Training Data Selection for Optimal Generalization in Trigonometric Polynomial Networks

Tokyo Institute of Technology

Masashi Sugiyama

Hidemitsu Ogawa

Supervised Learning

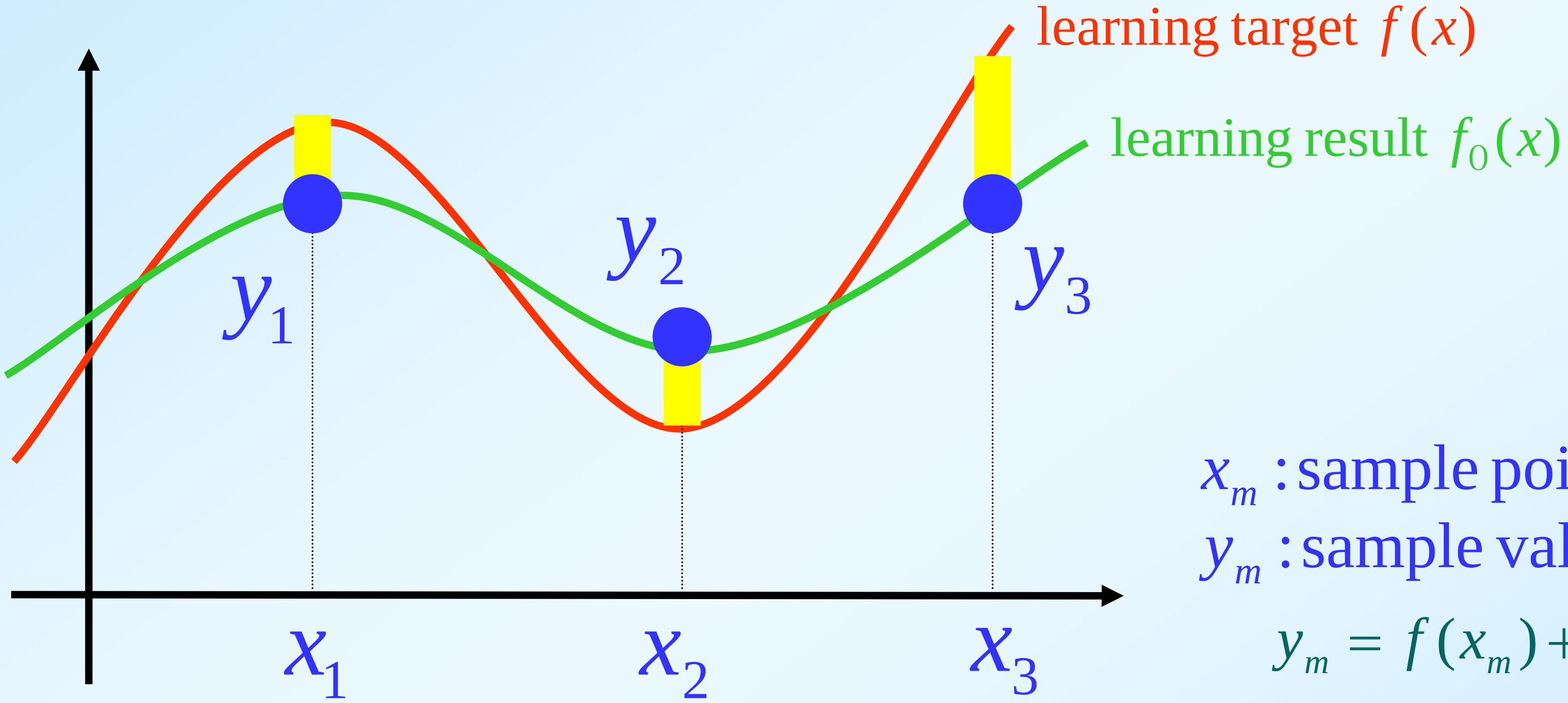
Estimating underlying rule from training examples

By using the acquired rule,

we can give appropriate output to unknown input

This ability is called generalization capability

Function Approximation Problem



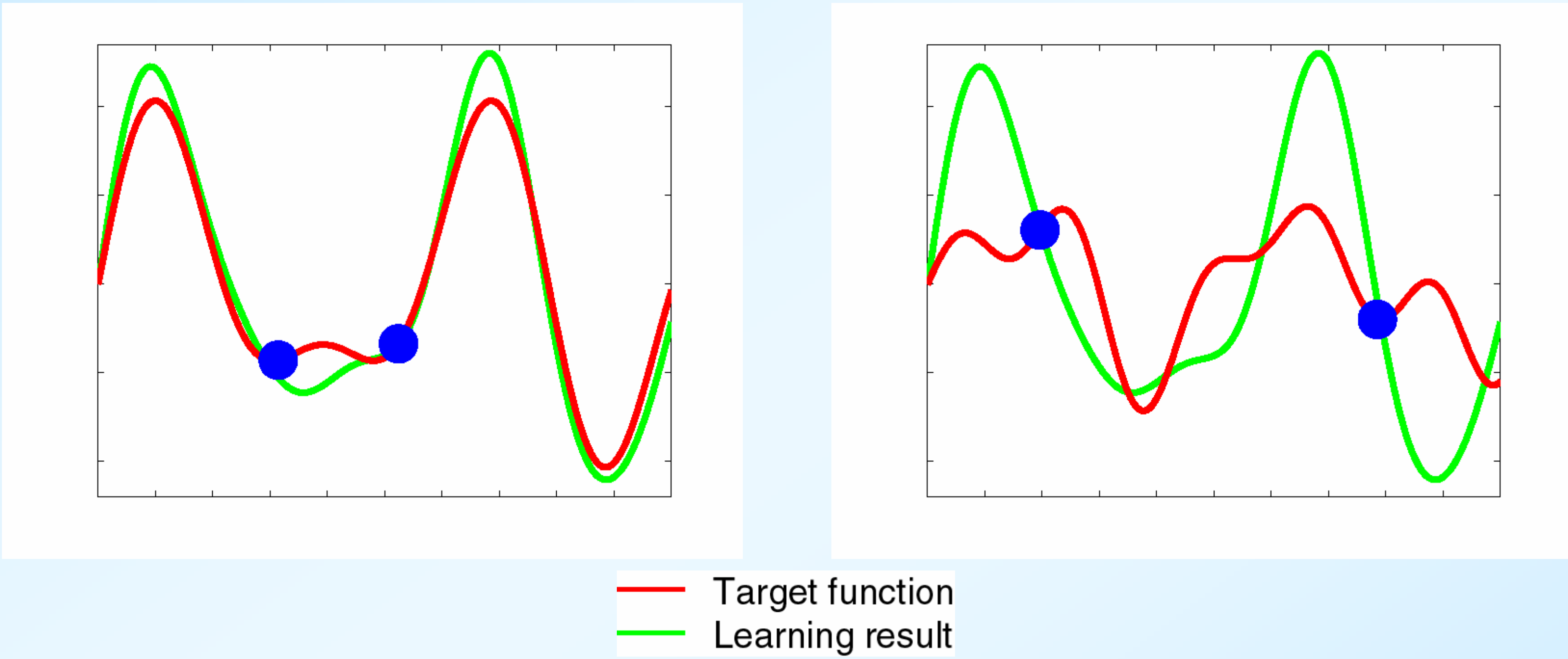
x_m : sample point

y_m : sample value

$$y_m = f(x_m) + n_m$$

Obtain the optimal approximation to $f(x)$ from training examples $\{(x_m, y_m)\}_{m=1}^{M}$

Active Learning (1)



The level of generalization depends heavily on the choice of sample points.

Active Learning (2)

The problem of designing sample points for optimal generalization is called active learning.

- Incremental active learning

  Optimize the next sample point

  (MacKay 1992, Cohn 1994, Fukumizu 1996, Sugiyama and Ogawa 1999)

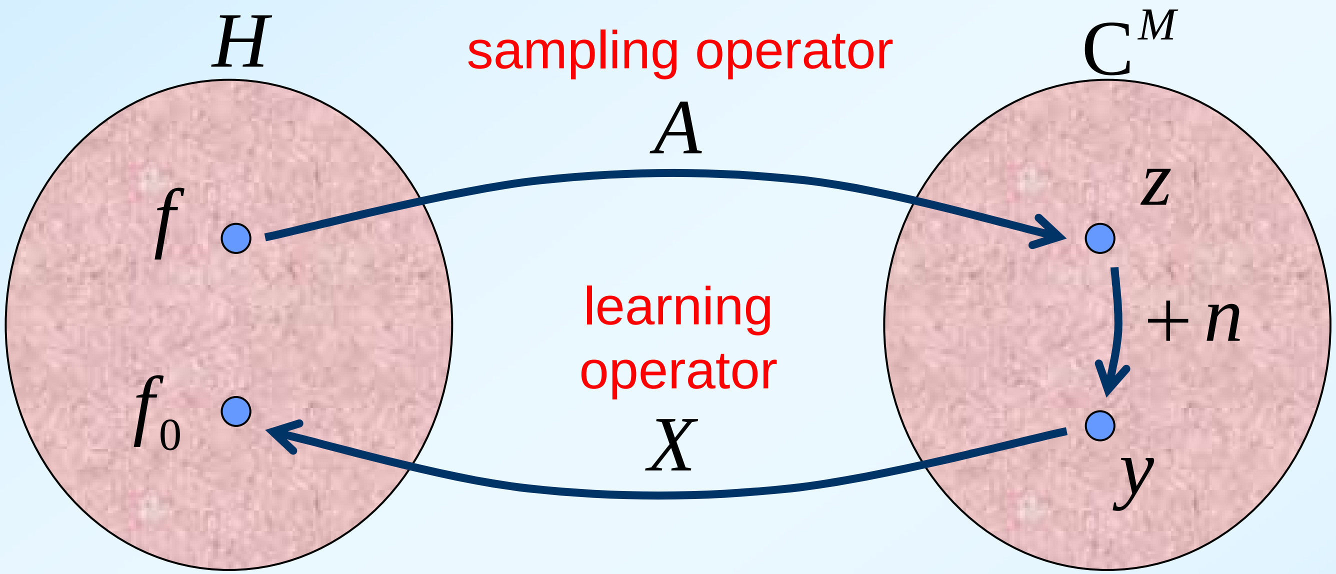
- Batch active learning

  Optimize the set of all sample points

  (Fedorov 1972)

← This research

Supervised Learning As an Inverse Problem



Subspace Information Criterion (SIC)
(Sugiyama and Ogawa 1999)

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

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

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

$(\cdot \otimes \overline{\cdot})$: Schatten product

$$y = Af + n$$

$$f_0 = Xy$$

$$z = \left(f(x_1) \quad f(x_2) \quad \cdots \quad f(x_M) \right)^T$$

$$n = \left(n_1 \quad n_2 \quad \cdots \quad n_M \right)^T$$

$$y = \left(y_1 \quad y_2 \quad \cdots \quad y_M \right)^T$$

Projection Learning

$$f_0 = \underbrace{XAf}_{\text{Signal component}} + \underbrace{Xn}_{\text{Noise component}}$$

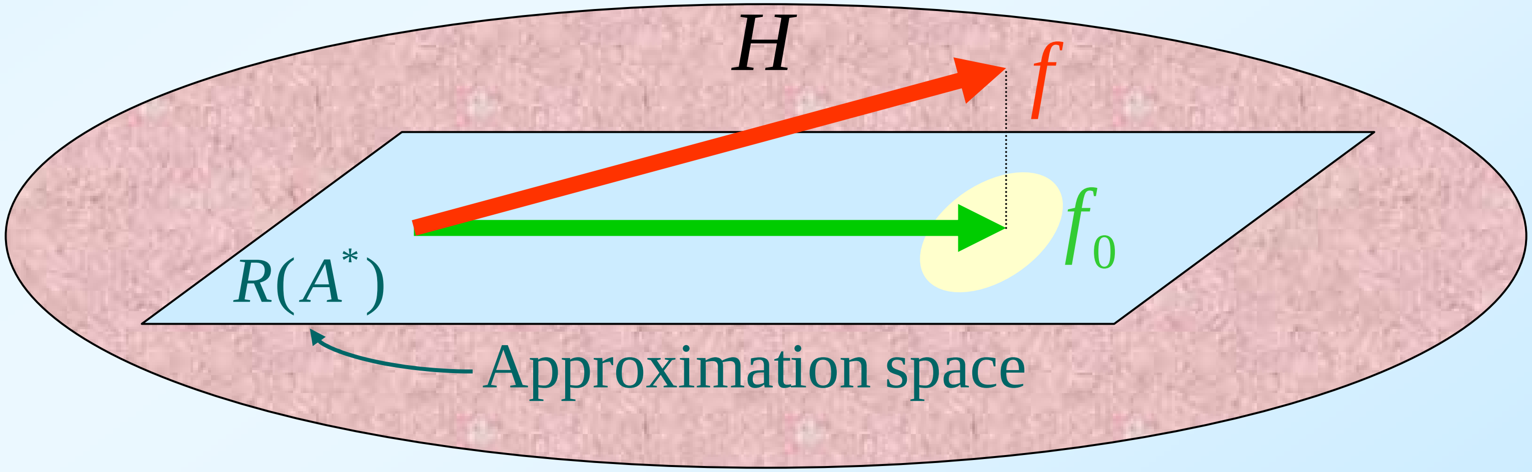
Minimize	$E_n \|Xn\|^2$
under the constraint	$XAf = P_{R(A^*)} f$

E_n : Noise average

A^* : Adjoint operator of A

$R(A^*)$: The range of A^*

$P_{R(A^*)}$: Orthogonal projection onto $R(A^*)$



Projection Learning Operator

We assume that the noise covariance matrix Q is

$$Q = \sigma^2 I.$$

Then, the projection learning operator X is given as

$$X = A^+.$$

A^+ : Moore - Penrose generalized inverse of A

Trigonometric Polynomial Space (1)

Let $x = \left(\xi^{(1)}, \xi^{(2)}, \cdots, \xi^{(L)}\right).$

A function space H is called

a trigonometric polynomial space of order $\mathrm{N} = \left(N_1, N_2, \cdots, N_L\right)$

if H is spanned by

$$\left\{\prod_{\mathrm{l}=1}^{\mathrm{L}} \exp(i n_l \xi^{(l)})\right\}_{n_1=-N_1, n_2=-N_2, \cdots, n_L=-N_L}^{N_1, N_2, \cdots, N_L}$$

and the inner product is defined as

$$\langle f, g \rangle = \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} f(x) \overline{g(x)} d\xi^{(1)} d\xi^{(2)} \cdots d\xi^{(L)}.$$

Trigonometric Polynomial Space (2)

The dimension μ of

a trigonometric polynomial space of order $\mathrm{N} = \left(N_1, N_2, \cdots, N_L\right)$ is

$$\mu = \prod_{\mathrm{l}=1}^{\mathrm{L}} \left(2N_l + 1\right)$$

and the reproducing kernel is

$$K(x, x') = \prod_{l=1}^{L} K_l(\xi^{(l)}, \xi^{(l)'})$$

$$K_l(\xi^{(l)}, \xi^{(l)'}) = \begin{cases} \sin \dfrac{(2N_l + 1)(\xi^{(l)} - \xi^{(l)'})}{2} \Big/ \sin \dfrac{\xi^{(l)} - \xi^{(l)'}}{2} & \text{if } \xi^{(l)} \neq \xi^{(l)'} \\ 2N_l + 1 & \text{if } \xi^{(l)} = \xi^{(l)'} \end{cases}$$

Generalization Measure

$$J_G = E_n \| f_0 - f \|^2$$

$$= \underbrace{\| P_{R(A^*)} f - f \|^2}_{\text{Bias}} + \underbrace{E_n \| A^+ n \|^2}_{\text{Variance}}$$

The bias is zero for all $f \in H$ if and only if $R(A^*) = H$.

Our strategy

Find $\{x_m\}_{m=1}^{M}$ minimizing $E_n \| A^+ n \|^2$

under the constraint $R(A^*) = H$.

Main Theorem

J_G is minimized if and only if

$$A^*A = MI.$$

The minimum value of J_G is $\dfrac{\sigma^2\mu}{M}$.

$A = \sum_{m=1}^{M} (e_m^{(M)} \otimes \overline{\psi_m})$ $\psi_m(x) = K(x, x_m)$ $K(x, x')$: Reproducing kernel

σ^2 : noise variance μ : dimension of H M :# of training examples

$A^*A = MI$ is equivalent to that $\{\frac{1}{\sqrt{\mathrm{M}}}\psi_{\mathrm{m}}\}_{\mathrm{m}=1}^{M}$ forms

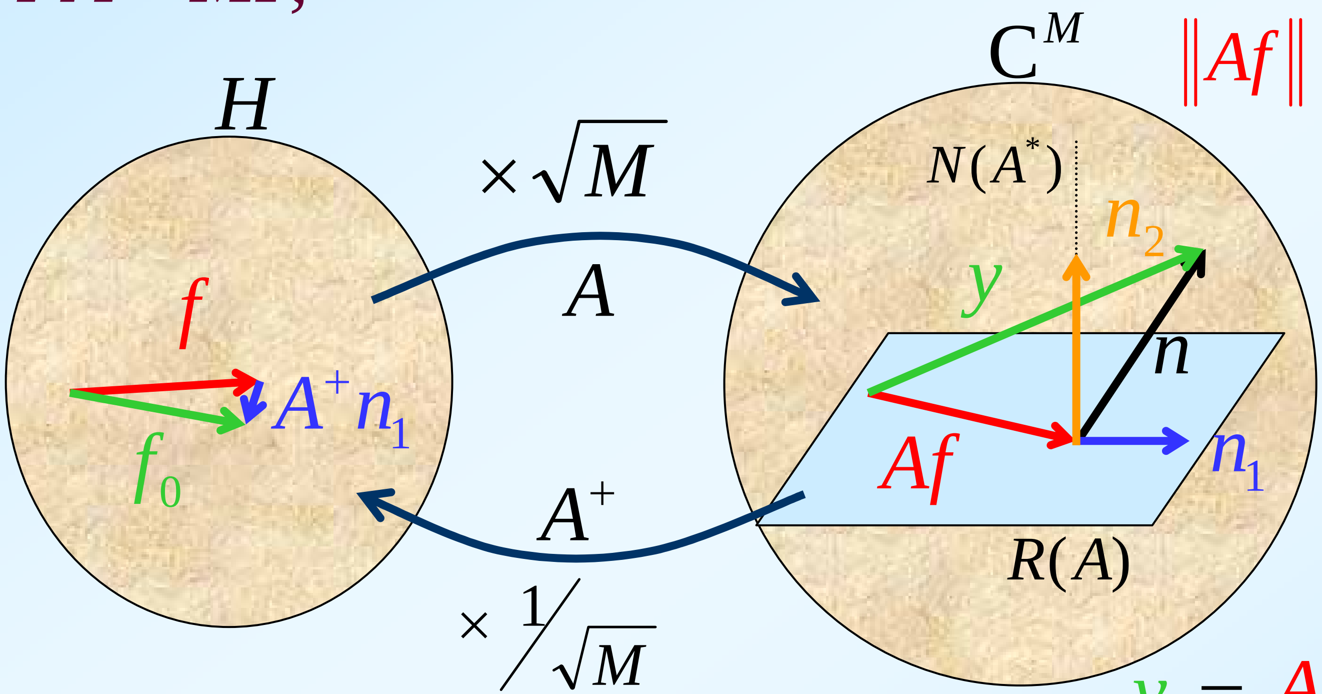
a pseudo orthonormal basis,

which is an extension of orthonormal basis.

Interpretation

When $A^* A = MI$,

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



$$y = Af + n$$
$$= Af + n_1 + n_2$$

$$f_0 = A^+ y = A^+ A f + A^+ n_1 + A^+ n_2$$

$R(A^*) = H$ $\;\;$ $A^+Af \to f$

$$\|A^+ n_1\| = \frac{1}{\sqrt{M}}\|n_1\|$$

$n_2 \in N(A^*)$ $\;\;$ $A^+ n_2 \to 0$

Example of Sample Points (1)

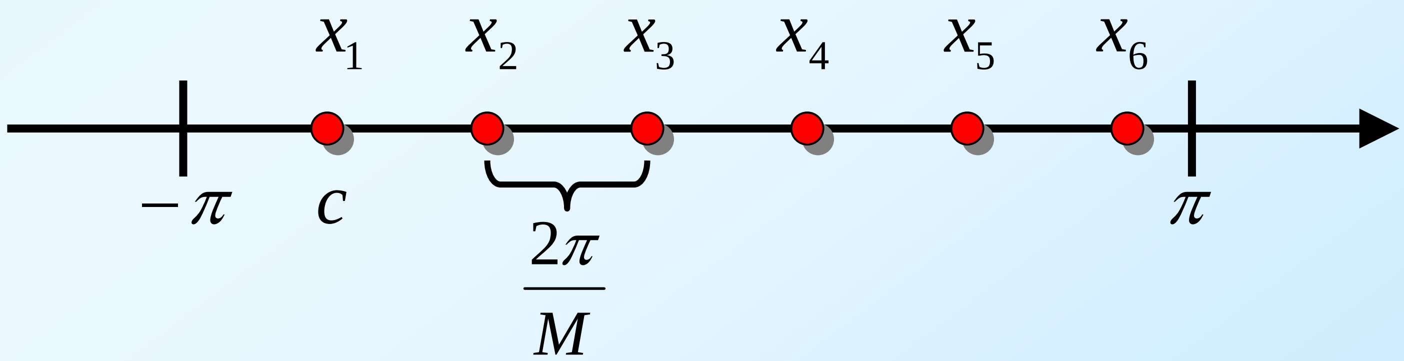
When the dimension of x is 1,

$$M \geq \mu, \qquad c : -\pi \leq c \leq -\pi + {}^{2\pi}\!/_{M}$$

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

$\mu = \dim(H)$

$N = 1,\ \mu = 3,\ M = 6$

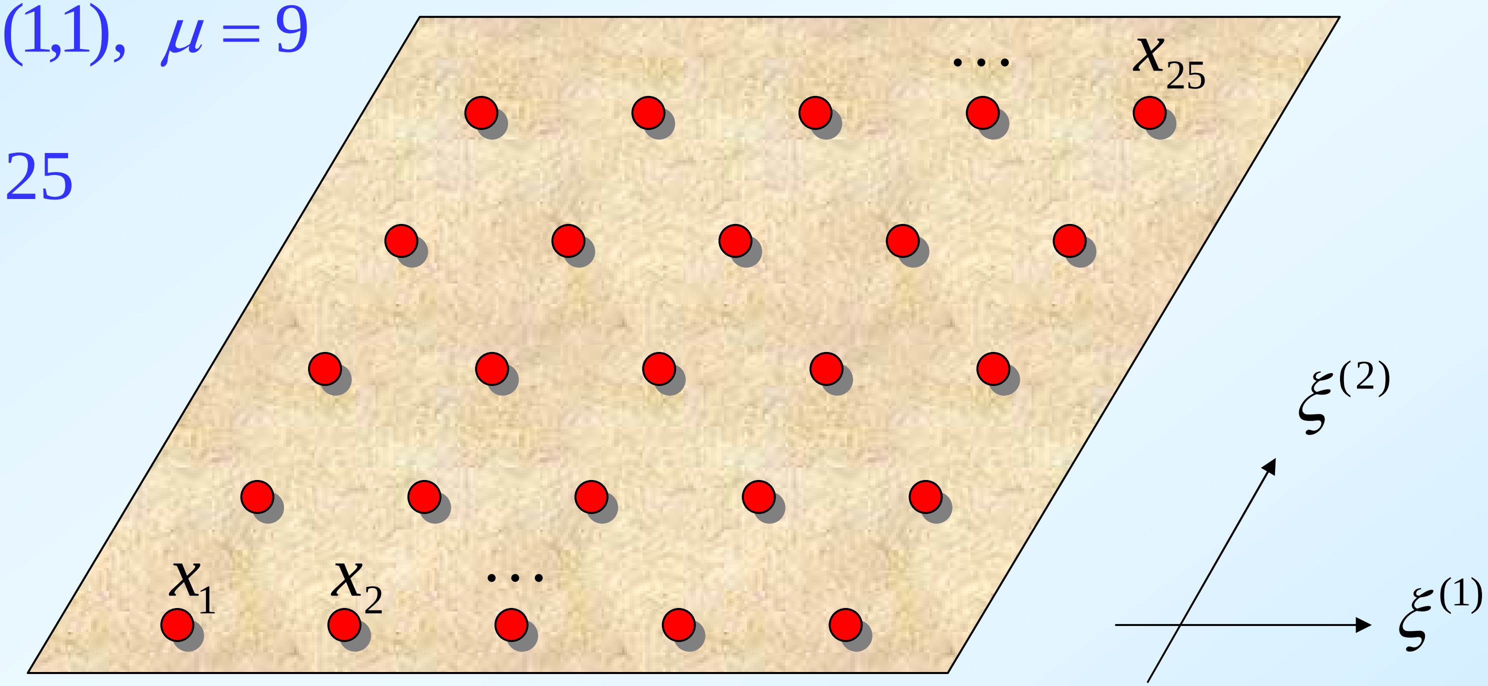


When the dimension of x is 2,

$$x = \left(\xi^{(1)}, \xi^{(2)}\right)$$

$N = (1,1), \ \mu = 9$

$M = 25$



M sample points are fixed to regular intervals in the domain

Example of Sample Points (2)

When the dimension of x is 1,

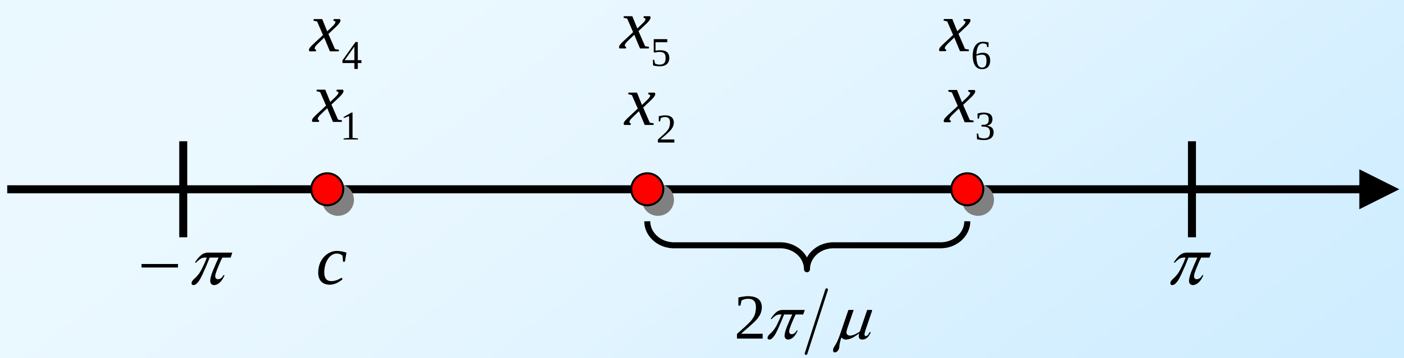
k : a positive integer

$M = k\mu, \qquad c : -\pi \le c \le -\pi + 2\pi/\mu$

$$x_m = c + \frac{2\pi}{\mu} p \; : \; p = m - 1 \pmod{\mu}$$

$\mu = \dim(H)$

$N = 1, \; \mu = 3, \; M = 6$

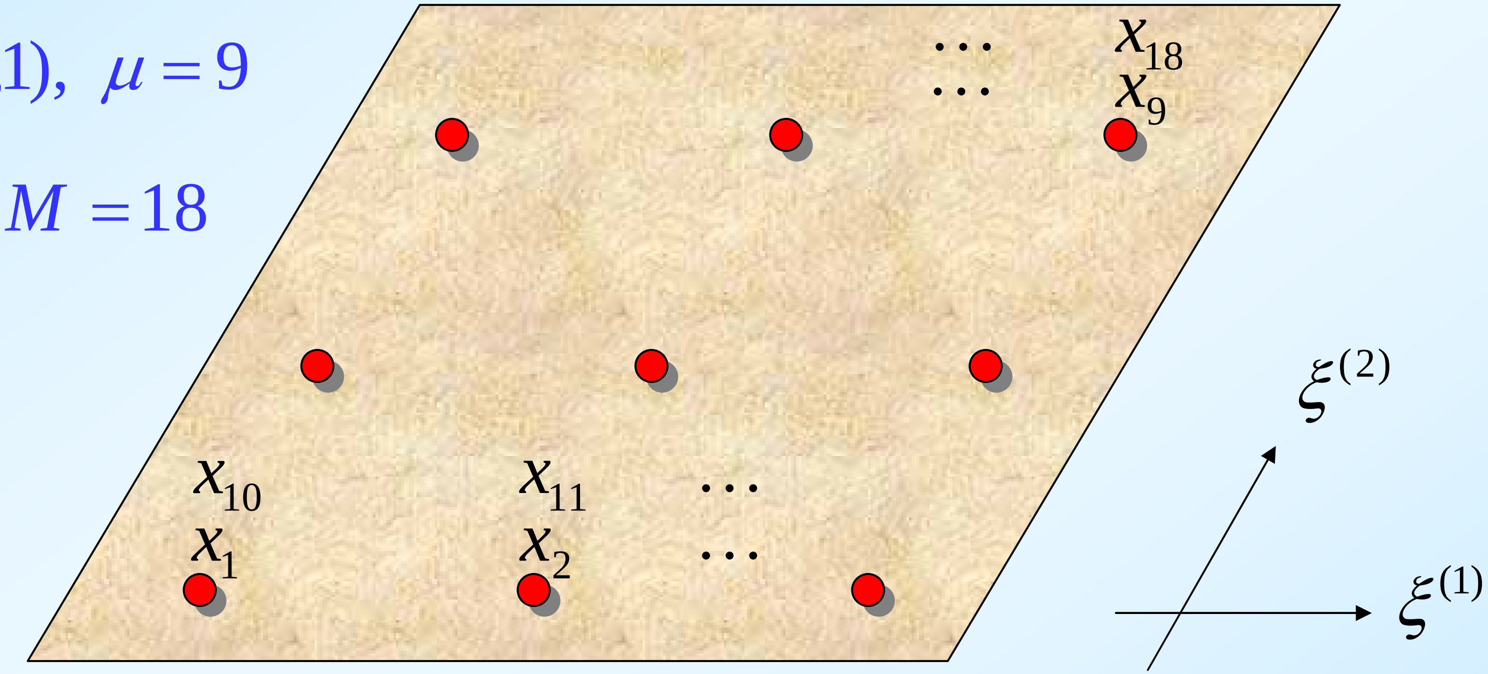


When the dimension of x is 2,

$$x = \left(\xi^{(1)}, \xi^{(2)}\right)$$

$N = (1,1),\ \mu = 9$

$k = 2,\ M = 18$



μ sample points are fixed to regular intervals in the domain

sample values are gathered k times at each point

Calculation of Learning Results

Sample Points	Expression of Learning Result	Computational Complexity	Memory
General	$\sum_{m=1}^{M} \langle y, h_m \rangle \psi_m(x)$	$O(M^2)$	$O(M^2)$
$A^* A = MI$	$\frac{1}{M}\sum_{m=1}^{M} y_m \psi_m(x)$	$O(M)$	$O(M)$
Example (2)	$\frac{1}{\mu}\sum_{p=1}^{\mu} \bar{y}_p \psi_p(x)$	$O(\mu)$	$O(\mu)$ $(M = k\mu)$

Our method provides:

- Optimal generalization
- Complexity reduction
- Memory reduction

M : number of training examples

h_m : m-th column vector of $(AA^*)^+$

$\bar{y}_p$: average of sample values at x_p

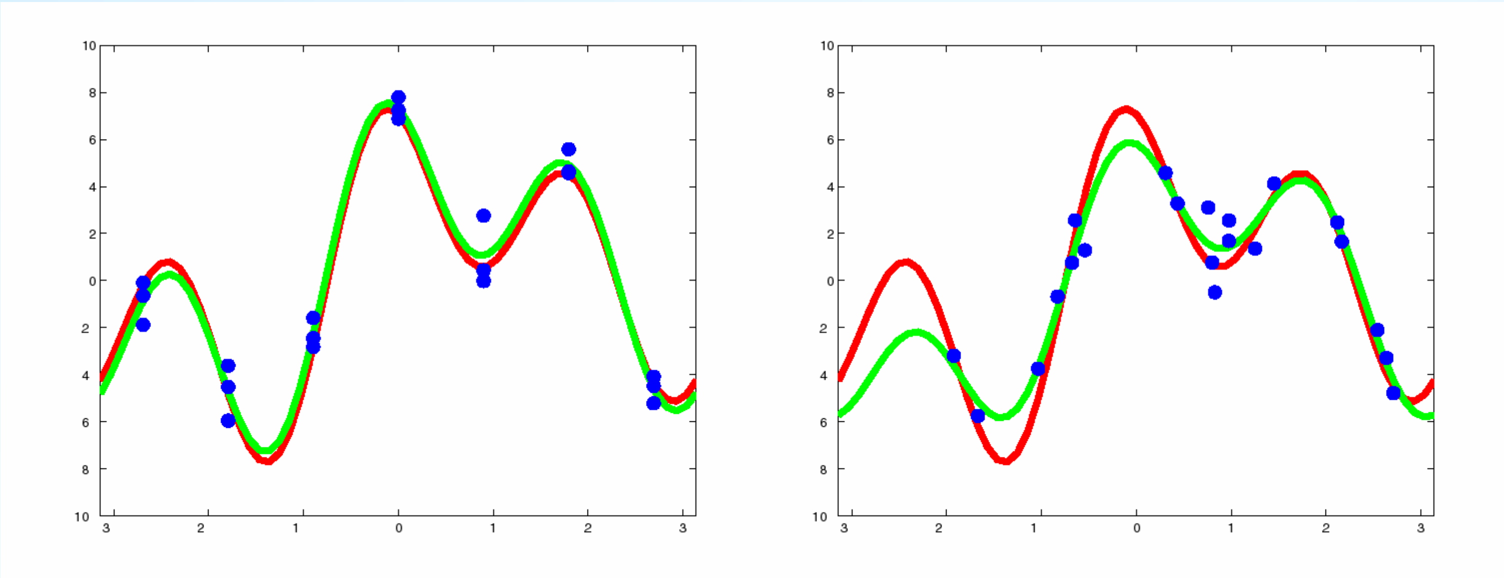
μ : dimension of H

Simulation (1)

H : trigonometric polynomial space of order 3 $(\dim(H) = 7)$

\# of training examples is 21

(a) Optimal sampling (b) Random sampling



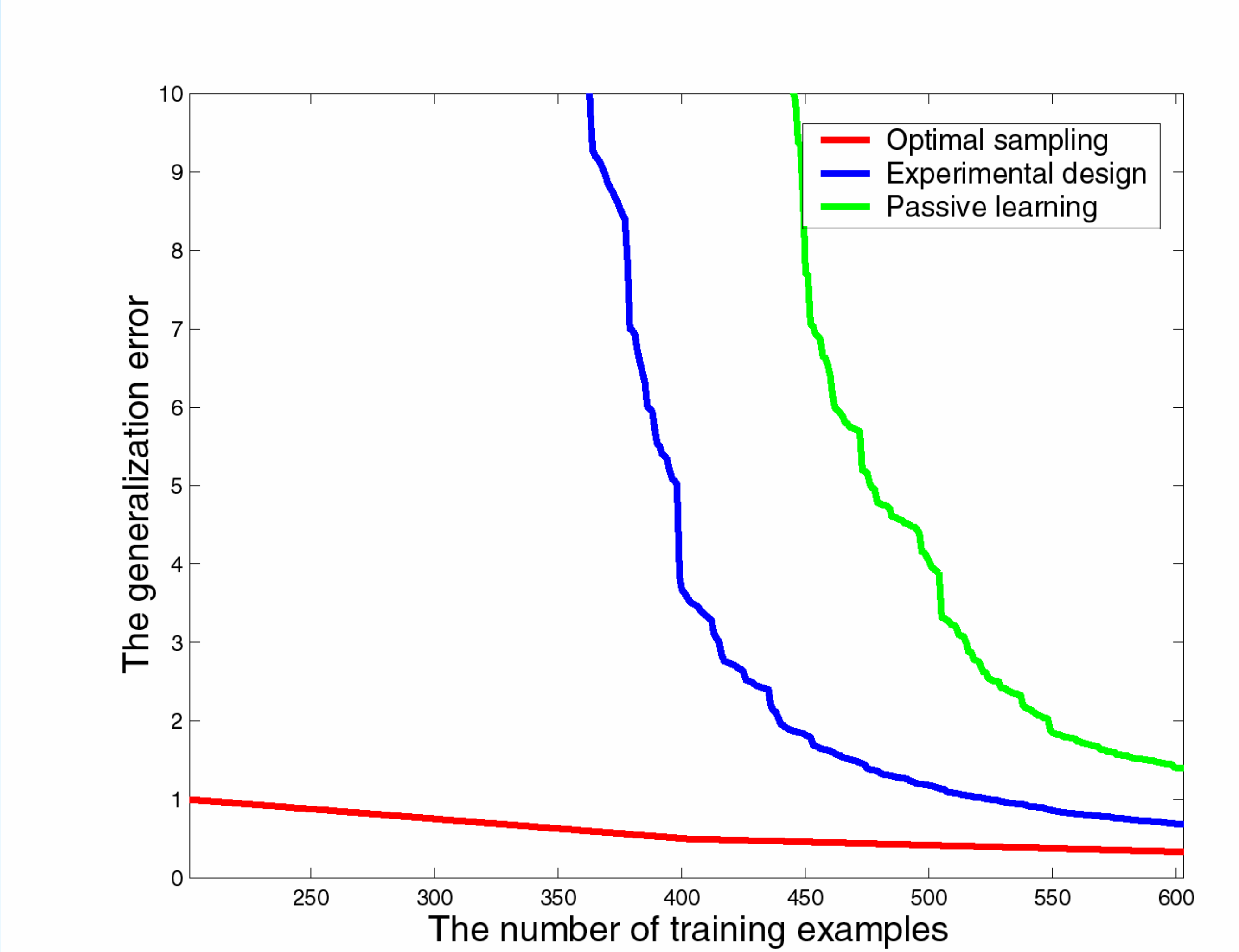
$J_G = 0.333$ $J_G = 1.202$

Target function
Learning result

Our method gives a 72.3% reduction in generalization error

Simulation (2)

H : trigonometric polynomial space of order 100 $(\dim(H) = 201)$



Conclusion

- A necessary and sufficient condition of sample points to provide the optimal generalization capability was given
- The mechanism of achieving the optimal generalization was clarified
- An efficient calculation method of learning results was given