Pattern Information Processing: Input-Dependent Estimation of Generalization Error

Masashi Sugiyama

(Department of Computer Science)

Contact: W8E-505

sugi@cs.titech.ac.jp

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

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

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

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

- Regularization learning:

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

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

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

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

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

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

Decomposition of Generalization Error

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

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

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

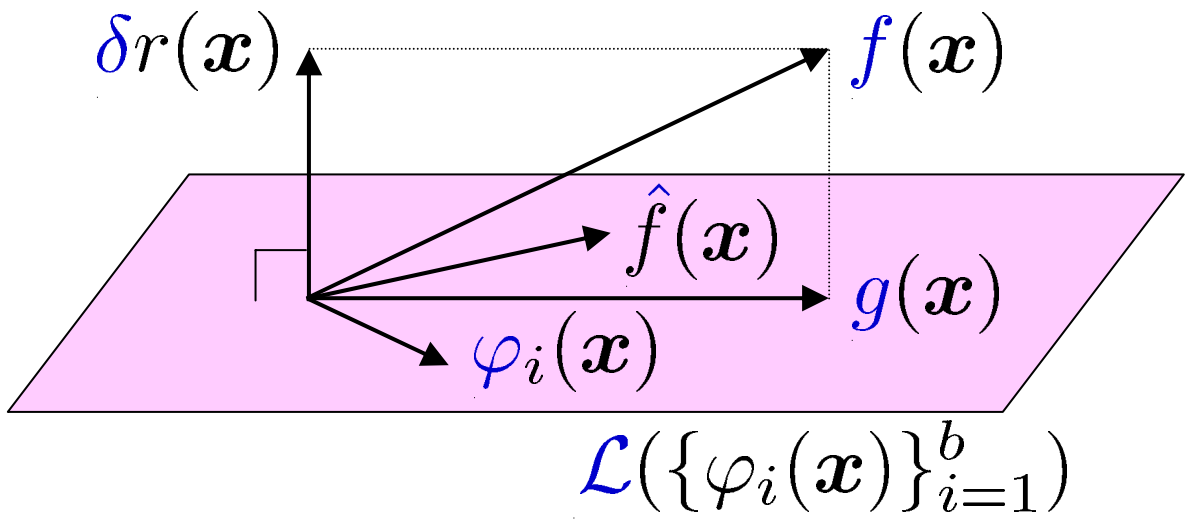
$$+\int f(\boldsymbol{x})^2 q(\boldsymbol{x}) d\boldsymbol{x} \qquad \text{(constant: ignored)}$$

Decomposition of Target Function

$$f(\boldsymbol{x}) = g(\boldsymbol{x}) + \delta r(\boldsymbol{x}) \qquad \int r(\boldsymbol{x})^2 q(\boldsymbol{x}) d\boldsymbol{x} = 1$$

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

$\delta \geq 0$



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

Estimation of Generalization Error

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

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

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

$$\mathbb{E}_{\boldsymbol{\epsilon}} \int \hat{f}(\boldsymbol{x}) g(\boldsymbol{x}) q(\boldsymbol{x}) d\boldsymbol{x}$$
$$= \mathbb{E}_{\boldsymbol{\epsilon}} \langle \boldsymbol{U} \hat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}^* \rangle$$
$$= \mathbb{E}_{\boldsymbol{\epsilon}} [\langle \boldsymbol{U} \boldsymbol{L} \boldsymbol{y}, \boldsymbol{L}_u \boldsymbol{y} \rangle - \sigma_u^2 \mathrm{tr}(\boldsymbol{U} \boldsymbol{L} \boldsymbol{L}_u^\top)]$$

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

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

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

$\widehat{\boldsymbol{L}}_u$ satisfies

$$\mathbb{E}_{\boldsymbol{\epsilon}} \widehat{\boldsymbol{L}}_u \boldsymbol{y} = \boldsymbol{\alpha}^* + \mathcal{O}_p(\delta n^{-\frac{1}{2}})$$

Proof

- $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha}^* + \delta \boldsymbol{z}_r + \boldsymbol{\epsilon}$

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

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

- $\widehat{\boldsymbol{L}}_u \boldsymbol{X}\boldsymbol{\alpha}^* = \boldsymbol{\alpha}^*$

- $[\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{z}_r]_k = \frac{1}{n}\sum_{i=1}^{n} \varphi_k(\boldsymbol{x}_i) r(\boldsymbol{x}_i)$

$$= \int_{\mathcal{D}} \varphi_k(\boldsymbol{x}) r(\boldsymbol{x}) q(\boldsymbol{x}) d\boldsymbol{x} + \mathcal{O}_p(n^{-\frac{1}{2}})$$

(Central limit theorem)

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

Estimation of Generalization Error (cont.)

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

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

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

$\widehat{\sigma_u^2}$ satisfies

$$\mathbb{E}_{\boldsymbol{\epsilon}}\widehat{\sigma_u^2} = \sigma^2 + \mathcal{O}(\delta^2)$$

Proof

- $G^2 = G$
- $GX\alpha^* = 0$
- $\|Gz_r\|^2 = \mathcal{O}(\delta^2)$

- $$\mathbb{E}_\epsilon \sigma_u^2 = \frac{\mathbb{E}_\epsilon \|Gy\|^2}{\mathrm{tr}(G)}$$
$$= \frac{\|GX\alpha^* + Gz_r\|^2 + \mathbb{E}_\epsilon \|G\epsilon\|^2}{\mathrm{tr}(G)}$$
$$= \sigma^2 + \mathcal{O}(\delta^2)$$

Generalization Error Estimator

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

Bias:

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

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

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

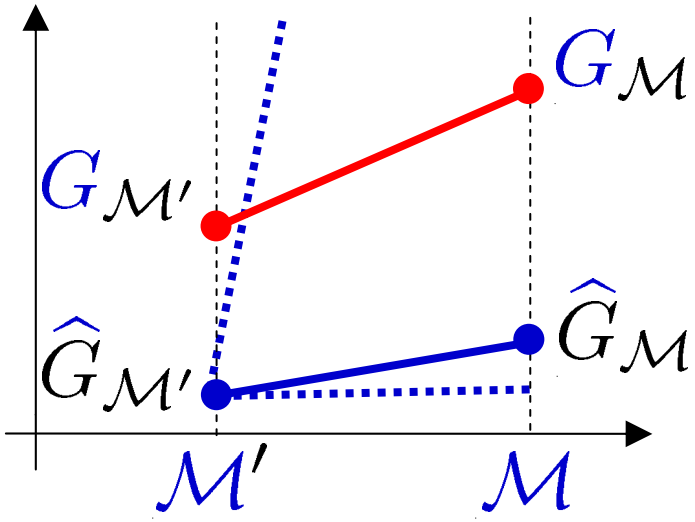
Model Comparison

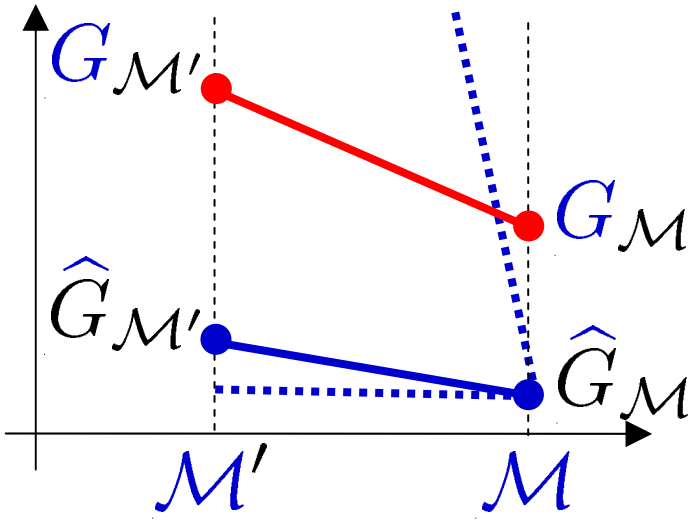
- A purpose of estimating generalization error is model selection.
- We want to know whether $\widehat{G}$ can distinguish good models from poor ones.

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

Model Comparison

- Difference in G : $\Delta G = G_{\mathcal{M}} - G_{\mathcal{M}'}$
- Difference in $\widehat{G}$: $\Delta \widehat{G} = \widehat{G}_{\mathcal{M}} - \widehat{G}_{\mathcal{M}'}$

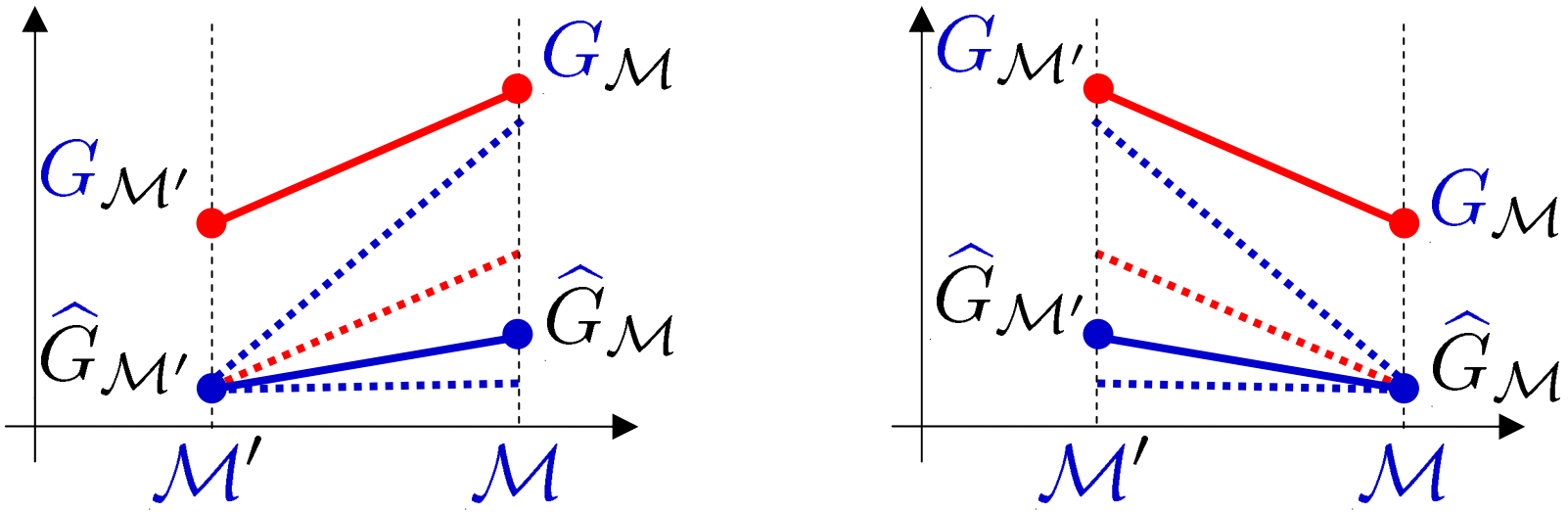




- If $\operatorname{sgn}(\mathbb{E}_\epsilon \Delta G) = \operatorname{sgn}(\mathbb{E}_\epsilon \Delta \widehat{G})$, better model can be selected on average.

Model Comparison

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



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

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

Model Comparison

- Difference in the bias B_ϵ:

$$\Delta B_\epsilon = \mathbb{E}_\epsilon[\Delta\widehat{G} - \Delta G]$$

- Effective in model comparison:

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

- Asymptotically effective in model comparison:

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

Effectiveness in Model Comparison

$\widehat{G}$ is

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

Schedule

- July 11th : regular lecture
  (active learning)
  workshop registration
- July 18th : no class
- July 25th : mini-workshop