Input Dependent Estimation of Generalization Error

- CV is very general and useful.
- Its unbiasedness holds with respect to both input points and output noise.
- However, input points are known.
- Is it possible to have an unbiased estimator of the generalization error only with respect to the noise?

Setting

- $p(\boldsymbol{x})$ is known.
- Linear model:

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

- Regularization learning:

$$\min_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{n} \left(\hat{f}(\boldsymbol{x}_i) - y_i \right)^2 + \lambda \|\boldsymbol{\alpha}\|^2 \right]$$

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{L}\boldsymbol{y}$$

$$\boldsymbol{L} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top$$

$$\boldsymbol{X}_{i,j} = \varphi_j(\boldsymbol{x}_i)$$

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_p)^\top$$

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$$

Decomposition of Generalization Error

$$J = \int \left(\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x}) \right)^2 p(\boldsymbol{x}) d\boldsymbol{x}$$

$$= \int \hat{f}(\boldsymbol{x})^2 p(\boldsymbol{x}) d\boldsymbol{x}$$ (accessible)

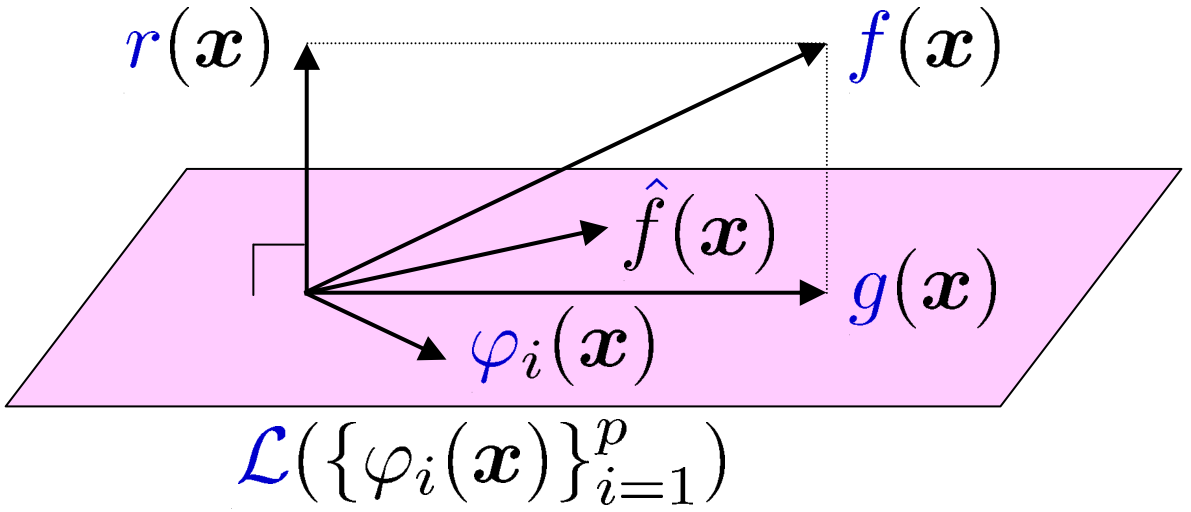
$$-2 \int \hat{f}(\boldsymbol{x}) f(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$$ (to be estimated)

$$+ \int f(\boldsymbol{x})^2 p(\boldsymbol{x}) d\boldsymbol{x}$$ (constant: ignored)

Decomposition of Target Function

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

$$g(\boldsymbol{x}) = \sum_{i=1}^{p} \alpha_i^* \varphi_i(\boldsymbol{x}) \qquad \int \varphi_i(\boldsymbol{x}) r(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} = 0$$



$$\int \hat{f}(\boldsymbol{x}) f(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} = \int \hat{f}(\boldsymbol{x}) g(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$$

Estimation of Generalization Error

Suppose we have $\boldsymbol{L}_u, \sigma_u^2$ such that

(i) $\mathbb{E}_\epsilon \boldsymbol{L}_u \boldsymbol{y} = \boldsymbol{\alpha}^*$ ($\boldsymbol{L}_u$ and $\boldsymbol{L}$ are irrelevant)

(ii) $\mathbb{E}_\epsilon \sigma_u^2 = \sigma^2$

$$\mathbb{E}_\epsilon \int \hat{f}(\boldsymbol{x}) g(\boldsymbol{x}) p_t(\boldsymbol{x}) d\boldsymbol{x}$$

$$= \mathbb{E}_\epsilon \langle \boldsymbol{U}\hat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}^* \rangle$$

$$= \mathbb{E}_\epsilon [\langle \boldsymbol{U}\boldsymbol{L}\boldsymbol{y}, \boldsymbol{L}_u \boldsymbol{y} \rangle - \sigma_u^2 \mathrm{tr}(\boldsymbol{U}\boldsymbol{L}\boldsymbol{L}_u^\top)]$$

$$\boldsymbol{U}_{i,j} = \int \varphi_i(\boldsymbol{x}) \varphi_j(\boldsymbol{x}) p_t(\boldsymbol{x}) d\boldsymbol{x}$$

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

- However, such $\boldsymbol{L}_u, \sigma_u^2$ are not available in practice, so we use approximations.

Estimation of Generalization Error (cont.)

(i) $\mathbb{E}_{\boldsymbol{\epsilon}} \boldsymbol{L}_u \boldsymbol{y} = \boldsymbol{\alpha}^*$

$$\widehat{\boldsymbol{L}}_u = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$$

$\widehat{\boldsymbol{L}}_u$ corresponds to least-squares, hence

- $\mathbb{E}_{\boldsymbol{\epsilon}} \widehat{\boldsymbol{L}}_u \boldsymbol{y} = \boldsymbol{\alpha}^*$ if $f(\boldsymbol{x})$ is realizable
- $\mathbb{E}_{\boldsymbol{\epsilon}} \widehat{\boldsymbol{L}}_u \boldsymbol{y} \to \boldsymbol{\alpha}^* \text{ as } n \to \infty$ o.w.

Estimation of Generalization Error (cont.)

(ii) $\mathbb{E}_{\boldsymbol{\epsilon}} \sigma_u^2 = \sigma^2$

$$\widehat{\sigma_u^2} = \frac{\|\boldsymbol{G}\boldsymbol{y}\|^2}{\mathrm{tr}(\boldsymbol{G})}$$

$$\boldsymbol{G} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$$

- $\mathbb{E}_{\boldsymbol{\epsilon}} \widehat{\sigma_u^2} = \sigma^2$ if $f(\boldsymbol{x})$ is realizable
- $\mathbb{E}_{\boldsymbol{\epsilon}} \widehat{\sigma_u^2} \not\to \sigma^2$ as $n \to \infty$ o.w.

Generalization Error Estimator

$$\hat{J} = \langle \boldsymbol{U}\boldsymbol{L}\boldsymbol{y}, \boldsymbol{L}\boldsymbol{y} \rangle - 2\langle \boldsymbol{U}\boldsymbol{L}\boldsymbol{y}, \widehat{\boldsymbol{L}}_u \boldsymbol{y} \rangle + 2\widehat{\sigma_u^2}\mathrm{tr}(\boldsymbol{U}\boldsymbol{L}\widehat{\boldsymbol{L}}_u^\top)$$

Bias：

$$B_{\boldsymbol{\epsilon}} = \mathbb{E}_{\boldsymbol{\epsilon}}[\hat{J} - J] + C$$

$$C = \int f(\boldsymbol{x})^2 p_t(\boldsymbol{x}) d\boldsymbol{x}$$

- If $r(\boldsymbol{x}_i) = 0$

$$B_{\boldsymbol{\epsilon}} = 0$$

- If $\delta = \max\{r(\boldsymbol{x}_i)\}$ is sufficiently small

$$B_{\boldsymbol{\epsilon}} = \mathcal{O}(\delta)$$

- In general,

$$B_{\boldsymbol{\epsilon}} = \mathcal{O}_p(n^{-\frac{1}{2}})$$

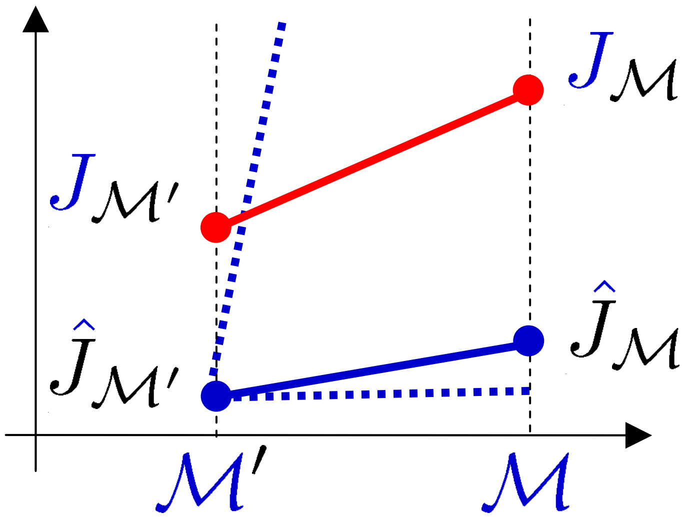
Model Comparison

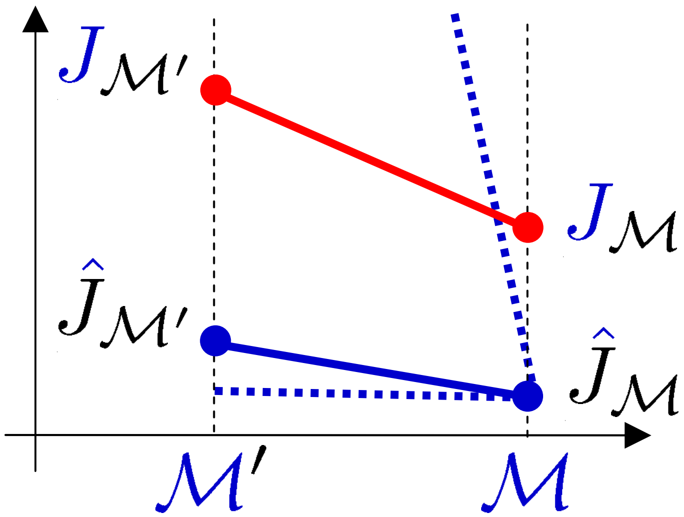
- A purpose of estimating generalization error is model selection.
- We want to know whether $\hat{J}$ can distinguish good models from poor ones.

$$\mathcal{M} = \{\{\varphi_i(\boldsymbol{x})\}_{i=1}^{p}, \lambda\}$$

Model Comparison

- Difference in J : $\Delta J = J_{\mathcal{M}} - J_{\mathcal{M}'}$
- Difference in $\hat{J}$: $\Delta \hat{J} = \hat{J}_{\mathcal{M}} - \hat{J}_{\mathcal{M}'}$

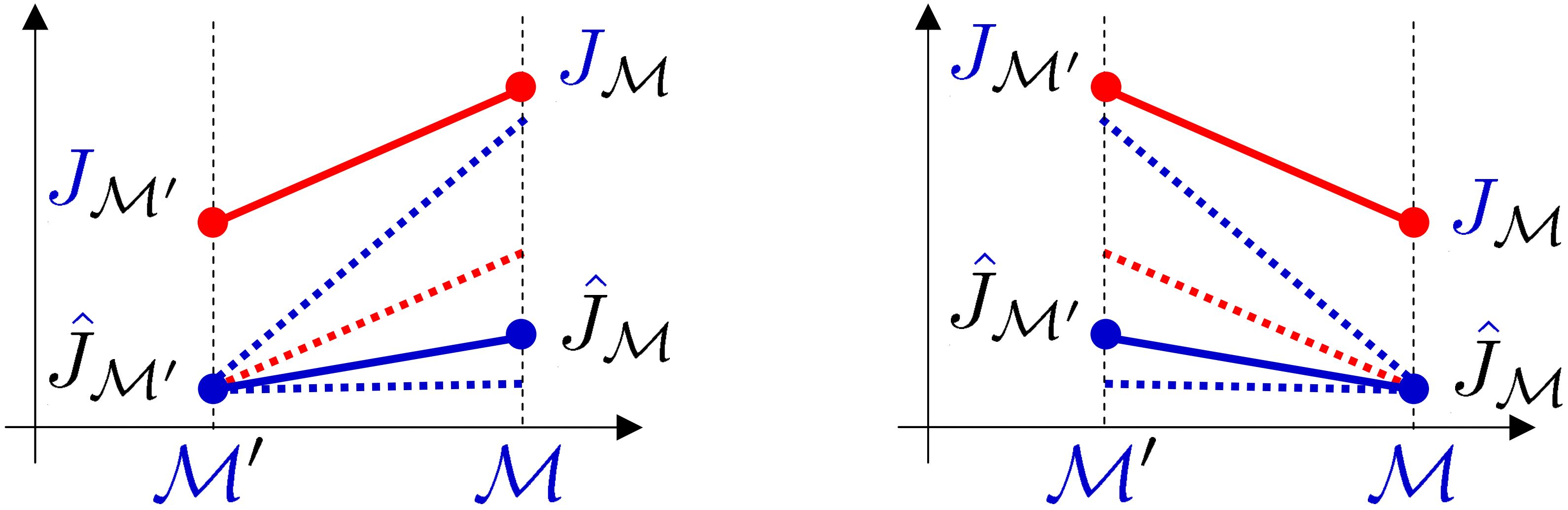




- If $\mathrm{sgn}(\mathbb{E}_{\epsilon} \Delta J) = \mathrm{sgn}(\mathbb{E}_{\epsilon} \Delta \hat{J})$, better model can be selected on average.

Model Comparison

- However, checking the sign is not easy, so we simplify the criterion.



- "Good" if $0 < \mathbb{E}_\epsilon \Delta \hat{J} < 2\mathbb{E}_\epsilon \Delta J \quad (\mathbb{E}_\epsilon \Delta J > 0)$

$$0 > \mathbb{E}_\epsilon \Delta \hat{J} > 2\mathbb{E}_\epsilon \Delta J \quad (\mathbb{E}_\epsilon \Delta J < 0)$$

Model Comparison

- Difference in the bias B_{ϵ}:

$$\Delta B_{\epsilon} = \mathbb{E}_{\epsilon}[\Delta \hat{J} - \Delta J]$$

- Effective in model comparison:

$$|\Delta B_{\epsilon}| < |\mathbb{E}_{\epsilon} \Delta J|$$

- Asymptotically effective in model comparison:

$$\Delta B_{\epsilon} = o_p(n^{-t}), \ \ \mathbb{E}_{\epsilon} \Delta J \neq o_p(n^{-t})$$

Effectiveness in Model Comparison

$\hat{J}$ is

- Effective in model comparison (if $f(\boldsymbol{x})$ is realizable)
- Asymptotically effective in model comparison (o.w.)

Final Report

1. Read one/both of the following articles and write your opinions.
   A) T. Carlo, Learning theory: Past performance and future results, *Nature* vol.428, p.378, 2004.
   B) E. Mjolsness and D. DeCoste, Machine learning for science: State of the art and future prospects, *Science*, vol.293, pp.2051-2055, 2001.
2. Write your opinions about:
   A) Drawbacks of current machine learning technologies
   B) Future research directions of machine learning (either theoretical studies or applications, either it is realistic or dreamy).
3. Evaluate this course, e.g.,
   - What was interesting/uninteresting and how should it be improved?
   - We covered only a fraction of machine learning field. What else do you want to learn?
   - Anything: questions, impressions, errata…

Deadline: Feb. 10, 2005
(E-mail to sugi@cs.titech.ac.jp)