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

Masashi Sugiyama
Hidemitsu Ogawa

Department of Computer Science,
Tokyo Institute of Technology, Japan

Pseudo Orthogonal Bases (POBs)

Definition

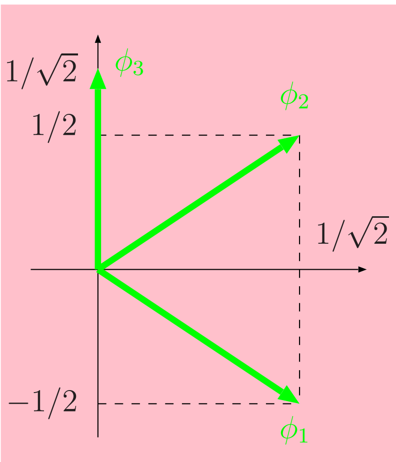
H : a finite dimensional Hilbert space

$M \geq \dim(H)$

A set $\{\phi_m\}_{m=1}^M$ of elements in H is called a POB if any f in H is expressed as

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

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



$H = \mathbf{R}^2$, $M = 3$

- If $M = \dim(H)$, a POB is reduced to an ONB.
- A POB is a tight frame with frame bound 1.

$$\|f\|^2 = \sum_{m=1}^{M} |\langle f, \phi_m \rangle|^2.$$

If $\|\phi_1\| = \|\phi_2\| = \cdots = \|\phi_M\|$,
then $\{\phi_m\}_{m=1}^M$ is called
a pseudo orthonormal basis (PONB).

Frame, POB, PBOB, $\cdots$

- Frame
  - Duffin and Shaeffer (1952)
  - Young (1980)

- Pseudo orthogonal basis (POB)
  - Ogawa and Iijima (1973)

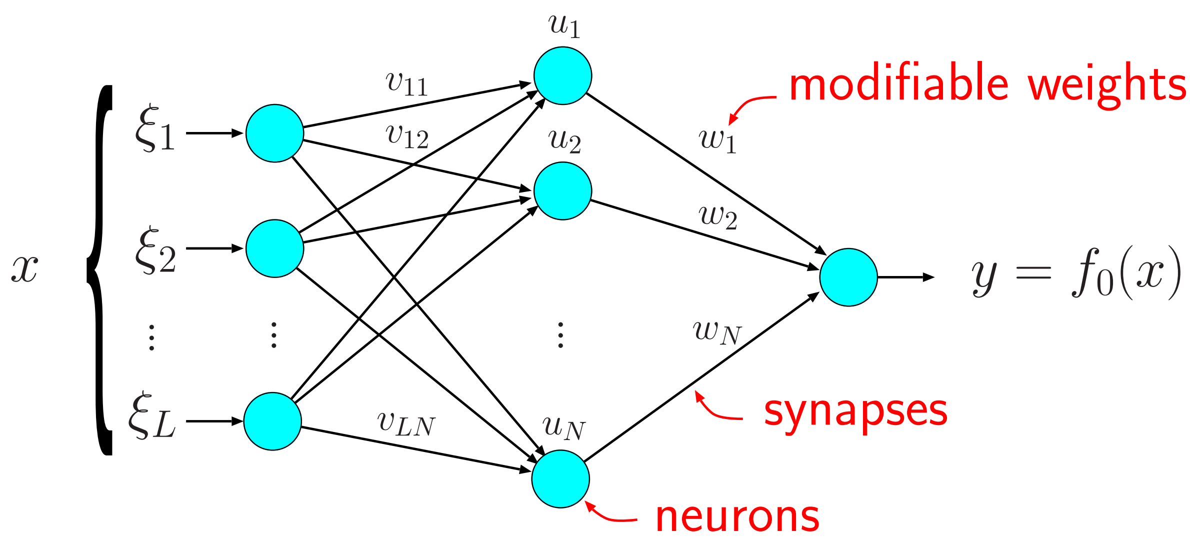
$$f = \sum_{m=1}^{M} \langle f, \phi_m \rangle \phi_m$$

- Pseudo biorthogonal basis (PBOB)
  - Ogawa (1978)

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

→ Signal restoration,
Computerized Tomography,
Neural Network Learning,
⋮

Learning in Neural Networks

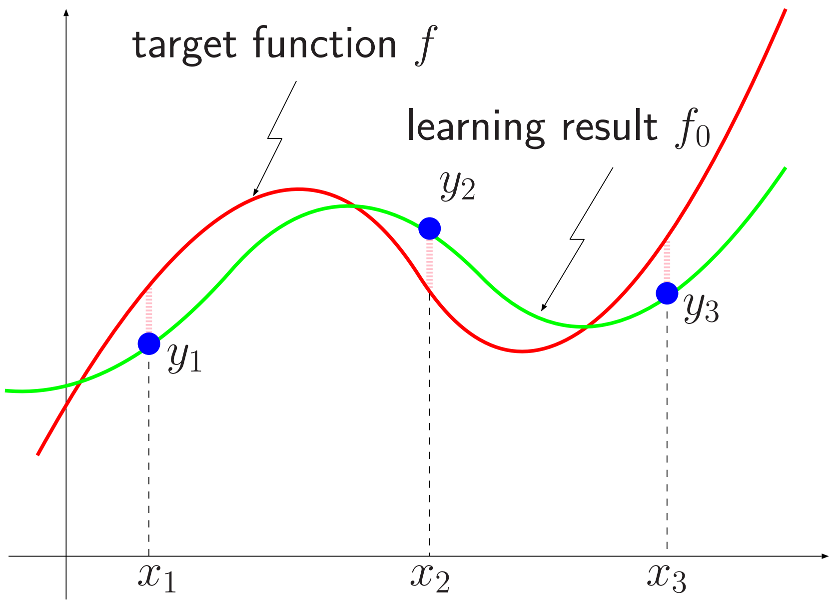


Purpose of NN Learning

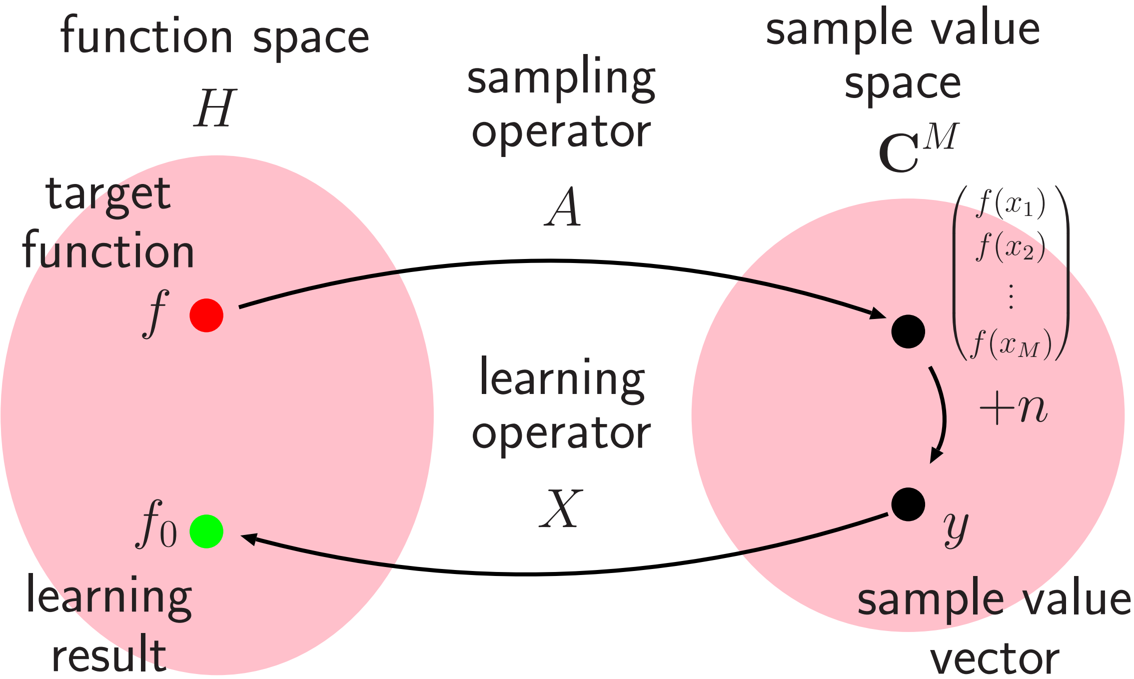
Modify weights by using training examples:

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

and obtain underlying input-output rule.



NN Learning as an Inverse Problem



$$\text{sampling} \ : \ y = \begin{pmatrix} y_1 \\ \vdots \\ y_M \end{pmatrix} = Af + n$$

$$\text{learning} \ : \ f_0 = Xy$$

representation of sampling operator A

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

$$\psi_m(x) = K(x, x_m)$$

$$K(x, x') \ : \ \text{reproducing kernel}$$

$$\langle f, \psi_m \rangle = f(x_m)$$

Trigonometric Polynomial Space

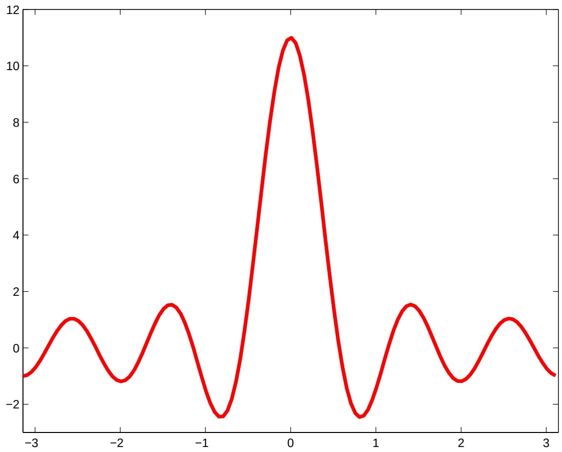
A Hilbert space H is called
a trigonometric polynomial space of order N
if H is spanned by

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

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

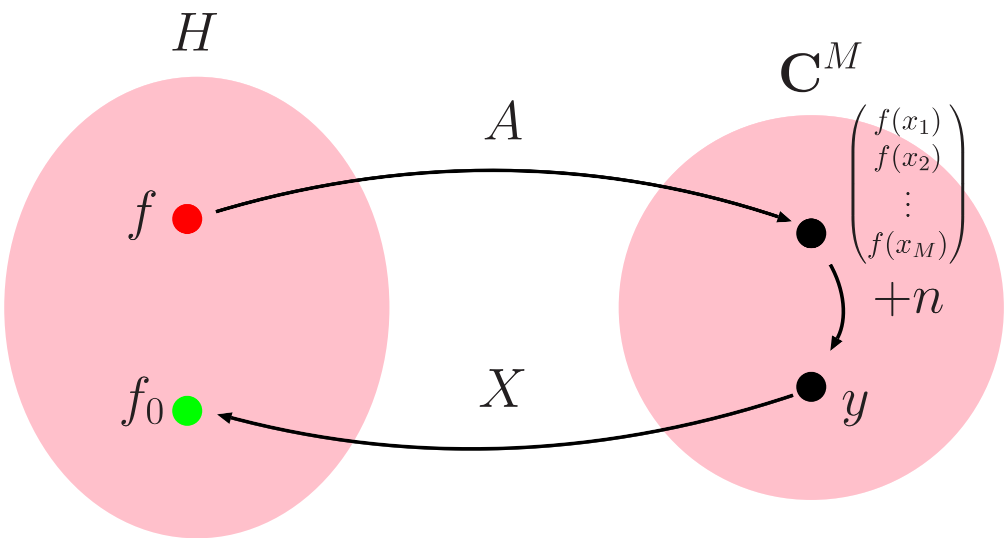
$$\langle f, g \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\overline{g(x)}dx.$$

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



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

Process of NN Learning



1. (Active Learning)
   Sample points $\{x_m\}_{m=1}^{M}$ are determined.
2. Sample values $\{y_m\}_{m=1}^{M}$ are gathered.
3. X and f_0 are calculated : Projection Learning
   When noise covariance matrix is $\sigma^2 I$,

$$X = A^{\dagger}.$$

$A^{\dagger}$ is the Moore-Penrose generalized inverse of A.

Our goal

We give the optimal solution to active learning.

Active Learning

Find a set $\{x_m\}_{m=1}^{M}$ of sample points which minimizes

$$J_G = E_n \| f_0 - f \|^2, \quad \text{Generalization error}$$

where E_n denotes the ensemble average over the noise.

If noise covariance matrix is $\sigma^2 I$,
then J_G yields

$$J_G = \underbrace{\| P_{\mathcal{N}(A)} f \|^2}_{\text{bias}} + \underbrace{\sigma^2 \mathrm{tr}((AA^*)^\dagger)}_{\text{variance}},$$

where $\mathcal{N}(A)$ denotes the null space of A.

Bias of f_0 is 0 $\iff$ $\mathcal{N}(A) = \{0\}$

$\Downarrow$

Strategy

Find a set $\{x_m\}_{m=1}^{M}$ of sample points which minimizes

$$J_G = \sigma^2 \mathrm{tr}((AA^*)^\dagger)$$

under the constraint of $\mathcal{N}(A) = \{0\}$.

Main Theorem

Suppose 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

$$\frac{\sigma^2(2N+1)}{M}.$$

$$f = \sum_{m=1}^{M} \langle f, \frac{1}{\sqrt{M}}\psi_m \rangle \frac{1}{\sqrt{M}}\psi_m \quad \text{for all } f \in H.$$

$$\|\psi_1\| = \|\psi_2\| = \cdots = \|\psi_M\|$$

$$\psi_m(x) = K(x, x_m)$$

$K(x,x')$: reproducing kernel

$$K(x,x') = \begin{cases} \sin\frac{(2N+1)(x-x')}{2} \Big/ \sin\frac{x-x'}{2} & (x \neq x') \\ 2N+1 & (x = x') \end{cases}$$

Interpretation

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

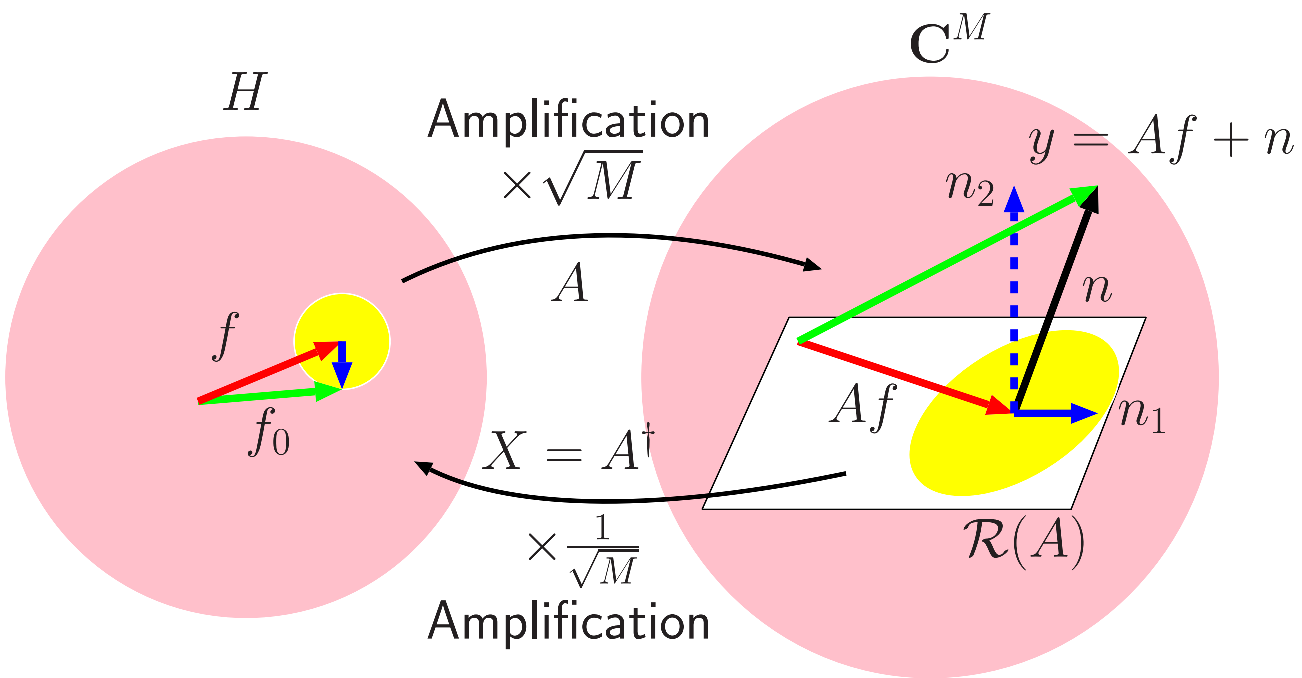
$$\|Af\| = \sqrt{M}\|f\|.$$

$$f_0 = Xy = A^\dagger Af + A^\dagger n_1 + A^\dagger n_2.$$

$$A^\dagger Af = f \quad \Longleftarrow \quad \mathcal{N}(A) = \{0\}$$

$$A^\dagger n_2 = 0 \quad \Longleftarrow \quad X : \text{Projection Learning}$$

$$\|A^\dagger n_1\| = \frac{1}{\sqrt{M}}\|n_1\| \quad \Longleftarrow \quad \{\frac{1}{\sqrt{M}}\psi_m\}_{m=1}^{M} : \text{ PONB}$$



Examples of PONB –1–

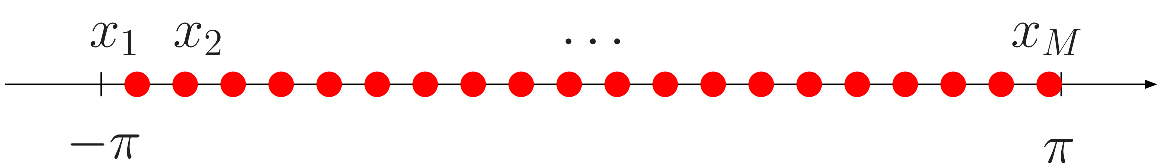
Example 1

$$M \geq 2N+1 \ (= \dim(H)),$$
$$c \ : \ -\pi \leq c \leq -\pi + \frac{2\pi}{M}.$$

If we put $\{x_m\}_{m=1}^{M}$ as

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

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



M sample points are fixed to $2\pi/M$ intervals and sample values are gathered once at each point.

$$\psi_m(x) = K(x, x_m)$$
$$K(x,x') \ : \ \text{reproducing kernel}$$

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

Examples of PONB –2–

$M = k(2N+1)$: k is a positive integer.

For a general finite dimensional Hilbert space H, $\{\phi_m\}_{m=1}^{M}$ becomes a PONB if $\{\sqrt{k}\phi_m\}_{m=1}^{M}$ consists of k sets of ONBs in H.

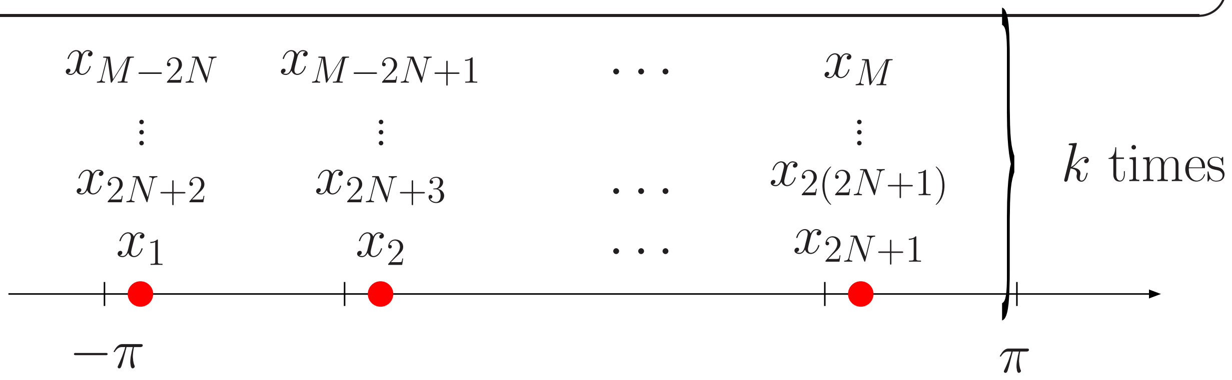
Example 2

$$c : -\pi \le c \le -\pi + \frac{2\pi}{2N+1}.$$

If we put $\{x_m\}_{m=1}^{M}$ as

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

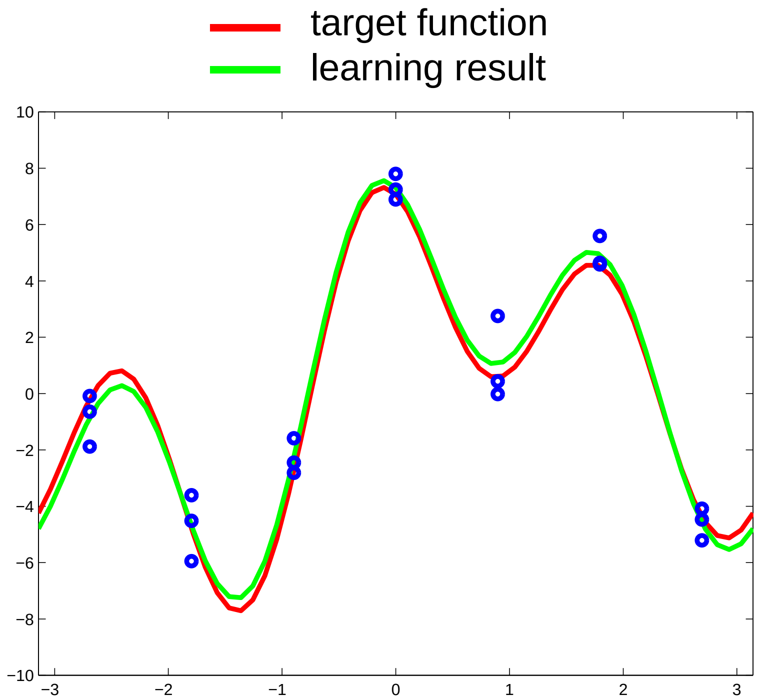
then $\{\frac{1}{\sqrt{M}}\psi_m\}_{m=1}^{M}$ forms a PONB in H.



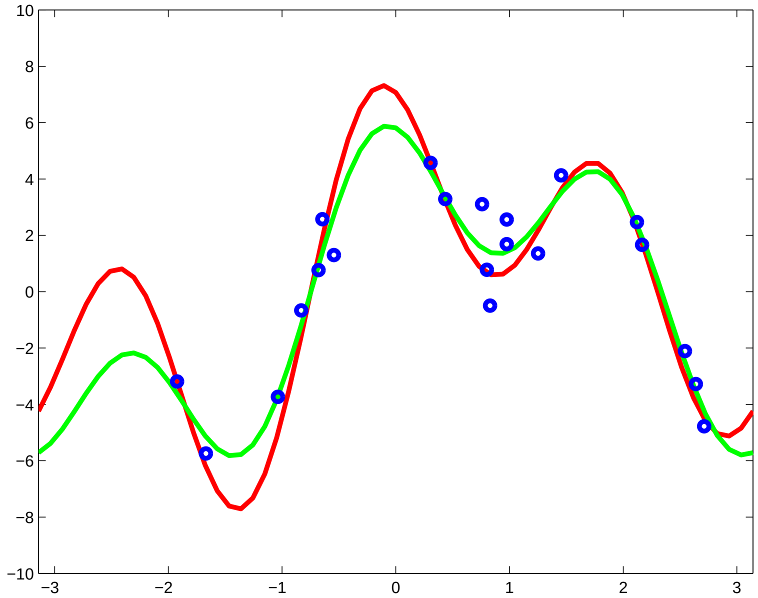
$(2N+1)$ sample points are fixed to $2\pi/(2N+1)$ intervals and sample values are gathered k times at each point.

Computer Simulation 1

$$N = 3\ (\dim(H) = 7),\ M = 21$$

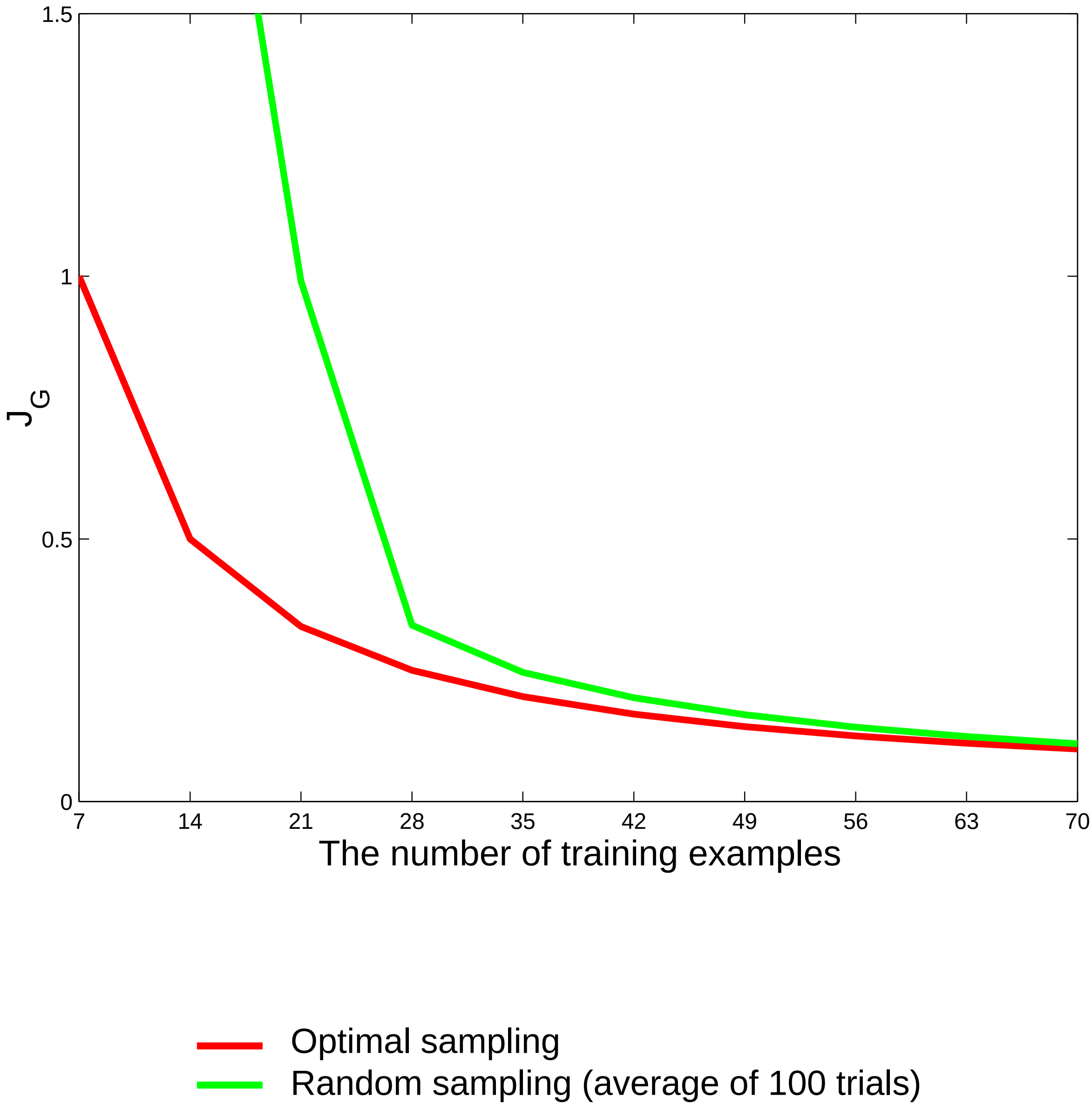


(A) Optimal sampling : $J_G = 0.333$



(B) Random sampling : $J_G = 1.202$

Computer simulation 2



Conclusions

1. We showed that pseudo orthogonal bases (POBs) give the optimal solution to active learning in neural networks.

2. By utilizing properties of POBs, we clarified the mechanism of achieving the optimal generalization.

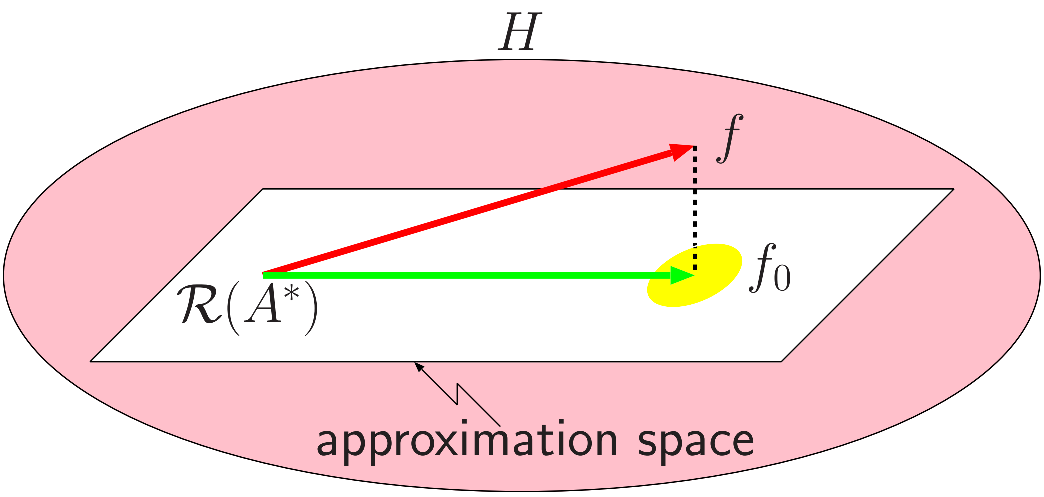
3. We gave two construction methods of PONBs.

Active Learning in Neural Networks

Projection learning

$$f_0 = \underbrace{XAf}_{\text{signal component}} + \underbrace{Xn}_{\text{noise component}}$$

minimize	$E_n\|Xn\|^2$
under the constraint of	$XAf = P_{\mathcal{R}(A^*)}f$



projection learning operator

$$X = V^\dagger A^* U^\dagger + Y(I - UU^\dagger)$$

Q : noise covariance matrix

$U = AA^* + Q$

$V = A^*U^\dagger A$

Y : arbitrary operator

A^* : adjoint operator of A

$U^\dagger$: Moore-Penrose generalized inverse of U