Pseudo Orthogonal Bases Give the Optimal Generalization Capability in Neural Network Learning

Masashi Sugiyama[a] and Hidemitsu Ogawa[a]

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

ABSTRACT

Pseudo orthogonal bases are a certain type of frames proposed in the engineering field, whose concept is equivalent to a tight frame with frame bound 1 in the frame terminology. This paper shows that pseudo orthogonal bases play an essential role in neural network learning. One of the most important issues in neural network learning is "what training data provides the optimal generalization capability?", which is referred to as active learning in the neural network community. We derive a necessary and sufficient condition of training data to provide the optimal generalization capability in the trigonometric polynomial space, where the concept of pseudo orthogonal bases is essential. By utilizing useful properties of pseudo orthogonal bases, we clarify the mechanism of achieving the optimal generalization.

Keywords: frame, pseudo orthogonal basis (POB), pseudo orthonormal basis (PONB), neural network, active learning, generalization capability, trigonometric polynomial space.

1. INTRODUCTION

Pseudo orthogonal bases (POBs) are a certain type of *frames*. This paper shows that POBs play an essential role when we work on *active learning* in neural networks (NNs).

The concept of frames was proposed by Duffin and Schaeffer[1] in 1952 in terms of *non harmonic Fourier series*. After a long period, Young's book[2] in 1980 again illuminated the concept of frames, also from the viewpoint of non harmonic Fourier series. Then, the concept attracted a great deal of attention and it has been continually gathering attention to the present time. In the engineering field, Ogawa and Iijima[3,4] presented the concept of POBs in 1973, independently of the Duffin and Schaeffer's work. A POB is a *tight frame with frame bound 1* in the frame terminology. Ogawa[5,6] extended POBs to *pseudo biorthogonal bases* (PBOBs) in 1978, and devoted himself to showing their properties in detail.[7,8] So far, PBOBs have been applied to many problems such as signal restoration,[9] computerized tomography,[10] and NN learning.[11] Especially in NN learning, PBOBs play a major role when we discuss the optimal generalization capability and the robustness of NNs.

One of the most important issues in NN learning is "what training data provides the optimal generalization capability?", which is referred to as *active learning* in the NN community. Active learning has been extensively studied in the fields of mathematical statistics,[12,13] machine learning,[14] and computational learning theory[15] as well as in the field of neural networks.[16,17] However, most of the studies do not directly aim for the optimal generalization. In this paper, we give a new method of active learning which provides exactly the optimal generalization capability in the trigonometric polynomial space, where the concept of POBs is crucial. By utilizing useful properties of POBs, we clarify the mechanism of achieving the optimal generalization capability.

2. FORMULATION OF NN LEARNING

In this section, the NN learning problem is formulated from the functional analytic point of view (see Ref. 18,11). Then, our learning criterion and model are described.

E-mail: sugi@cs.titech.ac.jp; URL: http://ogawa-www.cs.titech.ac.jp/~sugi.

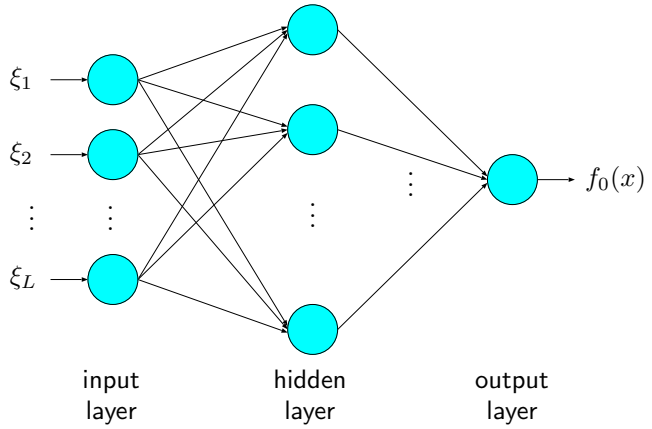


Figure 1. A three-layer feedforward neural network.

2.1. NN Learning as an Inverse Problem

Let us consider a learning problem of a three-layer feedforward NN whose numbers of input and output units are L and 1, respectively (see Fig. 1). The relationship between input $x = (\xi_1, \xi_2, \cdots, \xi_L)^\top$ and output of the network is expressed by using a function $f_0(x)$ of L variables. Therefore, the NN learning problem is equivalent to obtaining the optimal approximation to a target function f from a set of M training examples made up of input signals $x_m \in \mathbf{R}^L$ and corresponding output signals $y_m \in \mathbf{C}$:

$$\{(x_m, y_m) \mid y_m = f(x_m) + n_m\}_{m=1}^M,$$

where y_m is degraded by zero-mean additive noise n_m. Let n and y be M-dimensional vectors whose m-th elements are n_m and y_m, respectively. y is called a *sample value vector*, and a space which y belongs to is called a *sample value space*. In this paper, the target function f is assumed to belong to a reproducing kernel Hilbert space[19] H. Let $\mathcal{D}$ be the domain of f. The reproducing kernel is a bivariate function defined on $\mathcal{D} \times \mathcal{D}$ which satisfies the following conditions.

1. For any fixed x' in $\mathcal{D}$, $K(x, x')$ is a function of x in H.
2. For any function f in H and for any x' in $\mathcal{D}$, it holds that

$$\langle f(\cdot), K(\cdot, x') \rangle = f(x'),$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in H.

If a function $\psi_m(x)$ is defined as

$$\psi_m(x) = K(x, x_m), \tag{1}$$

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

$$f(x_m) = \langle f, \psi_m \rangle. \tag{2}$$

For this reason, ψ_m is called a *sampling function*. Let A be an operator which maps f to an M-dimensional vector whose m-th element is $f(x_m)$. We call A a *sampling operator*. Then, the relationship between f and y can be expressed as

$$y = Af + n. \tag{3}$$

Note that A is always a linear operator even when we are concerned with a non-linear function f. Indeed, A can be expressed as

$$A = \sum_{m=1}^{M} \left(e_m \otimes \overline{\psi_m} \right),$$

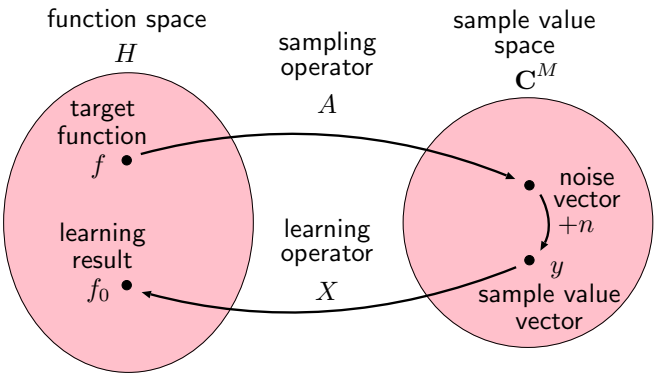


Figure 2. NN learning as an inverse problem.

where e_m is the m-th vector of the so-called standard basis in $\mathbf{C}^M$ and $(\cdot \otimes \bar{\cdot})$ stands for the *Neumann-Schatten product**. Let f_0 be a learning result. Then, the relationship between y and f_0 is denoted as

$$f_0 = Xy, \tag{4}$$

where X is called a *learning operator*. Consequently, the NN learning problem is reformulated as an inverse problem of obtaining X which provides the best approximation f_0 to f under a certain criterion (Fig. 2).

2.2. Learning Criterion and Model

As mentioned above, the function approximation is performed on the basis of a learning criterion. Our purpose of learning in this paper is to minimize the generalization error of the learning result f_0, which is measured by

$$J_G = E_n \|f_0 - f\|^2. \tag{5}$$

Equation (5) can be decomposed as follows:

Proposition 2.1. *(Ref. 21) It holds that*

$$J_G = \|E_n f_0 - f\|^2 + E_n \|f_0 - E_n f_0\|^2. \tag{6}$$

The first and second terms of the right-hand side of eq.(6) is called the *bias* and *variance* of f_0, respectively. In this paper, we adopt the projection learning criterion. Let A^*, $\mathcal{R}(A^*)$, and $P_{\mathcal{R}(A^*)}$ be the adjoint operator of A, the range of A^*, and the orthogonal projection operator onto $\mathcal{R}(A^*)$, respectively. Then, projection learning is defined as follows:

Definition 2.2. *(Projection learning)(Ref. 22–25) An operator X is called the projection learning operator if X minimizes the functional*

$$J_P[X] = E_n \|Xn\|^2$$

under the constraint

$$XA = P_{\mathcal{R}(A^*)}. \tag{7}$$

*For any fixed g in a Hilbert space H_1 and any fixed f in a Hilbert space H_2, the *Neumann-Schatten product* $(f \otimes \overline{g})$ is an operator from H_1 to H_2, which is defined by using any $h \in H_1$ as (see Ref. 20)

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

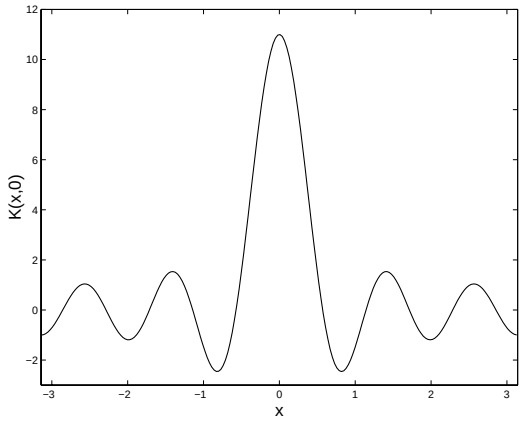


Figure 3. Profile of the reproducing kernel of a trigonometric polynomial space of order 5 ($x' = 0$).

From eqs.(4) and (3), the learning result f_0 can be decomposed as

$$f_0 = XAf + Xn. \tag{8}$$

The first and second terms of the right-hand side of eq.(8) are called the *signal* and *noise components* of f_0, respectively. Substituting eqs.(8) and (7) into eq.(6), we have

$$J_G = \|P_{\mathcal{R}(A^*)}f - f\|^2 + E_n\|Xn\|^2. \tag{9}$$

Hence, the projection learning criterion decreases the bias of f_0 to a certain level and minimizes the variance of f_0. Let us denote the projection learning operator by $A^{(P)}$, which comes from the notation of generalized inverse operators (see Ref. 26) since the role of learning operators is similar to the role of generalized inverse operators. Then, the following proposition holds.

Proposition 2.3. *If the noise covariance matrix is $\sigma^2 I$ with $\sigma^2 > 0$, then the projection learning operator $A^{(P)}$ is expressed as*

$$A^{(P)} = A^\dagger, \tag{10}$$

where $A^\dagger$ denotes the Moore-Penrose generalized inverse[†] of A.

Let us consider learning in the following function space.

Definition 2.4. *(Trigonometric polynomial space) A Hilbert space H is called a trigonometric polynomial space of order N if H is spanned by*

$$\{\exp(inx)\}_{n=-N}^{N}$$

which are defined on $[-\pi, \pi]$ and the inner product in H is defined as

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

In a trigonometric polynomial space of order N, the reproducing kernel is expressed as

$$K(x, x') = \begin{cases} \sin\dfrac{(2N+1)(x-x')}{2} \Big/ \sin\dfrac{x-x'}{2} & \text{if } x \neq x', \\ 2N+1 & \text{if } x = x'. \end{cases} \tag{11}$$

The profile of eq.(11) is illustrated in Fig. 3.

[†]An operator X is called the *Moore-Penrose generalized inverse* of an operator A if X satisfies the following four conditions (see Ref. 27).

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

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

3. PSEUDO ORTHOGONAL BASES

In this section, we describe the concept of POBs and show its fundamental properties. Let M be an integer larger than or equal to the dimension of a finite dimensional Hilbert space H.

Definition 3.1. *(Ref. 3,4) A set $\{\phi_m\}_{m=1}^M$ of elements in H is called a POB if any f in H is expressed as*

$$f = \sum_{m=1}^{M} \langle f, \phi_m \rangle \phi_m.$$

If M is equal to the dimension of H, a POB is reduced to an orthonormal basis (ONB) in H. The following proposition shows the characteristics of POBs.

Proposition 3.2. *(Ref. 3,4) Let f and g be any elements in H. Then, the following conditions are mutually equivalent.*

1. *A set $\{\phi_m\}_{m=1}^M$ is a POB in H.*
2. $\|f\|^2 = \sum_{m=1}^{M} |\langle f, \phi_m \rangle|^2$.
3. $\langle f, g \rangle = \sum_{m=1}^{M} \langle f, \phi_m \rangle \overline{\langle g, \phi_m \rangle}$.

The condition 2 implies that a POB is equivalent to a tight frame with frame bound 1. When M is equal to the dimension of H, the conditions 2 and 3 are reduced to *Parseval's equalities*.

Let H' be an M-dimensional Hilbert space and $\{\varphi_m\}_{m=1}^M$ be an arbitrary ONB in H'. Let U be an operator defined as

$$U = \sum_{m=1}^{M} \left(\varphi_m \otimes \overline{\phi_m} \right). \tag{12}$$

Then, the following proposition holds.

Proposition 3.3. *(Ref. 3,4) Let f and g be any elements in H. Then, the following conditions are mutually equivalent.*

1. *A set $\{\phi_m\}_{m=1}^M$ is a POB in H.*
2. *$U^*U = I$, where I is the identity operator.*
3. $\|Uf\| = \|f\|$.
4. $\langle Uf, Ug \rangle = \langle f, g \rangle$.

The condition 3 means that the operator U is an *isometry*[‡]. From this property, we have the following construction method of POBs.

Proposition 3.4. *(Ref. 3,4) Let U be an arbitrary isometry from H to H' and $\{\varphi_m\}_{m=1}^M$ be an arbitrary ONB in H'. If we put*

$$\phi_m = U^* \varphi_m$$

for $1 \leq m \leq M$, then $\{\phi_m\}_{m=1}^M$ becomes a POB in H. All POBs can be constructed by changing U with a fixed ONB $\{\varphi_m\}_{m=1}^M$ or by changing $\{\varphi_m\}_{m=1}^M$ with a fixed U.

If $\{\phi_m\}_{m=1}^M$ is a POB and all norms $\|\phi_m\|$ agree with each other, then $\{\phi_m\}_{m=1}^M$ is called a *pseudo orthonormal basis* (PONB). Since the concept of PONBs is essential in the following sections, we shall show some properties of PONBs. To begin with, we give a construction method of PONBs.

[‡]An operator U is called an *isometry* if $\|Uf\| = \|f\|$ for all f in H.

Theorem 3.5. *Let the dimension of H be μ and $M = k\mu$ where k is an arbitrary integer. Then, $\{\phi_m\}_{m=1}^M$ becomes a PONB if $\{\sqrt{k}\phi_m\}_{m=1}^M$ consists of k sets of ONBs in H.*

Proofs of all theorems and lemmas are given in Appendix A. The following theorem gives another construction method of PONBs for a trigonometric polynomial space of order N.

Theorem 3.6. *Let c be an arbitrary constant such that $-\pi \leq c \leq -\pi + \frac{2\pi}{M}$ and let x_m be*

$$x_m = c + \frac{2\pi}{M}(m-1) \tag{13}$$

for $1 \leq m \leq M$. If we put

$$\phi_m = \frac{1}{\sqrt{M}} K(x, x_m) \tag{14}$$

for $1 \leq m \leq M$ where $K(\cdot,\cdot)$ is defined by eq.(11), then $\{\phi_m\}_{m=1}^M$ becomes a PONB in a trigonometric polynomial space of order N.

Finally, we show an important characteristic of PONBs.

Theorem 3.7. *Let $\{\phi_m\}_{m=1}^M$ be elements in μ-dimensional Hilbert space H such that $\|\phi_m\| = \sqrt{\frac{\mu}{M}}$ for all m and $\{\phi_m\}_{m=1}^M$ spans H. Let n be an M-dimensional zero-mean random vector subject to the covariance matrix $\sigma^2 I$ with $\sigma^2 > 0$. Then, the variance of $U^\dagger n$*

$$E_n \|U^\dagger n\|^2 \tag{15}$$

is minimized if and only if $\{\phi_m\}_{m=1}^M$ forms a PONB in H, where E_n denotes the ensemble average over n. In this case, the minimum value is $\sigma^2\mu$.

By making use of the above theorems, we discuss the problem of active learning in the following sections.

4. ACTIVE LEARNING IN TRIGONOMETRIC POLYNOMIAL SPACE

The problem of active learning is to find a set $\{x_m\}_{m=1}^M$ of sample points which provides the optimal generalization capability. In this section, we give the optimal solution to the active learning problem in the trigonometric polynomial space.

From eq.(9), the bias of a learning result f_0 becomes zero for all f in H if and only if $\mathcal{N}(A) = \{0\}$, where $\mathcal{N}(\cdot)$ stands for the null space of an operator. For this reason, we consider the case that a set $\{x_m\}_{m=1}^M$ of sample points satisfies $\mathcal{N}(A) = \{0\}$. When the noise covariance matrix is $\sigma^2 I$ with $\sigma^2 > 0$, it follows from eq.(10) that eq.(9) is reduced to

$$J_G = E_n \|A^\dagger n\|^2, \tag{16}$$

which is equivalent to the noise variance in H. Consequently, the problem of active learning becomes a problem of finding a set $\{x_m\}_{m=1}^M$ of sample points which minimizes eq.(16) under the constraint of $\mathcal{N}(A) = \{0\}$. Then, we have the following theorem.

Theorem 4.1. *Assume that the noise covariance matrix is $\sigma^2 I$ with $\sigma^2 > 0$. J_G is minimized under the constraint of $\mathcal{N}(A) = \{0\}$ if and only if $\{\frac{1}{\sqrt{M}}\psi_m\}_{m=1}^M$ forms a PONB in H. In this case, the minimum value of J_G is $\sigma^2(2N+1)/M$.*

Theorem 4.1 states that a PONB gives the optimal generalization capability. Now we give an interpretation of Theorem 4.1 by utilizing useful properties of POBs. As shown in the beginning of this section, minimizing the generalization error J_G defined by eq.(5) is equivalent to minimizing the noise variance in H. Hence, we shall investigate the mechanism of noise suppression by $A^{(P)}$.

When $\{\frac{1}{\sqrt{M}}\psi_m\}_{m=1}^M$ forms a POB in H, $\frac{1}{\sqrt{M}}A$ becomes an isometry because of Proposition 3.3 with $U = \frac{1}{\sqrt{M}}A$. This implies

$$\|Af\| = \sqrt{M}\|f\|$$

for all f in H. Then, the following lemma holds.

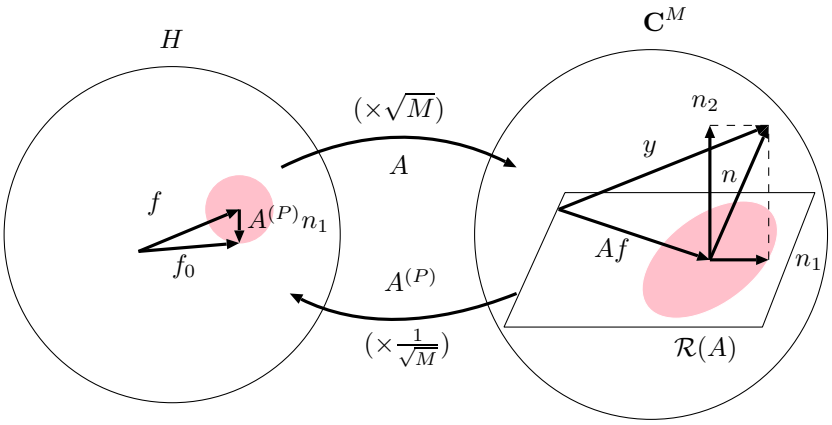


Figure 4. Mechanism of noise suppression by Theorem 4.1: The sample value vector y is decomposed as $y = Af + n_1 + n_2$. If $\{\frac{1}{\sqrt{M}}\psi_m\}_{m=1}^M$ forms a POB in H, then $A^{(P)}Af = f$, $\|A^{(P)}n_1\| = \frac{1}{\sqrt{M}}\|n_1\|$, and $A^{(P)}n_2 = 0$.

Lemma 4.2. *When $\{\frac{1}{\sqrt{M}}\psi_m\}_{m=1}^M$ forms a PONB in H, it holds that*

$$A^{(P)}Au = u \quad \textit{for all } u \in H, \tag{17}$$

$$\|A^{(P)}v\| = \begin{cases} \frac{1}{\sqrt{M}}\|v\| & \textit{for } v \in \mathcal{R}(A), \\ 0 & \textit{for } v \in \mathcal{R}(A)^\perp. \end{cases} \tag{18}$$

Equation (18) implies that $\sqrt{M}A^{(P)}$ becomes a *partial isometry*[§]. Let us decompose the noise n as

$$n = n_1 + n_2,$$

where $n_1 \in \mathcal{R}(A)$ and $n_2 \in \mathcal{R}(A)^\perp$. Then, the sample value vector y is rewritten as

$$y = Af + n_1 + n_2.$$

From eq.(17), it holds that

$$A^{(P)}Af = f,$$

which implies that the signal component Af is transformed into the original function f by $A^{(P)}$. For the noise component, it follows from eq.(18) that $A^{(P)}$ suppresses the magnitude of noise in $\mathcal{R}(A)$ by $\frac{1}{\sqrt{M}}$ and completely removes the noise n_2 in $\mathcal{R}(A)^\perp$. The above analysis is summarized in Fig. 4.

In Theorem 4.1, we have given a necessary and sufficient condition to minimize J_G. Now we give two examples of sample points which satisfy the condition in Theorem 4.1.

Theorem 4.3. *Let $M \geq 2N+1$ and c be an arbitrary constant such that $-\pi \leq c \leq -\pi + \frac{2\pi}{M}$. If we put $\{x_m\}_{m=1}^M$ as eq.(13), then $\{\frac{1}{\sqrt{M}}\psi_m\}_{m=1}^M$ forms a PONB in H.*

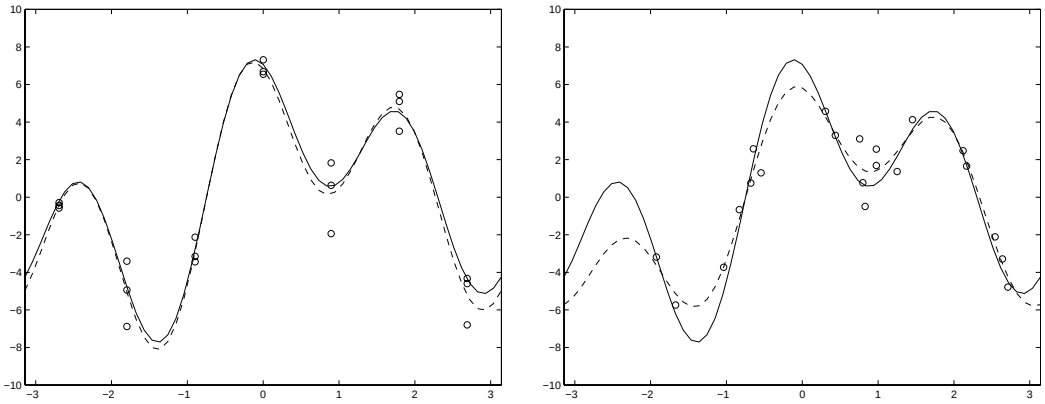
Theorem 4.4. *Let $M = k(2N+1)$ where k is a positive integer and c be an arbitrary constant such that $-\pi \leq c \leq -\pi + \frac{2\pi}{2N+1}$. If we put $\{x_m\}_{m=1}^M$ as*

$$x_m = c + \frac{2\pi p}{2N+1} \ : \ p = m - 1 \ \big(mod\ (2N+1)\big), \tag{19}$$

then $\{\frac{1}{\sqrt{M}}\psi_m\}_{m=1}^M$ forms a PONB in H.

[§]An operator A is called a *partial isometry* if it holds that

$$\|Af\| = \begin{cases} \|f\| & \text{if } f \in \mathcal{N}(A)^\perp, \\ 0 & \text{if } f \in \mathcal{N}(A). \end{cases}$$



$J_G = 0.333$ (a) Optimal sampling

$J_G = 1.202$ (b) Passive learning

Figure 5. Learning results in a trigonometric polynomial model of order 3 with noise covariance matrix I. The number of training examples is 21. The solid line denotes the target function f and the dashed line does the learning result. $\circ$ indicates a training example. The generalization error J_G is measured by eq.(5).

Theorem 4.3 is clear from Theorem 3.6. Hence, we omit its proof. Theorem 4.3 means that M sample points are fixed to $2\pi/M$ intervals in the domain $[-\pi, \pi]$ and sample values are gathered once at each point. In contrast, the sampling method shown in Theorem 4.4 means that $(2N + 1)$ sample points are fixed to $2\pi/(2N + 1)$ intervals in the domain and sampling is performed k times at each point.

5. SIMULATIONS

In this section, we demonstrate the effectiveness of the proposed active learning method through computer simulations. Let us consider the following two sampling schemes.

(A) Optimal sampling: Training examples are gathered following Theorem 4.4.

(B) Passive learning: Training examples are given unilaterally.

Let a target function f be

$$f(x) = 2\sqrt{2}\sin x + 2\sqrt{2}\cos x + \frac{1}{\sqrt{2}}\sin 2x + \sqrt{2}\cos 2x - 2\sqrt{2}\sin 3x + 2\sqrt{2}\cos 3x,$$

and H be a trigonometric polynomial space of order 3. Let the number M of training examples be 21 and the noise covariance matrix be I. Figure 5 shows the learning results where the solid line denotes the target function f and the dashed line does the learning result. $\circ$ indicates a training example. Figure 5 (a) displays the learning result of the scheme (A) with $k = 3$. The generalization error J_G measured by eq.(5) is 0.333. In contrast, Fig. 5 (b) shows the learning result of the scheme (B). The generalization error J_G is 1.202. This result shows that the scheme (A) gives a 72.3 percent reduction in generalization error compared with the scheme (B). Therefore, we can confirm that the proposed method is considerably effective in acquiring better generalization capability.

Figure 6 shows the relation between the number of training examples and the generalization error. This simulation is also performed under the condition that the order N of trigonometric polynomial is 3 and the noise covariance matrix is I. The horizontal and vertical axes display the number of training examples and the generalization error J_G measured by eq.(5), respectively. The solid line shows the scheme (A). The dashed line denotes the average of 100 trials of the scheme (B). This graph illustrates that the generalization error tends to decrease in both sampling schemes when the number of training examples increases. In all numbers of training examples, the scheme (A) gives better generalization capability than the scheme (B). Although the scheme (B) also provides good generalization capability with a large number of training examples, the difference between two sampling schemes is remarkable when it comes to a small number of training examples. Hence, our active learning method is shown to be effective especially when the number of training examples is small.

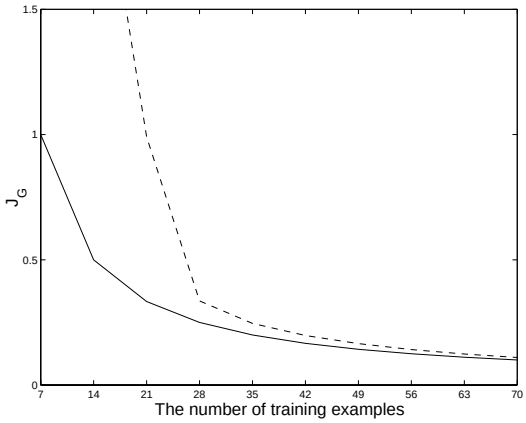


Figure 6. The relation between the number of training examples and the generalization error: The order N of trigonometric polynomial is 3 and the noise variance σ^2 is 1. The horizontal and vertical axes display the number of training examples and the generalization error J_G measured by eq.(5), respectively. The solid line shows the scheme (A), optimal sampling. The dashed line denotes the average of 100 trials of the scheme (B), passive learning.

6. CONCLUSION

This paper showed that pseudo orthogonal bases play an essential role when we work on active learning in neural networks. Our solution to active learning gives the optimal generalization capability. By utilizing useful properties of pseudo orthogonal bases, we clarified the mechanism of achieving the optimal generalization.

REFERENCES

1. R. Duffin and A. Schaeffer, "A class of non harmonic Fourier series," *Transactions on American Mathematical Society* **72**, pp. 341–366, 1952.
2. R. Young, *An Introduction to Nonharmonic Fourier Series*, Academic Press, 1980.
3. H. Ogawa and T. Iijima, "A theory of pseudo-orthogonal bases," Tech. Rep. PRL73-44, IECE Japan, July 1973.
4. H. Ogawa and T. Iijima, "A theory of pseudo orthogonal bases," *Transactions on IECE Japan* **J58-D**, pp. 271–278, May 1975.
5. H. Ogawa, "A theory of pseudo biorthogonal bases," Tech. Rep. PRL77-60, IECE Japan, Jan. 1978.
6. H. Ogawa, "A theory of pseudo biorthogonal bases," *Transactions on IECE Japan* **J64-D**, pp. 555–562, July 1981.
7. H. Ogawa, "Pseudo biorthogonal bases of type O," *Transactions on IECE Japan* **J64-D**, pp. 563–569, July 1981.
8. H. Ogawa, "Theory of pseudo biorthogonal bases and its application," in *Research Institute for Mathematical Science, RIMS Kokyuroku*, No. 1067 in Reproducing Kernels and their Applications, pp. 24–38, Oct. 1998.
9. H. Ogawa, "A unified approach to generalized sampling theorems," in *Proceedings of ICASSP'86, IEEE-IECEJ-ASJ International Conference on Acoustics, Speech, and Signal Processing*, pp. 1657–1660, Apr. 1986.
10. H. Ogawa and I. Kumazawa, "Radon transform and analog coding," in *Mathematical Methods in Tomography*, G. T. Herman, A. K. Louis, and F. Natterer, eds., vol. 1497 of *Lecture Notes in Mathematics*, pp. 229–241, Springer-Verlag, 1991.
11. H. Ogawa, "Neural network learning, generalization and over-learning," in *Proceedings of the ICIIPS'92, International Conference on Intelligent Information Processing & System*, pp. 1–6, (Beijing, China), Oct. 30–Nov. 1 1992.
12. V. V. Fedorov, *Theory of Optimal Experiments*, Academic Press, New York, 1972.
13. F. Pukelsheim, *Optimal Design of Experiments*, John Wiley & Sons, 1993.
14. L. P. Kaelbling, ed., *Machine Learning*, vol. 22, pp. 7–290. Kluwer Academic Publishers, Jan./Feb./Mar. 1996.
15. D. Angluin, "Queries and concept learning," *Machine Learning* **2**, pp. 319–342, 1988.

16. D. MacKay, "Information-based objective functions for active data selection," *Neural Computation* **4**(4), pp. 590–604, 1992.
17. K. Fukumizu, "Active learning in multilayer perceptrons," in *Advances in Neural Information Processing Systems*, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, eds., vol. 8, pp. 295–301, The MIT Press, 1996.
18. H. Ogawa, "Inverse problem and neural networks," in *Proceedings of IEICE 2nd Karuizawa Workshop on Circuits and Systems*, pp. 262–268, (Karuizawa, Japan), May 24–25 1989.
19. N. Aronszajn, "Theory of reproducing kernels," *Transactions on American Mathematical Society* **68**, pp. 337–404, 1950.
20. R. Schatten, *Norm Ideals of Completely Continuous Operators*, Springer-Verlag, Berlin, 1970.
21. A. Takemura, *Modern Mathematical Statistics*, Sobunsya, Tokyo, 1991.
22. N. Nakamura and H. Ogawa, "Optimal digital image restoration under additive noises," Tech. Rep. PRL82-32, IECE Japan, Oct. 1982.
23. N. Nakamura and H. Ogawa, "Optimum digital image restoration under additive noises," *Transactions on IECE Japan* **J67-D**, pp. 563–570, May 1984.
24. H. Ogawa and N. Nakamura, "Projection filter restoration of degraded images," in *IEEE Seventh International Conference of Pattern Recognition Proceedings*, pp. 601–603, 1984.
25. H. Ogawa, "Projection filter regularization of ill-conditioned problem," in *Proceedings of SPIE, Inverse Problems in Optics, 808*, pp. 189–196, 1987.
26. A. Ben-Israel and T. N. E. Greville, *Generalized Inverses: Theory and Applications*, John Wiley & Sons, New York, 1974.
27. A. Albert, *Regression and the Moore-Penrose Pseudoinverse*, Academic Press, New York and London, 1972.

APPENDIX A. PROOFS

A.1. Theorem 3.5

Let T be an operator defined as

$$T = U^*U = \sum_{m=1}^{M} \left(\phi_m \otimes \overline{\phi_m}\right). \tag{20}$$

For an ONB $\{\varphi'_j\}_{j=1}^{\mu}$ in H, it holds that

$$\sum_{j=1}^{\mu} \left(\varphi'_j \otimes \overline{\varphi'_j}\right) = I.$$

Hence, if $\{\sqrt{k}\phi_m\}_{m=1}^{M}$ consists of k sets of ONBs, it follows from eq.(20) that

$$T = \frac{1}{k} \sum_{m=1}^{M} \left(\sqrt{k}\phi_m \otimes \overline{\sqrt{k}\phi_m}\right) = I,$$

which implies that $\{\phi_m\}_{m=1}^{M}$ forms a POB in H because of Proposition 3.3. In this case, $\{\phi_m\}_{m=1}^{M}$ is a PONB in H since $\{\sqrt{k}\phi_m\}_{m=1}^{M}$ consists of k sets of ONBs. □

A.2. Theorem 3.6

Let f be an element in H denoted as

$$f(x) = \sum_{n=-N}^{N} c_n \exp(inx), \tag{21}$$

where c_n is a complex number. It follows from eqs.(14), (1), (2), (21), and (13) that

$$\begin{aligned}\sum_{m=1}^{M} |\langle f, \phi_m\rangle|^2 &= \frac{1}{M}\sum_{m=1}^{M} |\langle f, \psi_m\rangle|^2 = \frac{1}{M}\sum_{m=1}^{M} |f(x_m)|^2 = \frac{1}{M}\sum_{m=1}^{M}\left|\sum_{n=-N}^{N} c_n \exp(inx_m)\right|^2 \\ &= \frac{1}{M}\sum_{n=-N}^{N}\sum_{n'=-N}^{N} c_n\overline{c_{n'}} \exp\left(i(n-n')(c-\frac{2\pi}{M})\right)\sum_{m=1}^{M}\exp\left(i(n-n')\frac{2\pi m}{M}\right). \end{aligned} \tag{22}$$

For any integer n and n', it holds that

$$\sum_{m=1}^{M} \exp\left(i(n-n')\frac{2\pi m}{M}\right) = \begin{cases} M & \text{if } n = n', \\ 0 & \text{if } n \neq n'. \end{cases}$$

Hence, eq.(22) yields

$$\sum_{m=1}^{M} |\langle f, \phi_m \rangle|^2 = \sum_{n=-N}^{N} |c_n|^2 = \|f\|^2,$$

which implies that $\{\phi_m\}_{m=1}^M$ is a POB because of Proposition 3.2. It follows from eqs.(14), (1), (2), and (11) that

$$\|\phi_m\|^2 = \frac{1}{M}\langle \psi_m, \psi_m \rangle = \frac{1}{M}\psi_m(x_m) = \frac{1}{M}K(x_m, x_m) = \frac{2N+1}{M} \tag{23}$$

for $1 \leq m \leq M$. Hence, $\{\phi_m\}_{m=1}^M$ is a PONB in a trigonometric polynomial space of order N. □

A.3. Theorem 3.7

From eq.(20), eq.(15) is reduced to

$$E_n\|U^\dagger n\|^2 = \sigma^2 \mathrm{tr}\left(T^\dagger\right), \tag{24}$$

where $\mathrm{tr}\,(\cdot)$ denotes the trace of an operator. Since $\{\phi_m\}_{m=1}^M$ spans H and T is *positive semi-definite*[¶], T has μ positive eigenvalues $\{\lambda_k\}_{k=1}^\mu$ considering the *geometric multiplicity*. Then, it holds that

$$\mathrm{tr}\,(T) = \sum_{k=1}^{\mu} \lambda_k, \tag{25}$$

$$\mathrm{tr}\left(T^\dagger\right) = \sum_{k=1}^{\mu} \frac{1}{\lambda_k}. \tag{26}$$

It is well-known that the arithmetic and harmonic means have the following relation.

$$\frac{\sum_{k=1}^{\mu} \lambda_k}{\mu} \geq \frac{\mu}{\sum_{k=1}^{\mu} \frac{1}{\lambda_k}}, \tag{27}$$

where equality holds if and only if all λ_k agree with each other. From eqs.(24)–(27), we have

$$E_n\|U^\dagger n\|^2 \geq \frac{\sigma^2\mu^2}{\mathrm{tr}\,(T)}. \tag{28}$$

Since it holds from eq.(20) that

$$\mathrm{tr}\,(T) = \sum_{m=1}^{M} \|\phi_m\|^2 = \mu,$$

eq.(28) yields

$$E_n\|U^\dagger n\|^2 \geq \sigma^2\mu, \tag{29}$$

where equality holds if and only if $\lambda_k = 1$ for all k. From eq.(20), this condition is equivalent to

$$T = U^*U = I,$$

which implies that $\{\phi_m\}_{m=1}^M$ forms a POB in H because of Proposition 3.3. Since $\|\phi_m\| = \sqrt{\frac{\mu}{M}}$ for all m, $\{\phi_m\}_{m=1}^M$ is a PONB. □

[¶]An operator T is said to be *positive semi-definite* if $\langle Tf, f \rangle \geq 0$ for any f.

A.4. Theorem 4.1

Let an M-dimensional Hilbert space H' be $\mathbf{C}^M$. For $1 \leq m \leq M$, if we put

$$\begin{aligned} \varphi_m &= e_m, \\ \phi_m &= \frac{1}{\sqrt{M}}\psi_m, \end{aligned} \tag{30}$$

then U defined by eq.(12) becomes

$$U = \frac{1}{\sqrt{M}}A.$$

Hence, eq.(16) yields

$$J_G = E_n\|A^\dagger n\|^2 = \frac{1}{M}E_n\|U^\dagger n\|^2.$$

If we put $\mu = 2N+1$, then it follows from eq.(23) that

$$\|\phi_m\|^2 = \frac{\mu}{M}$$

for $1 \leq m \leq M$. Consequently, Theorem 4.1 is clear from Theorem 3.7. □

A.5. Lemma 4.2

It follows from eq.(10) and $\mathcal{N}(A) = \{0\}$ that for all u in H,

$$A^{(P)}Au = A^\dagger Au = P_{\mathcal{R}(A^*)}u = u,$$

which implies eq.(17). Since $\frac{1}{\sqrt{M}}A$ is an isometry, it holds that

$$\|\frac{1}{\sqrt{M}}Au\| = \|u\|. \tag{31}$$

If we put

$$v = Au,$$

then it follows from eqs.(17) and (31) that

$$\|A^{(P)}v\| = \|A^{(P)}Au\| = \|u\| = \|\frac{1}{\sqrt{M}}Au\| = \frac{1}{\sqrt{M}}\|v\|,$$

which implies the upper half of eq.(18). The bottom half is clear from eq.(10). □

A.6. Theorem 4.4

If we determine $\{x_m\}_{m=1}^M$ as eq.(19) and put ϕ_m as eq.(30) for $1 \leq m \leq M$, then it follows from eqs.(2), (1), and (11) that

$$\begin{aligned} \langle\sqrt{k}\phi_{p+(q-1)(2N+1)}, \sqrt{k}\phi_{p'+(q-1)(2N+1)}\rangle &= \frac{k}{M}\langle\psi_{p+(q-1)(2N+1)}, \psi_{p'+(q-1)(2N+1)}\rangle \\ &= \frac{k}{M}\psi_{p+(q-1)(2N+1)}(x_{p'+(q-1)(2N+1)}) \\ &= \frac{k}{M}K(x_{p'+(q-1)(2N+1)}, x_{p+(q-1)(2N+1)}) \\ &= \delta_{pp'} \end{aligned}$$

for $1 \leq p, p' \leq 2N+1$ and $1 \leq q \leq k$, where $\delta_{pp'}$ denotes *Kronecker's delta*. This implies that for each $q = 1, 2, \cdots, k$, $\{\sqrt{k}\phi_{p+(q-1)(2N+1)}\}_{p=1}^{2N+1}$ forms an ONB in H, and hence $\{\sqrt{k}\phi_m\}_{m=1}^M$ consists of k sets of ONBs in H. Consequently, Theorem 4.4 is clear from Theorem 3.5. □