Training Data Selection for Optimal Generalization with Noise Variance Reduction in Neural Networks

Sethu Vijayakumar
Lab for Information Synthesis, RIKEN Brain Science Institute
Wako, Saitama 351-0106, Japan. Email: sethu@brain.riken.go.jp

Masashi Sugiyama
Department of Computer Science, Tokyo Institute of Technology
Meguro-ku, Tokyo 152, Japan. Email: sugi@cs.titech.ac.jp

Hidemitsu Ogawa
Department of Computer Science, Tokyo Institute of Technology
Meguro-ku, Tokyo 152, Japan. Email: ogawa@cs.titech.ac.jp

Abstract

In this paper, we discuss the problem of active training data selection in the presence of noise. We formalize the learning problem in neural networks as an inverse problem using a functional analytic framework and use the Averaged Projection criterion as our optimization criterion for learning. Based on the above framework, we look at training data selection from two objectives, namely, improving the generalization ability and secondly, reducing the noise variance in order to achieve better learning results. The final result uses the *apriori* correlation information on noise characteristics and the original function ensemble to devise an efficient sampling scheme, which can be used in conjunction with the incremental learning schemes devised in our earlier work to achieve optimal generalization.

1 Introduction

It has been well known that supervised learning in non-asymptotic cases is highly data dependent. The level of generalization, i.e., the ability to correctly respond to novel inputs, achievable using a fixed number of training data is heavily dependent on the quality of the data used[9]. It is also interesting to note that many natural learning systems are not simply passive but make use of at least some form of active learning to examine the problem domain. By *active* learning, we mean any form of learning in which the learning program has some control over the inputs over which it trains.

The problem of "active learning" has been extensively studied in economic theory and statistics [1]. Optimal data selection within the Bayesian framework for interpolation have been studied by Luttrell[4] and MacKay[5]. It has been shown that a smaller training set gathered by an active learner produces

generalization performance equal to or better than a much larger data set containing redundant examples [7].Plutowski and White [6] assume that a large amount of data has been collected and work on principles of selecting a subset of that data for efficient training; the entire data sets (inputs and outputs) is consulted at each iteration to decide which example to add, an option that is not permitted in this work.

Here, we look at the learning problem from a functional analytic perspective and define an optimization measure which decides on the usefulness of the training data. Works based on the Shannon entropy and Fisher's information criterion [2] already exist. We use the Averaged Projection criterion described in Section 3.1, a criterion which enforces a trade-off between expanding the approximation space and reducing the noise variance. The training data selection scheme developed here works in two phases. At first, a batch selection of m data to optimize generalization is carried as shown in Section 4.1. Since, selection of the data incorporates additive noise, as a second stage, more data is added incrementally with a view to reduce the effect of noise by exploiting the apriori information on the noise correlation matrix as shown in Section 4.2. The effectiveness of this sampling scheme is demonstrated through a simulation.

2 Functional analytic framework for learning

Let us consider a three-layer feedforward neural network whose number of input, hidden, and output units are L, N, and 1, respectively. It can be easily shown that the input-output relationship of such a network is equivalent to a real valued function of L variables. Based on this interpretation, it follows that the learning in Neural Networks(NNs) is analogous to obtaining an optimal approximation f_m to a desired function f from the set of m training data made up of the inputs $x_i \in R^L$ and the corresponding outputs $y_i \in R$:

$$\{(x_i, y_i) | y_i = f(x_i) + n_i : i = 1, \cdots, m\},$$

where n_i is the additive noise. Let a Hilbert space H, with a reproducing kernel $K(x, x')$, represent the space of all functions to be approximated by the NN. Let D be the domain of the functions to be approximated, which is a subset of the L-dimensional Euclidean space R^L. The reproducing kernel $K(x, x')$ is a bivariate function defined on $D \times D$ which satisfies the following two conditions:

1. For any fixed x' in D , $K(x, x')$ is a function in H.

2. For any function f in H and for any x' in D, it holds that

$$(f(x), K(x, x')) = f(x'), \tag{1}$$

where the left hand side of eq.(1), represented by the notation $(\cdot, \cdot)$, denotes the inner product in H.

In the theory of Hilbert space, arguments are developed by regarding a function as a point in that space. Thus, things such as 'value of a function at a point'

cannot be discussed under the general framework of Hilbert space. However, if the Hilbert space has a reproducing kernel, then it is possible to deal with the value of a function at a point. Indeed, if we define functions $\psi_i(x)$ as

$$\psi_i(x) = K(x, x_i) : 1 \leq i \leq m, \tag{2}$$

then, the value of f at a sample point x_i is expressed in Hilbert space language as the inner product of f and ψ_i as

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

Let $\{y_i\}_{i=1}^m$ and $\{n_i\}_{i=1}^m$ form the elements of the m-dimensional vectors $y^{(m)}$ and $n^{(m)}$, respectively. Once the training set $\{x_i\}_{i=1}^m$ is fixed, we can introduce an operator A_m such that

$$y^{(m)} = A_m f + n^{(m)}. \tag{4}$$

The operator A_m, called the *sampling operator*, becomes a linear operator even when we are concerned with nonlinear neural networks. It is expressed by using the Schatten product[1] as

$$A_m = \sum_{i=1}^{m} e_i \otimes \overline{\psi_i}, \tag{5}$$

where $\{e_i\}_{i=1}^m$ is the so-called natural basis [2] in R^m. Now, the learning problem can be reformulated as an *inverse problem* (See Fig.1) of obtaining an operator X_m which provides an optimal approximation f_m of the true function f from the noisy sample values $y^{(m)}$:

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

The generalization ability of the NN, which corresponds to the closeness of the original function f and the approximated function f_m, can be measured using various criterion. In this work, we will restrict ourselves to the Averaged Projection criterion, which will be discussed in the next section. Here, X_m is referred to as the learning operator. Results on obtaining the optimal X_m for a given sampling scheme, both as a batch as well as incremental procedure, will be reviewed in Section 3.2.

In the active learning problem, we have the task of selecting the optimal training data, which is analogous to deciding on the optimal sampling operator A_m under this framework.

3 Optimization criterion for training data selection

As a measure of deciding the usefulness of the training data as well as for obtaining an optimal approximation using the selected training data, we make use of the Averaged Projection criterion, described in the next subsection.

[1] The Schatten product denoted by $(. \otimes .)$ is defined by $(e_i \otimes \overline{\psi_i})f = (f, \psi_i)e_i$.

[2] The vector e_i is the m-dimensional vector consisting of zero elements except the element i equal to 1.

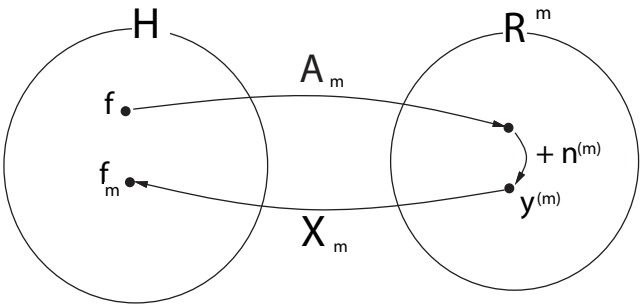

Figure 1: NN learning as an inverse problem

3.1 Averaged Projection criterion

The Averaged Projection optimization corresponds to finding an optimal sampling operator A_m and learning operator X_m such that it minimizes the functional

$$\min_{A_m} J_{AP}^{(0)}[A_m], \tag{7}$$

where

$$J_{AP}^{(0)}[A_m] = \min_{X_m} J_{AP}^{(2)}[X_m] \text{ under the constraint } \min_{X_m} J_{AP}^{(1)}[X_m], \tag{8}$$

$$J_{AP}^{(1)}[X_m] = E_f \|X_m A_m f - f\|^2, \tag{9}$$

$$J_{AP}^{(2)}[X_m] = E_n \|X_m n^{(m)}\|^2. \tag{10}$$

$J_{AP}^{(1)}$ in the functional corresponds to optimizing the generalization ability of the approximated function while $J_{AP}^{(2)}$ corresponds to reducing the effect of noise by reducing the noise variance.

3.2 Optimal learning operator for given sampling scheme

Methods of obtaining the optimal learning operators X_m for a given sampling scheme, i.e. for fixed A_m, have already been obtained based on [3]. This corresponds to minimizing $J_{AP}^{(2)}$ under the constraint of minimizing $J_{AP}^{(1)}$. Here, we provide both the batch and incremental versions of this solution, a result which we will use in optimization of the functional of eq.(7).

3.2.1 Batch solution

Let R represent the correlation operator of the function ensemble, i.e., $R = E_f(f \otimes \overline{f})$ and Q_m represent the correlation matrix of the noise vectors, i.e.,

$Q_m = E_n(n^{(m)} \otimes \overline{n^{(m)}})$. These are apriori information about the learning problem that we assume to possess. Let $\dagger$ be the Moore-Penrose generalized inverse.

Lemma 3.1 *[3] A learning operator $X_m^{(AP)}$ which satisfies the Averaged Projection criterion is given as*

$$X_m^{(AP)} = R^{\frac{1}{2}} V_m^{\dagger} R^{\frac{1}{2}} A_m^* U_m^{\dagger} + Y_m(I_m - U_m U_m^{\dagger}), \tag{11}$$

where Y_m is an arbitrary operator from R^m to H, and

$$U_m = A_m R A_m^* + Q_m, \tag{12}$$
$$V_m = R^{\frac{1}{2}} A_m^* U_m^{\dagger} A_m R^{\frac{1}{2}}. \tag{13}$$

3.2.2 Incremental solution

The batch solution provided in the previous section can be computed in an incremental manner using the learning results of the previous stage and the newly added training data. The batch and the incremental solutions result in exactly the same optimal learning operator X_m after the m iterations. We define a few notations used for this purpose here.

$$T_m = A_m R^{\frac{1}{2}} \tag{14}$$
$$\Gamma_m = \sum\nolimits_{n=1}^{m} (e_n^{(m+1)} \otimes \overline{e_n^{(m)}}) \tag{15}$$
$$\phi_{m+1} = R^{\frac{1}{2}} \psi_{m+1} \tag{16}$$
$$\tilde{\phi}_{m+1} = P_{\mathcal{N}(T_m)} \phi_{m+1} \tag{17}$$
$$\xi_{m+1} = \phi_{m+1} - T_m^* U_m^{\dagger} s_{m+1} \tag{18}$$
$$q_{m+1} = E_n(n_{m+1} n^{(m)}) \tag{19}$$
$$s_{m+1} = T_m \phi_{m+1} + q_{m+1} \tag{20}$$
$$\tau_{m+1} = E_n(n_{m+1}^2) \tag{21}$$
$$\alpha_{m+1} = \|\phi_{m+1}\|^2 + \tau_{m+1} - (U_m^{\dagger} s_{m+1}, s_{m+1}) \tag{22}$$

Lemma 3.2 *By using a prior learning operator and new training data, a posterior learning operator satisfying Averaged Projection criterion can be obtained as follows.*

$$X_{m+1}^{(AP)} = \left(X_m^{(AP)} - \zeta_{m+1} \otimes \overline{U_m^{\dagger} s_{m+1} + U_m^{\dagger} T_m V_m^{\dagger} \xi_{m+1}} \right) \Gamma_m^* + \zeta_{m+1} \otimes \overline{e_{m+1}^{(m+1)}}, \tag{23}$$

where ζ_{m+1} is defined as follows.

(a) When $\alpha_{m+1} > 0$ and $\phi_{m+1} \notin \mathcal{R}(T_m^*)$, then

$$\zeta_{m+1} = \frac{R^{\frac{1}{2}} \tilde{\phi}_{m+1}}{\|\tilde{\phi}_{m+1}\|^2}. \tag{24}$$

(b) When $\alpha_{m+1} > 0$ and $\phi_{m+1} \in \mathcal{R}(T_m^*)$, then

$$\zeta_{m+1} = \frac{R^{\frac{1}{2}} V_m^\dagger \xi_{m+1}}{\alpha_{m+1} + (\xi_{m+1}, V_m^\dagger \xi_{m+1})}. \tag{25}$$

For the remaining conditions, it can be shown that the new training data is redundant from the learning problem perspective and hence, need not be used for training.

4 Active learning

Active learning involves using the apriori information available about the learning problem and devising a sampling scheme to achieve good learning results with limited training data. We will look at the active learning problem from two separate but interrelated objectives. We will first devise a scheme for selecting m training data (batch operation) which will concentrate on improving the generalization ability of the NN by selecting data which extends our projection approximation space in the direction with higher function ensemble probability. Then, in addition to using these m data, we incrementally select the next set of data so as to reduce the noise variance. The second stage of operations is proved to maintain the generalization capability while reducing the effect of additive noise, resulting in a net improvement in the approximation results. Details of each of the stages are provided in the next sections.

4.1 Training data selection for optimal generalization

In this section, we propose a training data selection scheme which selects m data points in a batch operation to optimize the generalization ability, i.e., minimizes the functional

$$\min_{A_m} J_{10}[A_m] = \min_{A_m} E_f \|X_m^{(AP)} A_m f - f\|^2. \tag{26}$$

This corresponds to a modified version of the functional of eq.(7) without including the noise variance minimization criterion. Here, $X_m^{(AP)}$ is the optimal learning operator for a given sampling scheme derived in Section 3.2.1.

4.1.1 Conditions for optimizing generalization capability

Solving the functional of eq.(26),

$$\begin{aligned} J_{10}[A_m] &= \underset{f}{E} \|X_m^{(AP)} A_m f - f\|^2 \\ &= \underset{f}{E}\, tr\{(X_m^{(AP)} A_m - I)(f \otimes \overline{f})(X_m^{(AP)} A_m - I)^*\} \\ &= tr\{(X_m^{(AP)} A_m - I) R (X_m^{(AP)} A_m - I)^*\} \\ &= tr(X_m^{(AP)} A_m R A_m^* X_m^{(AP)*} - X_m^{(AP)} A_m R - R A_m^* X_m^{(AP)*} + R). \end{aligned} \tag{27}$$

For a learning operator $X_m^{(AP)}$ satisfying the Averaged Projection criterion, it can be shown that the following relation holds.

$$X_m^{(AP)} A_m R A_m^* = R A_m^*. \tag{28}$$

Using the relation of eq.(28) in eq.(27), we have

$$J_{10}[A_m] = \{tr(R) - tr(X_m^{(AP)} A_m R)\}. \tag{29}$$

Since, R is a fixed apriori information, we can now consider optimizing an equivalent optimization criterion J_{11}.

$$\min_{A_m} J_{10}[A_m] = \max_{A_m} J_{11}[A_m], \tag{30}$$

where

$$J_{11}[A_m] = tr(X_m^{(AP)} A_m R). \tag{31}$$

Based on [3], it is known that[3]

$$X_m^{(AP)} A_m R^{\frac{1}{2}} = R^{\frac{1}{2}} P_{\mathcal{R}(R^{\frac{1}{2}} A_m^*)} = R^{\frac{1}{2}} P_{\mathcal{R}(T_m^*)}. \tag{32}$$

Using this relation in eq.(31),

$$J_{11}[A_m] = tr(R^{\frac{1}{2}} P_{\mathcal{R}(T_m^*)} R^{\frac{1}{2}}) = (P_{\mathcal{R}(T_m^*)}, R). \tag{33}$$

From now on, we will represent the orthogonal projection onto $\mathcal{R}(T_m^*)$ as P without referring to the subspace of projection. Let $K = dim(\mathcal{R}(T_m^*))$ denote the dimension of the projection space. Using the property of the projection operator P, namely, $PP^* = P$ and $tr(PP^*) = K$, we can convert the variational maximization problem of eq.(33) to a Lagrange maximization problem without constraints :

$$J_{11}[P] = (PR, P) + 2Re(C, P^*P - P) + \lambda[tr(P^*P) - K], \tag{34}$$

where C and λ are the Lagrange multiplier operator and multiplier, respectively. The maximization of the above functional can be done based on the following lemmas.

Lemma 4.1 *[8] The functional represented by eq.(34) is maximized only if*

$$PR = RP. \tag{35}$$

Lemma 4.2 *[8] $PR = RP$ if and only if*

$$R\mathcal{R}(R^{\frac{1}{2}} A_m^*) \subseteq \mathcal{R}(R^{\frac{1}{2}} A_m^*). \tag{36}$$

Lemma 4.3 *[8] $R\mathcal{R}(R^{\frac{1}{2}} A_m^*) \subseteq \mathcal{R}(R^{\frac{1}{2}} A_m^*)$ if and only if $\mathcal{R}(R^{\frac{1}{2}} A_m^*)$ is a Karhunen-Loève subspace of the kernel R.*

[3] P_S refers to the orthogonal projection onto a subspace S.

Lemmas 4.1, 4.2 and 4.3 mean that when the relation $PR = RP$ holds, the subspace $\mathcal{R}(R^{\frac{1}{2}}A^*_{m+1})$ is spanned by the eigenfunctions of R. Let λ_n be the n-th eigenvalue of the correlation operator R arranged in decreasing order and φ_n be the corresponding eigenfunction, i.e.,

$$R\varphi_n = \lambda_n \varphi_n \qquad (\lambda_1 \geq \lambda_2 \geq ...). \tag{37}$$

Then, P can be represented, due to Lemma 4.3, as

$$P = \sum_{n=1}^{K} (\varphi_{m_n} \otimes \overline{\varphi_{m_n}}). \tag{38}$$

where $\{m_n : 1 \leq n \leq K\}$ is a set of indices. Hence, our problem has been reduced to that of obtaining a set of eigenfunctions $\{\varphi_{m_n} : 1 \leq n \leq K\}$ which maximizes eq.(33). Since the functional being maximized, $J_{11}[P]$, is given as

$$J_{11}[P] = (P, R), \tag{39}$$

from eq.(38), we have,

$$J_{11}[P] = (R, P) = \sum_{n=1}^{K} (R\varphi_{m_n}, \varphi_{m_n}) = \sum_{n=1}^{K} \lambda_{m_n}. \tag{40}$$

Since λ_n are arranged in decreasing order, eq.(33) is maximized if and only if we take

$$\{\lambda_{m_n} : 1 \leq n \leq K\} = \{\lambda_n : 1 \leq n \leq K\}. \tag{41}$$

Based on the above analysis, we can write a theorem summarizing the necessary and sufficient condition for selecting the optimal training set, which is analogous to choosing the sampling operator A_m.

Theorem 4.1 *The necessary and sufficient condition for the optimization of the functional for optimal training data selection to maximize generalization capability, represented by eq.(33), is that $\mathcal{R}(R^{\frac{1}{2}}A^*_m)$ is the subspace spanned by $\mathcal{L}\{\varphi_n\}_{n=1}^{K}$ where $K = dim(\mathcal{R}(R^{\frac{1}{2}}A^*_m))$ and φ_n are the eigenfuctions corresponding to the K largest eigenvalues λ_n of the correlation operator R.*

4.1.2 An illustrative artificial example

Let us consider learning in a Hilbert space H spanned by the functions $\{\sin 6x,\ \sin 10x,\ \sin 15x\}$. Let the correlation operator R and the noise correlation matrix Q_2 be given as shown in Table 1-(1). The function to be learned, $f = 9\sin 6x + 4\sin 10x + \sin 15x$, is shown by a solid line in Fig.2. At first we consider selecting two training data from the optimal generalization perspective. For comparison, we look at two sampling schemes (a) and (b) as shown in Table 1-(2). The eigenvalues and eigenfunctions of R in vector representation[4] are

[4] A vector $(a\ b\ c)^T$ denotes a function $a\sin 6x + b\sin 10x + c\sin 15x$.

Table 1: Learning conditions and results : sampling for optimal generalization

(1) Correlation operators and their eigen decompositions

$$R = \begin{pmatrix} 9 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad Q_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Eigenvalue	Eigenfunction[3]
$\lambda_1^{(R)} = 9$	$\varphi_1^{(R)} = (1 \;\; 0 \;\; 0)^T$
$\lambda_2^{(R)} = 4$	$\varphi_2^{(R)} = (0 \;\; 1 \;\; 0)^T$
$\lambda_3^{(R)} = 1$	$\varphi_3^{(R)} = (0 \;\; 0 \;\; 1)^T$

(2) Sampling scheme (a) and (b)

Optimal scheme (a)	Non-optimal scheme (b)
$x_1^{(a)} = \frac{\pi}{5}, \;\; \psi_1^{(a)} = (-0.59 \;\; 0 \;\; 0)^T$	$x_1^{(b)} = \frac{\pi}{2}, \;\; \psi_1^{(b)} = (0 \;\; 0 \;\; -1)^T$
$x_2^{(a)} = \frac{\pi}{3}, \;\; \psi_2^{(a)} = (0 \;\; -0.87 \;\; 0)^T$	$x_2^{(b)} = \frac{\pi}{8}, \;\; \psi_2^{(b)} = (0.71 \;\; -0.71 \;\; -0.38)^T$
$\|f - f_2^{(a)}\|^2 = 3.91$	$\|f - f_2^{(b)}\|^2 = 60.74$

given in Table 1-(1). In Fig.2, the learning result due to sampling scheme (a) is shown by a dashed line ($f_2^{(a)}$) while the result due to sampling scheme (b) is represented by a dotted line ($f_2^{(b)}$). The comparison of the normed generalization error is shown in the bottom half of Table 1-(2). The space spanned by the sampling functions of scheme (a), i.e., $\mathcal{L}(\psi_1^{(a)}, \psi_2^{(a)})$ forms a K-L subspace of R and it is equivalent to the space spanned by the eigenfunctions corresponding to the two largest eigenvalues of R. On the other hand, the space spanned by the sampling functions of scheme (b), i.e., $\mathcal{L}(\psi_1^{(b)}, \psi_2^{(b)})$, does not form a K-L subspace of R. Based on Theorem 4.1, we predict that sampling scheme (a) will provide a better generalization result, a fact that is supported by the results of the simulation (refer Fig.2 and Table 1-(2)).

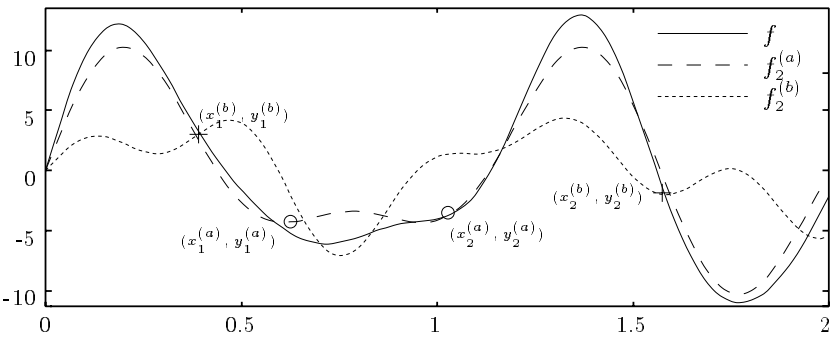

Figure 2: Learning from the generalization perspective: results using optimal (dashed) and non-optimal (dotted) training sets.

4.2 Training data selection for noise variance reduction

In this section, we focus on building upon the results of the previous section and devising a scheme for incrementally finding what training data to add to minimize noise variance of the learned function.

4.2.1 Conditions for maximal noise variance reduction

The criterion to be minimized incrementally for reducing the noise variance at each data selection stage can be written as

$$\min_{x_{m+1}} J_{20}[x_{m+1}], \tag{42}$$

where

$$J_{20} = E_n\|X_{m+1}^{(AP)} n^{(m+1)}\|^2 - E_n\|X_m^{(AP)} n^{(m)}\|^2. \tag{43}$$

By using the incremental solution for $X_{m+1}^{(AP)}$ from Lemma 3.2, J_{20} can be evaluated as follows.

(a) When $\alpha_{m+1} > 0$ and $\phi_{m+1} \notin \mathcal{R}(T_m^*)$, then

$$\begin{aligned} J_{20} &= E_n\big(n_{m+1} - (n^{(m)}, U_m^\dagger T_m V_m^\dagger \xi_{m+1})\big)^2 \frac{\|R^{\frac{1}{2}}\tilde{\phi}_{m+1}\|^2}{\|\tilde{\phi}_{m+1}\|^4} \\ &\geq 0. \end{aligned} \tag{44}$$

(b) When $\alpha_{m+1} > 0$ and $\phi_{m+1} \in \mathcal{R}(T_m^*)$, then

$$\begin{aligned} J_{20} &= -\frac{\|R^{\frac{1}{2}} V_m^\dagger \xi_{m+1}\|^2}{\alpha_{m+1} + (\xi_{m+1}, V_m^\dagger \xi_{m+1})} \\ &\leq 0. \end{aligned} \tag{45}$$

Since the case (a) always results in an increase of the noise variance (as is evident from the sign of eq.(44)), the training data which minimizes eq.(45) is the one which should be chosen for noise variance reduction.

In the case where the noise variance of the new training data is always constant and the noise on the new training data is not correlated with the noise on all training data sampled so far, that is, τ_{m+1} is a constant and $q_{m+1} = 0$, eq.(45) can be reduced to a simple form as

$$J_{20}[x_{m+1}] = -\frac{(B_m u_{m+1}, u_{m+1})}{(u_{m+1}, u_{m+1}) + \tau_{m+1}}, \tag{46}$$

where

$$B_m = Q_m^{\frac{1}{2}} X_m^{(AP)*} X_m^{(AP)} Q_m^{\frac{1}{2}}, \tag{47}$$

$$u_{m+1} = Q_m^{\frac{1}{2}} X_m^{(AP)*} \psi_{m+1}. \tag{48}$$

Table 2: Learning conditions and results : sampling for noise variance reduction

(1) Eigenvalues and eigenvectors of matrix $B_2^{(a)}$

$$B_2^{(a)} = Q_2^{\frac{1}{2}} X_2^{(a)*} X_2^{(a)} Q_2^{\frac{1}{2}} = \begin{pmatrix} 2.89 & 0 \\ 0 & 1.33 \end{pmatrix}$$

Eigenvalue	Eigenvector
$\lambda_1^{(B)} = 2.89$	$\varphi_1^{(B)} = (-1 \quad 0)^T$
$\lambda_2^{(B)} = 1.33$	$\varphi_2^{(B)} = (0 \quad -1)^T$

(2) Sampling scheme (c) and (d)

Optimal scheme (c)	Non-optimal scheme (d)
$x_3^{(c)} = \frac{\pi}{5}$, $u_3^{(c)} = (1 \quad 0)^T$	$x_3^{(d)} = \frac{\pi}{3}$, $u_3^{(d)} = (0 \quad 1)^T$
$\psi_3^{(c)} = (-0.59 \quad 0 \quad 0)^T$	$\psi_3^{(d)} = (0 \quad -0.87 \quad 0)^T$
noise variance reduction$= -2.89$	noise variance reduction$= -1.33$

By using the condition $\tau_{m+1} \geq 0$, we have

$$J_{20}[x_{m+1}] \geq -\frac{(B_m u_{m+1}, u_{m+1})}{(u_{m+1}, u_{m+1})}. \tag{49}$$

This form is known as Rayleigh's quotient. The minimum value of eq.(49) is $-\lambda_1$, where λ_1 is the maximum eigenvalue of B_m. Eq.(49) is minimized if $u_{m+1} = c\ \varphi_1$, where φ_1 is the eigenvector corresponding to λ_1 and c is an non-zero scalar. Thus, when $\tau_{m+1} = 0$, eq.(46) is minimized if ψ_{m+1} satisfies the condition

$$u_{m+1} = Q_m^{\frac{1}{2}} X_m^{(AP)*} \psi_{m+1} = c\ \varphi_1. \tag{50}$$

In this case, the minimum value of eq.(46) is $-\lambda_1$.

4.2.2 Noise variance reduction in the illustrative example

We continue with the artificial learning problem considered in Section 4.1.2 and assume that we have completed training with the two optimal training data $(x_1^{(a)}, y_1^{(a)})$ and $(x_2^{(a)}, y_2^{(a)})$ based on the optimal sampling scheme (a). Now, we consider the problem of selecting an additional training data with the view of reducing noise variance. We can compute $B_2^{(a)}$ corresponding to eq.(47) for the sampling scheme (a) as shown in Table 2-(1). We assume that q_3 and τ_3, corresponding to eqs.(19) and (21), are both zero and independent of sampling location. The eigenvalues and eigenvectors of $B_2^{(a)}$ are shown in Table 2-(1). As a candidate for the next training data x_3, we consider two locations corresponding to schemes (c) and (d), as shown in Table 2-(2). The value of u_3 corresponding to eq.(48) computed for each sampling scheme is shown alongside. Based on our results from Section 4.2, the sampling scheme (c) should provide a greater noise variance reduction since the vector $u_3^{(c)}$ is a scalar multiple of the eigenvector corresponding to the largest eigenvalue of

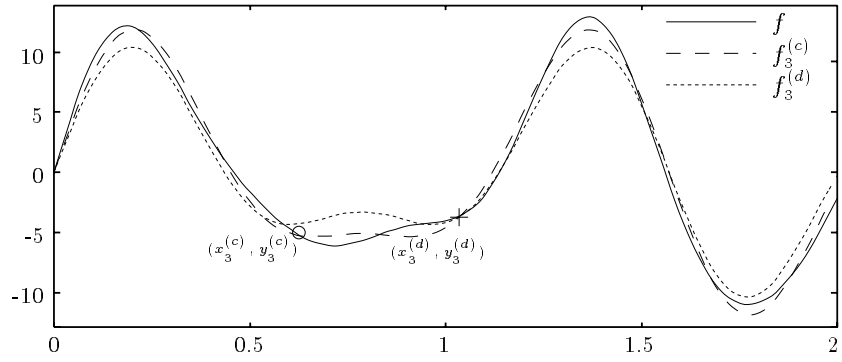

Figure 3: Learning from the noise variance reduction perspective: results using optimal (dashed) and non-optimal (dotted) training sets.

$B_2^{(a)}$ (refer eq.(50)), while scheme (d) does not satisfy this condition. The plot of the learned results in Fig.3 and the noise variance reduction shown in the bottom half of Table 2-(2) substantiate this prediction.

5 Empirical evaluations

We have demonstrated the mechanism by which the active learning scheme shows selectivity in the training data location through a simple artificial example in the previous section. Here, we demonstrate that the technique scales well with the complexity of the problem by considering a learning problem in high dimensional spaces (infinite dimensional original function space).

We consider learning a function within the band-limited Paley-Wiener space $H = \mathcal{L}(\{\frac{10}{\pi}\mathrm{sinc}(\frac{10}{\pi}x - i)\}_{i=-\infty}^{\infty})$, whose reproducing kernel is written as

$$K(x, x_i) = \psi_i(x) = \frac{10}{\pi}\mathrm{sinc}(\frac{10}{\pi}(x - x_i)). \tag{51}$$

To monitor the learning results, we use a decomposition of the generalization error into the noise and the signal component:

$$\begin{aligned} J_{gen} &= E_f E_n \|f_m - f\|^2 && (52)\\ &= E_f \|X_m A_m f - f\|^2 + E_n \|X_m n^{(m)}\|^2 && (53)\\ &= \underbrace{tr\{R - X_m A_m R\}}_{\text{signal component } J_s} + \underbrace{tr\{X_m Q_m X_m^*\}}_{\text{noise component } J_n} && (54) \end{aligned}$$

At first, we look at a task of learning using a set of 40 training points. The result of learning clearly shows that we achieve a better generalization error with the data points selected on the basis of our active learning scheme (see Fig.5) as compared to the results achieved using passive uniform sampling(refer Fig.4). The generalization error for the active sampling works out to be $J_{gen} =$

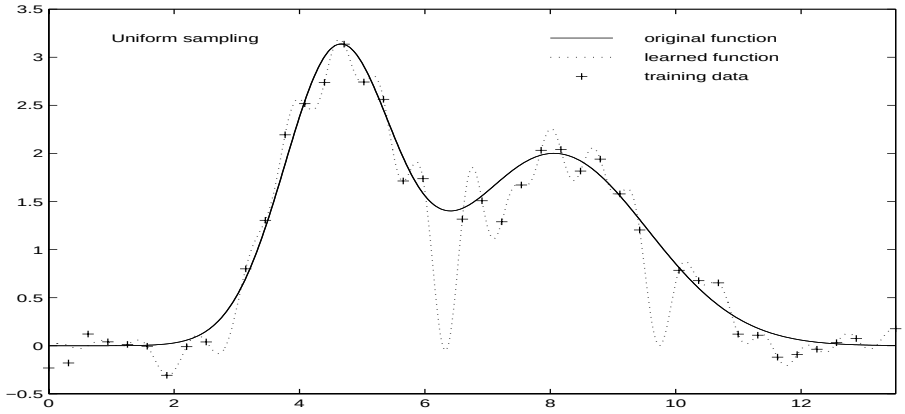

Figure 4: Sampling for improving generalization (Non-optimal)

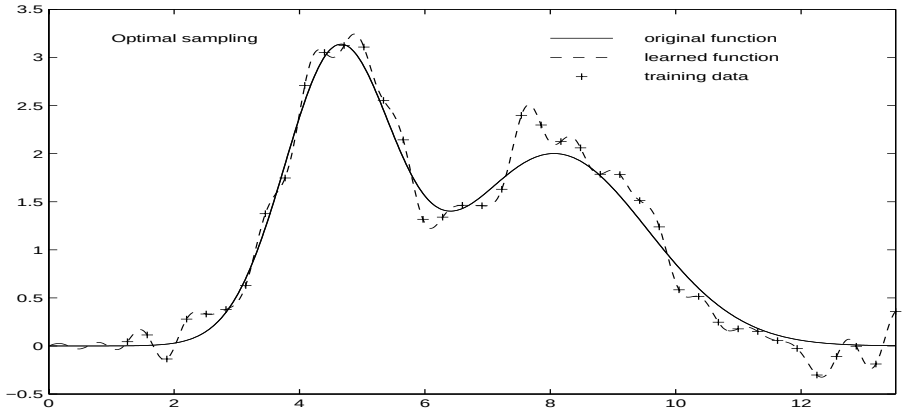

Figure 5: Sampling for improving generalization (Optimal)

$J_s + J_n = 3.18 + 2.36 = \mathbf{5.54}$ as compared to $J_{gen} = J_s + J_n = 12.37 + 2.17 = \mathbf{14.90}$ for the passive uniform sampling.

In the second part of the evaluation, we consider selecting incrementally, an additional 10 data points with an aim of reducing the noise variance while using the 40 optimal training data selected from the first stage. Again, the plots of the learned results, shown in Figs.6 and 7, and analysis of the generalization error ($J_{gen} = J_s + J_n = 3.18 + 1.89 = \mathbf{5.07}$ for the optimal case as against $J_{gen} = J_s + J_n = 3.18 + 2.05 = \mathbf{5.23}$ for the non-optimal case) shows that there is an effective reduction in the noise variance component while maintaining the generalization ability when we use the active sampling scheme devised here.

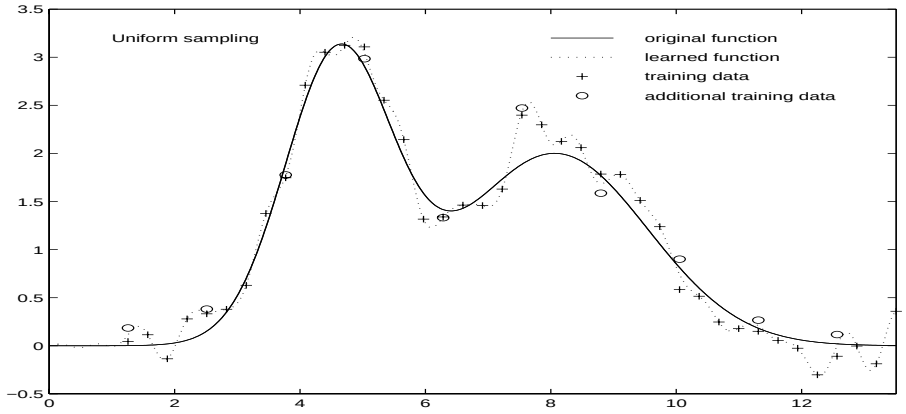

Figure 6: Sampling for noise variance reduction (Non-optimal)

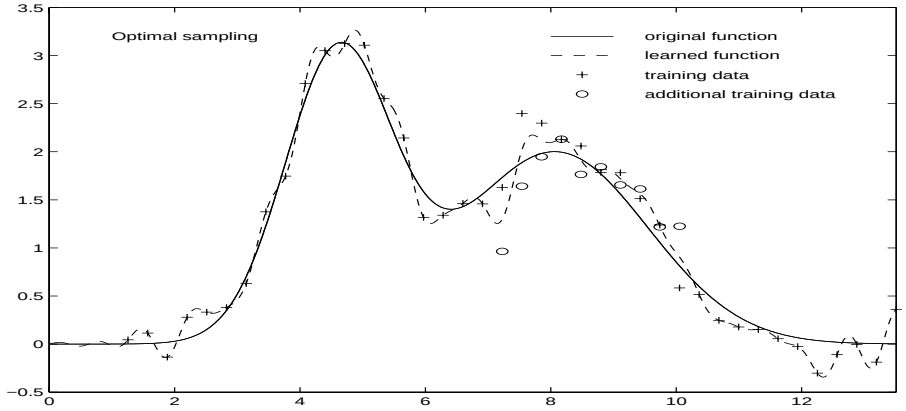

Figure 7: Sampling for noise variance reduction (Optimal)

6 Conclusion

The generalization ability of a learning system in a noisy environment is a delicate balance on how well it can select data to enlarge the approximation space and at the same time, reduce noise variance by redundant sampling. The framework provided here provides an effective mechanism of incorporating apriori information about the function ensemble and the noise correlation matrix to select training data in accordance with the goals of optimal generalization and noise variance reduction.

References

[1] V.V. Federov. *Theory of optimal experiments*. Academic Press, New York, 1972.

[2] K. Fukumizu. Active learning in multilayer perceptrons. In *Advances in Neural Information Processing Systems*, volume 8, pages 295–301. MIT Press, 1996.

[3] Y. Koide, Y. Yamashita, and H. Ogawa. A unified theory of the family of projection filters for signal and image estimation. *Transactions of the IEICE Japan*, J-77 D-II(7):1293–1301, 1994. In Japanese.

[4] S.P. Luttrell. The use of transinformation in the design of data sampling schemes for inverse problems. *Inverse Problems*, 1(1):199–218, 1985.

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

[6] M. Plutowski and H. White. Selecting concise training sets from clean data. *IEEE Transactions on Neural Networks*, 4(2):305–318, 1993.

[7] P. Sollich and D. Saad. Learning from queries for maximum information gain in imperfectly learnable problems. In *Advances in Neural Information Processing System*, volume 7, pages 287–294. MIT Press, 1995.

[8] S. Vijayakumar. *Computational theory of incremental and active leanrning for optimal generalization*. PhD thesis, Tokyo Institute of Technology, 1998.

[9] M. Wann, T. Hediger, and N.N. Greenbaun. The influence of training sets on generalization in feedforward neural networks. In *Proceedings, International Joint Conf. on Neural Networks*, volume 3, pages 137–142, 1990.