PAPER

Improving Generalization Ability through Active Learning

Sethu VIJAYAKUMAR[†] *and* **Hidemitsu OGAWA**[††], *Members*

SUMMARY In this paper, we discuss the problem of active training data selection for improving the generalization capability of a neural network. We look at the learning problem from a function approximation perspective and formalize it as an inverse problem. Based on this framework, we analytically derive a method of choosing a training data set optimized with respect to the Wiener optimization criterion. The final result uses the *apriori* correlation information on the original function ensemble to devise an efficient sampling scheme which, when used in conjunction with the learning scheme described here, is shown to result in optimal generalization. This result is substantiated through a simulated example and a learning problem in high dimensional function space.

key words: *active learning, wiener optimization criterion, generalization, inverse problem, training data selection*

1. Introduction

The level of generalization, i.e., the ability to correctly respond to novel inputs, achievable in supervised learning using a fixed number of training data is heavily dependent on the quality of the data used [14]. 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. In natural systems (such as humans), this phenomenon is exhibited at both high levels (e.g. active examination of objects) and low, subconscious level (e.g. Fernald and Kulh's [3] work on infant reactions to 'Motherese' speech). 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 [11].

The problem of "active learning" has been extensively studied in economic theory and statistics [2]. Optimal data selection within the Bayesian framework for interpolation have been studied by Luttrell [7] and MacKay [8]. Recent studies of active learning in neural network literature can be divided into groups based on the following distinctions: query selection algorithms can be *heuristic* or obtained by optimizing an *objective* function. Moreover, they can either *construct* training queries, i.e. calculate the next optimal training location or *filter* queries from a collection. Heuristic studies as in Baum [1] or Hwang et al.[6], although powerful in specific instances, do not allow a systematic study of improvements of query selection algorithms. We, therefore, restrict our attention to active learning by optimization of objective functions. Plutowski and White [10] 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 set (inputs and outputs) is consulted at each iteration to decide which example to add, an option that is not permitted in this work.

We look at the learning in NNs as an inverse 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 [4] already exist. However, in these approaches, the objective function considered mainly the *information gain* or entropy reduction per training example; moreover, the algorithms were chosen in such a way that they directly reflected the posterior distribution of teachers and hence, were optimally matched to the learning problem at hand. In contrast, here, we use the Wiener objective criterion which directly reflect the *generalization* error. The Wiener criterion described in Sect. 3, enforces a bias towards approximating, with higher precision, the functions with higher probability of occurance as opposed to functions that seldom show up. Apriori knowledge on the function ensemble distribution and other invariances can be easily incorporated into the framework. Section 2 provides the basic framework for dealing with learning from a functional analytic perspective and formalizes it as an inverse problem. Section 3 describes the Wiener optimization criterion, on the basis of which an efficient sampling scheme is formulated in Sect. 4 with the view to achieve maximum generalization ability or equivalently, reduce the number of training examples necessary to attain a certain level of generalization. The effectiveness of this sampling scheme is demonstrated through empirical evaluations in Sect. 5.

Manuscript received March 2, 1998.
Manuscript revised August 29, 1998.

[†] The author is with the Information Synthesis Lab, RIKEN Brain Science Institute, Wako-shi, 351–0106 Japan.

[††] The author is with the Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Tokyo, 152–8550 Japan.

2. Functional Analytic Framework for NN 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 $y_i \in R$:

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

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

$$(f(x), K(x, x')) = f(x'), \tag{2}$$

where the left hand side of Eq. (2), 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

$$\psi_i(x) = K(x, x_i) : 1 \le i \le m, \tag{3}$$

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

$$f(x_i) = (f, \psi_i). \tag{4}$$

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

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

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 as

$$A_m = \sum_{i=1}^{m} e_i \otimes \overline{\psi_i}, \tag{6}$$

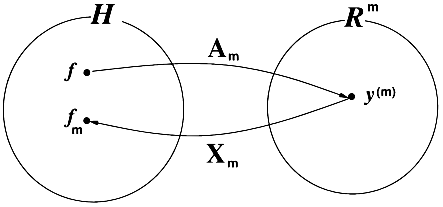

Fig. 1 NN learning as an inverse problem.

where $\{e_i\}_{i=1}^m$ is the so-called natural basis [†] in R^m. The Schatten product denoted by $(. \otimes .)$ is defined by

$$(e_i \otimes \overline{\psi_i}) f = (f, \psi_i) e_i. \tag{7}$$

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 sample value vector $y^{(m)}$:

$$f_m = X_m y^{(m)}. \tag{8}$$

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 criteria. In this work, we will restrict ourselves to the Wiener criterion, which will be discussed in Sect. 3.1. Here, X_m is referred to as the learning operator. Results on obtaining the optimal X_m for a given sampling scheme has been dealt with in our previous work. The results will be reviewed in Sect. 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 Wiener criterion, described in the next subsection.

3.1 Wiener Optimization Criterion

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

$$\min_{X_m, A_m} J_W[X_m, A_m] \tag{9}$$

where

$$J_W[X_m, A_m] = E_f \|X_m A_m f - f\|^2 \tag{10}$$

As can be seen from Eq. (9), actually we ought to

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

Table 1 Examples of RKHS Kernels and the decision surfaces they define.

Kernel Functions	**Approximation Scheme**
$K(x,x') = \frac{\sigma}{\pi} sinc[\frac{\sigma}{\pi}(x-x')]$	Band-limited Paley Wiener space
$K(x,x') = exp(-\|x-x'\|^2)$	Gaussian RBF
$K(x,x') = (1+x*x')^d$	Polynomial of degree d
$K(x,x') = tanh(x*x'-\theta)$	Multi layer perceptrons (Only for some values of θ)
$K(x,x') = B_{2n+1}(x-x')$	B-Splines (B_n: piecewise polynomials of degree n)
$K(x,x') = \frac{\sin(d+1/2)(x-x')}{\sin(x-x')/2}$	Trigonometric polynomials of degree d

minimize the functional by changing both the sampling operator A_m and the learning operator X_m simultaneously. However, here we use a substitute or modified criterion due to ease of solution. The modified criterion is as given below.

$$\min_{A_m} J_W^{(0)}[A_m], \tag{11}$$

where

$$J_W^{(0)}[A_m] = \min_{X_m} J_W[X_m], \tag{12}$$

$$J_W[X_m] = E_f\|X_m A_m f - f\|^2. \tag{13}$$

This involves a two-stage optimization in which we first optimize the learning operator X_m for a fixed sampling scheme and then, use the result of this optimization to minimize a functional involving a variable sampling operator A_m.

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, has already been obtained based on [16]. This corresponds to minimizing the functional $J_W[X_m]$ of Eq. (13) with respect to the learning operator X_m.

Let R represent the correlation operator of the function ensemble, i.e.,

$$R = E_f(f \otimes \overline{f}).$$

This is an apriori information about the learning problem that we assume to possess. This apriori information is analogous to the Bayesian prior, i.e., the covariance matrix in Gaussian processes [15]. If the original function space H is finite dimensional, then, essentially there is no restriction on the type of correlation operator used. In the event that we consider an infinite dimensional H, only such R that has a finite trace is allowed. Then, we have the following lemma.

Lemma 3.1: [16] The necessary and sufficient condition for minimizing the functional of Eq. (13) is that the following relation holds.

$$X_m A_m R A_m^* = R A_m^* \tag{14}$$

Solving Eq. (14) for X_m, we obtain the general solution of the optimal Wiener learning operator $X_m^{(W)}$ for a fixed sampling scheme as shown in the next lemma.

Lemma 3.2: The general form of the optimal Wiener learning operator $X_m^{(W)}$ for a given sampling scheme, i.e., for fixed A_m is given as

$$X_m^{(W)} = R A_m^*(A_m R A_m^*)^\dagger + Y(I - U_m U_m^\dagger), \tag{15}$$

where Y is an arbitrary operator, † refers to the Moore-Penrose generalized inverse and

$$U_m = A_m R A_m^*. \tag{16}$$

This framework might seem very constrained at the outset. However, it can be shown that kernel based approximation, a general form of which has been used here, is a formal way of representing many of the traditional approximation schemes [5]. Of course, the choice of the function space and their corresponding kernels for a given problem should reflect the prior knowledge on data like, for example, the smoothness properties of the desired solution. An important question is: which Hilbert space and kernels correspond to standard approximation schemes used conventionally. Table 1 gives examples of kernel functions corresponding to some RKHS and the type of decision surface they describe, recovering some well known approximation schemes like Gaussian RBF, MLP under constraint etc.

4. Active Learning

Active learning involves using the apriori information available about the learning problem and devising a sampling scheme to achieve good generalization ability with limited training data. We will separate the problems of selecting an optimal search space and selecting optimal training data within that search space. Various approaches to the former problem, corresponding to model selection in this framework, has been dealt upon in Vijayakumar [12] and Vijayakumar & Ogawa [13]. Here, we only deal with the question: how do we efficiently sample the function space to obtain maximum generalization.

4.1 Training Data Selection for Optimal Generalization

Here, we propose a training data selection scheme which selects m optimal data points in a batch operation to optimize the generalization ability, i.e., minimizes the functional of Eq. (11).

To begin with, we rewrite the functional $J_W[X_m]$ of Eq. (13) as

$$\begin{aligned} J_W[X_m] &= \underset{f}{E} \|X_m A_m f - f\|^2 \\ &= \underset{f}{E}\, tr\{(X_m A_m - I)(f \otimes \overline{f})(X_m A_m - I)^*\} \\ &= tr\{(X_m A_m - I)R(X_m A_m - I)^*\} \\ &= tr(X_m A_m R A_m^* X_m^* - X_m A_m R \\ &\quad - R A_m^* X_m^* + R) \end{aligned} \tag{17}$$

Based on Eq. (17) and Lemma 3.1, the functional $J_W^{(0)}[A_m]$ of Eq. (12) can be written as

$$\begin{aligned} J_W^{(0)}[A_m] &= tr(X_m^{(W)} A_m R A_m^* X_m^{(W)^*} - X_m^{(W)} A_m R \\ &\quad - R A_m^* X_m^{(W)^*} + R) \\ &= tr(R A_m^* X_m^{(W)^*} - X_m^{(W)} A_m R \\ &\quad - R A_m^* X_m^{(W)^*} + R) \\ &= tr(R) - tr(X_m^{(W)} A_m R). \end{aligned} \tag{18}$$

Since R is a fixed apriori information specific to the learning problem being solved, $tr(R)$ can be considered as a fixed quantity. Therefore, we can convert the problem of minimizing the functional described in Eq. (11) to the following equivalent maximization problem:

$$\min_{A_m} J_W^{(0)}[A_m] = \max_{A_m} J_W^{'}[A_m], \tag{19}$$

where

$$J_W^{'}[A_m] = tr(X_m^{(W)} A_m R). \tag{20}$$

Using the general form of solution of $X_m^{(W)}$ given in Eq. (15) with $Y = 0$ in the functional of Eq. (20), we have

$$\begin{aligned} J_W^{'}[A_m] &= tr(X_m^{(W)} A_m R) \\ &= tr\{R A_m^* (A_m R A_m^*)^\dagger A_m R\} \\ &= (R A_m^* (A_m R A_m^*)^\dagger A_m R, I) \\ &= ((A_m R A_m^*)^\dagger A_m R, A_m R). \end{aligned} \tag{21}$$

Let us substitute $T = A_m R^{\frac{1}{2}}$ in the above functional. Then, we have

$$\begin{aligned} J_W^{'}[A_m] &= ((TT^*)^\dagger T R^{\frac{1}{2}}, T R^{\frac{1}{2}}) \\ &= (T^*(TT^*)^\dagger T, R) \\ &= (T^\dagger T, R) \end{aligned}$$

$$J_W^{'}[A_m] = (P_{\mathcal{R}(T^*)}, R), \tag{22}$$

where $P_{\mathcal{R}(T^*)}$ is the orthogonal projection operator onto the range of $\mathcal{R}(R^{\frac{1}{2}} A_m^*)$.

From now on, in the intermediate equations, we will represent this projection operator as P without specifying the subspace of projection. Now, the functional of Eq. (22) can be written as

$$J_W^{'}[P] = (P, R) = (PR, P) \tag{23}$$

based on the property of orthogonal projection operators, i.e., $P^* = P$ and $P^2 = P$. Again, based on these properties, we have

$$P^* P = P. \tag{24}$$

Furthermore, if we assume the dimensionality of the subspace $R^{\frac{1}{2}} A_m^*$ to be equal to K, then, we have

$$tr(P^* P) = K. \tag{25}$$

Hence, our optimization problem of Eq. (19) can be restated as **"Obtain an operator P which maximizes Eq. (23) under constraints of Eqs. (24) and (25)."**

Let C and λ be the Lagrange multiplier operator and the Lagrange multiplier, respectively. Then, the above variational problem with constraints is reduced to the following variational problem without constraints with respect to P, C and λ:

$$\begin{aligned} J_W^{'}[P, C, \lambda] &= (PR, P) + 2(C, P^* P - P) \\ &\quad + \lambda[tr(P^* P) - K]. \end{aligned} \tag{26}$$

Since we have to maximize this functional, equating the partial derivative of $J_W^{'}$ in Eq. (26) with respect to P to zero yields

$$(P(R + C + C^* + \lambda P) - C, \delta P) = 0, \tag{27}$$

which in turn gives the relationship

$$P(R + C + C^* + \lambda P) = C. \tag{28}$$

Using the properties of the projection operator P and the fact that $R^* = R$, we can prove the following lemma, as shown in Appendix A.

Lemma 4.1: The functional represented by Eq. (26) is maximized only if

$$PR = RP. \tag{29}$$

We can convert the result of the above lemma into more useful form through the following results. The proof of these, on the lines of analysis in [9], is fairly straightforward.

Lemma 4.2: [12] $PR = RP$ if and only if

$$R\mathcal{R}(R^{\frac{1}{2}} A_m^*) \subseteq \mathcal{R}(R^{\frac{1}{2}} A_m^*). \tag{30}$$

Lemma 4.3: [12] $R\mathcal{R}(R^{\frac{1}{2}}A_m^*) \subseteq \mathcal{R}(R^{\frac{1}{2}}A_m^*)$ if and only if $\mathcal{R}(R^{\frac{1}{2}}A_m^*)$ is a Karhunen-Loéve subspace of the kernel R.

Lemma 4.2 and Lemma 4.3 mean that when the relation $PR = RP$ holds, the subspace $\mathcal{R}(R^{\frac{1}{2}}A_m^*)$ 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 the corresponding eigenfunction with unit norm, i.e.,

$$R\varphi_n = \lambda_n\varphi_n \qquad (\lambda_1 \geq \lambda_2 \geq \ldots). \tag{31}$$

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

$$P = \sum_{n=1}^{K}(\varphi_{m_n} \otimes \overline{\varphi_{m_n}}). \tag{32}$$

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. (23). Since the functional being maximized, $J_W^{'}[P]$, is given as

$$J_W^{'}[P] = (P, R), \tag{33}$$

from Eq. (32), we have,

$$\begin{aligned} J_W^{'}[P] &= (R, P) && (34)\\ &= \sum_{n=1}^{K}(R\varphi_{m_n}, \varphi_{m_n}) && (35)\\ &= \sum_{n=1}^{K}\lambda_{m_n}. && (36) \end{aligned}$$

Since λ_n are arranged in decreasing order, Eq. (23) is maximized if and only if we take

$$\{\lambda_{m_n} : 1 \leq n \leq K\} = \{\lambda_n : 1 \leq n \leq K\}. \tag{37}$$

Hence, to maximize the functional and select the optimal set of training data, we should use a sampling operator A_m such that $\mathcal{R}(R^{\frac{1}{2}}A_m^*)$ is the subspace spanned by $\mathcal{L}(\{\varphi_n\}_{n=1}^K)$, where φ_n are the eigenfunctions corresponding to the K largest eigenvalues λ_n of the correlation operator R.

Based on the above analysis, we can write a theorem summarizing the necessary and sufficient condition for selecting the optimal training set.

Theorem 4.1 (Optimal selection of training data): The necessary and sufficient condition for the optimization of the functional for optimal training data selection, represented by Eq. (23) 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 eigenfunctions corresponding to the K largest eigenvalues λ_n of the correlation operator R.

4.2 An Alternate Interpretation of the Optimal Sampling Result

In this section, we will further analyze the implications of the Theorem 4.1 on optimal training data selection. Since the sampling operator A_m is given by Eq. (6), we have

$$\mathcal{R}(A_m^*) = \mathcal{L}(\{\psi_k\}_{k=1}^m) \tag{38}$$

and

$$\mathcal{R}(R^{\frac{1}{2}}A_m^*) = \mathcal{L}(\{R^{\frac{1}{2}}\psi_k\}_{k=1}^m). \tag{39}$$

Also based on the eigenvalue decomposition of R, we can express the correlation operator and it's root as

$$\begin{aligned} R &= \sum_n \lambda_n(\varphi_n \otimes \overline{\varphi_n}),\\ R^{\frac{1}{2}} &= \sum_n \sqrt{\lambda_n}(\varphi_n \otimes \overline{\varphi_n}). && (40) \end{aligned}$$

Hence, using the result of Eq. (40), we can show that

$$\begin{aligned} R^{\frac{1}{2}}\psi_k(x) &= \sum_n \sqrt{\lambda_n}(\varphi_n \otimes \overline{\varphi_n})\psi_k(x)\\ &= \sum_n \sqrt{\lambda_n}(\varphi_n, \psi_k)\varphi_n(x)\\ &= \sum_n \sqrt{\lambda_n}\varphi_n(x_k)\varphi_n(x). && (41) \end{aligned}$$

Now, turning our attention to the condition to be satisfied for optimal training data selection, we know from Theorem 4.1 that the necessary and sufficient condition for optimal sampling is that the following relation holds:

$$\mathcal{R}(R^{\frac{1}{2}}A_m^*) = \mathcal{L}(\{\varphi_n\}_{n=1}^K), \tag{42}$$

where $K = dim(\mathcal{R}(R^{\frac{1}{2}}A_m^*))$ and $\mathcal{L}(\{\varphi_n\}_{n=1}^K)$ is the maximum variance subspace of the correlation operator. Based on Eq. (39), we can write the relation of Eq. (42) as

$$\mathcal{L}(\{R^{\frac{1}{2}}\psi_k\}_{k=1}^m) = \mathcal{L}(\{\varphi_n\}_{n=1}^K), \tag{43}$$

which, due to Eq. (41), is equivalent to

$$\mathcal{L}\left(\left\{\sum_n \sqrt{\lambda_n}\varphi_n(x_k)\varphi_n\right\}_{k=1}^m\right) = \mathcal{L}(\{\varphi_n\}_{n=1}^K). \tag{44}$$

Whether the condition specified in Eq. (44) can always be satisfied will depend on the nature of the correlation operator and the function space being used. A more general solution can be obtained by changing the step where we substituted the minimization of the functional of Eq. (22) with Eq. (23). The point to be noted here is that we solved for a general projection operator P whereas in the original function, the projection

Table 2 Learning conditions and results : sampling for optimal generalization.

(1) Eigenvalues and eigenfunctions of R

Eigen value	Vector expression of eigenfunctions[†]
$\lambda_1^{(R)} = 9$	$\varphi_1^{(R)} = (1\ \ 0\ \ 0)^T$
$\lambda_2^{(R)} = 4$	$\varphi_2^{(R)} = (0\ \ 1\ \ 0)^T$
$\lambda_3^{(R)} = 1$	$\varphi_3^{(R)} = (0\ \ 0\ \ 1)^T$

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

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

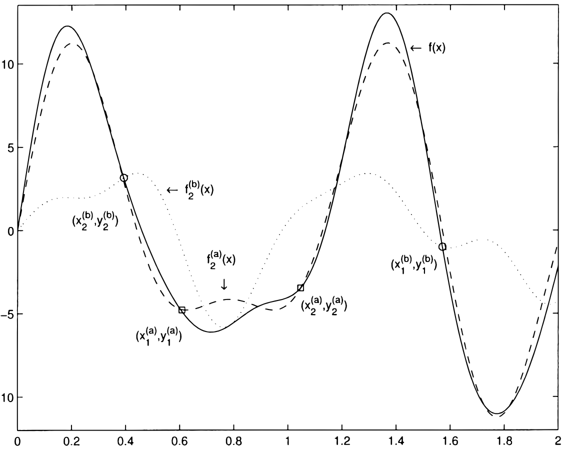

Fig. 2 Learning results under two different sampling scheme: Optimal (dashed) and non-optimal (dotted).

operator was a projection onto a particular subspace $\mathcal{R}(R^{\frac{1}{2}}A_m^*)$.

At this stage, for many function spaces and objective functions the active learning scheme is more like a *filter* (differentiates between an optimal set and a non-optimal set of training data) since the closed form equations of the central theorem can be analytically solved to derive the optimal samplng scheme only for particular function spaces. However, ongoing research in the field aims at generalizing this to the *construct* kind of querying scheme for all objective functions and functions spaces.

5. Empirical Evaluations

5.1 An Illustrative Artificial Example

In this section, we demonstrate the effectiveness of the sampling scheme designed here using an 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 be given as

$$R = \begin{pmatrix} 9 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{45}$$

The function to be learned, $f = 9\sin 6x + 4\sin 10x + \sin 15x$, is shown by a solid line in Fig. 2. We consider selecting a set of two training data from an optimal generalization perspective. For comparison, we look at two sampling schemes (a) using training set $\{x_k^{(a)}, y_k^{(a)}\}_{k=1}^2$ and (b) using training set $\{x_k^{(b)}, y_k^{(b)}\}_{k=1}^2$ as shown in Table 2-(2).

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 2-(2). The eigenvalues and eigenfunctions of R in vector representation[†] are given in Table 2-(1). 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, i.e., $\mathcal{L}(\varphi_1^{(R)}, \varphi_2^{(R)})$ (refer Table 2). 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 the normed generalization error shown at the bottom of Table 2-(2)).

5.2 Active Learning in High Dimensional Space

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 a real world problem of approximating the contours of a macroscopic surface in the production line from a finite number of ultrasonically sampled heights. In these problems, we usually have some apriori knowledge about the smoothness levels of the contours[††] and the probability of distribution of peaks, i.e., the contours are likely to be more peaked at the center than at the edges etc. A sample contour is shown using a bold line in Fig. 3.

We consider learning this contour by using a band-limited Paley-Wiener space $H = \mathcal{L}(\{\frac{\sigma}{\pi}\mathrm{sinc}(\frac{\sigma}{\pi}x -$

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

[††] Strictly speaking, we need only to know a rough upper bound on the frequency content of the resulting contours.

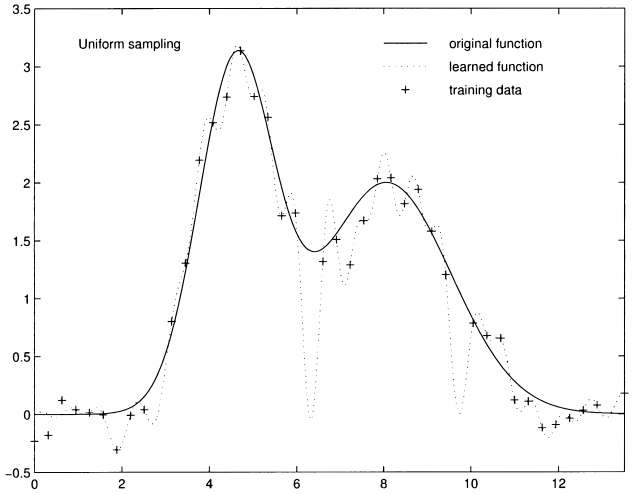

Fig. 3 Passive uniform sampling of contour data.

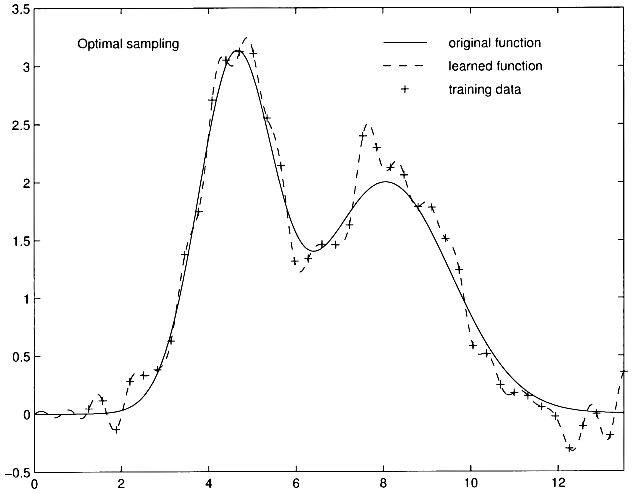

Fig. 4 Optimal sampling of contour data based on active learning scheme.

$i)\}_{i=-\infty}^{\infty})$, where σ represents our apriori knowledge on the frequency content of the function space implying the Fourier transform $F(\omega)$ of the function $f(x)$ is zero for $|\omega| > \sigma$. If we define the inner product in this Hilbert space as

$$(f, g) = \int_{-\infty}^{+\infty} f(x)g(x)dx, \tag{46}$$

the sampling functions, which are instantiations of the reproducing kernel at the training data points are written as

$$K(x, x_i) = \psi_i(x) = \frac{\sigma}{\pi}\mathrm{sinc}\left(\frac{\sigma}{\pi}(x - x_i)\right). \tag{47}$$

In this problem, to take the sample noise into account, we slightly modify our problem formulation. The training data labels are now defined as:

$$\{(x_i, y_i)|y_i = f(x_i) + n_i : i = 1, \cdots, m\}. \tag{48}$$

where n_i is the additive noise.

In this setting, we use a decomposition of the generalization error into the noise and the signal component in order to monitor the learning results:

$$\begin{aligned} J_{gen} &= E_f E_n \|f_m - f\|^2 && (49)\\ &= E_f \|X_m A_m f - f\|^2 + E_n \|X_m n^{(m)}\|^2 && (50)\\ &= \underbrace{tr\{(X_m A_m - I)R(X_m A_m - I)^*\}}_{\text{signal component } J_s} \\ &\quad + \underbrace{tr\{X_m Q X_m^*\}}_{\text{noise component } J_n}, \end{aligned}$$

where $Q_m = E_n(n^{(m)} \otimes n^{(m)})$ represents the correlation matrix of the noise vectors. Selection of training data to reduce the signal component error, i.e. the first half of the R.H.S. of Eq. (50), corresponds to the active data selection method devised in this work.

We look at a task of approximating the contour using a set of 40 training points. We choose a function space with a band limitation corresponding to $\sigma = 10$, which has been found to be suitable in these problems based on the smoothness properties and SNR of the sampling scheme. Although we use an infinite dimensional function space, we are able to compute the apriori information about the correlation operator R by restricting the range of input values x_i. Apriori that peaks are most likely in the center are reflected by higher weighting of the *sinc* functions in these locations which are translated and summed to form the correlation operator R.

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. 4) as compared to the results achieved using passive uniform sampling (refer Fig. 3). This observation is supported by theoretical computations of the generalization error for the active sampling which works out to be $J_{gen} = J_s + J_n = 3.18 + 2.36 = \mathbf{5.54}$ as compared to $J_{gen} = J_s + J_n = 12.37 + 2.17 = \mathbf{14.90}$ for the passive uniform sampling.

6. Conclusion

The generalization ability of a learning system depends not only on the learning methods used but also on the quality of the training data used. The framework developed here provides an effective mechanism of incorporating apriori information about the function ensemble to select training data, which in conjunction with the optimal learning framework developed in our previous work, is shown to ensure optimal generalization ability.

References

[1] E. Baum, "Neural net algorithms that can learn in polynomial time from examples and queries," IEEE Trans. Neural

Networks, vol.2, no.1, pp.5–19, 1991.

[2] V.V. Federov, "Theory of optimal experiments," Academic Press, New York, 1972.

[3] A. Fernald and P. Kuhl, "Acoustic determinants of infant preferences for Motherese speech," Infant Behavior and Development, vol.10, pp.279–293, 1987.

[4] K. Fukumizu, "Active learning in multilayer perceptrons," Advances in Neural Information Processing Systems, vol.8, pp.295–301, MIT Press, 1996.

[5] F. Girosi, "An equivalence between sparse approximation and support vector machines," Neural Computation, vol.10, no.6, pp.1445–1454, 1988.

[6] J-N. Hwang, J.J. Choi, S. Oh, and R.J. Marks, "Query based learning applied to partially trained multilayer perceptrons," IEEE Trans. Neural Networks, vol.2, no.1, pp.131–136, 1991.

[7] S.P. Luttrell, "The use of transinformation in the design of data sampling schemes for inverse problems," Inverse Problems, vol.1, no.1, pp.199–218, 1985.

[8] D. Mackay, "Information-based objective functions for active data selection," Neural Computation, vol.4, no.4, pp.590–604, 1992.

[9] H. Ogawa, "Karhunen-Loéve subspace," Proc. 11th International Conference on Pattern Recognition (The Netherlands), vol.2, pp.75–78, 1992.

[10] M. Plutowski and H. White, "Selecting concise training sets from clean data," IEEE Trans. Neural Networks, vol.4, no.2, pp.305–318, 1993.

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

[12] S. Vijayakumar, "Computational theory of incremental and active learning for optimal generalization," Ph.D. thesis, Tokyo Institute of Technology, 1998.

[13] S. Vijayakumar and H. Ogawa, "RKHS based functional analysis for exact incremental learning," Neurocomputing, 1999. (in Press).

[14] M. Wann, T. Hediger, and N.N. Greenbaun, "The influence of training sets on generalization in feedforward neural networks," Proc. International Joint Conf. on Neural Networks, vol.3, pp.137–142, 1990.

[15] C.K.I. Williams, "Prediction with Gaussian processes: From linear regression to linear prediction and beyond," In Learning and Inference in Graphical Models, ed. M.I. Jordan, Kluwer Academic Press, 1998.

[16] Y. Yamashita and H. Ogawa, "Mutual relations among optimum image restoration filters," IEICE Trans., vol.J75-D-II, no.5, pp.890–898, 1992.

Appendix A: Proof of Lemma 4.1

The maximization of the functional Eq. (26) is equivalent to solving for P in Eq. (28),

$$P(R+C+C^*+\lambda P)=C. \tag{A$\cdot$1}$$

Multiplying Eq. (A·1) with P from the left, we get the relation

$$PR+PC^*+\lambda P=0. \tag{A$\cdot$2}$$

Taking conjugate of Eq. (A·2), we have

$$RP+CP+\lambda P=0. \tag{A$\cdot$3}$$

Multiplying Eq. (A·1) with P from the right, we get the relation

$$P(R+C+C^*+\lambda P)P=CP, \tag{A$\cdot$4}$$

which in turn implies that

$$CP=(CP)^*=PC^*. \tag{A$\cdot$5}$$

Combining Eqs. (A·2), (A·3), and (A·5), we have

$$PR=RP,$$

which completes the proof.

Sethu Vijayakumar was born in 1970 in Kerala, India. He received the B.E.(Comp.Sc.) degree from the Regional Engineering College, Tiruchirapalli, India in 1992 and the M.E. and Ph.D. degrees from the Tokyo Institute of Technology, Japan in 1995 and 1998, respectively. He is currently a researcher with the Information Synthesis Laboratory of the RIKEN Brain Science Institute, Japan and holds a part time affiliation with the CLMC Lab, USC, California. His research interests includes statistical and machine learning, neural networks and computational neuroscience. He received the ICNN'95 Best Student Paper Award in 1995, the IEEE Vincent Bendix Award in 1991 and the IEEE R.K. Wilson RAB Award in 1996. Dr. Vijayakumar is also a member of the International Neural Network Society, and the IEEE.

Hidemitsu Ogawa was born in 1942 in Hiroshima Prefecture, Japan. He received the B.E. and D.E. degrees from the Tokyo Institute of Technology (TIT), Japan, in 1965 and 1977, respectively. From 1965 to 1972 he was with the Electrotechnical Laboratory. In 1972 he joined the faculty of the TIT, where he is now a professor of the Department of Computer Science, Graduate School of Information Science and Engineering. During the academic year 1984–1985 he was a visiting professor of the Department of Technical Physics at the Helsinki University of Technology, Finland. His research interests include pattern recognition, neural networks, and image processing. He received the Yonezawa Memorial Award in 1969 and Paper Awards in 1976, 1985, 1990, 1993, and 1994, respectively, from the Institute of Electronics, Information and Communication Engineers of Japan. Dr. Ogawa is also a member of the Information Processing Society of Japan, the Japanese Society for Artificial Intelligence, the American Mathematical Society, the International Neural Network Society, and the IEEE.