Pattern Information Processing: Properties of Least-Squares

Masashi Sugiyama

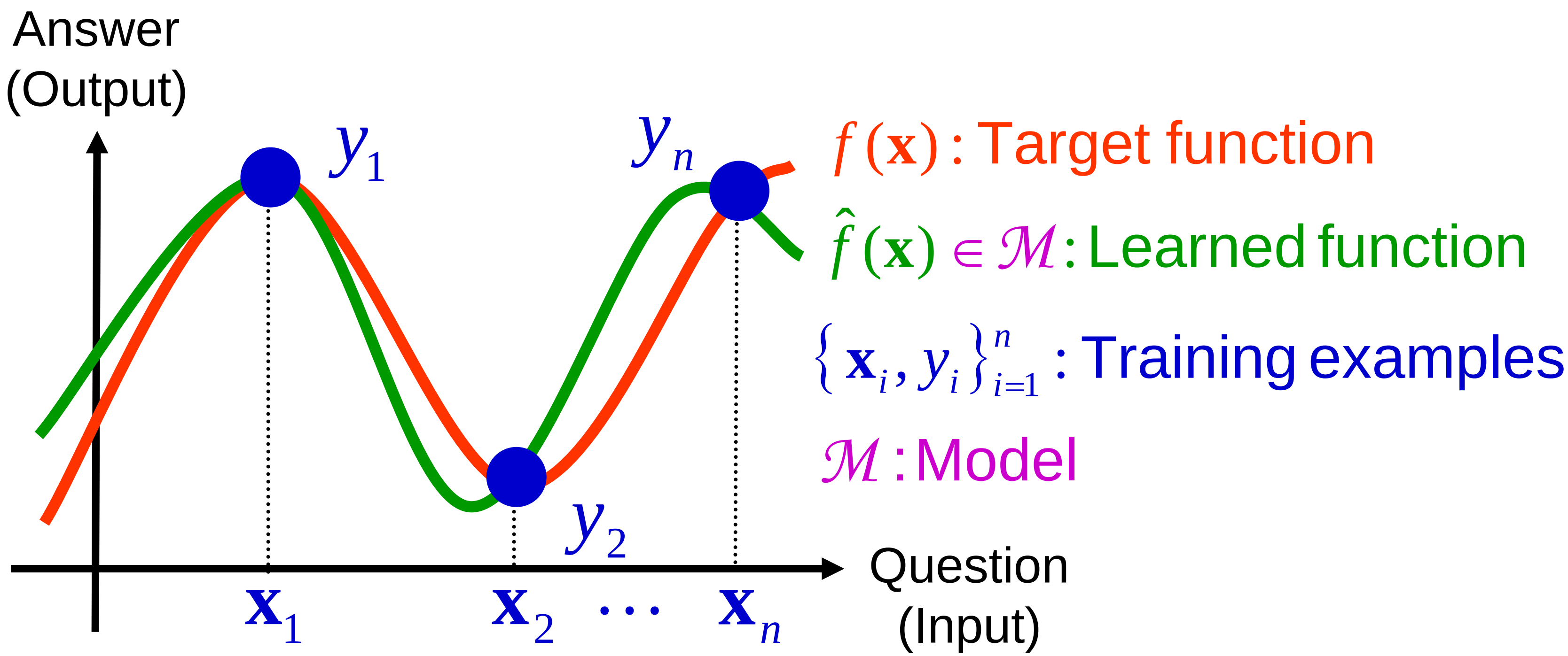
(Department of Computer Science)

Contact: W8E-505

sugi@cs.titech.ac.jp

http://sugiyama-www.cs.titech.ac.jp/~sugi/

Supervised Learning As Function Approximation



Using training examples $\{\mathbf{x}_i, y_i\}_{i=1}^n$,
find a function $\hat{f}(\mathbf{x})$ from a model $\mathcal{M}$
that well approximates the target function $f(\mathbf{x})$.

Assumptions

- Training examples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$
  - Training inputs $\boldsymbol{x}_i$: i.i.d. from a probability distribution with density $q(\boldsymbol{x})$
  - Training outputs y_i : additive noise contained

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i$$

  - Output noise ϵ_i : i.i.d. with mean zero

$$\mathbb{E}_{\boldsymbol{\epsilon}}[\epsilon_i] = 0 \qquad \mathbb{E}_{\boldsymbol{\epsilon}}[\epsilon_i \epsilon_j] = \begin{cases} \sigma^2 & (i = j) \\ 0 & (i \neq j) \end{cases}$$

Reviews

- Linear models / kernel models:

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{b} \alpha_i \varphi_i(\boldsymbol{x})$$

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i)$$

- Least-squares learning:

$$\hat{\boldsymbol{\alpha}}_{LS} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}}\, J_{LS}(\boldsymbol{\alpha})$$

$$J_{LS}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \left(\hat{f}(\boldsymbol{x}_i) - y_i \right)^2$$

Today's Plan

- How does LS contribute to reducing the generalization error?

$$G = \int_{\mathcal{D}} \left(\hat{f}(\boldsymbol{t}) - f(\boldsymbol{t}) \right)^2 q(\boldsymbol{t}) d\boldsymbol{t}$$

- Justification of LS for linear models:
  - Realizable cases
  - Unrealizable cases

Realizability

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{b} \alpha_i \varphi_i(\boldsymbol{x})$$

- Realizable: Learning target function $f(\boldsymbol{x})$ can be expressed by the model, i.e., there exists a parameter vector $\boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*, \ldots \alpha_b^*)^\top$ such that

$$f(\boldsymbol{x}) = \sum_{i=1}^{b} \alpha_i^* \varphi_i(\boldsymbol{x})$$

- Unrealizable: $f(\boldsymbol{x})$ is not realizable

Justification in Realizable Cases

- In realizable cases, generalization error is expressed as

$$G = \int_{\mathcal{D}} \left(\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x}) \right)^2 q(\boldsymbol{x}) d\boldsymbol{x}$$

$$= \|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_{\boldsymbol{U}}^2$$

$$\|\boldsymbol{\alpha}\|_{\boldsymbol{U}}^2 = \langle \boldsymbol{U}\boldsymbol{\alpha}, \boldsymbol{\alpha} \rangle$$

$$U_{i,j} = \int_{\mathcal{D}} \varphi_i(\boldsymbol{x}) \varphi_j(\boldsymbol{x}) q(\boldsymbol{x}) d\boldsymbol{x}$$

Bias/Variance Decomposition

$\mathbb{E}_{\boldsymbol{\epsilon}}$:Expectation over noise

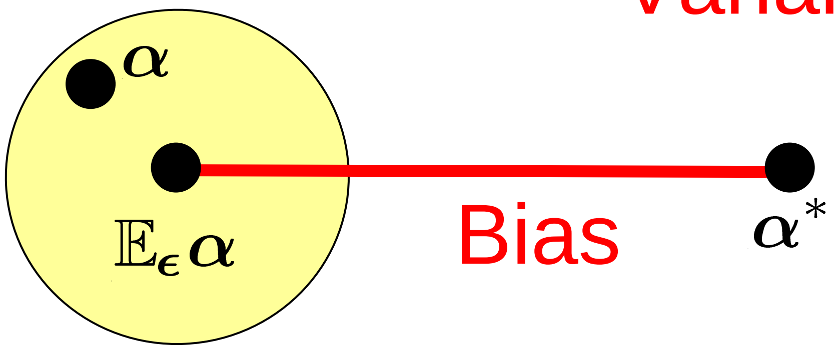
- Expected generalization error:

$$\mathbb{E}_{\boldsymbol{\epsilon}}[G] = \mathbb{E}_{\boldsymbol{\epsilon}}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_{\boldsymbol{U}}^2$$

$$= \mathbb{E}_{\boldsymbol{\epsilon}}\|\boldsymbol{\alpha} - \mathbb{E}_{\boldsymbol{\epsilon}}\boldsymbol{\alpha} + \mathbb{E}_{\boldsymbol{\epsilon}}\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_{\boldsymbol{U}}^2$$

$$= \underbrace{\mathbb{E}_{\boldsymbol{\epsilon}}\|\boldsymbol{\alpha} - \mathbb{E}_{\boldsymbol{\epsilon}}\boldsymbol{\alpha}\|_{\boldsymbol{U}}^2}_{\text{Variance}} + \underbrace{\|\mathbb{E}_{\boldsymbol{\epsilon}}\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_{\boldsymbol{U}}^2}_{\text{Bias}}$$

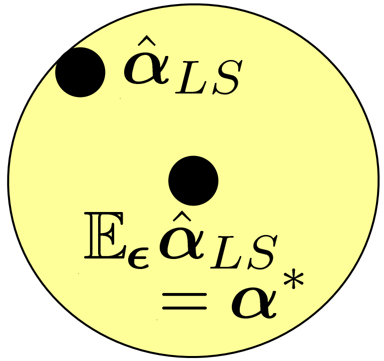
Variance



Unbiasedness

- When $f(\boldsymbol{x})$ is realizable, $\hat{\boldsymbol{\alpha}}_{LS}$ is an unbiased estimator:

$$\mathbb{E}_{\boldsymbol{\epsilon}}[\hat{\boldsymbol{\alpha}}_{LS}] = \boldsymbol{\alpha}^*$$



- Proof: In realizable cases,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha}^* + \boldsymbol{\epsilon}$$

$$\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^\top$$

Then

$$\mathbb{E}_{\boldsymbol{\epsilon}}[\hat{\boldsymbol{\alpha}}_{LS}] = \mathbb{E}_{\boldsymbol{\epsilon}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

$$= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\alpha}^* + \mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}])$$

$$= \boldsymbol{\alpha}^*$$

$$\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}] = \boldsymbol{0}$$

Best Linear Unbiased Estimator

- $\hat{\boldsymbol{\alpha}}_{LS}$ is the best linear unbiased estimator (BLUE, a linear estimator which has the smallest variance among all linear unbiased estimators)

$$\mathbb{E}_{\boldsymbol{\epsilon}} \|\hat{\boldsymbol{\alpha}}_{LS} - \mathbb{E}_{\boldsymbol{\epsilon}} \hat{\boldsymbol{\alpha}}_{LS}\|_{\boldsymbol{U}}^2$$

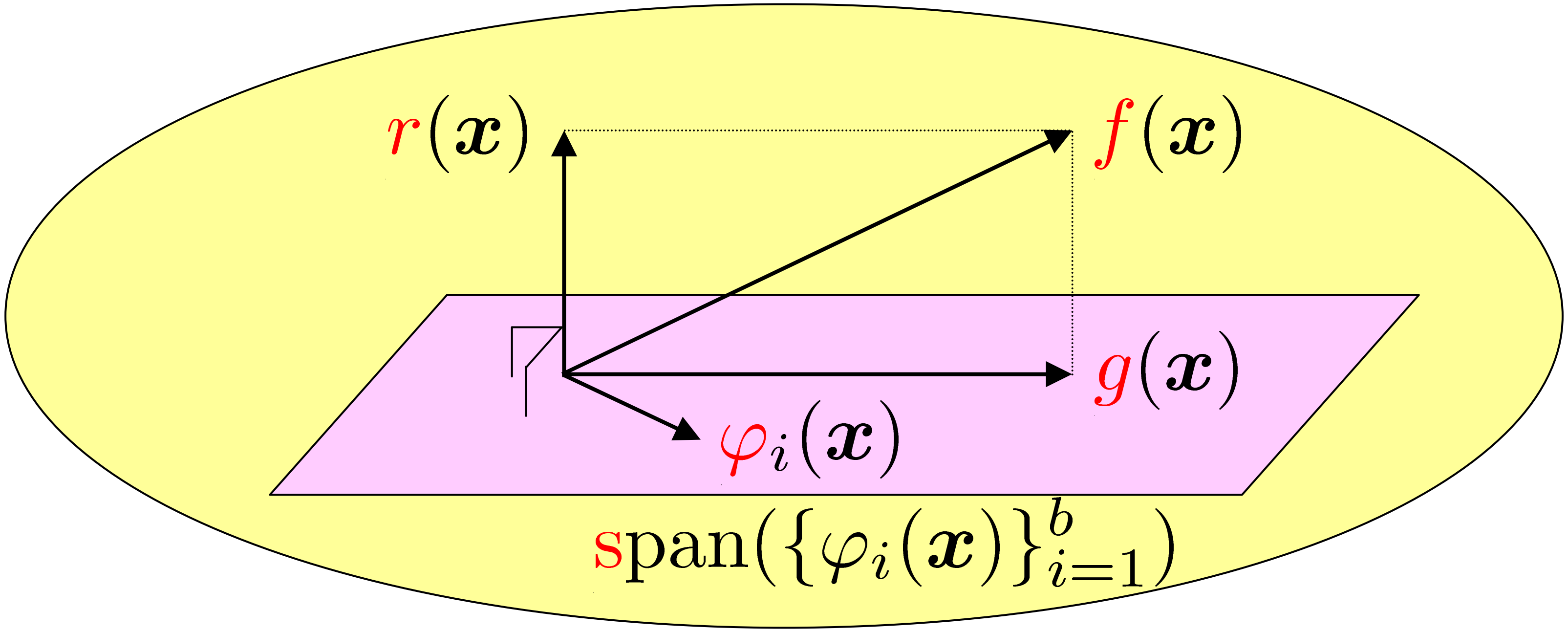
$$\leq \mathbb{E}_{\boldsymbol{\epsilon}} \|\hat{\boldsymbol{\alpha}}_{LU} - \mathbb{E}_{\boldsymbol{\epsilon}} \hat{\boldsymbol{\alpha}}_{LU}\|_{\boldsymbol{U}}^2$$

$$\text{for any linear unbiased estimator } \hat{\boldsymbol{\alpha}}_{LU}$$

- Proof: Homework!

Justification of LS (Unrealizable Cases)

- Decomposition: $f(\boldsymbol{x}) = g(\boldsymbol{x}) + r(\boldsymbol{x})$



$$\int_{\mathcal{D}} \varphi_i(\boldsymbol{x}) r(\boldsymbol{x}) q(\boldsymbol{x}) d\boldsymbol{x} = 0$$

$$g(\boldsymbol{x}) = \sum_{i=1}^{b} \alpha_i^* \varphi_i(\boldsymbol{x})$$

Generalization Error Decomposition

$$G = \int_{\mathcal{D}} \left(\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x}) \right)^2 q(\boldsymbol{x}) d\boldsymbol{x}$$

$$= \int_{\mathcal{D}} \left(\hat{f}(\boldsymbol{x}) - g(\boldsymbol{x}) - r(\boldsymbol{x}) \right)^2 q(\boldsymbol{x}) d\boldsymbol{x}$$

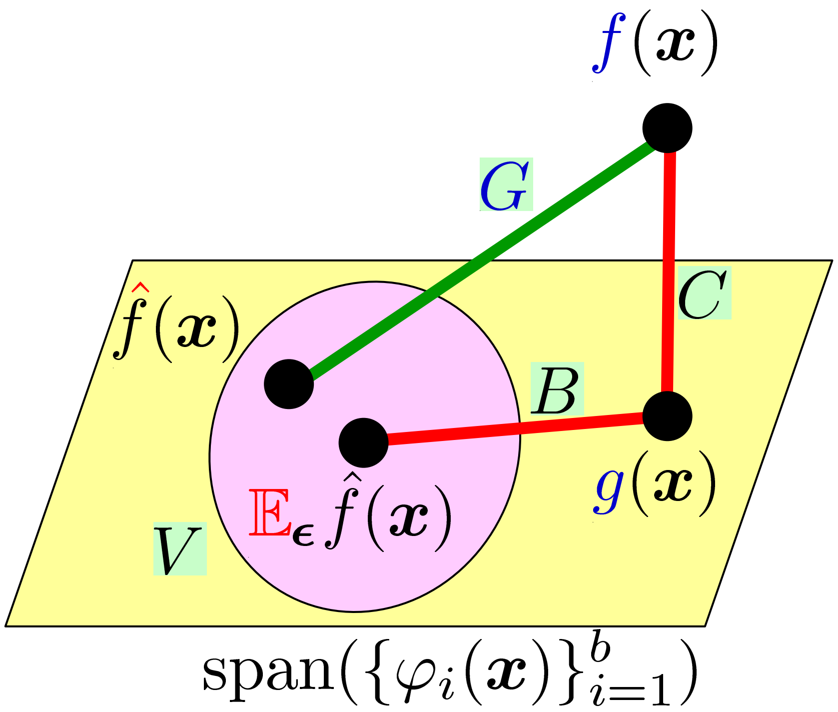
$$= \int_{\mathcal{D}} \left(\hat{f}(\boldsymbol{x}) - g(\boldsymbol{x}) \right)^2 q(\boldsymbol{x}) d\boldsymbol{x} + \int_{\mathcal{D}} r(\boldsymbol{x})^2 q(\boldsymbol{x}) d\boldsymbol{x}$$

$$= \|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_{\boldsymbol{U}}^2 + C$$

$$C = \int_{\mathcal{D}} r(\boldsymbol{x})^2 q(\boldsymbol{x}) d\boldsymbol{x}$$

Bias/Variance Decomposition

$$\mathbb{E}_{\epsilon}[G] = \underbrace{\mathbb{E}_{\epsilon}\|\boldsymbol{\alpha} - \mathbb{E}_{\epsilon}\boldsymbol{\alpha}\|_{\boldsymbol{U}}^2}_{\text{Variance}} + \underbrace{\|\mathbb{E}_{\epsilon}\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_{\boldsymbol{U}}^2}_{\text{Bias}} + \underbrace{C}_{\text{Model error}}$$



$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{b} \alpha_i \varphi_i(\boldsymbol{x})$$

$$g(\boldsymbol{x}) = \sum_{i=1}^{b} \alpha_i^* \varphi_i(\boldsymbol{x})$$

Asymptotic Unbiasedness

- $\hat{\boldsymbol{\alpha}}_{LS}$ is an asymptotically unbiased estimator of the optimal parameter $\boldsymbol{\alpha}^*$:

$$\mathbb{E}_{\boldsymbol{\epsilon}}[\hat{\boldsymbol{\alpha}}_{LS}] \rightarrow \boldsymbol{\alpha}^* \text{ as } n \rightarrow \infty$$

- Proof:
  - $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha}^* + \boldsymbol{z}_r + \boldsymbol{\epsilon}$ $\qquad \boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^\top$

$$\boldsymbol{z}_r = (r(\boldsymbol{x}_1), r(\boldsymbol{x}_2), \ldots, r(\boldsymbol{x}_n))^\top$$

  - $\mathbb{E}_{\boldsymbol{\epsilon}}[\hat{\boldsymbol{\alpha}}_{LS}] = \mathbb{E}_{\boldsymbol{\epsilon}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$

$$= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\alpha}^* + \boldsymbol{z}_r + \mathbb{E}_{\boldsymbol{\epsilon}}\boldsymbol{\epsilon})$$

$$= \boldsymbol{\alpha}^* + (\tfrac{1}{n}\boldsymbol{X}^\top\boldsymbol{X})^{-1}\tfrac{1}{n}\boldsymbol{X}^\top\boldsymbol{z}_r$$

Proof (cont.)

- By the law of large numbers,
  - $$[\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}]_{i,j} = \frac{1}{n}\sum_{k=1}^{n} \varphi_i(\boldsymbol{x}_k)\varphi_j(\boldsymbol{x}_k)$$
    $$\to \int_{\mathcal{D}} \varphi_i(\boldsymbol{x})\varphi_j(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} = U_{i,j}$$
  - $$[\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{z}_r]_i = \frac{1}{n}\sum_{k=1}^{n} \varphi_i(\boldsymbol{x}_k)r(\boldsymbol{x}_k)$$
    $$\to \int_{\mathcal{D}} \varphi_k(\boldsymbol{x})r(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} = 0$$
- Thus, $\mathbb{E}_{\boldsymbol{\epsilon}}[\hat{\boldsymbol{\alpha}}_{LS}] \to \boldsymbol{\alpha}^* \text{ as } n \to \infty$

(Q.E.D.)

Efficiency

- The Cramér-Rao lower bound: Lower bound of the variance of all (possibly non-linear) unbiased estimators.
- Efficient estimator: An unbiased estimator whose variance attains Cramér-Rao bound.
- For linear model with LS and $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, Cramér-Rao bound is

$$\sigma^2 \mathrm{tr}(\boldsymbol{U}(\boldsymbol{X}^\top \boldsymbol{X})^{-1})$$

Asymptotic Efficiency

- Asymptotically efficient estimator: An unbiased estimator that attains Cramér-Rao's lower bound asymptotically.
- When $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ and $\boldsymbol{x}_i \overset{i.i.d.}{\sim} q(\boldsymbol{x})$, LS estimator is asymptotically efficient.
- Proof: LS estimator is asymptotically unbiased and

$$\mathbb{E}_{\boldsymbol{\epsilon}} \|\hat{\boldsymbol{\alpha}}_{LS} - \mathbb{E}_{\boldsymbol{\epsilon}} \hat{\boldsymbol{\alpha}}_{LS}\|_{\boldsymbol{U}}^2 = \mathbb{E}_{\boldsymbol{\epsilon}} \|\boldsymbol{L}_{LS} \boldsymbol{\epsilon}\|_{\boldsymbol{U}}^2$$
$$= \sigma^2 \mathrm{tr}(\boldsymbol{U}(\boldsymbol{X}^\top \boldsymbol{X})^{-1})$$

which is Cramér-Rao's lower bound.