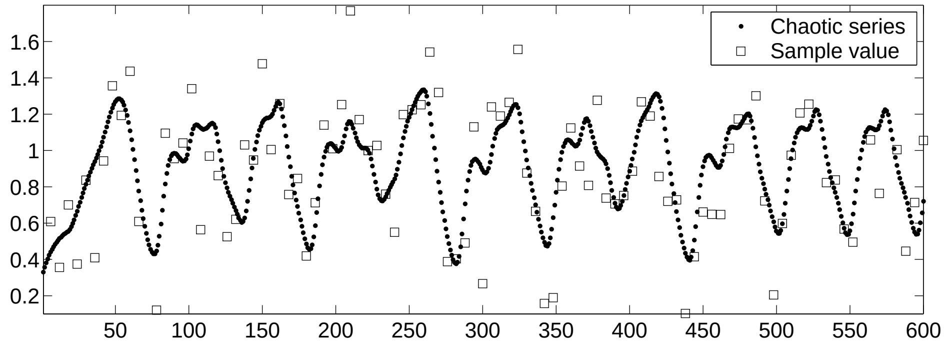


Figure 4: Mackey-Glass chaotic series of 600 points and 100 sample values ($p(m) = 6m$ and $\sigma^2 = 0.07$).

We are given M degraded sample values $\{y_m\}_{m=1}^{M}$:

$$y_m = h_{r(m)} + \epsilon_m, \tag{41}$$

where $r(m)$ is an integer such that $1 \leq r(m) \leq 600$ which indicates the sampling location, and the noise ϵ_m is independently drawn from the normal distribution with mean 0 and variance σ^2.

The task is to obtain the best estimates $\{\hat{h}_t\}_{t=1}^{600}$ of $\{h_t\}_{t=1}^{600}$ that minimize the error:

$$\text{Error} = \frac{1}{600}\sum_{t=1}^{600}\left|\hat{h}_t - h_t\right|^2. \tag{42}$$

In this simulation, we consider the following four cases:

$$\begin{aligned}(M, \sigma^2) \quad = \quad & (100, 0.07), (300, 0.07), \\ & (100, 0.04), (300, 0.04).\end{aligned} \tag{43}$$

Figure 4 displays the original chaotic series $\{h_t\}_{t=1}^{600}$ (shown by '•') and an example of 100 sample values $\{y_m\}_{m=1}^{100}$ (shown by '□') with the noise variance $\sigma^2 = 0.07$.

We shall obtain the estimates $\{\hat{h}_t\}_{t=1}^{600}$ as follows. Let us consider sample points $\{x_m\}_{m=1}^{M}$ corresponding to the sample values $\{y_m\}_{m=1}^{M}$:

$$x_m = -\pi + \frac{2\pi}{600}\left(r(m) - 1\right). \tag{44}$$

By using the training examples $\{(x_m, y_m)\}_{m=1}^{M}$, LS learning is carried out. Then the estimates $\{\hat{h}_t\}_{t=1}^{600}$ are given by

$$\hat{h}_t = \hat{f}\left(-\pi + \frac{2\pi}{600}(t-1)\right). \tag{45}$$

We adopt S_{40} as the largest model, i.e., $N = 40$. Note that the 600 chaotic series can not be expressed by the functions in S_{40}. This means that we consider the learning target function which is not included in S_{40}. Let the set $\mathcal{M}$ of model candidates be

$$\mathcal{M} = \{S_0, S_1, S_2, \ldots, S_{40}\}. \tag{46}$$

Similar to the previous experiment, we compare the performance of the following two sampling schemes.

(i) Optimal sampling: Sample points are fixed to regular intervals, i.e.,

$$r(m) = \frac{600m}{M}. \tag{47}$$

In this case, Eqs.(44) and (47) yield Eq.(34) with $c = -\pi + \frac{2\pi}{M} - \frac{2\pi}{600}$.

(ii) Random sampling: Sample points are randomly created in the domain, i.e., $r(m)$ randomly gives an integer such that $1 \leq r(m) \leq 600$.

Figure 5 depicts the results of the active learning simulation. The horizontal axis denotes the order n of the model while the vertical axis denotes the error measured by Eq.(42). The solid and dashed lines show the mean errors over 100 trials by (i) Optimal sampling and (ii) Random sampling, respectively. These graphs show that (i) Optimal sampling outperforms (ii) Random sampling even in an unrealizable case. Especially, when M is small and σ^2 is large, its effectiveness is remarkable.

By using the optimal sample points $\{x_m\}_{m=1}^{M}$ designed by Eqs.(44) and (47), we will perform a model selection simulation. Here we attempt the following model selection criteria.

(a) Subspace information criterion (SIC) [22],

(b) Leave-one-out cross-validation (CV) [15],

(c) Akaike's information criterion (AIC) [1],

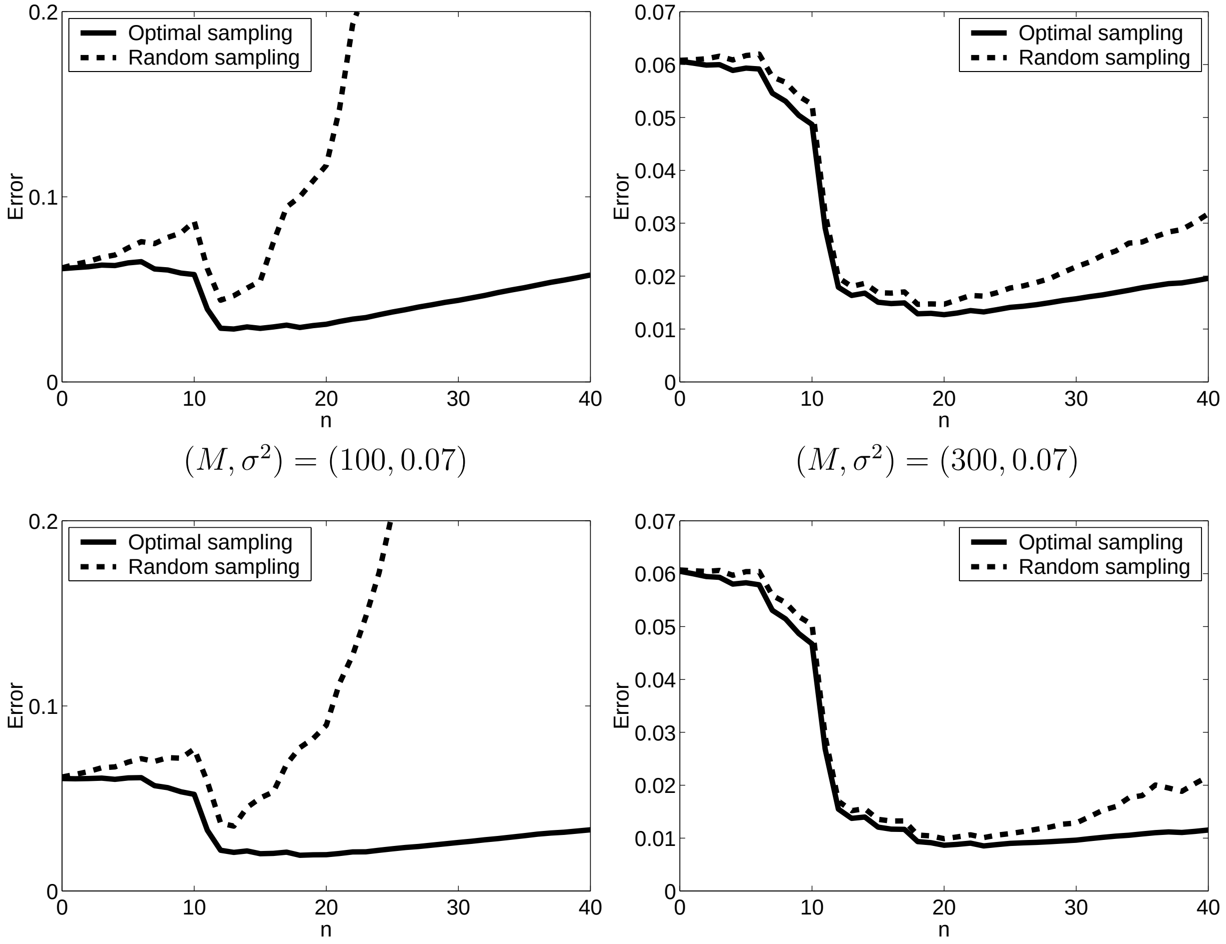
(d) Corrected AIC (cAIC) [19],

(e) Bayesian information criterion (BIC) [18],

(f) Vapnik's measure (VM) [5].

Note that for optimal sampling, SIC essentially agrees with Mallows's C_L [12].

Figures 6, 7, 8, and 9 depict the simulation results. The left seven graphs show the values of the error and model selection criteria as a function of the order n of the model S_n (see Eq.(46)). The box plot notation specifies marks at 95, 75, 50, 25, and 5 percentiles of values. The solid line denotes the mean values. The middle seven graphs show the distributions of the selected order n of models. 'OPT' indicates the optimal model that



$(M, \sigma^2) = (100, 0.07)$ $(M, \sigma^2) = (300, 0.07)$

$(M, \sigma^2) = (100, 0.04)$ $(M, \sigma^2) = (300, 0.04)$

Figure 5: Results of active learning simulation with Mackey-Glass data.

minimizes the error defined by Eq.(42). The right seven graphs show the distributions of the obtained error obtained by each model selection criterion.

Figure 6 corresponds to the case with $(M, \sigma^2) = (300, 0.04)$, i.e., the easiest case with large samples and small noise variance. In this case, all model selection criteria give good estimates of the error. As a consequence, all methods select reasonable models and result in smaller errors.

Figure 7 corresponds to the case with $(M, \sigma^2) = (100, 0.04)$, i.e., the case with small samples and small noise variance. In this case, SIC, CV, and cAIC give reasonable estimates of the error and therefore work well. In contrast, the curves of AIC, BIC, and VM seem to be rather corrupted. AIC tends to select larger models, and BIC and VM are inclined to select smaller models. Because BIC and VM almost always select the smallest model, they yield large errors.

Figure 8 corresponds to the case with $(M, \sigma^2) = (300, 0.07)$, i.e., the case with large

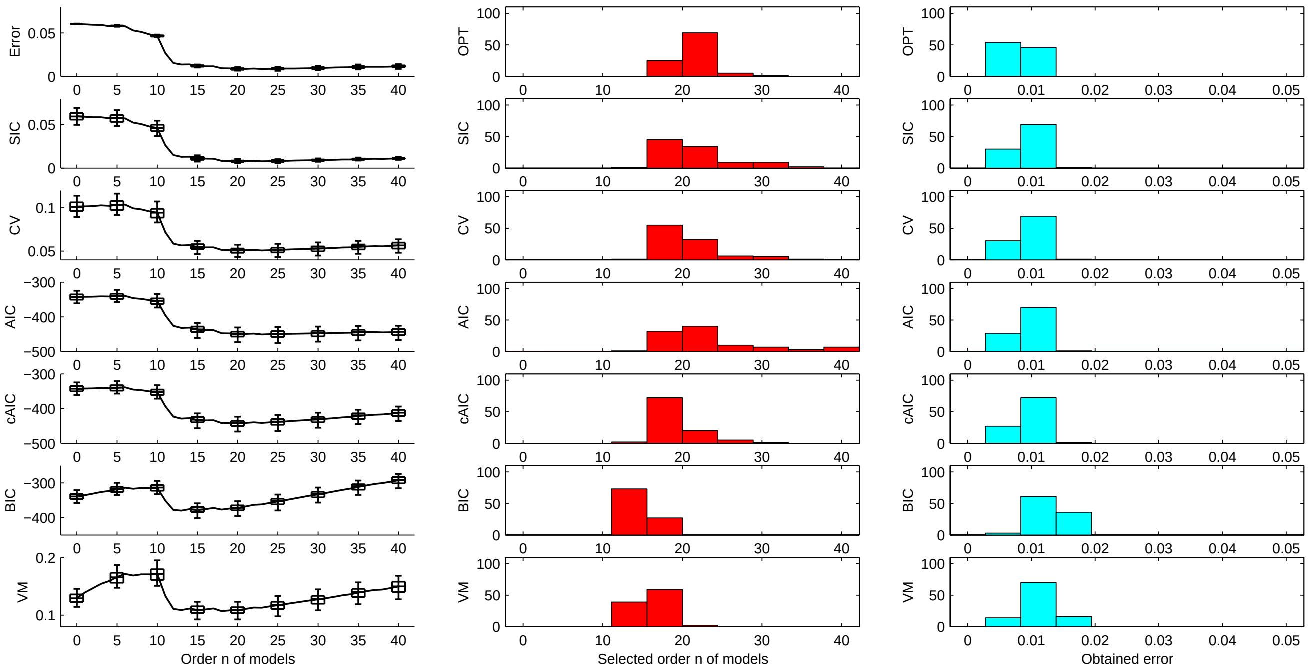


Figure 6: Results of model selection simulation with Mackey-Glass data. $(M, \sigma^2) = (300, 0.04)$.

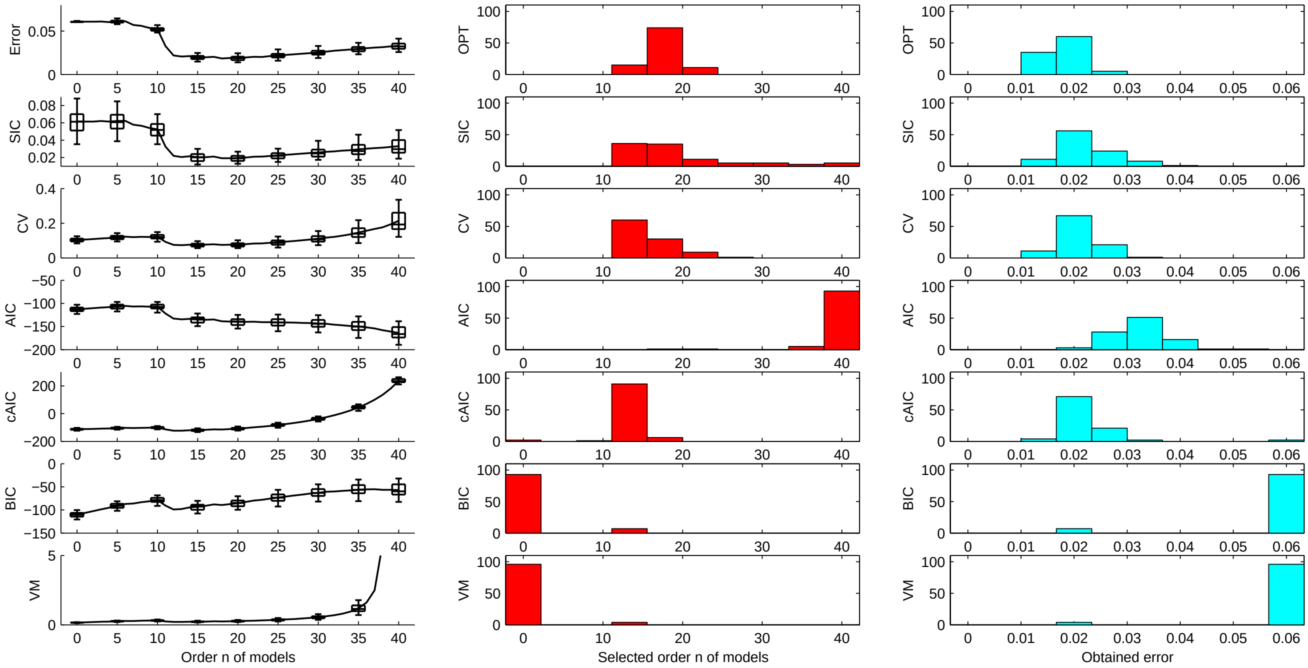


Figure 7: Results of model selection simulation with Mackey-Glass data. $(M, \sigma^2) = (100, 0.04)$.

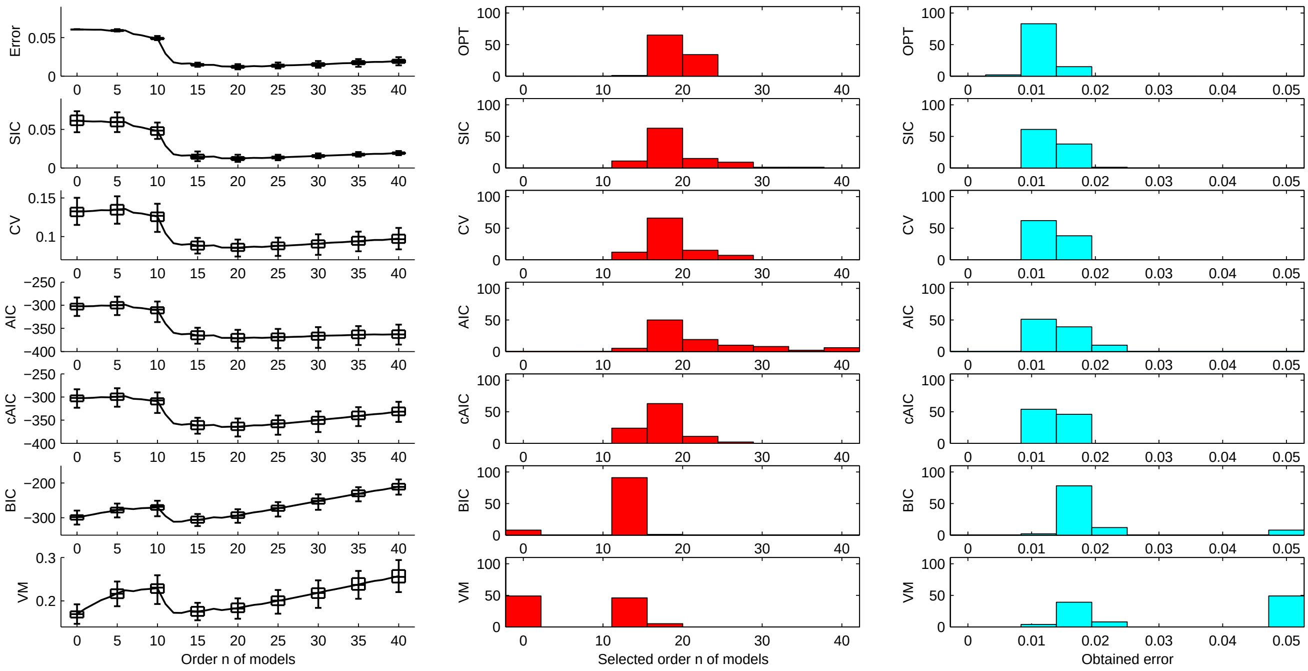


Figure 8: Results of model selection simulation with Mackey-Glass data. $(M, \sigma^2) = (300, 0.07)$.

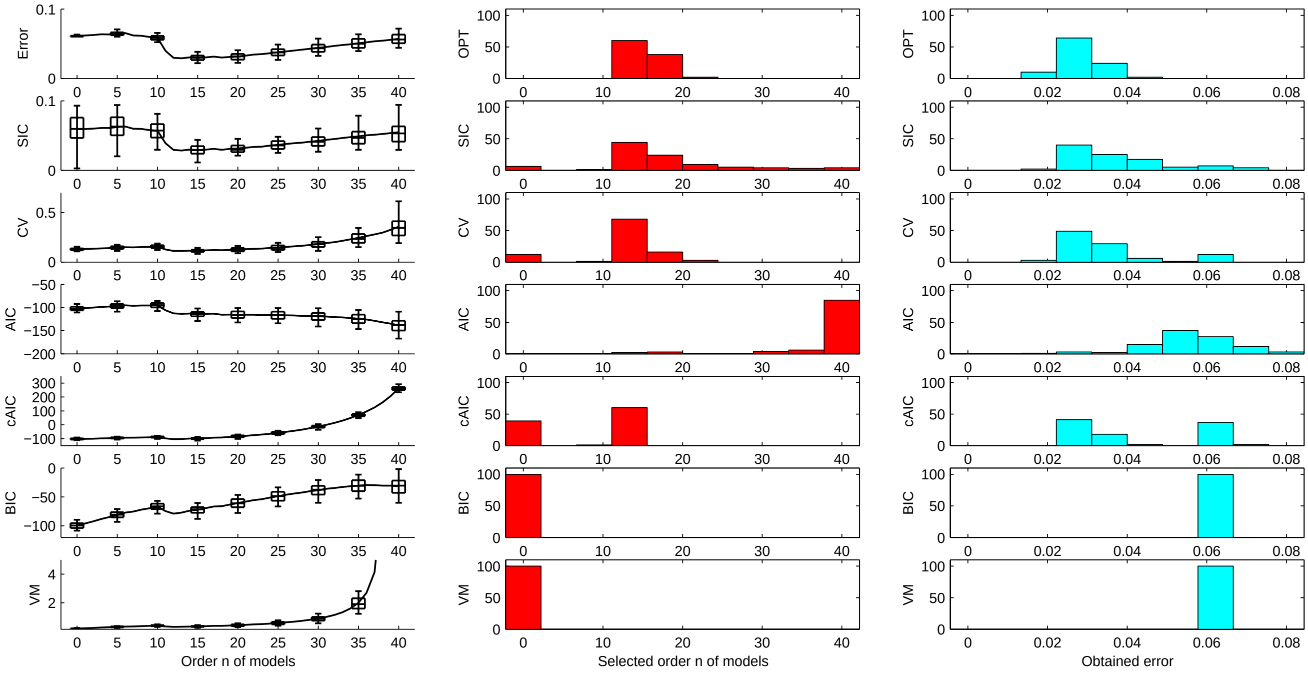


Figure 9: Results of model selection simulation with Mackey-Glass data. $(M, \sigma^2) = (100, 0.07)$.

samples and large noise variance. In this case, SIC, CV, AIC, and cAIC estimate the error fairly well and therefore result in small errors. In contrast, BIC and VM sometimes select the smallest model and give large errors.

Finally, Figure 9 corresponds to the case with $(M, \sigma^2) = (100, 0.07)$, i.e., the hardest case with small samples and large noise variance. The estimates of the error by all model selection criteria seem rather inaccurate. However, even so SIC and CV still show a tendency to select reasonable models, and result in comparatively small errors. On the other hand, AIC tends to select larger models, and cAIC, BIC, and VM show a tendency to select smaller models. As a result, they yield large errors.

To sum up, the simulation results show that the proposed sampling method with SIC or CV is an effective method for aquiring a higher level of the generalization capability.

7 Discussions and conclusions

We discussed the problem of optimizing sample points and models at the same time. We first pointed out that the problem can not be generally solved, in a batch manner, by simply combining existing active learning and model selection methods because of the *active learning / model selection dilemma*: the model should be fixed for selecting sample points and conversely the sample points should be fixed for selecting models. The main contribution of this paper was the basic strategy for dissolving the dilemma (Section 3), where the commonly optimal sample points play essential roles (see Figure 1).

Although it seemed that the strategy is rather idealistic, we showed that the strategy can be practically realized for trigonometric polynomial models (Section 5). Computer simulations shown in Section 6 demonstrated that the proposed procedure shows good performance even when the learning target function is *not* included in the trigonometric polynomial models. This fact implies that the proposed procedure may be useful for any learning target functions.

In the derivation of our active learning procedure, we imposed a rather strong assumption on the number of training examples (Section 4.4). This assumption may be practical only when the input dimension is rather small. Our important future direction is to investigate whether similar discussions can be possible even when the input dimension is large. We expect that the equidistance sampling still has intersting properties even in this scenario.

Since the properties of trigonometric polynomials is considerably used for showing that the equidistance sampling is commonly optimal, the equidistance sampling may not be optimal for other classes of models. Our important future work is to find the commonly optimal sample points for other classes of models. It was shown in [21] that Proposition 1 is valid for any finite dimensional reproducing kernel Hilbert spaces such that $K(\boldsymbol{x}, \boldsymbol{x})$ is a constant. We expect that following the disucssion in [21] would be a promissing approach to solving this problem.

Finally, we would like to devise a method for active learning with model selection even for the classes of models such that the commonly optimal sample points do *not* exist. We expect that the basic strategy proposed in this paper still plays an important role in this challenging scenario, e.g., finding approximately commonly optimal sample points would be promissing.

Acknowledgements

The authors would like to anonymous reviewers for their useful comments.

Appendix: Proof of Theorem 1

According to the reference [14], Eq.(32) is equivalent to

$$\|\frac{1}{\sqrt{M}}A_n g_n\|^2 = \|g_n\|^2 \tag{48}$$

for all $S_n \in \mathcal{M}$ and for all $g_n \in S_n$. Hence, we shall prove Eq.(48). A function $g_n(\boldsymbol{x})$ in S_n is expressed as

$$\begin{aligned} g_n(\boldsymbol{x}) \;=\; & \sum_{p_1=-n_1}^{n_1}\sum_{p_2=-n_2}^{n_2}\cdots\sum_{p_L=-n_L}^{n_L} \\ & \qquad a_{p_1,p_2,\ldots,p_L}\prod_{l=1}^{L}\exp\left(ip_l\xi^{(l)}\right), \end{aligned} \tag{49}$$

where $a_{p_1,p_2,\ldots,p_L}$ is a scalar. From Eqs.(25), (49), and (31), it holds for all S_n and for all g_n in S_n that

$$\begin{aligned} & \|\frac{1}{\sqrt{M}}A_n g_n\|^2 \\ & = \sum_{m_1=1}^{M_1}\sum_{m_2=1}^{M_2}\cdots\sum_{m_L=1}^{M_L}\left|\langle g_n(\cdot), \frac{1}{\sqrt{M}}K_n(\cdot,\boldsymbol{x}_m)\rangle\right|^2 \\ & = \frac{1}{M}\sum_{m_1=1}^{M_1}\sum_{m_2=1}^{M_2}\cdots\sum_{m_L=1}^{M_L}|g_n(\boldsymbol{x}_m)|^2 \\ & = \frac{1}{M}\sum_{m_1=1}^{M_1}\sum_{m_2=1}^{M_2}\cdots\sum_{m_L=1}^{M_L} \\ & \quad \left|\sum_{p_1=-n_1}^{n_1}\sum_{p_2=-n_2}^{n_2}\cdots\sum_{p_L=-n_L}^{n_L}a_{p_1,p_2,\ldots,p_L}\prod_{l=1}^{L}\exp(ip_l\xi_m^{(l)})\right|^2 \\ & = \frac{1}{M}\sum_{m_1=1}^{M_1}\sum_{m_2=1}^{M_2}\cdots\sum_{m_L=1}^{M_L}\sum_{p_1=-n_1}^{n_1}\sum_{p_2=-n_2}^{n_2}\cdots\sum_{p_L=-n_L}^{n_L} \end{aligned}$$

$$
\begin{aligned}
&\sum_{p'_1=-n_1}^{n_1}\sum_{p'_2=-n_2}^{n_2}\cdots\sum_{p'_L=-n_L}^{n_L} \\
&a_{p_1,p_2,\ldots,p_L}\overline{a_{p'_1,p'_2,\ldots,p'_L}}\prod_{l=1}^{L}\exp\left(i(p_l-p'_l)\xi_m^{(l)}\right) \\
= \ & \frac{1}{M}\sum_{p_1=-n_1}^{n_1}\sum_{p_2=-n_2}^{n_2}\cdots\sum_{p_L=-n_L}^{n_L} \\
&\sum_{p'_1=-n_1}^{n_1}\sum_{p'_2=-n_2}^{n_2}\cdots\sum_{p'_L=-n_L}^{n_L} a_{p_1,p_2,\ldots,p_L}\overline{a_{p'_1,p'_2,\ldots,p'_L}} \\
&\times\prod_{l=1}^{L}\left[\sum_{m_l=1}^{M_l}\exp\left(i(p_l-p'_l)\frac{2\pi m_l}{M_l}\right)\right] \\
&\times\prod_{l=1}^{L}\exp\left(i(p_l-p'_l)(c_l-\frac{2\pi}{M_l})\right). \qquad (50)
\end{aligned}
$$

Since it holds for any integers p_l and p'_l that

$$
\sum_{m_l=1}^{M_l}\exp\left(i(p_l-p'_l)\frac{2\pi m_l}{M_l}\right)=\begin{cases}0 & \text{if } p_l\neq p'_l,\\ M_l & \text{if } p_l=p'_l,\end{cases} \qquad (51)
$$

Eq.(50) yields

$$
\begin{aligned}
&\|\frac{1}{\sqrt{M}}A_n g_n\|^2 \\
&\quad=\frac{1}{M}\sum_{p_1=-n_1}^{n_1}\sum_{p_2=-n_2}^{n_2}\cdots\sum_{p_L=-n_L}^{n_L}|a_{p_1,p_2,\ldots,p_L}|^2 \\
&\quad\quad\times\prod_{l=1}^{L}M_l\times\prod_{l=1}^{L}\exp(0) \\
&\quad=\sum_{p_1=-n_1}^{n_1}\sum_{p_2=-n_2}^{n_2}\cdots\sum_{p_L=-n_L}^{n_L}|a_{p_1,p_2,\ldots,p_L}|^2 \\
&\quad=\|g_n\|^2, \qquad (52)
\end{aligned}
$$

which concludes the proof. ■

References

[1] H. Akaike "A new look at the statistical model identification," IEEE Trans. Automatic Control, vol. AC-19(6), pp. 716–723, 1974.

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

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

[4] A. C. Atkinson and D. R. Cox, "Planning experiments for discriminating between models," J. Royal Statistical Society, Series B, vol. 36, pp. 321–348, 1974.

[5] V. Cherkassky, X. Shao, X., F. M. Mulier, and V. N. Vapnik, "Model complexity control for regression using VC generalization bounds," IEEE Trans. Neural Networks, vol. 10, no. 5, pp. 1075–1089, 1999.

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

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

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

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

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

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

[12] C. L. Mallows, "Some comments on C_P," Technometrics, vol. 15, no. 4, pp. 661–675, 1973.

[13] N. Murata, S. Yoshizawa, and S. Amari, "Network information criterion—Determining the number of hidden units for an artificial neural network model," IEEE Trans. Neural Networks, vol. 5, no. 6, pp. 865–872, 1994.

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

[15] M. J. L. Orr, "Introduction to radial basis function networks," Technical Report, Center for Cognitive Science, University of Edinburgh, 1996.

[16] J. Platt, "A resource-allocating network for function interpolation," Neural Computation, vol. 3, no. 2, pp. 213–225, 1991.

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

[18] G. Schwarz, “Estimating the dimension of a model,” Annals of Statistics, vol. 6, pp. 461–464, 1978.

[19] N. Sugiura “Further analysis of the data by Akaike’s information criterion and the finite corrections,” Communications in Statistics. Theory and Methods, vol. 7, no. 1, pp. 13–26, 1978.

[20] M. Sugiyama and H. Ogawa, “Incremental active learning for optimal generalization,” Neural Computation, vol. 12, no. 12, pp. 2909–2940, 2000.

[21] M. Sugiyama and H. Ogawa, “Active learning for optimal generalization in trigonometric polynomial models,” IEICE Trans. Fundamentals, vol. E84-A, no. 9, pp. 2319–2329, 2001.

[22] M. Sugiyama and H. Ogawa, “Subspace information criterion for model selection,” Neural Computation, vol. 13, no. 8, pp. 1863–1889, 2001.