Incremental Active Learning with Bias Reduction*

Masashi Sugiyama Hidemitsu Ogawa

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

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

Abstract

The problem of designing input signals for optimal generalization in supervised learning is called active learning. In many active learning methods devised so far, the bias of the learning results is assumed to be zero. In this paper, we remove this assumption and propose a new active learning method with the bias reduction. The effectiveness of the proposed method is demonstrated through computer simulations.

1 Introduction

Supervised learning is obtaining an underlying rule from sampled training examples and can be formulated as a function approximation problem. If sample points are actively designed, then learning can be performed more efficiently. In this paper, we will discuss the problem of designing sample points, referred to as *active learning*, for optimal generalization.

So far, active learning has been studied from two different standpoints depending on the optimality: *global optimal* where a set of all sample points is optimal (e.g. Fedorov [2], Sugiyama and Ogawa [11]) and *greedy optimal* where the next sample point to add is optimal in each step (e.g. MacKay [4], Cohn [1], Fukumizu [3]). Generally, the global optimal methods give better generalization capability than the greedy optimal methods. However, the global optimal results have been obtained only for restricted cases. In contrast, the greedy optimal methods have been derived under general conditions. Even so, the greedy optimal methods devised so far are still restricted since the *bias* of the learning result is assumed to be zero, which sometimes prevents us from applying active learning to real world problems.

In this paper, we focus on the greedy optimal case and propose a new incremental active learning method with the bias reduction. The proposed method does not require the assumption of zero-bias. Our computer simulations show that the proposed method works better than usual 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 [6]).

Let us consider the 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 $\mathbf{C}$:

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

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. In this paper, the target function $f(x)$ is assumed to belong to a reproducing kernel Hilbert space H. If H is unknown, then it can be estimated by model selection methods (e.g. Sugiyama and Ogawa [10]). Let $K(x, x')$ be the reproducing kernel of H. If a function $\psi_j(x)$ is defined as $\psi_j(x) = K(x, x_j)$, then the value of f at a sample point x_j is expressed as $f(x_j) = \langle f, \psi_j \rangle$. Let A_m be an operator defined as $A_m = \sum_{j=1}^m \left(e_j^{(m)} \otimes \overline{\psi_j} \right)$, where $e_j^{(m)}$ is the j-th vector of the so-called standard basis

*The complete version of this paper is available at "ftp://ftp.cs.titech.ac.jp/pub/TR/99/TR99-0010.ps.gz".

in the m-dimensional unitary space $\mathbf{C}^m$ and $(\cdot \otimes \bar{\cdot})$ stands for the *Neumann-Schatten product*[1]. Then, the relationship between f and $y^{(m)}$ can be expressed as

$$y^{(m)} = A_m f + n^{(m)}. \tag{2}$$

Let us denote a mapping from $y^{(m)}$ to a learning result f_m by X_m:

$$f_m = X_m y^{(m)}, \tag{3}$$

where X_m is called a *learning operator*. Consequently, the supervised learning problem can be reformulated as an inverse problem of obtaining X_m providing the best approximation f_m to f under a certain learning criterion.

3 Learning process

In this section, a general process for supervised learning is described.

Supervised learning is generally processed as follows.

(i) The learning criterion is determined.

(ii) What data to gather is decided and sample values are gathered at the decided locations. (Incremental active learning)

(iii) By using the gathered training examples, a learning procedure is carried out. (Incremental learning)

(iv) The 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. The purpose of this paper is to give an incremental active learning method corresponding to (ii).

As the learning criterion corresponding to (i), we adopt *projection learning* (Ogawa [5]). Let E_n, A_m^*, $\mathcal{R}(A_m^*)$, and $P_{\mathcal{R}(A_m^*)}$ be the ensemble average over noise, the adjoint operator of A_m, the range of A_m^*, and the orthogonal projection operator onto $\mathcal{R}(A_m^*)$, respectively. Then, projection learning is defined as follows.

Definition 1 (Projection learning) *(Ogawa [5]) An operator X_m is called the projection learning operator if X_m minimizes the functional $J_P[X_m] = E_n\|X_m n^{(m)}\|^2$ under the constraint $X_m A_m = P_{\mathcal{R}(A_m^*)}$.*

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

Proposition 1 *(Ogawa [5]) A general form of the projection learning operator is expressed as*

$$X_m = V_m^\dagger A_m^* U_m^\dagger + Y_m(I_m - U_m U_m^\dagger), \tag{4}$$

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

$$Q_m = E_n\left(n^{(m)} \otimes \overline{n^{(m)}}\right), \quad U_m = A_m A_m^* + Q_m, \quad \textit{and} \quad V_m = A_m^* U_m^\dagger A_m. \tag{5}$$

Note that the projection learning operator given by eq.(4) is linear. Since the projection learning result f_m obtained by eqs.(3) and (4) belongs to $\mathcal{R}(A_m^*)$, $\mathcal{R}(A_m^*)$ is called the *approximation space*.

As an incremental learning method corresponding to (iii), we adopt a method of *incremental projection learning* (IPL) (Sugiyama and Ogawa [7, 8]). In the rest of this section, IPL is reviewed.

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 $\{(x_j, y_j)\}_{j=1}^m$. Let the noise characteristics of (x_{m+1}, y_{m+1}) be

$$q_{m+1} = E_n(\overline{n_{m+1}} n^{(m)}), \quad \text{and} \quad \sigma_{m+1} = E_n|n_{m+1}|^2, \tag{6}$$

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

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 $\mathcal{N}(A_m)$ and $P_{\mathcal{N}(A_m)}$ be the null space of A_m and the orthogonal projection operator onto $\mathcal{N}(A_m)$, respectively, and the following notation is defined.

$$
\begin{array}{lrcl}
\text{Vectors:} & s_{m+1} & = & A_m\psi_{m+1} + q_{m+1}, \quad (7)\\
& t_{m+1} & = & U_m^\dagger s_{m+1}. \quad (8)\\
\text{Scalars:} & \alpha_{m+1} & = & \psi_{m+1}(x_{m+1}) + \sigma_{m+1} - \langle t_{m+1}, s_{m+1}\rangle, \quad (9)\\
& \beta_{m+1} & = & y_{m+1} - f_m(x_{m+1}) - \langle y^{(m)} - A_m f_m, t_{m+1}\rangle. \quad (10)\\
\text{Functions:} & \tilde{\psi}_{m+1} & = & P_{\mathcal{N}(A_m)}\psi_{m+1}, \quad (11)\\
& \xi_{m+1} & = & \psi_{m+1} - A_m^* t_{m+1}, \quad (12)\\
& \tilde{\xi}_{m+1} & = & V_m^\dagger \xi_{m+1}. \quad (13)
\end{array}
$$

As shown in Sugiyama and Ogawa [7, 9], the additional training examples such that $\xi_{m+1} = 0$ can be rejected since they have no effect on learning results. Hence, from here on, we focus on the training examples such that $\xi_{m+1} \neq 0$. Then, IPL is given as follows.

Proposition 2 (Incremental projection learning) *(Sugiyama and Ogawa [7, 8]) When ξ_{m+1} defined by eq.(12) 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.*

$$
f_{m+1} = f_m + \begin{cases} \beta_{m+1}\tilde{\psi}_{m+1}/\tilde{\psi}_{m+1}(x_{m+1}) & \text{if } \psi_{m+1} \notin \mathcal{R}(A_m^*), \\ \beta_{m+1}\tilde{\xi}_{m+1}/(\alpha_{m+1} + \langle \tilde{\xi}_{m+1}, \xi_{m+1}\rangle) & \text{if } \psi_{m+1} \in \mathcal{R}(A_m^*). \end{cases} \quad (14)
$$

Note that f_{m+1} obtained by Proposition 2 exactly agrees with the learning result obtained by batch projection learning with $\{(x_j, y_j)\}_{j=1}^{m+1}$. Namely, IPL provides the optimal learning result in the sense of projection learning. 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 $\mathcal{R}(A_m^*)$.

4 Active learning based on the two-stage sampling scheme

In this section, a new method of incremental active learning is given based on the basic sampling strategy called the *two-stage sampling scheme*.

Let us measure the *generalization error* of the learning result f_m by

$$
J_g = E_n\|f_m - f\|^2. \quad (15)
$$

It is well-known that eq.(15) can be decomposed into the *bias* and *variance*:

$$
J_g = \|P_{\mathcal{R}(A_m^*)}f - f\|^2 + E_n\|X_m n^{(m)}\|^2. \quad (16)
$$

Let ΔJ_b and ΔJ_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.,

$$
\begin{aligned}
\Delta J_b &= \|P_{\mathcal{R}(A_{m+1}^*)}f - f\|^2 - \|P_{\mathcal{R}(A_m^*)}f - f\|^2, \quad (17)\\
\Delta J_v &= E_n\|X_{m+1} n^{(m+1)}\|^2 - E_n\|X_m n^{(m)}\|^2. \quad (18)
\end{aligned}
$$

Then, the following proposition holds.

Proposition 3 *(Sugiyama and Ogawa, [7, 9]) For any additional training example (x_{m+1}, y_{m+1}) such that $\xi_{m+1} \neq 0$, the following relations hold.*

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

$$
\Delta J_b \leq 0 \text{ and } \Delta J_v \geq 0. \quad (19)
$$

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

$$\Delta J_b = 0 \text{ and } \Delta J_v < 0. \tag{20}$$

Proposition 3 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 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.(16) that the bias of learning results is zero for any f in H if and only if $\mathcal{N}(A_m) = \{0\}$. Based on this fact, we comply with the following *two-stage sampling scheme*.

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 that the additional training examples such that $\psi_{m+1} \in \mathcal{R}(A_m^*)$ maintain the bias (see Proposition 3 (b)), i.e., the bias remains zero throughout Stage 2.

Since the purpose of learning is to minimize the generalization error defined by eq.(15), our active learning problems in both stages become as follows.

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

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

Note that all additional training examples in Stage 2 satisfy $\psi_{m+1} \in \mathcal{R}(A_m^*)$ since $\mathcal{N}(A_m) = \{0\}$ has been attained at the end of Stage 1. This means that, in Stage 2, the constraint $\psi_{m+1} \in \mathcal{R}(A_m^*)$ does not have to be taken into account.

In the statistical active learning methods devised so far, the bias of the estimator is assumed to be zero (MacKay [4], Cohn [1], Fukumizu [3]). The assumption of zero-bias is equivalent to that f belongs to H and $E_n f_m$ agrees with f. In contrast, the condition assumed in our framework is only $f \in H$. The difference between f and $E_n f_m$ is explicitly evaluated in Stage 1 in spite of the fact that the bias is unknown.

Based on the two-stage sampling scheme described above, we shall give an incremental active learning method. The following theorem plays a central role in the derivation.

Theorem 1 ΔJ_v *defined by eq.(18) can be expressed as follows.*

$$\Delta J_v = \begin{cases} (\alpha_{m+1} + \langle \tilde{\xi}_{m+1}, \xi_{m+1} \rangle) / \tilde{\psi}_{m+1}(x_{m+1}) & \text{if } \psi_{m+1} \notin \mathcal{R}(A_m^*), \\ -\|\tilde{\xi}_{m+1}\|^2 / (\alpha_{m+1} + \langle \tilde{\xi}_{m+1}, \xi_{m+1} \rangle) & \text{if } \psi_{m+1} \in \mathcal{R}(A_m^*). \end{cases} \tag{21}$$

Theorem 1 implies that ΔJ_v can be calculated without y_{m+1}. Namely, the quality of additional training examples can be evaluated only by using their sampling locations. It should also be noted that when the noise covariance matrix Q_{m+1} is in the form $Q_{m+1} = \sigma^2 I_{m+1}$ with $\sigma^2 > 0$, the minimization of ΔJ_v can be performed without the value σ^2 of the noise variance. In this case, the lower half of eq.(21) is essentially equivalent to the criteria used in MacKay [4], Cohn [1], and Fukumizu [3].

In this paper, the minimization of ΔJ_v is performed by *multi-point-search*, i.e., c locations are created in the domain and the one minimizing ΔJ_v is selected. The algorithm of *two-stage active learning by multi-point-search* is described in Fig.1.

5 Computer simulations

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

Let us consider learning in a trigonometric polynomial space of order 100, i.e., H is spanned by $\{1, \sin nx, \cos nx\}_{n=1}^{100}$ and the inner product is defined as

$$\langle f, g \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \overline{g(x)} dx. \tag{22}$$

$m \leftarrow 0$;
while $\mathcal{N}(A_m) \neq \{0\}$ { % Stage 1
    `Generate` c `locations` $\{x_{m+1}^{(j)}\}_{j=1}^{c}$ `such that` $\psi_{m+1} \notin \mathcal{R}(A_m^*)$ `as candidates`;
    $j_0 \leftarrow \underset{j}{\operatorname{argmin}} \Delta J_v(x_{m+1}^{(j)})$;
    `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 $m < M$ { % Stage 2
    `Generate` c `locations` $\{x_{m+1}^{(j)}\}_{j=1}^{c}$ `as candidates`;
    $j_0 \leftarrow \underset{j}{\operatorname{argmin}} \Delta J_v(x_{m+1}^{(j)})$;
    `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 1: Algorithm of two-stage active learning by multi-point-search.

Let the total number M of training examples to add be 500, and the noise covariance matrix be $Q_M = I_M$. In this case, projection learning gives the same learning result as usual least squares learning minimizing the empirical error $\sum_{j=1}^{m}(y_j - f_m(x_j))^2$. We shall compare the performance of the following sampling schemes.

(A) Proposed method: Training examples are sampled following the two-stage active learning method shown in Fig.1. Let the number c of candidates be 3 and randomly generate them in the domain $[-\pi, \pi]$.

(B) Experimental design: Eq.(2) in Cohn [1] is adopted as the active learning criterion. The value of this criterion is evaluated by 30 reference points. The next sampling location is determined by multi-point-search with 3 candidates.

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

Note that the performance of sampling schemes can be fairly compared by this simulation since the common model, learning criterion, and incremental learning method are adopted.

The changes in the variance through the addition of training examples are shown in Fig.2. The horizontal axis denotes the number m of training examples while the vertical axis denotes the variance. The solid, dashed, and dotted lines denote the means of 10 trials by the sampling schemes (A)–(C), respectively. In the sampling scheme (A), it always holds that $\mathcal{N}(A_{201}) = \{0\}$ because the dimension of H is 201 (see Section 4). In the sampling schemes (B) and (C), $\mathcal{N}(A_{201}) = \{0\}$ was attained in all 10 trials in this simulation. Hence, it follows from eq.(16) that the vertical axis in Fig.2 can be regarded as the generalization error when $m \geq 201$.

This graph shows that, when $m \leq 201$, the variances of all sampling schemes increase, this phenomenon is in good agreement with Proposition 3. When $m = 201$, the generalization error of the sampling scheme (A) is 7.48 while the generalization errors of the sampling schemes (B) and (C) are 3.18×10^4 and 8.75×10^4, respectively. When $m > 201$, the variances of all sampling schemes decrease as shown in Proposition 3. This result shows that the sampling scheme (A) gives much better generalization capability than the sampling schemes (B) and (C).

6 Conclusion

In this paper, we proposed a new active learning method called two-stage active learning. In many active learning methods devised so far, the bias of the learning results is assumed to be zero. In contrast, the

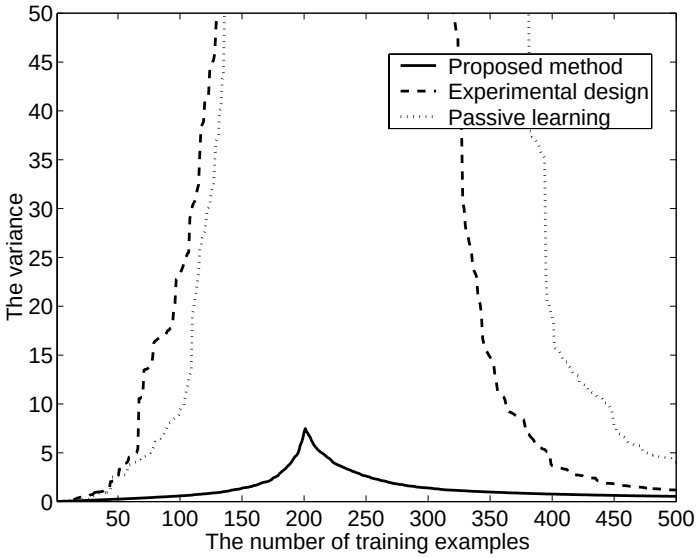

Figure 2: Relation between the number m of training examples and the variance in a trigonometric polynomial space of order 100 with the noise covariance matrix $Q_{500} = I_{500}$. The vertical axis can be regarded as the generalization error when $m \geq 201$.

proposed method did not require the assumption of zero-bias. Our simulation demonstrated the effectiveness of the proposed method.

References

[1] Cohn, D. A. (1994). Neural network exploration using optimal experiment design. In J. Cowan et al. (Eds.), *Advances in Neural Information Processing Systems 6* (pp. 679–686). San Mateo, CA: Morgan-Kaufmann Publishers, Inc.

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

[3] Fukumizu, K. (1996). Active learning in multilayer perceptrons. In D. Touretzky et al. (Eds.), *Advances in Neural Information Processing Systems 8* (pp. 295–301). Cambridge: MIT Press.

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

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

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

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

[8] 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 ftp://ftp.cs.titech.ac.jp/pub/TR/99/TR99-0007.ps.gz).

[9] 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 ftp://ftp.cs.titech.ac.jp/pub/TR/99/TR99-0008.ps.gz).

[10] Sugiyama, M., & Ogawa, H. (1999d). Functional analytic approach to model selection — Subspace information criterion. In *Proceedings of 1999 Workshop on Information-Based Induction Sciences (IBIS'99)* (pp. 93–98). Izu, Japan (Its complete version is available at ftp://ftp.cs.titech.ac.jp/pub/TR/99/TR99-0009.ps.gz).

[11] Sugiyama, M., & Ogawa, H. (1999e). Training data selection for optimal generalization in trigonometric polynomial networks. To be published in S. A. Solla et al. (Eds.), *Advances in Neural Information Processing Systems 12*.