Release from Active Learning/Model Selection Dilemma: Optimizing Sample Points and Models at the Same Time

Masashi Sugiyama Hidemitsu Ogawa
Department of Computer Science, Tokyo Institute of Technology
2-12-1, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan
sugi@og.cs.titech.ac.jp *http://ogawa-www.cs.titech.ac.jp/~sugi/*

Abstract— In 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, this problem 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 spite of the dilemma, we show that the problem of active learning with model selection can be straightforwardly solved if there is a set of sample points that is optimal for all models in consideration. Based on the idea, we give a procedure for active learning with model selection in trigonometric polynomial models.

I. Supervised Learning and Active Learning/Model Selection Dilemma

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 $\mathbb{C}$:

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

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.

In supervised learning, there are two factors we can control for optimal generalization: *sample points* and a *model*. The 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*, and the problem of determining the model is called *model selection*. Let us denote a set of M sample points $\{\boldsymbol{x}_m\}_{m=1}^M$ by $\mathcal{X}$, a *model* by S, and a set of models from which the model is selected by $\mathcal{M}$.

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*.

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:*

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

In general, the model should be fixed for active learning [4, 7, 3, 6, 5, 14, 15, 18] [1], and conversely the training examples gathered at fixed sample points are required for model selection [8, 1, 13, 12, 11, 2, 16, 17]. This implies 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*. In this paper, we suggest a basic strategy for solving this dilemma, and give a practical procedure for active learning with model selection in trigonometric polynomial models.

II. Basic Strategy

As we pointed out in Section I, the problem of 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 active learning and conversely sample points should be fixed for model selection.

However, if there is a set $\mathcal{X}$ of sample points that is optimal for all models in the set $\mathcal{M}$, the problem of

[1] Some of the methods are incremental active learning methods so it is possible to change the model through the incremental learning process. However, such active learning methods essentially work for a fixed model, i.e., the sample points are designed to be optimal for the current model.

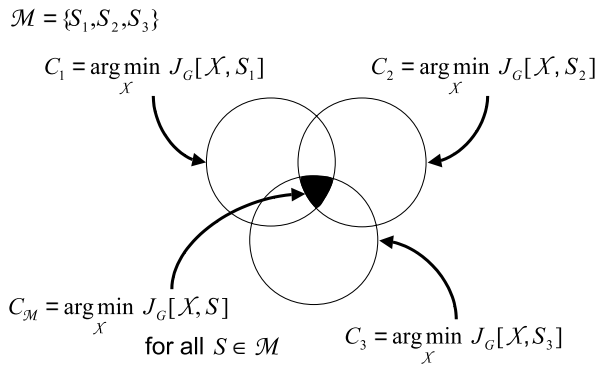


Fig. 1. Basic strategy for active learning with model selection. Let the set $\mathcal{M}$ of models be $\{S_1, S_2, S_3\}$. The top-left circle denotes a set C_1 of optimal $\mathcal{X}$ for the model S_1, i.e., an element in C_1 is a set $\mathcal{X}$ of sample points $\{\boldsymbol{x}_m\}_{m=1}^M$ 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.

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.

In the following sections, we give a procedure for active learning with model selection based on the above strategy.

III. Setting

In this section, the setting is described. From here on, we focus on the case where the dimension L of the input vector $\boldsymbol{x}$ is 1 for simplicity. All the theoretical results hold true for $L \geq 1$. However, note that the proposed method may be practically applicable only for a small L, say at most 3 (see Section III-D for detail).

A. Trigonometric Polynomial Space

We assume that the learning target function $f(x)$ belongs S_N, that is a *trigonometric polynomial space* of order N.

A trigonometric polynomial space of order n is a function space spanned by

$$\left\{ \exp(ipx) \;\middle|\; p = -n, -n+1, \ldots, n \right\} \tag{3}$$

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

$$\langle f, g \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \overline{g(x)} dx. \tag{4}$$

The dimension of a trigonometric polynomial space of order n is

$$\dim S_n = 2n + 1, \tag{5}$$

and the reproducing kernel[2] of this space is expressed as

$$K_n(x, x') = \begin{cases} \sin \frac{(2n+1)(x-x')}{2} \Big/ \sin \frac{x-x'}{2} & \text{if } x \neq x', \\ 2n+1 & \text{if } x = x'. \end{cases} \tag{6}$$

B. Least Mean Squares Learning

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

$$J_{TE} = \frac{1}{M} \sum_{m=1}^{M} \left| \hat{f}(x_m) - y_m \right|^2. \tag{7}$$

In the LMS learning case, a subspace S is the model. Since S_N has the reproducing kernel (see Section III-A), a subspace S also has the reproducing kernel. Let $K(x, x')$ be the reproducing kernel of S and A be a linear operator defined by

$$A = \sum_{m=1}^{M} \left(\boldsymbol{e}_m \otimes \overline{K(\cdot, x_m)} \right), \tag{8}$$

where $(\cdot \otimes \overline{\cdot})$ denotes the *Neumann-Schatten product*[3], 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 $Af = (f(x_1), f(x_2), \ldots, f(x_M))^\top$, where $\top$ denotes the transpose of a vector. Let $A^\dagger$ be the *Moore-Penrose generalized inverse* of A and $\boldsymbol{y}$ be an M-dimensional vector whose m-th element is the sample value y_m:

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_M)^\top. \tag{9}$$

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

$$\hat{f} = A^\dagger \boldsymbol{y}. \tag{10}$$

[2]The reproducing kernel, denoted by $K(x, x')$, is a bivariate function defined on $\mathcal{D} \times \mathcal{D}$ that satisfies the following conditions
- For any fixed x' in $\mathcal{D}$, $K(x, x')$ is a function of x in S.
- For any function f in S and for any x' in $\mathcal{D}$, it holds that $\langle f(\cdot), K(\cdot, x') \rangle = f(x')$.

Note that the reproducing kernel is unique if it exists.

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

C. Generalization Measure

We define the generalization error J_G by the expected squared norm in S_N:

$$J_G = \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f} - f\|^2 = \mathrm{E}_{\boldsymbol{\epsilon}}\frac{1}{2\pi}\int_{\mathcal{D}} |\hat{f}(x) - f(x)|^2 dx, \quad (11)$$

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

D. The Number of Training Examples

We assume that the number M of training examples satisfies

$$M \geq 2N + 1. \quad (12)$$

This assumption implies that the required number M exponentially increases as the dimension L of the input x increases. This is the reason why we practically focus on a small L, say at most 3.

E. 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.

F. Model Candidates

In the LMS learning case, a subset is the model (see Section III-B). 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:

$$\mathcal{M} = \{S_n \mid n = 0, 1, \ldots, N\}. \quad (13)$$

IV. Active Learning with Model Selection for Trigonometric Polynomial Models

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

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

$$\hat{f}_n = A_n^{\dagger}\boldsymbol{y}, \quad (14)$$

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

$$A_n = \sum_{m=1}^{M} \left(\boldsymbol{e}_m \otimes \overline{K_n(\cdot, x_m)}\right). \quad (15)$$

It is known that the generalization error of $\hat{f}_n$ defined by Eq.(11) is decomposed into the *bias* and *variance*:

$$J_G = \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f}_n - f\|^2 = \|\mathrm{E}_{\boldsymbol{\epsilon}}\hat{f}_n - f\|^2 + \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f}_n - \mathrm{E}_{\boldsymbol{\epsilon}}\hat{f}_n\|^2. \quad (16)$$

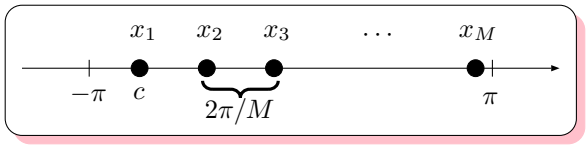


Fig. 2. Example of sample points such that Condition (20) holds.

Note that the bias of $\hat{f}_n$ can not be zero unless the learning target function f belongs to S_n. First, we show a necessary and sufficient condition for a set $\mathcal{X}$ of sample points $\{x_m\}_{m=1}^M$ that minimize the generalization error J_G for a fixed model S_n.

Proposition 1: *[15] 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 the bias being zero if and only if sample points $\mathcal{X}$ satisfy*

$$\frac{1}{M}A_n^* A_n = I_{S_n}, \quad (17)$$

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 (17) holds for a fixed model S_n [15]. Here, we give a design method of sample points that satisfy Condition (17) for all models in the set $\mathcal{M}$.

Theorem 1: [4] *Let $M \geq 2N + 1$ and c be an arbitrary constant such that*

$$-\pi \leq c \leq -\pi + \frac{2\pi}{M}. \quad (18)$$

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

$$x_m = c + \frac{2\pi}{M}(m - 1), \quad (19)$$

then

$$\frac{1}{M}A_n^* A_n = I_{S_n} \text{ for all } S_n \in \mathcal{M}. \quad (20)$$

Eq.(19) means that M sample points are fixed to regular intervals in the domain $\mathcal{D}$. An example of sample points designed by Eq.(19) is illustrated in Figure 2. When $L > 1$, sample points on the regular lattice satisfy Eq.(20).

Theorem 1 and Proposition 1 assert that the sample points designed by Eq.(19) 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.(19) minimize the variance under the constraint that the range of A_n^* agrees with S_n [15]. Computer simulations in Section V experimentally show that the sample points

[4] A proof of this theorem is given in the article available from `ftp://ftp.cs.titech.ac.jp/pub/TR/01/TR01-0012.pdf`.

designed by Eq.(19) 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.(19) can be practically regarded as a good design for all models in the set $\mathcal{M}$.

With training examples $\{(x_m, y_m)\}_{m=1}^{M}$ gathered at the optimal sample points $\{x_m\}_{m=1}^{M}$ designed by Eq.(19), 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.(19) is that LMS learning result functions can be computed efficiently since $A_n^\dagger$ is given by $\frac{1}{M}A_n^*$ [15].

V. Computer Simulations

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

A. Illustrative Example

Let the order N of the largest trigonometric polynomial space be 100. Let the learning target function $f(x)$ be

$$f(x) = \frac{1}{10}\sum_{n=1}^{50}(\sin nx + \cos nx). \tag{21}$$

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

$$\mathcal{M} = \{S_0, S_1, S_2, \ldots, S_{100}\}. \tag{22}$$

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

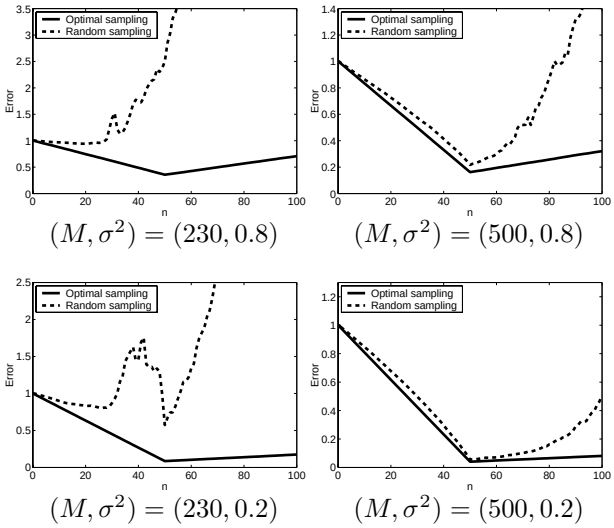
$$\text{Error} = \|\hat{f} - f\|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\hat{f}(x) - f(x)|^2 dx. \tag{23}$$

We compare the performance of the following two sampling schemes.

(i) Optimal sampling: Sample points $\{x_m\}_{m=1}^{M}$ are designed by Eq.(19).

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

Figure 3 shows the simulation results for $(M, \sigma^2) = (500, 0.2), (230, 0.2), (500, 0.8)$, and $(230, 0.8)$. The horizontal axis denotes the order n of the model and the vertical axis denotes the error measured by Eq.(23). 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.



$(M, \sigma^2) = (230, 0.8)$ $(M, \sigma^2) = (500, 0.8)$

$(M, \sigma^2) = (230, 0.2)$ $(M, \sigma^2) = (500, 0.2)$

Fig. 3. Results of illustrative simulation.

B. Unrealizable Case

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

Let us consider the chaotic series created by the *Mackey-Glass delay-difference equation* [10]:

$$g(t+1) = \begin{cases} (1-b)g(t) + \dfrac{a\, g(t-\tau)}{1+g(t-\tau)^{10}} & \text{for } t \geq \tau+1, \\ 0.3 & \text{for } 0 \leq t \leq \tau, \end{cases} \tag{24}$$

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

$$h_t = g(t + \tau + 1). \tag{25}$$

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

$$y_m = h_{r(m)} + \epsilon_m, \tag{26}$$

where $r(m)$ is an integer such that $1 \leq r(m) \leq 600$. $r(m)$ indicates the sampling location. 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:

$$\text{Error} = \frac{1}{600}\sum_{t=1}^{600}\left|\hat{h}_t - h_t\right|^2. \tag{27}$$

In this simulation, we consider the cases that

$$(M, \sigma^2) = (300, 0.04), (100, 0.07). \tag{28}$$

Figure 4 depicts 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 $\sigma^2 = 0.07$.

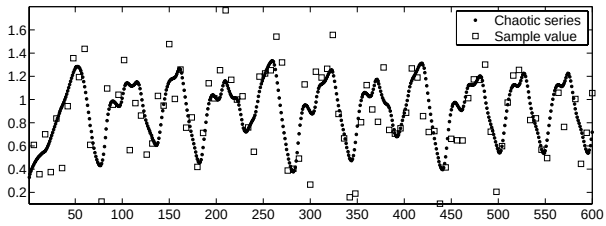
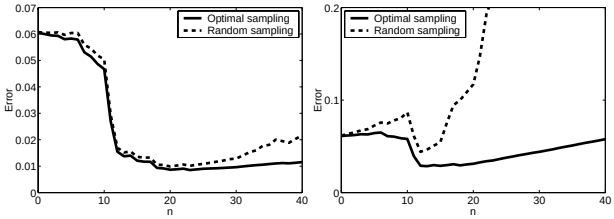


Fig. 4. Mackey-Glass chaotic series of 600 points and 100 sample values ($p(m) = 6m$ and $\sigma^2 = 0.07$).



$(M, \sigma^2) = (300, 0.04)$ $(M, \sigma^2) = (100, 0.07)$

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

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 $\{y_m\}_{m=1}^{M}$:

$$x_m = -\pi + \frac{2\pi}{600}(r(m) - 1). \tag{29}$$

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

$$\hat{h}_t = \hat{f}\left(-\pi + \frac{2\pi}{600}(t-1)\right). \tag{30}$$

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

$$\mathcal{M} = \{S_0, S_1, S_2, \ldots, S_{40}\}. \tag{31}$$

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.,

$$r(m) = 600m/M. \tag{32}$$

(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 and the vertical axis denotes the error measured by Eq.(27). 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 when the learning target function does not belong to S_N. Especially, when $(M, \sigma^2) = (100, 0.07)$, its effectiveness is remarkable.

By using the optimal sample points $\{x_m\}_{m=1}^{M}$ designed by Eqs.(29) and (32), we perform a model selection simulation. Here we use the following model selection criteria:

- Subspace information criterion (SIC) [16],
- Leave-one-out cross-validation (CV) [9],
- Akaike's information criterion (AIC) [1],
- Corrected AIC (cAIC) [13],
- Bayesian information criterion (BIC) [12],
- Vapnik's measure (VM) [2].

Note that for optimal sampling with $\frac{1}{M}A_n^* A_n = I_{S_n}$, SIC essentially agrees with Mallows's C_L [8].

Figure 6 depicts the simulation results. The left column corresponds to $(M, \sigma^2) = (300, 0.04)$ and the right column corresponds to $(M, \sigma^2) = (100, 0.07)$. The top seven graphs show the values of the error and model selection criteria corresponding to the order n of the model S_n (see Eq.(31)). 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 minimizes the error defined by Eq.(27). The bottom seven graphs show the distributions of the error obtained by the model that each model selection criterion selects.

When $(M, \sigma^2) = (300, 0.04)$, all model selection criteria work well. When $(M, \sigma^2) = (100, 0.07)$, SIC and CV almost always select reasonable models, so they provide small errors. In contrast, AIC tends to select larger models, and cAIC, BIC, and VM tend to select smaller models. As a result, they give large errors.

The simulation results show that the proposed sampling method with SIC (C_L) or CV works excellently.

VI. 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 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. In this paper, we gave a basic strategy for avoiding the dilemma, and devised a practical procedure for active learning with model selection in trigonometric polynomial models. Computer simulations demonstrated that the proposed procedure

shows exceedingly good performance irrespective of the the number of training examples and the noise variance, and whether the learning target function is realizable or unrealizable.

The range of application of the proposed procedure is restricted to trigonometric polynomial models, but the basic strategy can be applied to any models. Devising a general procedure for active learning with model selection is our important and prospective future work.

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

References

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

[2] V. Cherkassky, X. Shao, F. M. Mulier, and V. N. Vapnik. Model complexity control for regression using VC generalization bounds. *IEEE Transactions on Neural Networks*, 10(5):1075–1089, 1999.

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

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

[5] K. Fukumizu. Statistical active learning in multilayer perceptrons. *IEEE Transactions on Neural Networks*, 11(1):17–21, 2000.

[6] K. Fukumizu and S. Watanabe. Optimal training data and predictive error of polynomial approximation. *IEICE Transactions*, J79-A(5):1100–1108, 1996. (In Japanese).

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

[8] C. L. Mallows. Some comments on C_P. *Technometrics*, 15(4):661–675, 1973.

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

[10] J. Platt. A resource-allocating network for function interpolation. *Neural Computation*, 3(2):213–225, 1991.

[11] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[12] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[13] N. Sugiura. Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics. Theory and Methods*, 7(1):13–26, 1978.

[14] M. Sugiyama and H. Ogawa. Incremental active learning for optimal generalization. *Neural Computation*, 12(12):2909–2940, 2000.

[15] M. Sugiyama and H. Ogawa. Active learning for optimal generalization in trigonometric polynomial models. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E84-A(9):2319–2329, 2001.

[16] M. Sugiyama and H. Ogawa. Subspace information criterion for model selection. *Neural Computation*, 13(8):1863–1889, 2001.

[17] M. Sugiyama and H. Ogawa. Theoretical and experimental evaluation of subspace information criterion. *Machine Learning*, 48(1/2/3):25–50, 2002.

[18] S. Tong and D. Koller. Active learning for parameter estimation in bayesian networks. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 647–653. MIT Press, 2001.

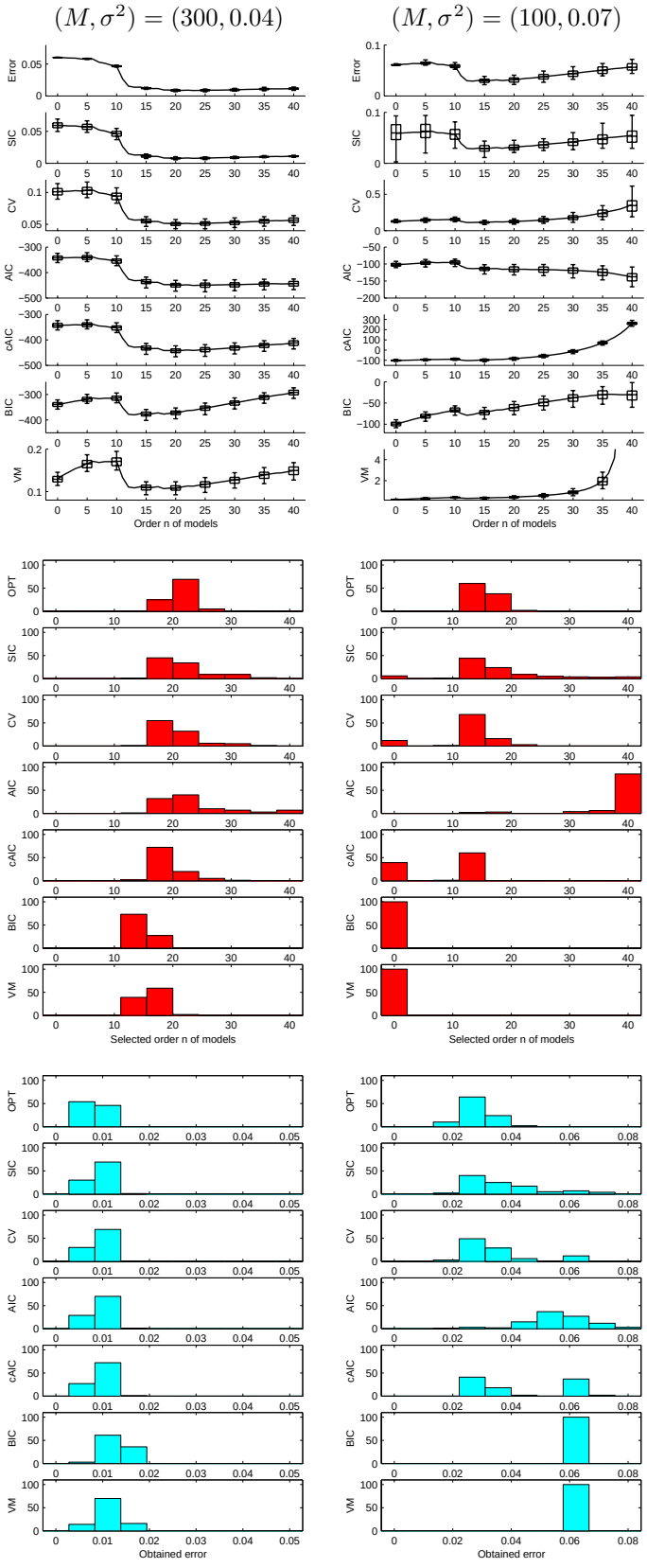


Fig. 6. Results of model selection simulation with Mackey-Glass data.