Partial Oblique Projection Learning for Optimal Generalization

LIU Benyong\(^1\)   ZHANG Jing\(^2\)
(1. School of Electronic Engineering, UESTC  Chengdu  610054  China;
2. The Information Center of Sichuan Radio and Television University  Chengdu  610073  China)

Abstract  In practice, it is necessary to implement an incremental and active learning for a learning method. In terms of such implementation, this paper shows that the previously discussed S-L projection learning is inappropriate to constructing a family of projection learning, and proposes a new version called partial oblique projection (POP) learning. In POP learning, a function space is decomposed into two complementary subspaces, so that functions belonging to one of the subspaces can be completely estimated in noiseless case; while in noisy case, the dispersions are set to be the smallest. In addition, a general form of POP learning is presented and the results of a simulation are given.

Key words  supervised learning; generalization; S-L projection learning; partial oblique projection learning

1 Projection Learning and a Family of Projection Learning

1.1  PL, PTPL, and APL

In some viewpoints, supervised learning is discussed in the framework of function approximation, which means that different criteria result in learning methods of different abilities in generalization \[^1\]. From the standpoint of the original space to which the desired function belongs, projection-based criterion aims directly at the generalization ability \[^2\]. The projection concept is applied to projection learning (PL), partial projection learning (PTPL), and averaged projection learning (APL) \[^2~4\]. Based on the common properties of PL, PTPL, and APL, a family of projection learning called S-L projection learning is devised to discuss infinite kinds of learning in a unified way \[^5\]. However, comparing with batch learning, incremental and active learning implementation has more intensive importance in practice \[^4\]. Such work has been done upon PL and APL directly \[^4, 6\]. In terms of such implementation, this paper shows that S-L projection learning is inappropriate to constructing a family of projection learning, and a new one should be designed. Based on an analysis of the application extent of S-L projection learning under novel data, this paper proposes a new definition called partial oblique projection (POP) learning.

In some viewpoints, supervised learning is discussed in the framework of function approximation, which means that different criteria result in learning methods of different abilities in generalization \[^1\]. From the standpoint of the original space to which the desired function belongs, projection-based criterion aims directly at the generalization ability \[^2\]. The projection concept is applied to projection learning (PL), partial projection learning (PTPL), and averaged projection learning (APL) \[^2~4\]. Based on the common properties of PL, PTPL, and APL, a family of projection learning called S-L projection learning is devised to discuss infinite kinds of learning in a unified way \[^5\]. However, comparing with batch learning, incremental and active learning implementation has more intensive importance in practice \[^4\]. Such work has been done upon PL and APL directly \[^4, 6\]. In terms of such implementation, this paper shows that S-L projection learning is inappropriate to constructing a family of projection learning, and a new one should be designed. Based on an analysis of the application extent of S-L projection learning under novel data, this paper proposes a new definition called partial oblique projection (POP) learning.

1.1  PL, PTPL, and APL

In the viewpoint of function approximation, the goal of supervised learning is to obtain an optimal approximation to a desired function \(f(x)\), a point of a reproducing kernel Hilbert space (RKHS) \(H\), by using training data \[^1\]. Let \(A\) be a sampling operator, \(y\) be the teacher vector consisted of the sampled values, and \(n\) be noise contained in \(y\). The previous discussion shows that \[^7\]

\[
y = Af + n
\]

From the standpoint of an inverse problem, a kind of inverse operator \(X\) of \(A\) needs to be found to obtain an optimal approximation \(f_0\) to \(f\) from \(y\), as follows

\[
f_0 = Xy
\]

where \(X\) is called a learning operator, and the process of obtaining \(f_0\) from \(y\) is called supervised learning.

Different criteria lead to learning methods of different abilities. In the subsequent discussion, we briefly review the projection-based criteria aiming directly at generalization ability in supervised learning.

Received  2003-04-08
From Eqs.(1) and (2), the estimated function $f_0$ can be decomposed into two components: the signal component and the noise component, as follows

$$f_0 = XAf + Xn$$  \hspace{1cm} (3)

The common purpose of PL, PTPL, and APL is to minimize the following functional defined by the noise component

$$J_n[X] = E_n \| Xn \|^2$$  \hspace{1cm} (4)

where $E_n$ stands for average over noise ensemble and $\| \cdot \|$ is the norm of $H$. Different constraints result in different definitions of learning.

**Definition 1** An operator $X$ is called a projection leaning operator and denoted by $A^{(PL)}$ if it minimizes $J_n$ in Eq.(4) under the constraint

$$XA = P_{\mathcal{R}(A')} \hspace{1cm} \text{(5)}$$

where $\mathcal{R}(\cdot)$ denotes the range of an operator, $A^*$ is the adjoint operator of $A$, and $P_{\mathcal{R}(A')}^*$ is the orthogonal projection operator onto $\mathcal{R}(A')$. The process of obtaining $f_0$ by an $A^{(PL)}$ is called PL \cite{2}.

**Definition 2** An operator $X$ is called a partial projection learning operator and denoted by $A^{(PTPL)}$ if it minimizes $J_n$ in Eq.(4) under the constraint

$$XAP_s = P_{\mathcal{R}(P_sA')} P_s \hspace{1cm} \text{(6)}$$

where $S$ denotes a subspace to which the concerned function belongs, $P_S$ is the orthogonal projection operator onto $S$, $P_{\mathcal{R}(P_sA')}^*$ is the orthogonal projection operator onto the subspace $\mathcal{R}(P_sA')$. The process of obtaining $f_0$ by an $A^{(PTPL)}$ is called PTPL \cite{3}.

**Definition 3** An operator $X$ is called an averaged projection leaning operator and denoted by $A^{(APL)}$ if it minimizes $J_n$ in Eq.(4) under the constraint minimizing

$$J_{\text{AV}}[X] = E_f \| XAf - f \|^2$$  \hspace{1cm} (7)

where $E_f$ stands for average over $\{ f \}$. The process of obtaining $f_0$ by an $A^{(APL)}$ is called APL, since in this case $XA$ is a projection operator onto the subspace $\mathcal{R}(RA')$ along the subspace $\{ \mathcal{R}(R) \cap N(A) \}$ when the domain of $XA$ is restricted to $\mathcal{R}(R)$, the closure of $\mathcal{R}(R)$ \cite{4}. $N(\cdot)$ denotes the null space of an operator, and $R$ denotes the correlation operator of $\{ f \}$.

### 1.2 A Family of Projection Learning Derived from S-L Projection Learning

It is pointed out that the following common properties are shared by the above three kinds of projection learning: $XA$ is a projection operator if its domain is restricted to a subspace of $H$. $J_n$ in Eq.(4) is minimized under the first condition \cite{5}.

Based on these properties, S-L projection learning is given in the following definition, and a family of projection learning is correspondingly defined \cite{5}.

**Definition 4** An operator $X$ is called an S-L projection learning operator and denoted by $A^{(SL)}$ if it minimizes $J_n$ in Eq.(4) under the constraint

$$XAP_s = P_s \hspace{1cm} \text{(8)}$$

where $S$ is a closed subspace of $H$ to which the concerned target function belongs and it has the following direct sum decomposition:

$$S = L + \{ S \cap N(A) \} \hspace{1cm} \text{(9)}$$

where $P$ is a linear operator, it becomes projection operator onto the subspace $L$ along the subspace $S \cap N(A)$, when it is restricted to $S$. The process of obtaining $f_0$ by an $A^{(SL)}$ is called S-L projection learning. The totality of S-L projection learning for all $S$ and all $L$ is called a family of projection learning derived from $S$ and $L$ \cite{5}.

### 1.3 Application Extent of S-L Projection Learning under Novel Data

When an incremental and active learning is implemented for a learning method, the conditions on which the method is based should be retained to ensure the same learning. As for S-L projection learning, once the subspace $S$ and $L$ are selected, they should remain unchanged under novel data, otherwise it will be changed to other projection learning, say S-L’ projection learning when $L$ is changed to $L$’ by novel data. A necessary condition to ensure the same S-L projection learning is provided by the following theorem, proof of which can be finished by applying Theorems 4.5 and 5.2 of Ref.[8] to Eq.(9).

**Theorem 1** The same S-L projection learning retains for a given subspace $S$, under novel data and different sampling operator $A$, only if the multiplicity for the value $\lambda = 1$ to the following equation remains...
to be a constant
\[ P_\lambda^* P_{\lambda(i)} \phi_i = \lambda \phi_i \] (10)
where \( \phi_i \) is the eigen function corresponds to the eigenvalue \( \lambda \).

Notice that Theorem 1 provides a necessary but not sufficient condition to guarantee that the same S-L projection learning retains under novel data, since the subspace \( L \) of the given space \( S \) can be selected freely, even if the condition is still met under novel data. Therefore Theorem 1 describes the application extent of S-L projection learning under novel data.

2 Partial Oblique Projection (POP) Learning

Since the same S-L projection learning cannot always be retained under novel data, it is necessary to provide a family of projection learning with a novel definition so that incremental and active learning can be implemented. Based on the properties mentioned in Section 1.2, this section proposes a new definition. In our discussion, it is required that \( XA \) becomes an oblique projection operator onto a subspace of \( H \).

Before the new definition is given in detail, it is necessary to briefly introduce the concerning theory about oblique projection operator.

2.1 Oblique Projection Operator

Let \( S \) and \( S_i \) \((i=1,2)\) be closed subspaces of an RKHS \( H \) such that \( S \) has a direct sum decomposition as
\[ S = S_1 + S_2 \] (11)
and \( P_S \) and \( P_i \) are the orthogonal projection operators onto \( S \) and \( S_i \), respectively. Then for any function of \( H \), \( P_S f \) is an element of \( S \) and it can be uniquely decomposed as
\[ P_S f = f_1 + f_2 \] (12)
with \( f_i \) an element of \( S_i \) \((i=1,2)\). The map that transforms \( f \) to \( f_1 \) determines a unique linear idempotent operator \( P_{S_i, S_{ij}} \), called an oblique projection operator, onto \( S_i \) along \( S_{ij} \), with
\[ S_{ij} = S^\perp \oplus S_j \quad i, j = 1, 2 \quad j \neq i \] (13)
where \( \oplus \) denotes the orthogonal direct sum of subspaces. The range and null space of \( P_{S_i, S_{ij}} \) are \( S_i \) and \( S_{ij} \), respectively. From the definition of oblique projection operator, we have
\[ P_S = \sum_{i=1}^2 P_{S_i, S_{ij}} \] (14)
a method to construct the oblique projection operator is provided by the following proposition \([8,9]\).

Proposition The oblique projection operator and the orthogonal projection operators have the following relation
\[ P_{S_i, S_{ij}} = (I - P_{S_j})^{-1} P_i (I - P_i) \] (15)

After the above introduction of the concerning theory about oblique projection operator, the novel definition of a family of projection learning, POP learning, is given in the next subsection.

2.2 Partial Oblique Projection Learning

Assume that the desired target function belongs to the subspace \( S \) defined in Section 2.1, \( S_1 \) is a certain subspace of \( S \), to which in noiseless case the learned result of the target function belongs, \( S_2 \) is the complementary subspace of \( S_1 \) in \( S \). As for the learning operator \( X \) in Eq.(2), it is required that
\[ XA f = P_{S_1, S_{ij}} f \quad f \in S \] (16)
with \( P_{S_1, S_{ij}} \) an oblique projection operator onto \( S_1 \) along \( S_{ij} \), as defined in Section 2.1. In this case, Eq.(16) can be expressed by the following operator equation if the orthogonal projection operator onto the subspace \( S \) is denoted by \( P_S \)
\[ XAP_S = P_{S_1, S_{ij}} P_S \] (17)
notice that the above condition is different from that in PTPL, firstly, \( S_1 \) is a subspace of \( S \), while there is no such a requirement in PTPL; secondly, \( XA \) is required to be only an oblique projection operator, while in PTPL it is required to be an orthogonal projection operator.

From the definition of an oblique projection operator and the assumption that \( S_1 \) is a certain subspace of \( S \), Eq.(17) can be simplified as
\[ XAP_S = P_{S_1, S_{ij}} \] (18)
If we define
\[ A_5 = AP_5 \]  \tag{19} 

then, from Lemma 2 of Ref.[7], it is easy to prove that the existence of a solution \( X \) to Eq.(18) is guaranteed by the following lemma.

**Lemma 1** Eq.(18) has a solution, \( X \), if and only if

\[ N(A_5) \subset S_{[1]} \]  \tag{20} 

The above discussion leads us to the following new definition of a family of projection learning.

**Definition 5** An operator \( X \) is called a partial oblique projection learning operator and denoted by \( A(\text{POP}) \) if it minimizes the functional

\[ J_\lambda[X] = E_\lambda \| Xn \| \]  \tag{21} 

under the constraints

\[ XAP_5 = P_{S_1,S_{[1]}} \]  \tag{22} 

and

\[ N(A_5) = N(AP_5) \subset S_{[1]} = S_2 \oplus S^\perp \]  \tag{23} 

where \( S \) is a closed subspace of an RKHS \( H \) to which a target function belongs, \( P_5 \) is the orthogonal projection operator onto \( S \), \( P_{S_1,S_{[1]}} \) is an oblique projection operator onto a certain subspace \( S_1 \) of \( S \) along \( S_{[1]} \), and \( S_2 \) is the complementary subspace of \( S_1 \) in \( S \). The process of obtaining the approximation to a target function by an \( A(\text{POP}) \) is called POP learning. The totality of POP learning for all such kind of oblique projection operators is called a family of projection learning.

**Remarks:** 1) Notice that Definition 5 retains the traits of Definition 4 that vividly reflects the two properties mentioned in Section 1.2, and at the same time avoided requiring that the subspaces \( S \) and \( S_1 \) be fixed from the start. In other words, the definition always holds once \( P_{S_1,S_{[1]}} \) becomes a suitable oblique projection operator, no matter what concrete \( S \) and \( S_1 \) will be selected or changed into. 2) Eq.(22) provides such kind of direct sum decomposition of the subspace \( S \) that all functions in \( S_1 \) can be completely estimated in noiseless case, while the dispersion is set to be the smallest by minimizing Eq.(21) in noisy case. In this meaning, POP learning provides a learning approach with optimal generalization ability. Eq.(23) guarantees the existence of a POP learning operator. 3) When the target function does not belong to \( S_1 \), it cannot be completely estimated even without noise, but its oblique projection onto \( S_1 \) can be estimated in noiseless case. In noisy case, this projection is estimated with the smallest dispersion. This property has been applied to signal separating from inter-channel interference (ICI)[10].

### 2.3 POP Learning and PL, PTPL, APL, and S-L Projection Learning

Since the sampling operator \( A \) is uniquely decided by the set of inputs, the oblique projection operator \( P_{S_1,S_{[1]}} \) is determined by the subspace \( S \), which can be selected by model selection, and a certain subspace \( S_1 \) of \( S \), which can be obtained by a proper direct sum decomposition of \( S \). If this decomposition is carried out in such a special manner that Eq.(23) takes the extreme case of

\[ S_{[1]} = S_2 \oplus S^\perp = N(A_5) = \{ \Re(A_5') \}^\perp \]  \tag{24} 

which yields

\[ S = S_2 \oplus \Re(A_5') \]  \tag{25} 

then we can obtain

\[ S_2 = S \cap N(A) \]  \tag{26} 

because \( S \) can always be decomposed as

\[ S = \{ S \cap N(A) \} \oplus \Re(A_5') \]  \tag{27} 

an thus

\[ S = S_1 + \{ S \cap N(A) \} \]  \tag{28} 

Eqs.(28) and (9) shows that the subspace \( S_1 \) can be chosen as the subspace \( L \) in S-L projection learning. If a linear operator \( P \) is defined, as in Definition 4, to transform an element in \( S \) to an element in \( L \) parallel to \( \Re(A_5') \), then the oblique projection operator \( P_{S_1,S_{[1]}} \) becomes specially as \( PP_5 \), with \( S_1=L \). In fact, for any element \( f \) of the space \( H \), we have

\[ (PP_5)^2 f = (PP_5)(PP_5^2) f = P(PP_5) f = PP_5 f \]

because \((PP_5 f)\) belongs to \( L \). Therefore

\[ (PP_5)^2 = PP_5 \]  \tag{29} 

In addition, it is easy to show that the range and the null space of \( PP_5 \) are \( L \) and \( S_{[1]} \), respectively [12].
Therefore, \( PP_S \) is an oblique projection operator. And thus S-L projection learning is included in POP learning as a special case. Furthermore, in Ref.[5] it has been shown that S-L projection learning includes PL, PTPL, and APL as special cases. Henceforth, PL, PTPL, APL, and S-L projection learning are all included in POP learning as special cases.

2.4 A General Form of POP Learning

A general form of POP learning is given by the following theorem, proof of which is similar to that of an equivalent form of S-L projection learning[7].

**Theorem 2** A general form of the POP learning operator \( A^{(POP)} \) is given by

\[
A^{(POP)} = P_{S_{i-1}} V^* A^* + Y (I - UU^*)
\]  \hspace{1cm} (30)

where \( Y \) is an arbitrary operator from \( C^m \) to \( H \), \( I \) is the identity operator in \( C^m \), and

\[
U = AP^*_A + Q
\]  \hspace{1cm} (31)

\[
V = A^*_s U^* A_s
\]  \hspace{1cm} (32)

with \( Q \) is the correlation matrix of the noise ensemble. In this case, the minimum value, \( J_{n0} \), of \( J_n \) in Eq.(21), is given by

\[
J_{n0} = -\text{tr}\left(P_{s_i} P^*_{s_i} V^* P^*_{s_{i-1}} P_{s_{i-1}} - \text{tr}\left(P_{s_i} P^*_{s_i} P^*_{s_{i-1}} P_{s_{i-1}}\right)\right)
\]  \hspace{1cm} (33)

and the learned function from the degraded sample value vector \( y \) is given by

\[
f_{POP} = P_{s_i} V^* A^* y
\]  \hspace{1cm} (34)

Eqs.(30) and (33) show that POP learning is a transformed version of PTPL; indeed, the following corollary holds.

**Corollary 1** The learned functions by POP learning and by partial projection learning have the following relation

\[
f_{POP} = P_{s_i} f_{PTP}
\]  \hspace{1cm} (35)

where

\[
f_{PTP} = V^* A^* y = A^{(PTP)} y
\]  \hspace{1cm} (36)

with[2]

\[
A^{(PTP)} = V^* A^* + Y (I - UU^*)
\]  \hspace{1cm} (37)

3 Simulation

In simulation, we consider a learning problem in band-limited signal space, a space with the following reproducing kernel

\[
K(x,x') = \frac{\Omega}{\pi} \sin c \frac{\Omega}{\pi}(x - x')
\]  \hspace{1cm} (38)

with \( \Omega \) the bandwidth of signals and

\[
\sin c(x) = \begin{cases} \frac{\sin(\pi x)}{\pi x} & x \neq 0 \\ 1 & x = 0 \end{cases}
\]  \hspace{1cm} (39)

Let \( \{x_j\}_{j=1}^{512} = \{\pi n / \Omega\}_{n=-512} \), \( \Omega = \pi / 2 \) and correspondingly, \( \{y_j\}_{j=1}^{512} = \{K(x,x')\}_{j=1}^{512} \). Furthermore, let \( H = \text{span}\{\{y_j\}_{j=1}^{512}\} \), \( S = \text{span}\{y_1,y_2,\ldots,y_{15}\} \), and \( S_i = \text{span}\{y_{i-7},y_{i-6},\ldots,y_{i+6}\} \). Two cases are considered for \( S_2 \):

\[
S_2 = \text{span}\{y_1 + \frac{1}{2}y_2 + \frac{1}{2}y_7, y_7 + \frac{1}{2}y_2 + \frac{1}{2}y_1\}
\]  \hspace{1cm} (40)

Assume that a target function \( f \) belonging to the subspace \( S \) has the following coordinate

\[
f = \{0, 0, 7.89, -5.54, -0.13, -5.55, 3.75, 0\}
\]  \hspace{1cm} (41)

that is, in effect \( f \) belongs to \( S_1 \). The sampling points is set as \( \{x_j\}_{j=1}^{512} = \{\pi n / \Omega\}_{n=-512} \) to meet Eq.(23). The adding noise is Gaussian type with zero-mean and 1-variance, therefore the correlation operator \( Q \) over noise ensemble becomes the identity operator \( I \).

Tab.1 lists the coordinates of the learned functions and the variances between the learned functions and the target function.

| \( f \) & \( f_0 \) & Variance |
|---|---|---|
| \( f^{\text{PTPL}} \) | \{0.00,0.00, 6.77, -4.69, -0.43, -4.93, 4.33, 0.00\} | 0.13 |
| \( f^{\text{POP}} \) | \{0.00,0.00, 6.77, -4.69, -0.43, -4.93, 4.33, 0.00\} | 0.06 |
| \{0.00,0.00, 6.77, -3.94, -0.43, -4.93, 4.33, 0.00\} | 0.11 |
Fig. 1 depicts the target function, the teacher signal, and the functions learned by PTPL and POP learning (POPL) with two choices of $S_2$, respectively.

![Target function, teacher signal, and learned functions](image)

4 Conclusions

This paper proposed a novel definition of a family of projection learning, partial oblique projection learning, in which the common properties of the specific methods of projection learning are integrated, and at the same time, the necessity that the subspaces should remain unchanged to guarantee the same category of learning is avoided. A general form of POP learning is provided and a computer simulation is taken.

Acknowledgement

The author is grateful to Professor Hidemitsu Ogawa of Tokyo Institute of Technology (TIT) for helpful discussions on a Family of Projection Learning for Incremental and Active Implementation, the research subject for him during stay in TIT.

References


Brief Introduction to Author(s)

LIU Benyong (刘本永) was born in 1966. He is now an Associate Professor at UESTC. His research interests include: pattern recognition, signal processing, and computational intelligence. E-mail: byliu@uestc.edu.cn

ZHANG Jing (张晶) was born in 1972. Her research interests include computer network and signal processing. E-mail: zh.j@scrtvu.net