Properties of Incremental Projection Learning

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

We proposed a method of incremental projection learning which provides exactly the same generalization capability as that obtained by batch projection learning in the previous paper. However, properties of the method have not yet been investigated. In this paper, we analyze its properties from the following aspects: First, it is shown that some of the training examples regarded as redundant in most incremental learning methods have potential effectiveness, i.e., they will contribute to better generalization capability in the future learning process. Based on this fact, an improved criterion for the redundancy of additional training examples is derived. Second, the relationship between prior and posterior learning results is investigated where effective training examples are classified into two categories from the viewpoint of improving generalization capability. Finally, a simpler form of incremental projection learning under certain conditions is given. The size of memory required for storing prior results in the simple form is fixed and independent of the total number of training examples.

Keywords

Multilayer feedforward neural networks; Generalization capability; Incremental projection learning; Projection learning; Reproducing kernel Hilbert space (RKHS).

Nomenclature

$\mathbf{C}^m$	m-dimensional unitary space
H	Hilbert space
$\langle \cdot, \cdot \rangle$	inner product in H or $\mathbf{C}^m$
$\Vert \cdot \Vert$	norm in H or $\mathbf{C}^m$
$\cdot \otimes \bar{\cdot}$	Neumann-Schatten product
$K(\cdot, \cdot)$	reproducing kernel of H
x_i	input of neural networks
y_i	output of neural networks
(x_i, y_i)	training example
$y^{(m)}$	m-dimensional vector consisting of $\{y_i\}_{i=1}^m$
n_i	additive noise
$n^{(m)}$	m-dimensional vector consisting of $\{n_i\}_{i=1}^m$
f	function of learning target
f_m	learning result from a set of m training examples
A_m	sampling operator for a set of m training examples
X_m	learning operator for a set of m training examples
ψ_i	sampling function of the i-th training example
Q_m	noise correlation matrix of a set of m training examples
q_{m+1}	noise covariance of the $(m+1)$-st training example
σ_{m+1}	noise variance of the $(m+1)$-st training example
I_m	identity matrix on $\mathbf{C}^m$
$e_i^{(m)}$	i-th vector of the standard basis in $\mathbf{C}^m$
E_n	ensemble average over noise
A^*	adjoint operator of A
$A^\dagger$	Moore-Penrose generalized inverse of A
$\mathcal{R}(A)$	range of A
$\mathcal{N}(A)$	null space of A
P_S	orthogonal projection operator onto a subspace S
α_{m+1}, β_{m+1}, β'_{m+1}	scalars
$\tilde{\psi}_{m+1}$, ξ_{m+1}, $\tilde{\xi}_{m+1}$, ζ_{m+1}, ζ'_{m+1}	functions in H
s_{m+1}, t_{m+1}	elements in $\mathbf{C}^m$
U_m	operator from $\mathbf{C}^m$ to $\mathbf{C}^m$
V_m, V'_m	operators from H to H
Y_m	operator from $\mathbf{C}^m$ to H

1 Introduction

Incremental learning has practically extensive importance in many applications. However, the optimality of the generalization capability is not guaranteed in many incremental learning methods (Platt, 1991; Kadirkamanathan & Niranjan, 1993; Zhang, 1994; Vyšniauskas *et al.*, 1995; Yamauchi & Ishii, 1995; Jutten & Chentouf, 1995; Molina & Niranjan, 1996; Yingwei *et al.*, 1997, 1998; Vijayakumar & Schaal, 1998). Amari (1998) proposed an incremental learning method which gives asymptotically the same generalization capability as that obtained by batch learning (see also Murata, 1999). Still, the optimal generalization capability in the non-asymptotic case is not guaranteed in the method. In practice, the number of training examples is always finite. Sugiyama and Ogawa (2001) proposed a method of incremental projection learning (IPL). The learning result obtained by IPL exactly agrees with that obtained by batch projection learning (Ogawa, 1987) even in the non-asymptotic case. Thus, the optimal generalization capability can be acquired by IPL. Properties of IPL, however, have not been investigated yet.

In this paper, IPL is analyzed from the following aspects: First, we discuss the redundancy of additional training examples. In usual incremental learning methods, the additional training example is regarded as redundant if the posterior learning result of incremental learning is exactly the same as the prior learning result, Against the claim, we show that some of such training examples have potential effectiveness, i.e., they will contribute to better generalization capability in the future learning process. Based on the fact, an improved criterion for the redundancy of additional training examples is derived. Second, the relationship between prior and posterior learning results is studied where effective training examples are classified into two categories from the viewpoint of improving generalization capability: One category consists of training examples which contribute to reducing the bias of the learning results. The other category consists of training examples which reduce the variance of the learning results. Finally, a simpler form of IPL under certain conditions is given. The size of memory required for storing prior results in the simple form is fixed and independent of the number total of training examples.

2 Formulation of supervised learning problem

In the following sections, we show properties of IPL. As arrangements, the supervised learning problem is first formulated (see Ogawa, 1992). Then, the projection learning criterion (Ogawa, 1987) and a method of IPL (Sugiyama & Ogawa, 2001) are reviewed.

2.1 Supervised learning as an inverse problem

Let us consider a function approximation problem of obtaining the optimal approximation to a target function $f(x)$ of L variables from a set of m training examples. Training examples are made up of inputs $x_i \in \mathbf{R}^L$ and corresponding outputs $y_i \in \mathbf{C}$:

$$\{(x_i, y_i) | y_i = f(x_i) + n_i\}_{i=1}^{m}, \tag{1}$$

where y_i is degraded by additive noise n_i. Let $n^{(m)}$ and $y^{(m)}$ be m-dimensional vectors whose i-th elements are n_i and y_i, 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 underlying function $f(x)$ is assumed to belong to a reproducing kernel Hilbert space H (Aronszajn, 1950; Bergman, 1970; Saitoh, 1988, 1997). Let $K(x, x')$ be the reproducing kernel of H. If a function $\psi_i(x)$ is defined as

$$\psi_i(x) = K(x, x_i), \tag{2}$$

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

$$f(x_i) = \langle f, \psi_i \rangle. \tag{3}$$

For this reason, ψ_i is called a *sampling function*. Let A_m be an operator mapping f to an m-dimensional vector whose i-th element is $f(x_i)$. A_m is called a *sampling operator*, and it is expressed by using the *Neumann-Schatten product* [1] as

$$A_m = \sum_{i=1}^{m} \left(e_i^{(m)} \otimes \overline{\psi_i} \right), \tag{4}$$

where $e_i^{(m)}$ is an m-dimensional vector where all elements are zero except the i-th element which is equal to one. Then, the relationship between f and $y^{(m)}$ can be expressed as

$$y^{(m)} = A_m f + n^{(m)}. \tag{5}$$

Let us denote a learning result obtained from m training examples by f_m, and the relationship between $y^{(m)}$ and f_m as

$$f_m = X_m y^{(m)}, \tag{6}$$

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

[1] For any g in a Hilbert space H_1 and 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)

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

2.2 Incremental projection learning

As mentioned above, function approximation is performed on the basis of a learning criterion. In this paper, we adopt the projection learning criterion. We shall start from reviewing the definition of projection learning and a general form of the projection learning operator obtained in a batch manner.

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, 1987) An operator X_m is called the projection learning operator if X_m minimizes the functional*

$$J_P[X_m] = E_n \|X_m n^{(m)}\|^2 \tag{7}$$

under the constraint

$$X_m A_m = P_{\mathcal{R}(A_m^*)}. \tag{8}$$

Let I_m and Y_m be the identity matrix on $\mathbf{C}^m$ and an arbitrary operator from $\mathbf{C}^m$ to H, respectively, and

$$\begin{aligned} Q_m &= E_n \left(n^{(m)} \otimes \overline{n^{(m)}} \right), & (9) \\ U_m &= A_m A_m^* + Q_m, & (10) \\ V_m &= A_m^* U_m^\dagger A_m, & (11) \end{aligned}$$

where † stands for the *Moore-Penrose generalized inverse*[2]. Then, we have the following proposition.

Proposition 1 *(Ogawa, 1987) A general form of the projection learning operator is expressed as*

$$X_m = V_m^\dagger A_m^* U_m^\dagger + Y_m (I_m - U_m U_m^\dagger). \tag{12}$$

Since the projection learning operator is linear, it follows from eqs.(6) and (5) that the learning result f_m can be decomposed as

$$f_m = X_m A_m f + X_m n^{(m)}. \tag{13}$$

The first and second terms of eq.(13) are called the *signal* and *noise components* of f_m, respectively. The projection learning criterion requires the signal component to coincide

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

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

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

with the orthogonal projection of f onto $\mathcal{R}(A_m^*)$ and the noise component to minimize its variance.

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, the Euclidean inner product is adopted without loss of generality.

Now let us consider the case where the $(m+1)$-st training example (x_{m+1}, y_{m+1}) is added to a projection learning result f_m obtained from $\{(x_i, y_i)\}_{i=1}^m$. Let the noise characteristics of (x_{m+1}, y_{m+1}) be

$$\begin{aligned} q_{m+1} &= E_n(\overline{n_{m+1}}n^{(m)}), & (14)\\ \sigma_{m+1} &= E_n|n_{m+1}|^2, & (15) \end{aligned}$$

where $\overline{n_{m+1}}$ denotes the complex conjugate of $n^{(m)}$. Let $\mathcal{N}(A_m)$ be the null space of A_m and the following notation is defined for introducing a method of incremental projection learning (IPL).

Matrix:

$$\Gamma_{m+1} = \sum_{i=1}^{m}\left(e_i^{(m+1)} \otimes \overline{e_i^{(m)}}\right). \qquad (16)$$

Vectors:

$$\begin{aligned} s_{m+1} &= A_m\psi_{m+1} + q_{m+1}, & (17)\\ t_{m+1} &= U_m^\dagger s_{m+1}. & (18) \end{aligned}$$

Scalars:

$$\begin{aligned} \alpha_{m+1} &= \psi_{m+1}(x_{m+1}) + \sigma_{m+1} - \langle t_{m+1}, s_{m+1}\rangle, & (19)\\ \beta_{m+1} &= y_{m+1} - f_m(x_{m+1}) - \langle y^{(m)} - A_m f_m, t_{m+1}\rangle. & (20) \end{aligned}$$

Functions:

$$\begin{aligned} \tilde{\psi}_{m+1} &= P_{\mathcal{N}(A_m)}\psi_{m+1}, & (21)\\ \xi_{m+1} &= \psi_{m+1} - A_m^* t_{m+1}, & (22)\\ \tilde{\xi}_{m+1} &= V_m^\dagger \xi_{m+1}. & (23) \end{aligned}$$

Γ_{m+1} expands an m-dimensional vector s into an $(m+1)$-dimensional vector, while Γ_{m+1}^* removes the $(m+1)$-th element as follows:

$$\begin{pmatrix} s \\ 0 \end{pmatrix} = \Gamma_{m+1}s, \qquad s = \Gamma_{m+1}^*\begin{pmatrix} s \\ a \end{pmatrix}, \qquad (24)$$

where a is a scalar.

It has been shown in Sugiyama and Ogawa (2001) that α_{m+1} is always non-negative. Based on the fact, IPL is given as follows:

Proposition 2 (Incremental projection learning) *(Sugiyama & Ogawa, 2001) When $\alpha_{m+1} > 0$, 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*

$$f_{m+1} = f_m + \beta_{m+1}\zeta_{m+1}, \qquad (25)$$

where ζ_{m+1} *is given as follows:*

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

$$\zeta_{m+1} = \frac{\tilde{\psi}_{m+1}}{\tilde{\psi}_{m+1}(x_{m+1})}. \tag{26}$$

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

$$\zeta_{m+1} = \frac{\tilde{\xi}_{m+1}}{\alpha_{m+1} + \langle \tilde{\xi}_{m+1}, \xi_{m+1} \rangle}. \tag{27}$$

When $\alpha_{m+1} = 0$*, it holds that*

$$f_{m+1} = f_m. \tag{28}$$

Note that f_{m+1} incrementally obtained by eq.(25) exactly agrees with the learning result obtained in a batch manner as

$$f_{m+1} = X_{m+1} y^{(m+1)} \tag{29}$$

where X_{m+1} is given in the form of eq.(12).

The purpose of this paper is to clarify properties of IPL given in Proposition 2.

3 Effectiveness of additional training examples

In this section, it is pointed out that some of the training examples regarded as redundant in most incremental learning methods have, as a matter of fact, potential effectiveness. Based on this fact, an improved criterion for the redundancy of additional training examples is derived.

3.1 Potentially effective training examples

In many incremental learning methods such as the *resource allocating network (RAN)* (Platt, 1991), its derivatives (Kadirkamanathan & Niranjan, 1993; Molina & Niranjan, 1996; Yingwei *et al.*, 1997, 1998), and natural gradient on-line learning (Amari, 1998), their learning criteria are aimed at minimizing the training error at x_{m+1}, i.e.,

$$y_{m+1} - f_m(x_{m+1}). \tag{30}$$

If the training error at x_{m+1} is zero before adding (x_{m+1}, y_{m+1}), then the above methods do not extract any information from the training example and yield $f_{m+1} = f_m$. Therefore, such training examples are regarded as redundant and rejected. In contrast, we can extract valuable information from such training examples in some cases. Here, we show a simple example:

Let the function space H be spanned by

$$\{\sin 6x, \sin 10x, \sin 15x\}, \tag{31}$$

and the inner product in H be defined as

$$\langle f, g \rangle = \frac{2}{\pi} \int_0^{\frac{\pi}{2}} f(x)\overline{g(x)}\, dx. \tag{32}$$

Let the target function be $f(x) = 9\sin 6x + 5\sin 15x$. For the sake of simplicity, the learning takes place in the absence of noise in this example. The target function f and the learning result f_1 obtained by using $(x_1, y_1) = (\frac{\pi}{5}, f(\frac{\pi}{5}))$ are shown as the solid and dotted lines, respectively, in Fig.1 (a). The second training example is sampled as $(x_2, y_2) = (\frac{\pi}{3}, f(\frac{\pi}{3}))$. Note that the training error at x_2 is zero, i.e.,

$$y_2 - f_1(x_2) = 0. \tag{33}$$

Now let us consider two cases: One is adding (x_2, y_2) to f_1 without rejection and the other is complying with the usual criterion for the redundancy, i.e., reject (x_2, y_2) since the training error at x_2 is zero. If (x_2, y_2) is added to f_1 without rejection, we obtain the learning result f_2 which agrees with f_1. Then, we shall add $(x_3, y_3) = (\frac{\pi}{9}, f(\frac{\pi}{9}))$ to both f_1 and f_2. The learning results f_2' and f_3 obtained by adding (x_3, y_3) to f_1 and f_2 are shown as the dotted and solid lines, respectively, in Fig.1 (b). f_3 agrees with the target function f while f_2' does not. This example shows that f_3 acquires higher generalization capability than f_2', which implies that (x_2, y_2) is effective.

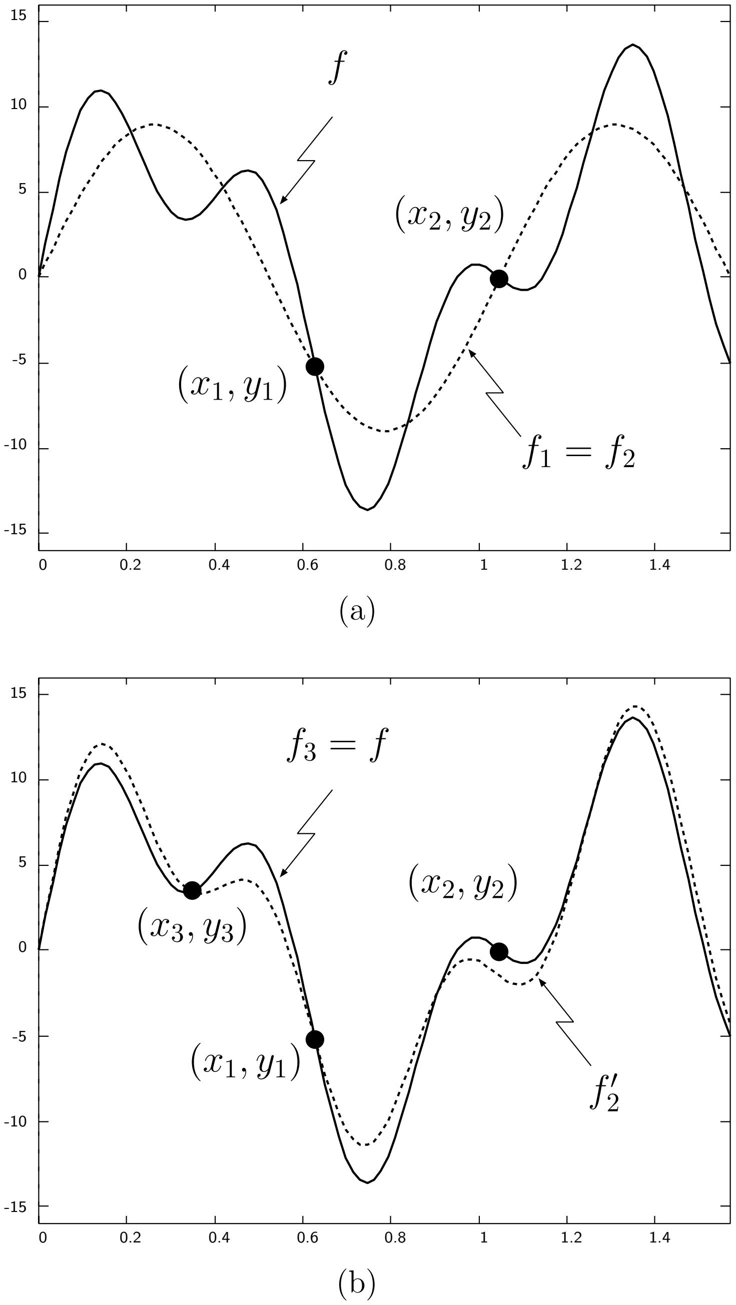
The reason why (x_2, y_2) had potential effectiveness can be understood from the functional analytic point of view. The geometrical relations between the target function f, learning results f_1, f_2, f_2', and f_3 in the function space H are shown in Fig.2. In the absence of noise, the projection learning result f_m is coincident with the orthogonal projection of f onto $\mathcal{R}(A_m^*)$. $\mathcal{R}(A_m^*)$ is called the *approximation space for* f_m. Since f belongs to $\mathcal{R}(A_1^*) + \mathcal{N}(A_2)$ in this example, we have

$$f_2 = P_{\mathcal{R}(A_2^*)} f = P_{\mathcal{R}(A_1^*)} f = f_1. \tag{34}$$

This implies that f_2 agrees with f_1 though $\mathcal{R}(A_2^*)$ properly includes $\mathcal{R}(A_1^*)$ as shown in Fig.2 (a). Rejecting (x_2, y_2) and adding (x_3, y_3) to f_1, we obtain f_2' (see Fig.2 (b)). In this case, the approximation space $\mathcal{R}(A_2'^*)$ for f_2' becomes a two-dimensional subspace. Since f does not belong to $\mathcal{R}(A_2'^*)$, f_2' does not agree with f. On the other hand, if we use (x_2, y_2) without rejection and add (x_3, y_3) to f_2, we obtain f_3. In this case, $\mathcal{R}(A_3^*)$ becomes a three-dimensional subspace which coincides with H. Since f belongs to $\mathcal{R}(A_3^*)$, f_3 agrees with f. After all, the difference between f_3 and f_2' is caused by the difference in approximation spaces, which can not be judged by simply comparing learning results.

3.2 Identification of redundant training examples

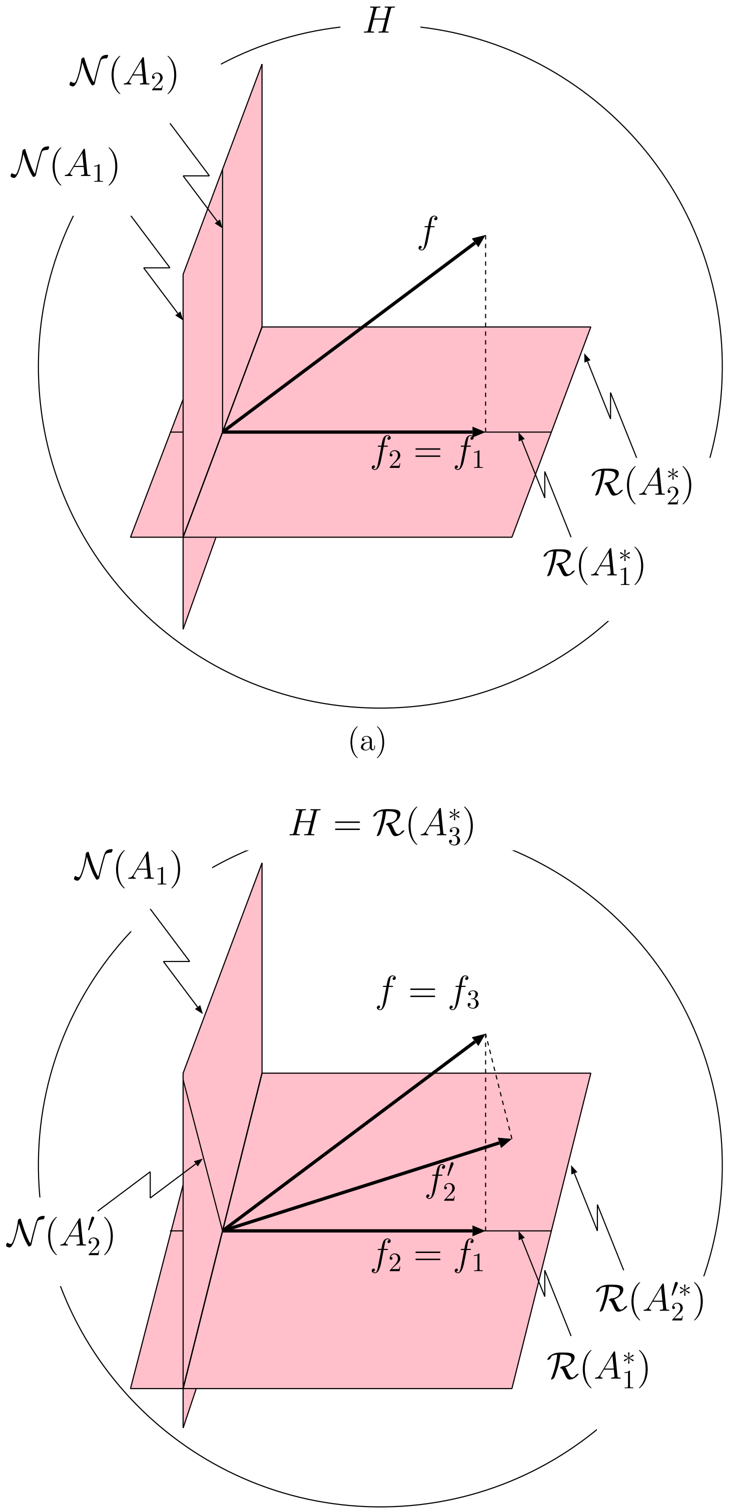
As mentioned in the previous subsection, the redundancy of additional training examples can not be simply judged by the training error at the location of the additional examples.



(a)

(b)

Figure 1: Example of the training example regarded as redundant in usual incremental learning methods but it is effective. Bullets denote training examples. (a) The target function f is shown as the solid line. The learning result f_1 obtained by using (x_1, y_1) is shown as the dotted line. The learning result f_2 obtained by adding (x_2, y_2) to f_1 is exactly the same as f_1. (b) The learning result f_2' obtained by adding (x_3, y_3) to f_1 is shown as the dotted line. The learning result f_3 obtained by adding (x_3, y_3) to f_2 is shown as the solid line, which is exactly the same as the target function f.



(a)

(b)

Figure 2: Geometrical interpretation of the example in Fig.1. (a) f_2 agrees with f_1 though $\mathcal{R}(A_2^*)$ properly includes $\mathcal{R}(A_1^*)$. (b) f_3 agrees with f while f_2' does not. Namely, f_3 acquires higher generalization capability than f_2'. This is caused by the difference in approximation spaces, i.e., $\mathcal{R}(A_3^*)$ properly includes $\mathcal{R}(A_2'^*)$.

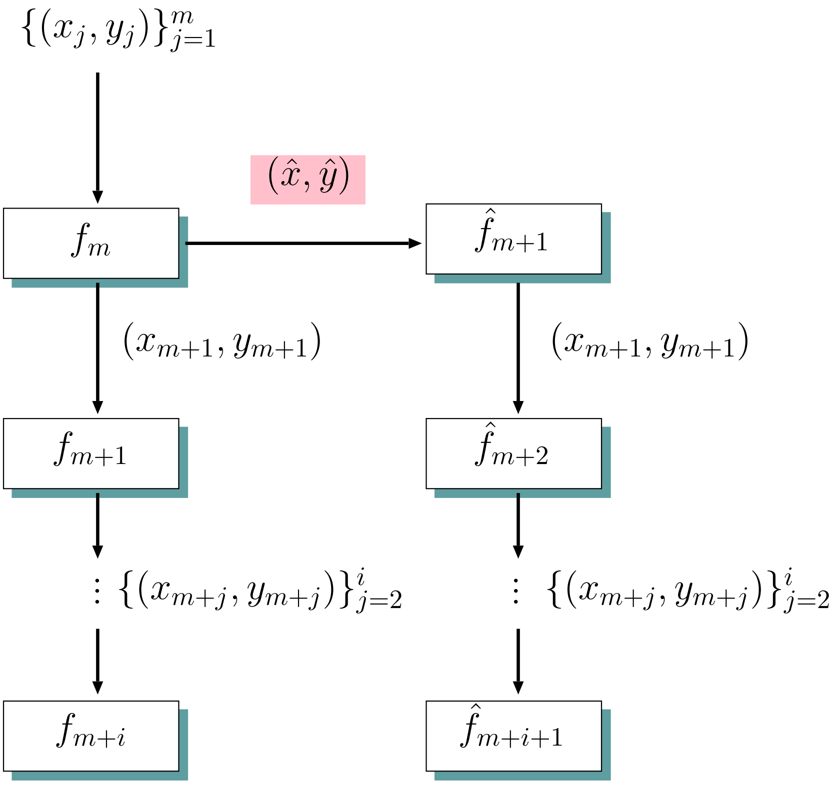


Figure 3: Definition of the effective training examples. Let f_m be a learning result obtained by using $\{(x_i, y_i)\}_{i=1}^{m}$, and $\hat{f}_{m+1}$ be a learning result obtained by adding $(\hat{x}, \hat{y})$ to f_m. Let f_{m+i} and $\hat{f}_{m+i+1}$ be learning results obtained by adding $\{(x_{m+j}, y_{m+j})\}_{j=1}^{i}$ to f_m and $\hat{f}_{m+1}$, respectively.

Here, we give proper definitions of the effectiveness and redundancy of additional training examples. Let f_m be a learning result obtained from $\{(x_j, y_j)\}_{j=1}^{m}$, and $\hat{f}_{m+1}$ be a learning result obtained by adding $(\hat{x}, \hat{y})$ to f_m. Let f_{m+i} and $\hat{f}_{m+i+1}$ be learning results obtained by adding i training examples $\{(x_{m+j}, y_{m+j})\}_{j=1}^{i}$ to f_m and $\hat{f}_{m+1}$, respectively (Fig.3).

Definition 2 $(\hat{x}, \hat{y})$ *is said to be effective if there exists at least one set of training examples which yield* $f_{m+i} \neq \hat{f}_{m+i+1}$ *for a non-negative integer* i. *Conversely,* $(\hat{x}, \hat{y})$ *is said to be redundant if it is not effective.*

Note that the above definition depends on f, f_m, A_m, and $U_m^{\dagger}$. Based on the definition, a criterion for the redundancy of additional training examples is given as follows:

Theorem 1 (Redundancy criterion) (x_{m+1}, y_{m+1}) *is redundant if* $\xi_{m+1} = 0$, *where* ξ_{m+1} *is the function defined by eq.(22)*

A proof of the theorem is given in Appendix A. As we mentioned above, additional training examples are rejected in usual redundancy criterion if $f_{m+1} = f_m$. In IPL, $f_{m+1} = f_m$ holds if and only if one of the following four conditions holds.

(a) $\alpha_{m+1} = 0$,

(b) $\alpha_{m+1} > 0$, $\psi_{m+1} \notin \mathcal{R}(A_m^*)$, and $\beta_{m+1} = 0$,

(c) $\alpha_{m+1} > 0$, $\psi_{m+1} \in \mathcal{R}(A_m^*)$, $\zeta_{m+1} \neq 0$, and $\beta_{m+1} = 0$,

(d) $\alpha_{m+1} > 0$, $\psi_{m+1} \in \mathcal{R}(A_m^*)$, and $\zeta_{m+1} = 0$,

where α_{m+1}, ψ_{m+1}, β_{m+1}, and ζ_{m+1} are given by eqs.(19), (2), (20), and (27), respectively. Among these conditions, $\xi_{m+1} = 0$ if and only if (a) or (d) holds. The conditions (a) and (d) do not depend on the value of y_{m+1} while (b) and (c) do, which implies that additional training examples are judged to be redundant if it causes $f_{m+1} = f_m$ independently of y_{m+1}. Note that the additional training example (x_2, y_2) in Fig.1 corresponds to the condition (b).

4 Improving generalization capability through IPL

The previous section discussed the redundancy of additional training examples. In this section, the properties of effective additional examples are studied from the viewpoint of improving generalization capability. The mean noise is assumed to be zero through this section.

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

$$J_G = E_n\|f_m - f\|^2. \tag{35}$$

Eq.(35) can be decomposed as follows:

Proposition 3 *(Takemura, 1991) It holds that*

$$J_G = \|E_n f_m - f\|^2 + E_n\|f_m - E_n f_m\|^2. \tag{36}$$

The first and second terms of eq.(36) are called the *bias* and *variance* of f_m, respectively. Substituting eqs.(13) and (8) into eq.(36), we have

$$J_G = \|P_{\mathcal{R}(A_m^*)} f - f\|^2 + E_n\|X_m n^{(m)}\|^2. \tag{37}$$

Eq.(37) implies that the projection learning criterion reduces the bias of f_m to a certain level and minimizes the variance of f_m.

Let J_b and J_v be the changes in the bias and variance of the learning results through the addition of a training example, respectively, i.e.,

$$\begin{aligned} J_b &= \|E_n f_{m+1} - f\|^2 - \|E_n f_m - f\|^2, & (38)\\ J_v &= E_n\|f_{m+1} - E_n f_{m+1}\|^2 - E_n\|f_m - E_n f_m\|^2. & (39)\end{aligned}$$

Then, we have the following theorem.

Theorem 2 *For any additional training example (x_{m+1}, y_{m+1}) such that $\xi_{m+1} \neq 0$, the following relations hold.*

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

$$J_b \leq 0 \text{ and } J_v \geq 0. \tag{40}$$

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

$$J_b = 0 \text{ and } J_v < 0. \tag{41}$$

A proof of the theorem is given in Appendix B. Theorem 2 states that additional training examples such that $\psi_{m+1} \notin \mathcal{R}(A_m^*)$ reduce or maintain the bias of the learning results while they increase or maintain the variance. On the other hand, additional training examples such that $\psi_{m+1} \in \mathcal{R}(A_m^*)$ maintain the bias while they reduce the variance. It seems that training examples which satisfy $\psi_{m+1} \notin \mathcal{R}(A_m^*)$ and yield $J_b = 0$ and $J_v = 0$ is redundant. However, it is not true since such training examples yield $\xi_{m+1} \neq 0$, and hence they are not always redundant as shown in Theorem 1. The additional training example (x_2, y_2) in Fig.1 is an example of such a training example.

Theorem 2 plays an important role when we work on active learning (Sugiyama & Ogawa, 2000). Generally, the bias of the learning results can not be evaluated, so that it is common to assume that the bias is zero or small enough to be neglected (MacKay, 1992; Cohn, 1996; Fukumizu, 1996). However, thanks to this theorem, the bias can be explicitly reduced by adding training examples such that $\psi_{m+1} \notin \mathcal{R}(A_m^*)$ though the bias is unknown.

5 Simple representation of IPL

In this section, a simple form of IPL under certain conditions is given.

Let an operator V'_m from H to H be

$$V'_m = A_m^* Q_m^\dagger A_m. \tag{42}$$

Then, we have the following proposition.

Proposition 4 *(Ogawa, 1987) If $\mathcal{R}(Q_m) \supset \mathcal{R}(A_m)$, then the projection learning operator is expressed as*

$$X_m = V_m'^\dagger A_m^* Q_m^\dagger. \tag{43}$$

Now, let us consider the case where the noise correlation matrix Q_{m+1} is positive definite and diagonal, i.e.,

$$Q_{m+1} = \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_{m+1}), \tag{44}$$

where $\sigma_i > 0$ for all i. In this case, V'_m becomes

$$V'_m = A_m^* Q_m^{-1} A_m. \tag{45}$$

Then, IPL can be reduced to as follows:

Theorem 3 *If Q_{m+1} is given by eq.(44) with $\sigma_i > 0$ for all i, a posterior projection learning result f_{m+1} can be obtained by using prior results f_m and $V_m'^{\dagger}$ as*

$$f_{m+1} = f_m + \beta'_{m+1}\zeta'_{m+1}, \tag{46}$$

where

$$\beta'_{m+1} = y_{m+1} - f_m(x_{m+1}), \tag{47}$$

and ζ'_{m+1} are given as follows:

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

$$\zeta'_{m+1} = \frac{\tilde{\psi}_{m+1}}{\tilde{\psi}_{m+1}(x_{m+1})}. \tag{48}$$

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

$$\zeta'_{m+1} = \frac{V_m'^{\dagger}\psi_{m+1}}{\sigma_{m+1} + \langle V_m'^{\dagger}\psi_{m+1}, \psi_{m+1}\rangle}. \tag{49}$$

A proof of the theorem is given in Appendix C.

Compared with Proposition 2, eq.(20) is replaced with eq.(47) in Theorem 3. In the case $\psi_{m+1} \notin \mathcal{R}(A_m^*)$, eq.(26) is the same as eq.(48). On the other hand, in the case $\psi_{m+1} \in \mathcal{R}(A_m^*)$, eq.(27) is replaced with eq.(49) where α_{m+1} does not appear. Although α_{m+1} played an important role in the derivation of Proposition 2 (see Sugiyama & Ogawa, 2001), it is not required in Theorem 3 since it is always positive when the noise correlation matrix is positive definite.

It is notable that Theorem 3 does not require $\{y_i\}_{i=1}^m$ for calculating f_{m+1}. Generally, the storage capacity for $V_m'^{\dagger}$ is $O(\mu^2)$ where μ is the dimension of H. In many practical situations, the number m of training examples is larger than μ. In such cases, the size of memory required for storing $V_m'^{\dagger}$ is independent of m.

Yoneda *et al.* (1992) and Yamakawa *et al.* (1993) used a temporary memory of fixed size to store some of the learned training examples in order to avoid forgetting the old training examples through the sequential process. However, the optimal generalization capability is not theoretically guaranteed in these methods. In contrast, Theorem 3 provides exactly the same generalization capability as that obtained by batch projection learning despite the fact that it also uses a memory of fixed size.

6 Conclusion

In this paper, incremental projection learning was analyzed from the following aspects: First, it was shown that some of the training examples regarded as redundant in most incremental learning methods have potential effectiveness, and an improved criterion for the redundancy of additional training examples was derived. Second, effective training examples were classified into two categories from the viewpoint of improving generalization capability: One category consists of training examples which contribute to reducing the bias of the learning results. The other category consists of training examples which reduce the variance of the learning results. Finally, a simpler form of IPL under certain conditions was given in which the size of memory required for storing prior results is fixed and independent of the number total of training examples.

Appendices

A Proof of Theorem 1

We employ the notation used in Definition 2. Namely, let us consider the case where $(\hat{x}, \hat{y})$ is added to f_m (see Fig.3). The following proof is also valid for the general case where (x_{m+1}, y_{m+1}) is added to f_m. The noise characteristics of $(\hat{x}, \hat{y})$ are denoted as

$$\begin{aligned} \hat{q}_{m+1} &= E_n(\hat{n}n^{(m)}), & (50) \\ \hat{\sigma} &= E_n(\hat{n}^2), & (51) \end{aligned}$$

where $\hat{n}$ is an additive noise in $\hat{y}$. All variables related to $(\hat{x}, \hat{y})$ are denoted by $\hat{\cdot}$, i.e.,

$$\begin{aligned} \hat{\psi} &= K(x, \hat{x}), & (52) \\ \hat{s}_{m+1} &= A_m\hat{\psi} + \hat{q}_{m+1}, & (53) \\ \hat{t}_{m+1} &= U_m^\dagger \hat{s}_{m+1}, & (54) \\ \hat{\alpha}_{m+1} &= \hat{\psi}(\hat{x}) + \hat{\sigma} - \langle \hat{t}_{m+1}, \hat{s}_{m+1} \rangle, & (55) \\ \hat{\xi}_{m+1} &= \hat{\psi} - A_m^*\hat{t}_{m+1}. & (56) \end{aligned}$$

Since learning results obtained by IPL and batch projection learning are always coincident with each other, learning results obtained by IPL do not depend on the order of training examples. This implies that $\hat{f}_{m+i+1}$ can also be obtained by adding $(\hat{x}, \hat{y})$ to f_{m+i}. By using this fact, we have the following lemma.

Lemma 1 *If $\hat{\alpha}_{m+1} = 0$, then $\hat{\alpha}_{m+i+1} = 0$ for any positive integer i.*

Proofs of all lemmas are given in Appendix D. Based on the above arrangements, we shall proof Theorem 1.

(Proof of Theorem 1) Suppose $\hat{\xi}_{m+1} = 0$. Since $\hat{\alpha}_{m+1} \geq 0$, we shall start from the case where $\hat{\alpha}_{m+1} = 0$. In this case, it follows from Proposition 2 that $\hat{f}_{m+1} = f_m$. Hence, Lemma 1 implies $\hat{f}_{m+i+1} = f_{m+i}$ for any positive integer i, which proves that $(\hat{x}, \hat{y})$ is redundant. Now, we show the case where $\hat{\alpha}_{m+1} > 0$. Let $\hat{X}_{m+1}$ be the projection learning operator obtained from $\{(x_i, y_i)\}_{i=1}^m \cup \{(\hat{x}, \hat{y})\}$. Since $\hat{\xi}_{m+1} = 0$, it follows from eq.(56) that

$$\hat{\psi} = A_m^*\hat{t}_{m+1}, \qquad (57)$$

which implies $\hat{\psi} \in \mathcal{R}(A_m^*)$. In this case, it follows from Lemma 8 in Sugiyama and Ogawa (2001) that

$$\hat{X}_{m+1} = X_m\Gamma_{m+1}^* + Y_{m+1}\Gamma_{m+1}(I_m - U_mU_m^\dagger)\Gamma_{m+1}^*. \tag{58}$$

Since the second term of the right-hand side of eq.(58) has no effect on learning results, $(\hat{x}, \hat{y})$ is clearly redundant. ■

B Proof of Theorem 2

For proving Theorem 2, the following lemma is used:

Lemma 2 *It holds that*

$$\begin{aligned} J_b &= \|P_{\mathcal{N}(A_{m+1})}f\|^2 - \|P_{\mathcal{N}(A_m)}f\|^2, && (59)\\ J_v &= tr(X_{m+1}Q_{m+1}X_{m+1}^*) - tr(X_mQ_mX_m^*), && (60)\end{aligned}$$

where $tr(\cdot)$ *stands for the trace of an operator.*

(Proof of Theorem 2) Assume that $\psi_{m+1} \notin \mathcal{R}(A_m^*)$. It follows from Lemma 1 in Sugiyama and Ogawa (2001) that

$$A_{m+1} = \Gamma_{m+1}A_m + e_{m+1}^{(m+1)} \otimes \overline{\psi_{m+1}}, \tag{61}$$

which implies $\mathcal{R}(A_{m+1}^*) \supset \mathcal{R}(A_m^*)$. Hence, from eq.(59), we immediately have $J_b \leq 0$. Since $n^{(m)} \in \mathcal{R}(U_m)$, it follows from Lemma 8 in Sugiyama and Ogawa (2001) that

$$\begin{aligned} \mathrm{tr}(X_{m+1}Q_{m+1}X_{m+1}^*) &= E_n\|X_{m+1}n^{(m+1)}\|^2 \\ &= E_n\|X_mn^{(m)} + \frac{\langle n^{(m+1)}, h_{m+1}\rangle}{\tilde{\psi}_{m+1}\,(x_{m+1})}\tilde{\psi}_{m+1}\|^2, && (62)\end{aligned}$$

where h_{m+1} is an $(m+1)$-dimensional vector. defined as

$$h_{m+1} = e_{m+1}^{(m+1)} - \Gamma_{m+1}(t_{m+1} + X_m^*\xi_{m+1}). \tag{63}$$

From Ogawa (1987), it holds that

$$\mathcal{R}(X_mU_m) = \mathcal{R}(A_m^*). \tag{64}$$

Since $n^{(m)} \in \mathcal{R}(U_m)$ because of eqs.(10) and (9), if follows from eq.(21) that

$$\langle X_mn^{(m)}, \tilde{\psi}_{m+1}\rangle = \langle P_{\mathcal{N}(A_m)}X_mn^{(m)}, \psi_{m+1}\rangle = 0. \tag{65}$$

Hence, it follows from eq.(62) that

$$\begin{aligned} \mathrm{tr}(X_{m+1}Q_{m+1}X_{m+1}^*) &= E_n\|X_mn^{(m)}\|^2 + \frac{E_n|\langle n^{(m+1)}, h_{m+1}\rangle|^2}{\|\tilde{\psi}_{m+1}\|^2} \\ &= \mathrm{tr}(X_mQ_mX_m^*) + \frac{E_n|\langle n^{(m+1)}, h_{m+1}\rangle|^2}{\|\tilde{\psi}_{m+1}\|^2}, && (66)\end{aligned}$$

which yields

$$J_v = \frac{E_n|\langle n^{(m+1)}, h_{m+1}\rangle|^2}{\|\tilde{\psi}_{m+1}\|^2} \geq 0. \tag{67}$$

Now, we shall prove the latter half. Assume that $\psi_{m+1} \in \mathcal{R}(A_m^*)$. From eq.(61), it holds that

$$\mathcal{R}(A_{m+1}^*) = \mathcal{R}(A_m^*). \tag{68}$$

Hence, from eq.(59), we have $J_b = 0$. It follows from eqs.(10), (12), (8), and (11) that

$$\begin{aligned} X_{m+1}Q_{m+1}X_{m+1}^* &= X_{m+1}U_{m+1}X_{m+1}^* - X_{m+1}A_{m+1}A_{m+1}^*X_{m+1}^* \\ &= V_{m+1}^\dagger A_{m+1}^* U_{m+1}^\dagger U_{m+1} U_{m+1}^\dagger A_{m+1} V_{m+1}^\dagger - P_{\mathcal{R}(A_{m+1}^*)} P_{\mathcal{R}(A_{m+1}^*)}^* \\ &= V_{m+1}^\dagger A_{m+1}^* U_{m+1}^\dagger A_{m+1} V_{m+1}^\dagger - P_{\mathcal{R}(A_{m+1}^*)} \\ &= V_{m+1}^\dagger V_{m+1} V_{m+1}^\dagger - P_{\mathcal{R}(A_{m+1}^*)} \\ &= V_{m+1}^\dagger - P_{\mathcal{R}(A_{m+1}^*)}. \end{aligned} \tag{69}$$

From Lemma 5 in Sugiyama and Ogawa (2001), it holds that $\alpha_{m+1} > 0$ since $\alpha_{m+1} \geq 0$ and $\xi_{m+1} \neq 0$. Hence, it follows from Lemma 7 in Sugiyama and Ogawa (2001) that

$$V_{m+1}^\dagger = V_m^\dagger - \frac{\tilde{\xi}_{m+1} \otimes \overline{\tilde{\xi}_{m+1}}}{\alpha_{m+1} + \langle \tilde{\xi}_{m+1}, \xi_{m+1}\rangle}. \tag{70}$$

Substituting eqs.(70) and (68) into eq.(69), we have

$$\begin{aligned} X_{m+1}Q_{m+1}X_{m+1}^* &= V_m^\dagger - \frac{\tilde{\xi}_{m+1} \otimes \overline{\tilde{\xi}_{m+1}}}{\alpha_{m+1} + \langle \tilde{\xi}_{m+1}, \xi_{m+1}\rangle} - P_{\mathcal{R}(A_m^*)} \\ &= X_m Q_m X_m^* - \frac{\tilde{\xi}_{m+1} \otimes \overline{\tilde{\xi}_{m+1}}}{\alpha_{m+1} + \langle \tilde{\xi}_{m+1}, \xi_{m+1}\rangle}. \end{aligned} \tag{71}$$

Eqs.(60) and (71) yield

$$J_v = -\mathrm{tr}(\frac{\tilde{\xi}_{m+1} \otimes \overline{\tilde{\xi}_{m+1}}}{\alpha_{m+1} + \langle \tilde{\xi}_{m+1}, \xi_{m+1}\rangle}) = -\frac{\|\tilde{\xi}_{m+1}\|^2}{\alpha_{m+1} + \langle \tilde{\xi}_{m+1}, \xi_{m+1}\rangle}. \tag{72}$$

Since $V_m^\dagger$ is a non-negative operator from eq.(11), it follows from eq.(23) that $\langle \tilde{\xi}_{m+1}, \xi_{m+1}\rangle \geq 0$. Since $\psi_{m+1} \in \mathcal{R}(A_m^*)$, it follows from eq.(22) that $\xi_{m+1} \in \mathcal{R}(A_m^*)$. Hence, $\xi_{m+1} \neq 0$ yields $\|\tilde{\xi}_{m+1}\|^2 > 0$ because $\mathcal{R}(V_m^\dagger) = \mathcal{R}(A_m^*)$ (Ogawa, 1987). Since it follows from Lemma 5 in Sugiyama and Ogawa (2001) that $\alpha_{m+1} > 0$ if $\xi_{m+1} \neq 0$, eq.(72) yields $J_v < 0$. ■

C Proof of Theorem 3

For proving Theorem 3, the following lemmas are prepared.

Lemma 3 *If Q_{m+1} is given by eq.(44) with $\sigma_i > 0$ for all i, then V'_{m+1} can be expressed by using V'_m as*

$$V'_{m+1} = V'_m + \frac{\psi_{m+1} \otimes \overline{\psi_{m+1}}}{\sigma_{m+1}}. \tag{73}$$

Lemma 4 *If Q_{m+1} is given by eq.(44) with $\sigma_i > 0$ for all i, then $V'^{\dagger}_{m+1}$ can be expressed by using $V'^{\dagger}_m$ as follows:*

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

$$\begin{aligned} V'^{\dagger}_{m+1} &= V'^{\dagger}_m + \frac{\sigma_{m+1} + \langle V'^{\dagger}_m \psi_{m+1}, \psi_{m+1}\rangle}{\tilde{\psi}_{m+1}(x_{m+1})^2} \tilde{\psi}_{m+1} \otimes \overline{\tilde{\psi}_{m+1}} \\ &\quad - \frac{V'^{\dagger}_m \psi_{m+1} \otimes \overline{\tilde{\psi}_{m+1}} + \tilde{\psi}_{m+1} \otimes \overline{V'^{\dagger}_m \psi_{m+1}}}{\tilde{\psi}_{m+1}(x_{m+1})}. \end{aligned} \tag{74}$$

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

$$V'^{\dagger}_{m+1} = V'^{\dagger}_m - \frac{V'^{\dagger}_m \psi_{m+1} \otimes \overline{V'^{\dagger}_m \psi_{m+1}}}{\sigma_{m+1} + \langle V'^{\dagger}_m \psi_{m+1}, \psi_{m+1}\rangle}. \tag{75}$$

Lemma 5 *If Q_{m+1} is given by eq.(44) with $\sigma_i > 0$ for all i, then X_{m+1} can be expressed by using X_m as*

$$X_{m+1} = X_m \Gamma^*_{m+1} + \zeta'_{m+1} \otimes \overline{h'_{m+1}}, \tag{76}$$

where h'_{m+1} is the $(m+1)$-dimensional vector defined as

$$h'_{m+1} = e^{(m+1)}_{m+1} - \Gamma_{m+1} X^*_m \psi_{m+1}, \tag{77}$$

and ζ'_{m+1} is given as shown in Theorem 3.

(Proof of Theorem 3) It follows from eq.(6) that

$$f_{m+1} = X_{m+1} y^{(m+1)}. \tag{78}$$

From eqs.(78), (76), (6), (77), (3), and (47), we have

$$\begin{aligned} f_{m+1} &= f_m + \langle y^{(m+1)}, h'_{m+1}\rangle \zeta'_{m+1} \\ &= f_m + \langle y^{(m+1)}, e^{(m+1)}_{m+1} - \Gamma_{m+1} X^*_m \psi_{m+1}\rangle \zeta'_{m+1} \\ &= f_m + (y_{m+1} - \langle X_m y^{(m)}, \psi_{m+1}\rangle) \zeta'_{m+1} \\ &= f_m + (y_{m+1} - f_m(x_{m+1})) \zeta'_{m+1} \\ &= f_m + \beta'_{m+1} \zeta'_{m+1}, \end{aligned} \tag{79}$$

which implies the theorem. ■

D Proofs of lemmas

(Proof of Lemma 1)
To begin with, we prepare some relations. It follows from eqs.(50) and (16) that

$$\hat{q}_{m+i+1} = \Gamma_{m+i}\hat{q}_{m+i} + E_n(\hat{n}n_{m+i})e_{m+i}^{(m+i)}. \tag{80}$$

From eqs.(53), (80), and (61), it holds that

$$\begin{aligned}\hat{s}_{m+i+1} &= (\Gamma_{m+i}A_{m+i-1} + e_{m+i}^{(m+i)} \otimes \overline{\psi_{m+i}})\hat{\psi} + \Gamma_{m+i}\hat{q}_{m+i} + E_n(\hat{n}n_{m+i})e_{m+i}^{(m+i)} \\ &= \Gamma_{m+i}(A_{m+i-1}\hat{\psi} + \hat{q}_{m+i}) + (\langle\hat{\psi}, \psi_{m+i}\rangle + E_n(\hat{n}n_{m+i}))e_{m+i}^{(m+i)} \\ &= \Gamma_{m+i}\hat{s}_{m+i} + c_{m+i}e_{m+i}^{(m+i)}, \end{aligned}\tag{81}$$

where c_{m+i} is a scalar defined as

$$c_{m+i} = \langle\hat{\psi}, \psi_{m+i}\rangle + E_n(\hat{n}n_{m+i}). \tag{82}$$

Based on the arrangements, we show that $\hat{\alpha}_{m+i+2} = 0$ if $\hat{\alpha}_{m+i+1} = 0$ for any fixed integer $i > 0$.

Since $\alpha_{m+i+1} \geq 0$, we shall start from the case where $\alpha_{m+i+1} > 0$. It follows from Lemma 4 in Sugiyama and Ogawa (2001) that

$$U_{m+i+1}^{\dagger} = \begin{pmatrix} U_{m+i}^{\dagger} + \dfrac{t_{m+i+1} \otimes \overline{t_{m+i+1}}}{\alpha_{m+i+1}} & -\dfrac{t_{m+i+1}}{\alpha_{m+i+1}} \\ -\dfrac{t_{m+i+1}^{*}}{\alpha_{m+i+1}} & \dfrac{1}{\alpha_{m+i+1}} \end{pmatrix}, \tag{83}$$

where t_{m+i+1}^{*} is the complex conjugate of the transpose of t_{m+i+1}. From eqs.(55), (54), (83), and (81), it holds that

$$\begin{aligned}\hat{\alpha}_{m+i+2} &= \hat{\psi}(\hat{x}) + \hat{\sigma} - \langle\hat{t}_{m+i+2}, \hat{s}_{m+i+2}\rangle \\ &= \hat{\psi}(\hat{x}) + \hat{\sigma} - \langle U_{m+i+1}^{\dagger}\hat{s}_{m+i+2}, \hat{s}_{m+i+2}\rangle \\ &= \hat{\psi}(\hat{x}) + \hat{\sigma} - \langle\hat{t}_{m+i+1}, \hat{s}_{m+i+1}\rangle - \frac{(\langle t_{m+i+1}, \hat{s}_{m+i+1}\rangle - c_{m+i+1})^2}{\alpha_{m+i+1}} \\ &= \hat{\alpha}_{m+i+1} - \frac{(\langle t_{m+i+1}, \hat{s}_{m+i+1}\rangle - c_{m+i+1})^2}{\alpha_{m+i+1}} \\ &= -\frac{(\langle t_{m+i+1}, \hat{s}_{m+i+1}\rangle - c_{m+i+1})^2}{\alpha_{m+i+1}}. \end{aligned}\tag{84}$$

Since $\hat{\alpha}_{m+i+2} \geq 0$, eq.(84) implies

$$\hat{\alpha}_{m+i+2} = 0. \tag{85}$$

Next, we show the case where $\alpha_{m+i+1} = 0$. Let a scalar γ_{m+i+1} and an $(m+i)$-dimensional matrix T_{m+i+1} be

$$\gamma_{m+i+1} = 1 + \|t_{m+i+1}\|^2, \tag{86}$$

$$T_{m+i+1} = I_{m+i} - \frac{t_{m+i+1} \otimes \overline{t_{m+i+1}}}{\gamma_{m+i+1}}. \tag{87}$$

It follows from Lemmas 2–4 in Sugiyama and Ogawa (2001) that

$$\begin{aligned} U_{m+i+1}U_{m+i+1}^{\dagger}\hat{s}_{m+i+2} &= \hat{s}_{m+i+2}, && (88)\\ U_{m+i+1} &= \begin{pmatrix} U_{m+i} & s_{m+i+1} \\ s_{m+i+1}^{*} & \psi_{m+i+1}(x_{m+i+1}) + \sigma_{m+i+1} \end{pmatrix}, && (89)\\ U_{m+i+1}^{\dagger} &= \begin{pmatrix} T_{m+i+1}U_{m+i}^{\dagger}T_{m+i+1} & \dfrac{T_{m+i+1}U_{m+i}^{\dagger}t_{m+i+1}}{\gamma_{m+i+1}} \\ \dfrac{(T_{m+i+1}U_{m+i}^{\dagger}t_{m+i+1})^{*}}{\gamma_{m+i+1}} & \dfrac{\langle U_{m+i}^{\dagger}t_{m+i+1}, t_{m+i+1}\rangle}{\gamma_{m+i+1}^{2}} \end{pmatrix}. && (90) \end{aligned}$$

From eqs.(88)–(90), and (81), we have

$$c_{m+i+1} = \langle t_{m+i+1}, \hat{s}_{m+i+1}\rangle. \quad (91)$$

Hence, it follows from eqs.(54), (90), and (81) that

$$\langle \hat{t}_{m+i+2}, \hat{s}_{m+i+2}\rangle = \langle \hat{t}_{m+i+1}, \hat{s}_{m+i+1}\rangle. \quad (92)$$

Substituting eq.(92) into eq.(55), we have

$$\hat{\alpha}_{m+i+2} = \langle \hat{\psi}, \hat{\psi}\rangle + \hat{\sigma} - \langle \hat{t}_{m+i+1}, \hat{s}_{m+i+1}\rangle = \hat{\alpha}_{m+i+1} = 0. \quad (93)$$

Eqs.(85) and (93) prove that $\hat{\alpha}_{m+i+2} = 0$ if $\hat{\alpha}_{m+i+1} = 0$, and hence $\hat{\alpha}_{m+i+1} = 0$ for any positive integer i. ■

(Proof of Lemma 2)
It holds from eqs.(13) and (8) that

$$\begin{aligned} \|E_n f_m - f\|^2 &= \|E_n X_m A_m f + E_n X_m n^{(m)} - f\|^2 \\ &= \|P_{\mathcal{R}(A_m^*)} f - f\|^2 \\ &= \|P_{\mathcal{N}(A_m)} f\|^2, && (94) \end{aligned}$$

which implies eq.(59). Similarly, it holds from eqs.(13) and (9) that

$$\begin{aligned} E_n\|f_m - E_n f_m\|^2 &= E_n\|X_m A_m f + X_m n^{(m)} - E_n X_m A_m f - E_n X_m n^{(m)}\|^2 \\ &= E_n\|X_m n^{(m)}\|^2 \\ &= E_n \mathrm{tr}(X_m \left(n^{(m)} \otimes \overline{n^{(m)}}\right) X_m^*) \\ &= \mathrm{tr}(X_m Q_m X_m^*), && (95) \end{aligned}$$

which implies eq.(60). ■

(Proof of Lemma 3)
It follows from eq.(45) that

$$V_{m+1}' = A_{m+1}^* Q_{m+1}^{-1} A_{m+1}. \quad (96)$$

From eq.(44), we have

$$Q_{m+1}^{-1} = \Gamma_{m+1} Q_m^{-1} \Gamma_{m+1}^* + \frac{e_{m+1}^{(m+1)} \otimes \overline{e_{m+1}^{(m+1)}}}{\sigma_{m+1}}. \tag{97}$$

Since $\Gamma_{m+1}^* e_{m+1}^{(m+1)} = 0$, eqs.(96), (97), and (61) yield eq.(73). ■

(Proof of Lemma 4)
Eqs.(74) and (75) hold from eq.(73) and Theorem 4.6 in Albert (1972). ■

(Proof of Lemma 5)
Since Q_{m+1} is positive definite, it holds that

$$\mathcal{R}(Q_{m+1}) \supset \mathcal{R}(A_{m+1}). \tag{98}$$

Hence, it follows from eq.(43) that

$$X_{m+1} = V'^{\dagger}_{m+1} A_{m+1}^* Q_{m+1}^{-1}. \tag{99}$$

Since $\Gamma_{m+1}^* e_{m+1}^{(m+1)} = 0$, eqs.(99), (61), and (97) yield

$$X_{m+1} = V'^{\dagger}_{m+1} A_m^* Q_m^{-1} \Gamma_{m+1}^* + \frac{V'^{\dagger}_{m+1} \psi_{m+1} \otimes \overline{e_{m+1}^{(m+1)}}}{\sigma_{m+1}}. \tag{100}$$

When $\psi_{m+1} \notin \mathcal{R}(A_m^*)$, it follows from eqs.(74) and (43) that

$$\begin{aligned} V'^{\dagger}_{m+1} A_m^* Q_m^{-1} &= V'^{\dagger}_m A_m^* Q_m^{-1} - \frac{\tilde{\psi}_{m+1} \otimes \overline{Q_m^{-1} A_m V'^{\dagger}_m \psi_{m+1}}}{\tilde{\psi}_{m+1}\left(x_{m+1}\right)} \\ &= X_m - \frac{\tilde{\psi}_{m+1} \otimes \overline{X_m^* \psi_{m+1}}}{\tilde{\psi}_{m+1}\left(x_{m+1}\right)}, \end{aligned} \tag{101}$$

$$\begin{aligned} V'^{\dagger}_{m+1} \psi_{m+1} &= V'^{\dagger}_m \psi_{m+1} + \frac{\sigma_{m+1} + \langle V'^{\dagger}_m \psi_{m+1}, \psi_{m+1} \rangle}{\tilde{\psi}_{m+1}(x_{m+1})} \tilde{\psi}_{m+1} - V'^{\dagger}_m \psi_{m+1} \\ &\quad - \frac{\langle V'^{\dagger}_m \psi_{m+1}, \psi_{m+1} \rangle}{\tilde{\psi}_{m+1}(x_{m+1})} \tilde{\psi}_{m+1} \\ &= \frac{\sigma_{m+1}}{\tilde{\psi}_{m+1}(x_{m+1})} \tilde{\psi}_{m+1}, \end{aligned} \tag{102}$$

since eq.(21) yields $A_m \tilde{\psi}_{m+1} = 0$, From eqs.(100), (101), (102), and (77), we have

$$\begin{aligned} X_{m+1} &= X_m \Gamma_{m+1}^* - \frac{\tilde{\psi}_{m+1} \otimes \overline{\Gamma_{m+1} X_m^* \psi_{m+1}}}{\tilde{\psi}_{m+1}(x_{m+1})} + \frac{\tilde{\psi}_{m+1} \otimes \overline{e_{m+1}^{(m+1)}}}{\tilde{\psi}_{m+1}(x_{m+1})} \\ &= X_m \Gamma_{m+1}^* + \frac{\tilde{\psi}_{m+1} \otimes \overline{h'_{m+1}}}{\tilde{\psi}_{m+1}\left(x_{m+1}\right)}. \end{aligned} \tag{103}$$

Similarly when $\psi_{m+1} \in \mathcal{R}(A_m^*)$, it follows from eqs.(75) and (43) that

$$\begin{aligned}
V_{m+1}'^{\dagger} A_m^* Q_m^{-1} &= V_m'^{\dagger} A_m^* Q_m^{-1} - \frac{V_m'^{\dagger}\psi_{m+1} \otimes \overline{Q_m^{-1} A_m V_m'^{\dagger}\psi_{m+1}}}{\sigma_{m+1} + \langle V_m'^{\dagger}\psi_{m+1}, \psi_{m+1}\rangle} \\
&= X_m - \frac{V_m'^{\dagger}\psi_{m+1} \otimes \overline{X_m^* \psi_{m+1}}}{\sigma_{m+1} + \langle V_m'^{\dagger}\psi_{m+1}, \psi_{m+1}\rangle}, \qquad (104) \\
V_{m+1}'^{\dagger}\psi_{m+1} &= V_m'^{\dagger}\psi_{m+1} - \frac{\langle V_m'^{\dagger}\psi_{m+1}, \psi_{m+1}\rangle V_m'^{\dagger}\psi_{m+1}}{\sigma_{m+1} + \langle V_m'^{\dagger}\psi_{m+1}, \psi_{m+1}\rangle} \\
&= \frac{\sigma_{m+1} V_m'^{\dagger}\psi_{m+1}}{\sigma_{m+1} + \langle V_m'^{\dagger}\psi_{m+1}, \psi_{m+1}\rangle}. \qquad (105)
\end{aligned}$$

From eqs.(100), (104), (105), and (77), we have

$$\begin{aligned}
X_{m+1} &= X_m \Gamma_{m+1}^* - \frac{V_m'^{\dagger}\psi_{m+1} \otimes \overline{\Gamma_{m+1} X_m^* \psi_{m+1}}}{\sigma_{m+1} + \langle V_m'^{\dagger}\psi_{m+1}, \psi_{m+1}\rangle} + \frac{V_m'^{\dagger}\psi_{m+1} \otimes \overline{e_{m+1}^{(m+1)}}}{\sigma_{m+1} + \langle V_m'^{\dagger}\psi_{m+1}, \psi_{m+1}\rangle} \\
&= X_m \Gamma_{m+1}^* + \frac{V_m'^{\dagger}\psi_{m+1} \otimes \overline{h_{m+1}'}}{\sigma_{m+1} + \langle V_m'^{\dagger}\psi_{m+1}, \psi_{m+1}\rangle}.
\end{aligned} \qquad (106)$$

Eqs.(103) and (106) prove Lemma 5. ■

References

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

[2] Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation, 10(2)*, 251–276.

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

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

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

[6] 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.

[7] Jutten, C., & Chentouf, R. (1995). A new scheme for incremental learning. *Neural Processing Letters, 2(1)*, 1–4.

[8] Kadirkamanathan, V., & Niranjan, M. (1993). A function estimation approach to sequential learning with neural networks. *Neural Computation, 5(6)*, 954–975.

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

[10] Molina, C., & Niranjan, M. (1996). Pruning with replacement on limited resource allocating networks by F-projections. *Neural Computation, 8(4)*, 855–868.

[11] Murata, N. (1999). Statistical study on on-line learning. In D. Saad (ed.), *On-line Learning in Neural Networks* (pp. 63–92). Cambridge University Press.

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

[13] 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.

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

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

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

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

[18] Sugiyama, M., & Ogawa, H. (2001). Incremental projection learning for optimal generalization. *Neural Networks 14(1)*, 53–66. (Its latest version is available at '`http://ogawa-www.cs.titech.ac.jp/~sugi/publications/2001/ipl.ps.gz`'.)

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

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

[21] Vijayakumar, S., & Schaal, S. (1998). Local adaptive subspace regression. *Neural Processing Letters, 7(3)*, 139–149.

[22] Vyšniauskas, V., Groen, F. C. A., & Kröse, B. J. A. (1995). Orthogonal incremental learning of a feedforward network. *Proceedings of International Conference on Artificial Neural Networks* (pp. 311–316). Paris, France.

[23] Yamakawa, H., Masumoto, D., Kimoto, T., & Nagata, S. (1993). Active data selection and subsequent revision for sequential learning. *IEICE Technical Report, NC92-99*, 33–40. (in Japanese)

[24] 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)

[25] Yamauchi, K., & Ishii, N. (1995). An incremental learning method with recalling interfered patterns. *Proceedings of IEEE International Conference on Neural Networks ICNN'95, 6* (pp. 3159–3164).

[26] Yingwei, L., Sundararajan, N., & Saratchandran, P. (1997). A sequential learning scheme for function approximation using minimal radial basis function neural networks. *Neural Computation, 9(2)*, 461–478.

[27] Yingwei, L., Sundararajan, N., & Saratchandran, P. (1998). Performance evaluation of a sequential minimal radial basis function neural network learning algorithm. *IEEE Transactions on Neural Networks, 9(2)*, 308–318.

[28] Yoneda, T., Yamanaka, M., & Kakazu, Y. (1992). Study on optimization of grinding conditions using neural networks — a method of additional learning —. *Journal of the Japan Society of Precision Engineering, 58(10)*, 1707–1712. (in Japanese)

[29] Zhang, B. T. (1994). An incremental learning algorithm that optimizes network size and sample size in one trial. *Proceedings of International Conference on Neural Networks* (pp. 215–220). Orlando, USA.