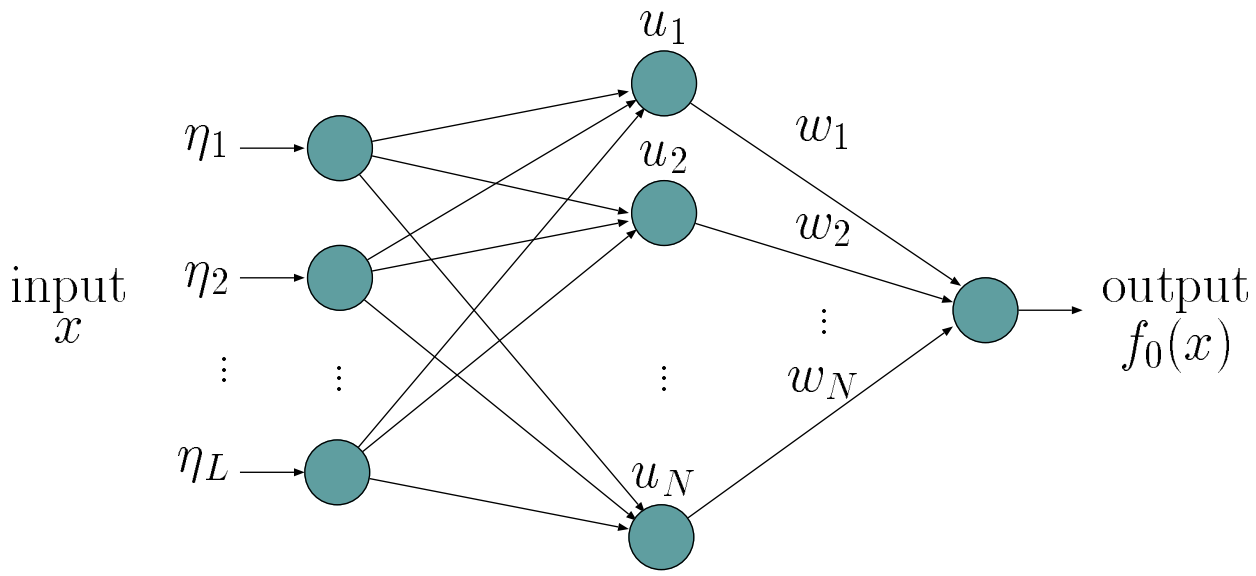
Learning in Neural Networks

Learning in feedforward neural networks with L-dimensional input, N hidden units and scalar output in the presence of noise.

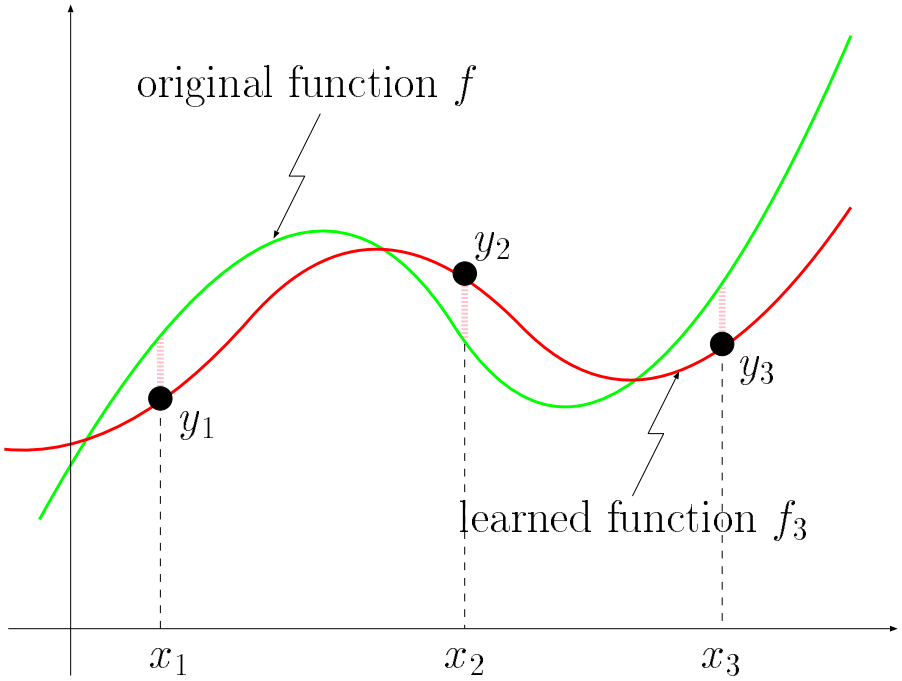


$$f_0(x) = \sum_{i=1}^{n} w_i u_i(x)$$

Networks with the existence of input to hidden layer weights and sigmoidal activation functions in the hidden units can be expressed as special cases of the general activation function $u_i(x)$:

$$u_i(x) = \sigma\left(\sum_{j=1}^{L} w_{ji}\eta_j\right)$$

Learning as a function approximation problem



$$y_i = f(x_i) + n_i$$

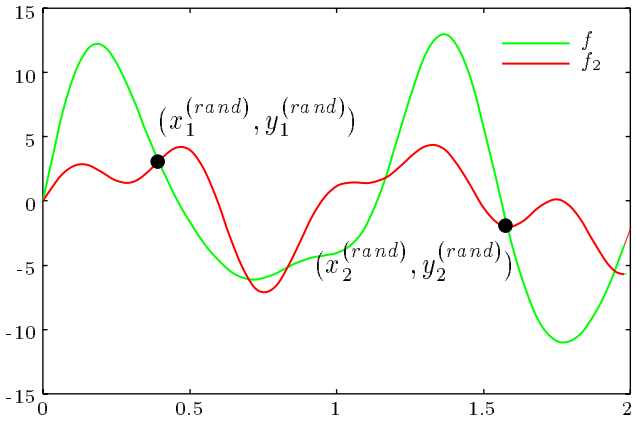
Noisy training data: $\{x_i, y_i\}_{i=1}^{m}$

Learning Problem

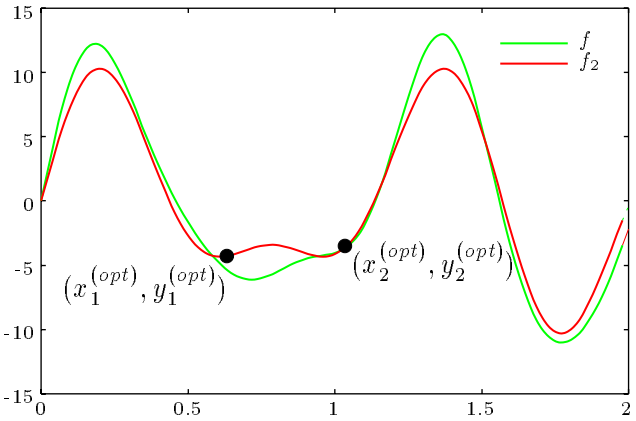
- To design an optimal sampling scheme for selecting training data.
- To construct a NN using the above training data such that it becomes the best approximation to a desired function f(x) based on some criterion.

Active Learning : What and Why ?

Non-optimal sampling



Optimal sampling



original function
learned function

- Learning results are generally highly dependent on the location of the training data.

$\Downarrow$

- Improve generalization ability through selection of optimal training sets.
- So, how do we select *good* training data ?

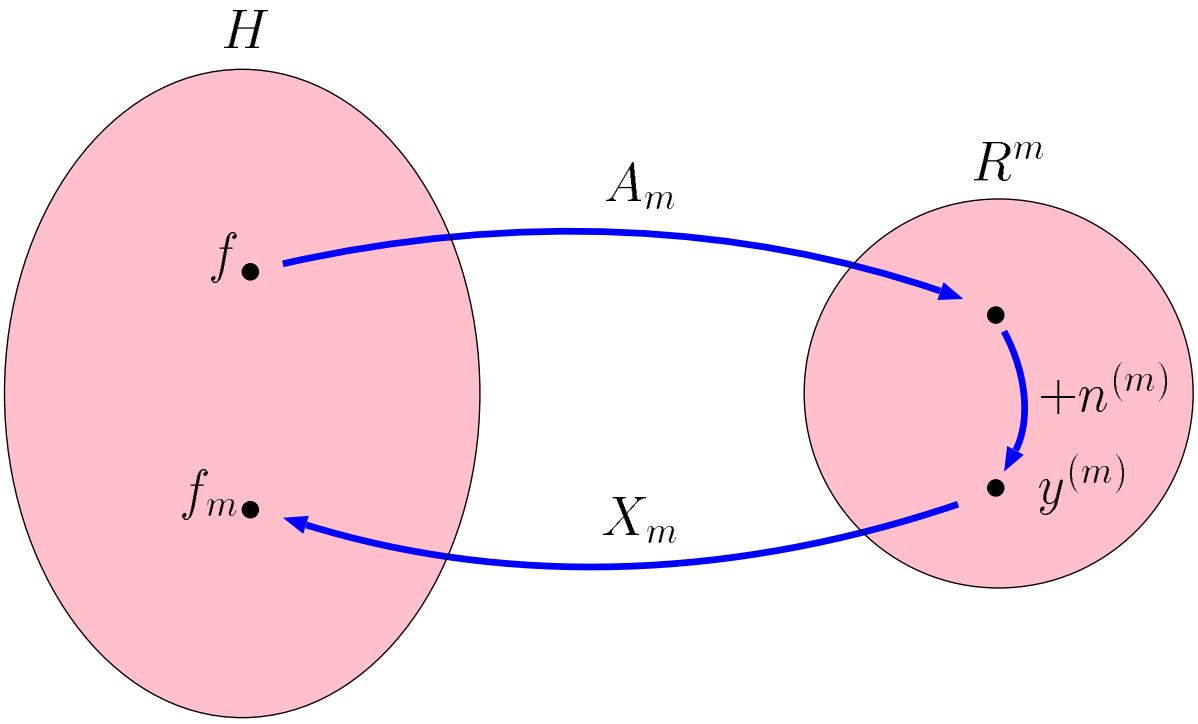
$\Downarrow$

- By utilizing apriori information such as
  1. Apriori knowledge of the function ensemble correlation
  2. Apriori knowledge of the noise characteristics

Inverse problem formalization of NN learning

H : Set of all functions to be approximated by the NN

R^m : Sampled space of dimension m.



Sampling: $y^{(m)} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = A_m f + n^{(m)}$

Learning: $f_m = X_m y^{(m)}$

Mathematical Formulations

$$A_m = \sum_{i=1}^{m} e_i \otimes \overline{\psi_i} \quad \text{(Sampling operator)}$$

$$\psi_i(x) = K(x, x_i) \quad \text{(Sampling function)}$$

$K(x, x_i)$: Reproducing kernel of H

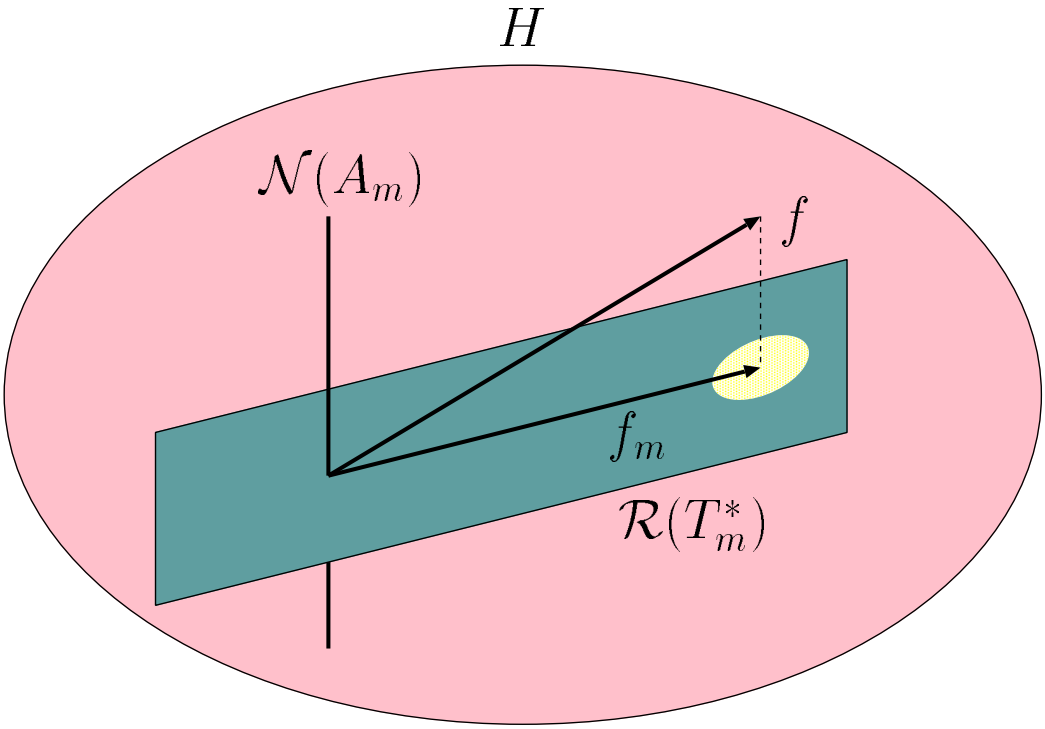
Optimization criterion for learning

Generalization Error Measure & it's S/N decomposition

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

Averaged Projection Criterion (for fixed sampling scheme)

$$\min_{X_m} J_n[X_m] \text{ under the constraint } \min_{X_m} J_s[X_m]$$



Optimal X_m for a fixed sampling scheme

Apriori Knowledge

$R = E_f(f \otimes \overline{f})$ (function ensemble correlation)

$Q_m = E_n(n^{(m)} \otimes \overline{n^{(m)}})$ (noise correlation)

Batch Solution

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

Symbol Definitions 1

$U_m = A_m R A_m^* + Q_m$

$V_m = R^{\frac{1}{2}} A_m^* U_m^{\dagger} A_m R^{\frac{1}{2}}$

$Y_m : R^m \to H$ is an arbitrary operator

Incremental Solution

$$X_{m+1}^{(AP)} = X_m^{(AP)} \Gamma_m^* + (\zeta_{m+1} \otimes \overline{h_{m+1}})$$

$$h_{m+1} = e_{m=1}^{(m+1)} - \Gamma_m^*(U_m^{\dagger} s_{m+1} + U_m^{\dagger} T_m V_m^{\dagger} \xi_{m+1})$$

$$\zeta_{m+1} = \begin{cases} \frac{R^{\frac{1}{2}} V_m^{\dagger} \xi_{m+1}}{\alpha_{m+1} + (\xi_{m+1}, V_m^{\dagger} \xi_{m+1})} & \text{when } \phi_{m+1} \in \mathcal{R}(T_m^*) \\ \frac{R^{\frac{1}{2}} \tilde{\phi}_{m+1}}{\|\tilde{\phi}_{m+1}\|^2} & \text{otherwise} \end{cases}$$

Symbol Definitions 2

$T_m = A_m R^{\frac{1}{2}}$ $\quad \Gamma_m = \Sigma_{n=1}^m (e_n^{(m+1)} \otimes \overline{e_n^{(m)}})$

$\phi_{m+1} = R^{\frac{1}{2}} \psi_{m+1}$ $\quad \tilde{\phi}_{m+1} = P_{\mathcal{N}(T_m)} \phi_{m+1}$

$\xi_{m+1} = \phi_{m+1} - T_m^* U_m^{\dagger} s_{m+1}$ $\quad q_{m+1} = E_n(n_{m+1} n^{(m)})$

$s_{m+1} = T_m \phi_{m+1} + q_{m+1}$ $\quad \tau_{m+1} = E_n(n_{m+1}^2)$

$\alpha_{m+1} = \|\phi_{m+1}\|^2 + \tau_{m+1} - (U_m^{\dagger} s_{m+1}, s_{m+1})$

Optimal training data selection : Active learning

- Batch selection of 'm' optimal data points from optimal generalization perspective

$$\Downarrow$$

$$\min_{x_1,\cdots,x_m} E_f \| X_m^{(AP)} A_m f - f \|^2$$

OBJECTIVE 1

- Incremental selection of additional training data to reduce noise variance

$$\Downarrow$$

$$\min_{x_{m+1}} (E_n \| X_{m+1}^{(AP)} n^{(m+1)} \|^2 - E_n \| X_m^{(AP)} n^{(m)} \|^2)$$

OBJECTIVE 2

Active learning for optimal generalization (Objective 1)

Theorem: Optimal training data selection (Obj 1)

The selected training data is optimal if and only if it satisfies the condition 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.

- Select $\{x_i\}_{i=1}^m$ such that

  $\mathcal{R}(R^{\frac{1}{2}}A_m^*)$ is the K-dimensional maximal variance subspace of R, where $K = dim(\mathcal{R}(R^{\frac{1}{2}}A_m^*))$.

$$\Downarrow$$

$$\mathcal{L}(\{R^{\frac{1}{2}}\psi_i\}_{i=1}^m) = \mathcal{L}(\{\varphi_j\}_{j=1}^K) : \lambda_1 \geq \lambda_2 \geq \cdots$$

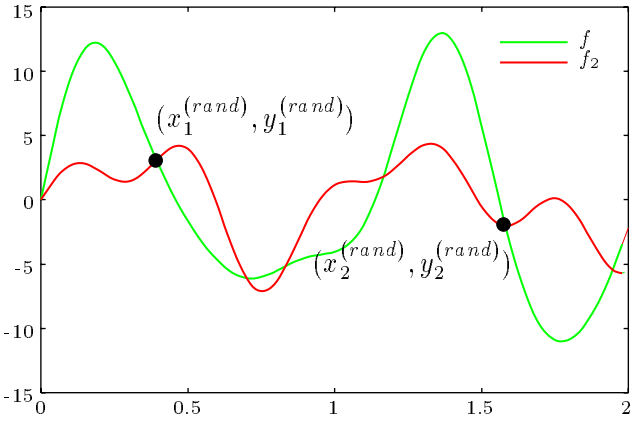
Empirical evaluations through artificial example –Part 1

$$H = \mathcal{L}\{\sin 6x, \sin 10x, \sin 15x\}$$
$$f = 9\sin 6x + 4\sin 10x + \sin 15x$$

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

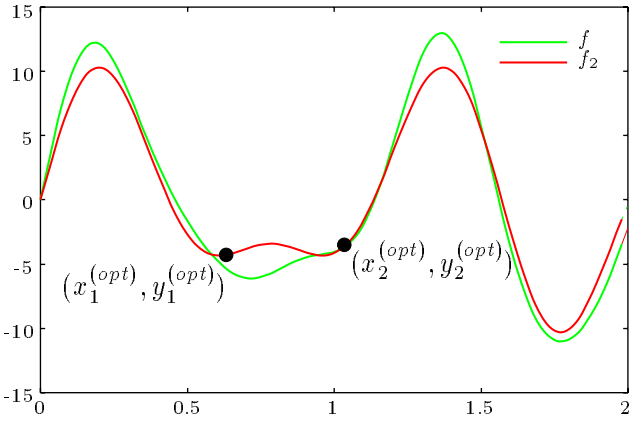
Eigenvalue	Eigenfunction
$\lambda_1^{(R)} = 9$	$\varphi_1^{(R)} = (1 \;\; 0 \;\; 0)^T$
$\lambda_2^{(R)} = 4$	$\varphi_2^{(R)} = (0 \;\; 1 \;\; 0)^T$
$\lambda_3^{(R)} = 1$	$\varphi_3^{(R)} = (0 \;\; 0 \;\; 1)^T$

Non-optimal training data



$$\| f_2^{(rand)} - f \|^2 = 60.74$$

Optimal training data



$$\| f_2^{(opt)} - f \|^2 = 3.91$$

Non-optimal scheme $\mathcal{L}(\{R^{\frac{1}{2}}\psi_i\}_{i=1}^2) \neq \mathcal{L}(\{\varphi_j^{(R)}\}_{j=1}^2)$
$x_1^{(rand)} = \frac{\pi}{8}$, $\psi_1^{(rand)} = \begin{pmatrix} 0.71 \\ -0.71 \\ -0.38 \end{pmatrix}$
$x_2^{(rand)} = \frac{\pi}{2}$, $\psi_2^{(rand)} = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix}$

Optimal scheme $\mathcal{L}(\{R^{\frac{1}{2}}\psi_i\}_{i=1}^2) = \mathcal{L}(\{\varphi_j^{(R)}\}_{j=1}^2)$
$x_1^{(opt)} = \frac{\pi}{5}$, $\psi_1^{(opt)} = \begin{pmatrix} -0.59 \\ 0 \\ 0 \end{pmatrix}$
$x_2^{(opt)} = \frac{\pi}{3}$, $\psi_2^{(opt)} = \begin{pmatrix} 0 \\ -0.87 \\ 0 \end{pmatrix}$

Active learning for noise variance reduction (Objective 2)

Change in noise variance with new training data

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

(a) $R^{\frac{1}{2}} \psi_{m+1} \notin \mathcal{R}(R^{\frac{1}{2}} A_m^*)$

$J_{m+1}^{(n)} \geq 0$ $\cdots$ (increase in noise variance)

(b) $R^{\frac{1}{2}} \psi_{m+1} \in \mathcal{R}(R^{\frac{1}{2}} A_m^*)$

$J_{m+1}^{(n)} \leq 0$ $\cdots$ (decrease in noise variance)

Hence, we should select training data satisfying condition (b) for Objective 2.

In case (b),

$$E_f \| X_{m+1}^{(AP)} A_{m+1} f - f \|^2 - E_f \| X_m^{(AP)} A_m f - f \|^2 = 0$$

$\Downarrow$

Maintains generalization ability
while reducing noise variance

Active learning for noise variance reduction (Objective 2)

Using case (b), the optimal 'next' training location which causes maximum noise variance reduction is determined as follows:

When $\cdots$

- the **noise variance** of new training data is **positive** and
- **noise** on new data is **uncorrelated** to data sampled so far,

Then $\cdots$

Theorem: Optimal training data selection (Obj 2)

Select the next training data x_{m+1} to satisfy

$$\psi_{m+1} = c \cdot \varphi_1$$

where φ_1 is the eigenfunction corresponding to the largest eigenvalue of $X_m^{(AP)} Q_m X_m^{(AP)*}$

Empirical evaluations through artificial example – Part 2

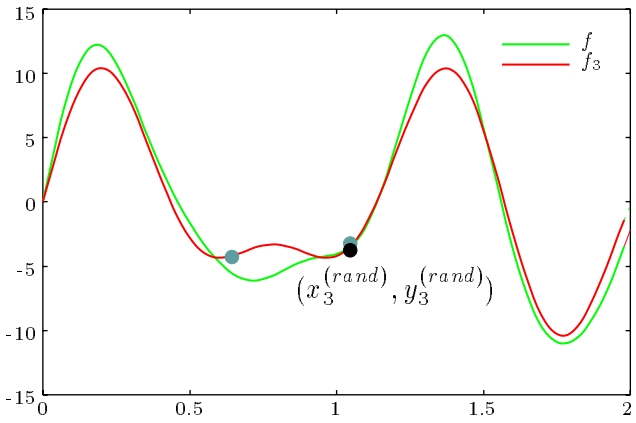
$(x_1^{(opt)}, y_1^{(opt)})$
$(x_2^{(opt)}, y_2^{(opt)})$
} **Use training data selected under the optimal generalization sampling scheme**

To select the third location for optimal noise variance reduction···

$$B_2 = X_2^{(AP)} Q_2 X_2^{(AP)^*}$$

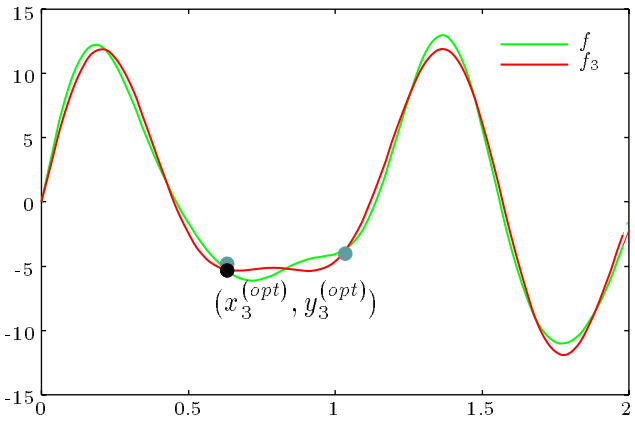
Eigenvalue	Eigenfunction
$\lambda_1^{(B)} = 2.89$	$\varphi_1^{(B)} = (1 \quad 0 \quad 0)^T$
$\lambda_2^{(B)} = 1.33$	$\varphi_2^{(B)} = (0 \quad 1 \quad 0)^T$

Non-optimal training data



noise variance reduction=-1.33

Optimal training data



noise variance reduction=-2.89

Non-optimal scheme

$\psi_3 \neq c \cdot \varphi_1^{(B)}$

$$x_3^{(rand)} = \frac{\pi}{3} \quad , \quad \psi_3^{(rand)} = \begin{pmatrix} 0 \\ -0.87 \\ 0 \end{pmatrix}$$

Optimal scheme

$\psi_3 = c \cdot \varphi_1^{(B)}$

$$x_3^{(opt)} = \frac{\pi}{5} \quad , \quad \psi_1^{(opt)} = \begin{pmatrix} -0.59 \\ 0 \\ 0 \end{pmatrix}$$

Comparison of learning results for Optimal / Non–optimal sampling schemes in high dimensional learning problems

Experimental Parameters

$$
\begin{aligned}
H &= \mathcal{L}(\{\frac{10}{\pi}\mathrm{sinc}(\frac{10}{\pi}x - i)\}_{i=-\infty}^{\infty}) \\
&: \text{Band-limited Paley-Wiener space} \\
K(x, x_i) &= \psi_i(x) = \frac{10}{\pi}\mathrm{sinc}(\frac{10}{\pi}(x - x_i)) \\
&: \text{Reproducing kernel of the function space } H
\end{aligned}
$$

Experimental error monitor parameters

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

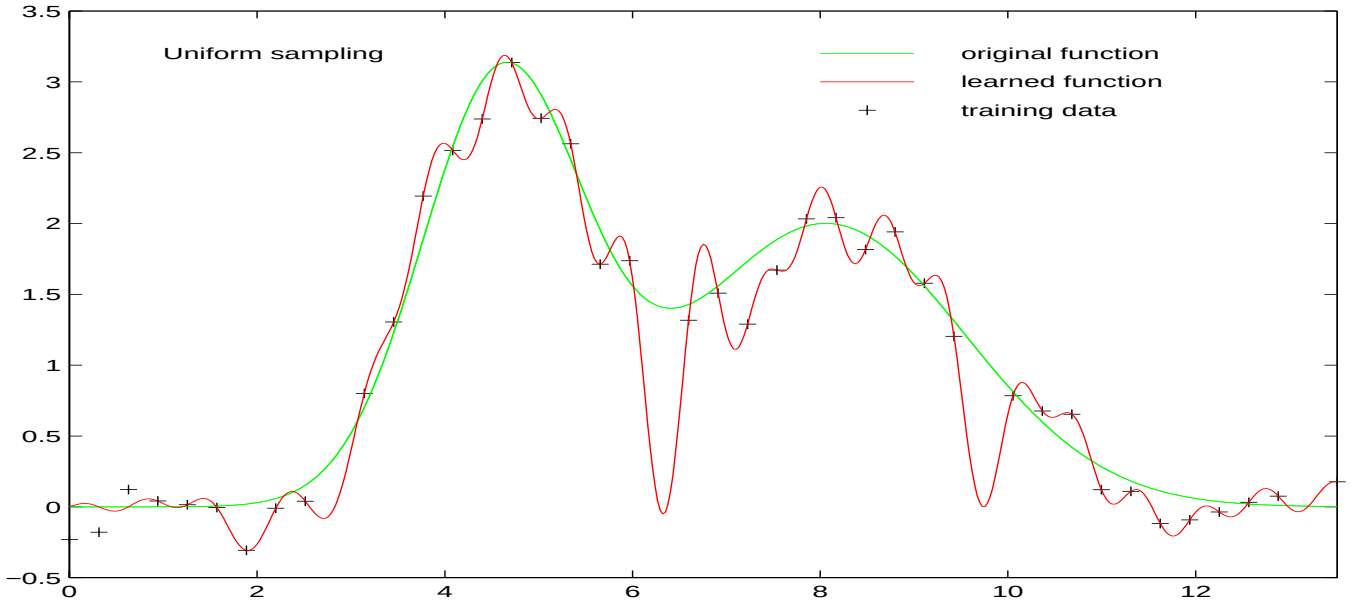
Gen.Error (J_{gen}) = signal bias error (J_s) + noise variance (J_n)

Experimental Objective

- Select 40 training data to optimally reduce the **signal bias error**.
- Use the previously selected data and additionally, select 10 more data points to maximally reduce **noise variance**.

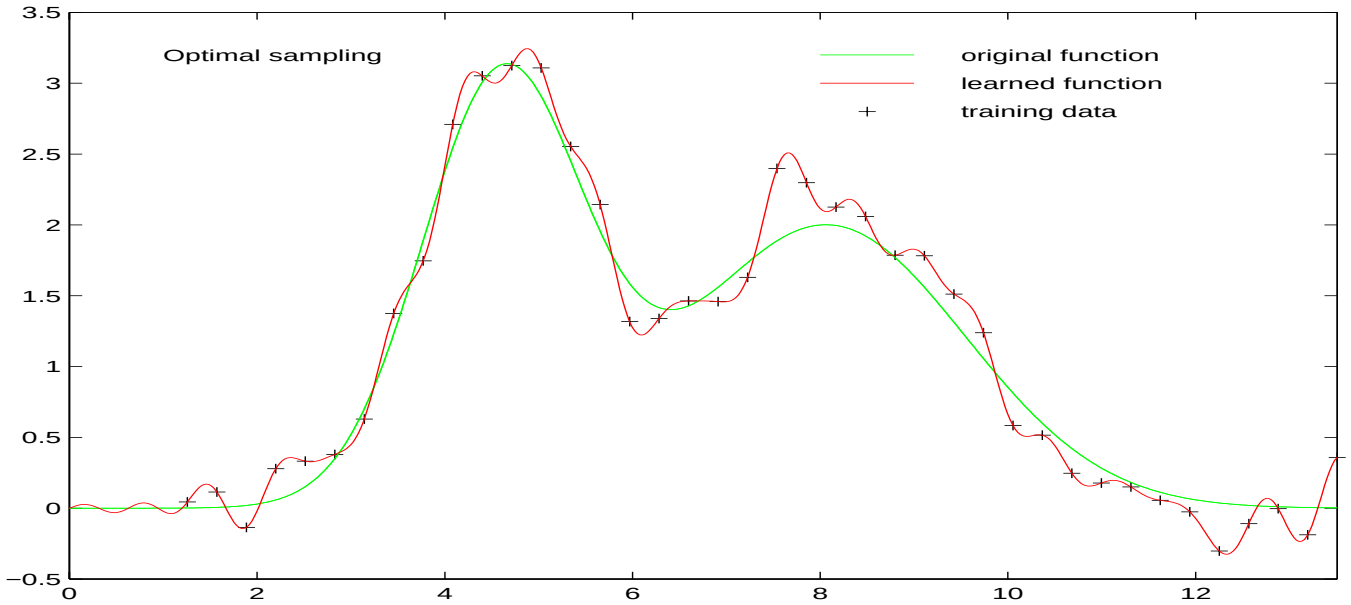
Batch selection for optimal generalization

Learning with uniformly sampled(**non-optimal**) training data



Gen. Error $= J_s + J_n = 12.73 + 2.17 =$**14.90**

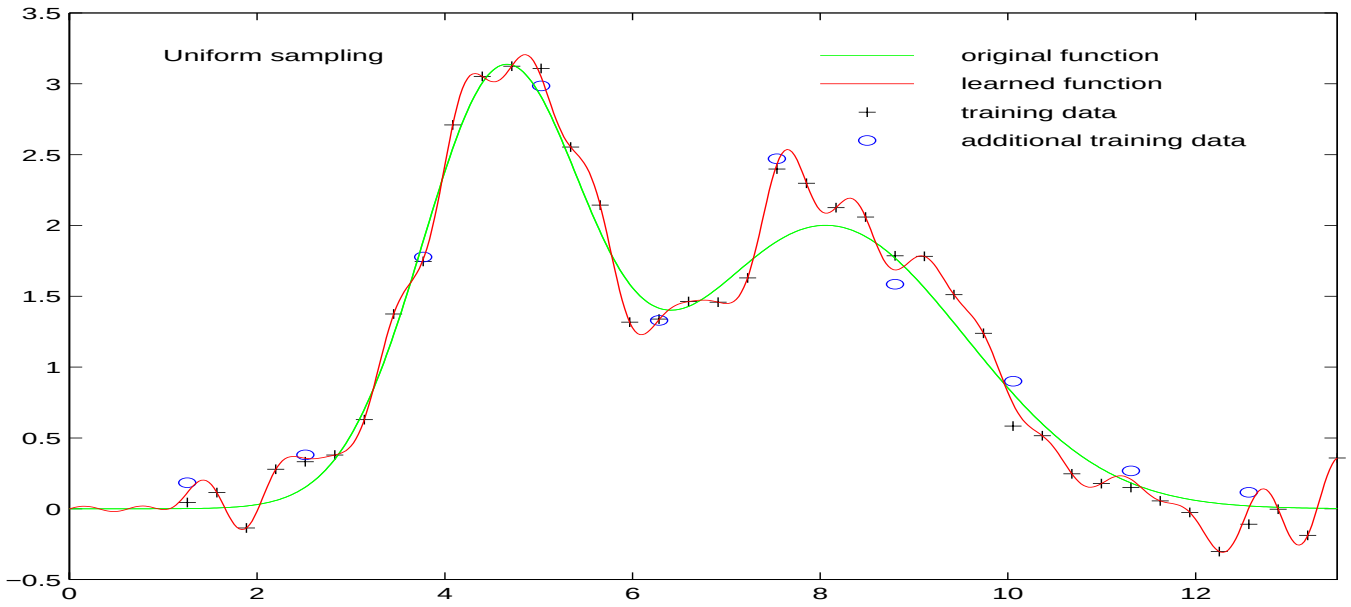
Learning with **optimally** sampled training data



Gen. Error $= J_s + J_n = 3.18 + 2.36 =$**5.54**

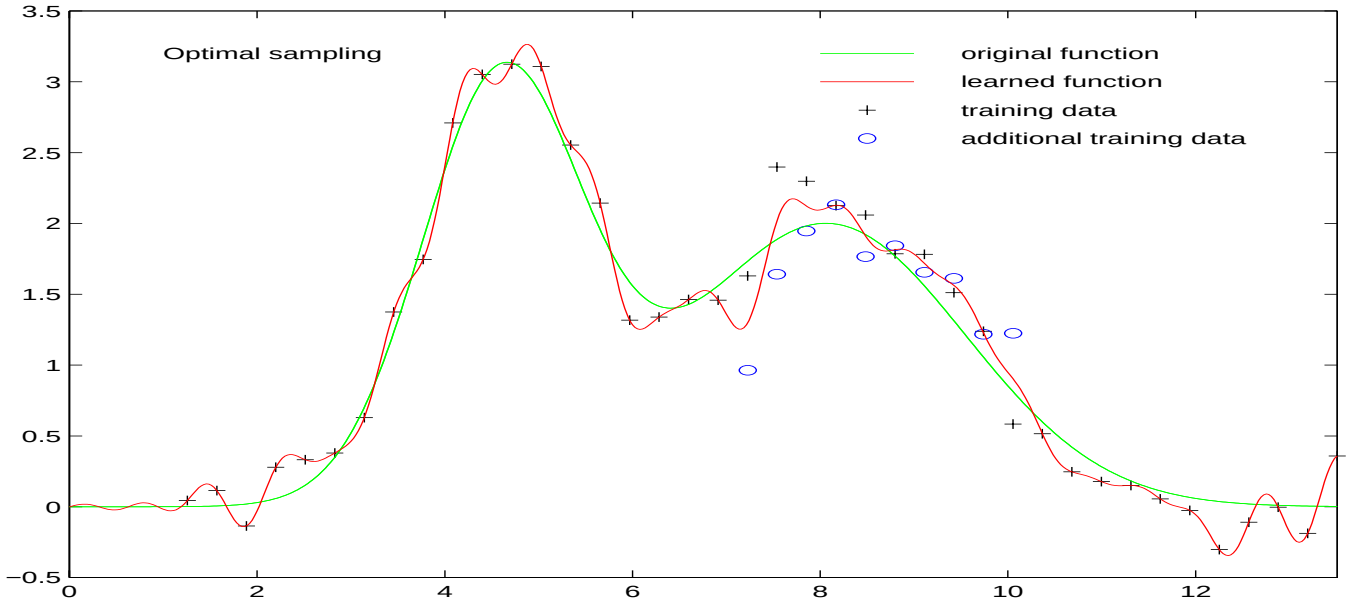
Incremental selection for noise variance reduction

Learning with uniformly sampled(**non-optimal**) additional data



Gen. Error $= J_s + J_n = 3.18 + 2.05 =$**5.23**

Learning with **optimally** sampled additional data



Gen. Error $= J_s + J_n = 3.18 + 1.89 =$**5.07**

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 described 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.