IASTED-AIA2004 Feb. 16-18, 2004

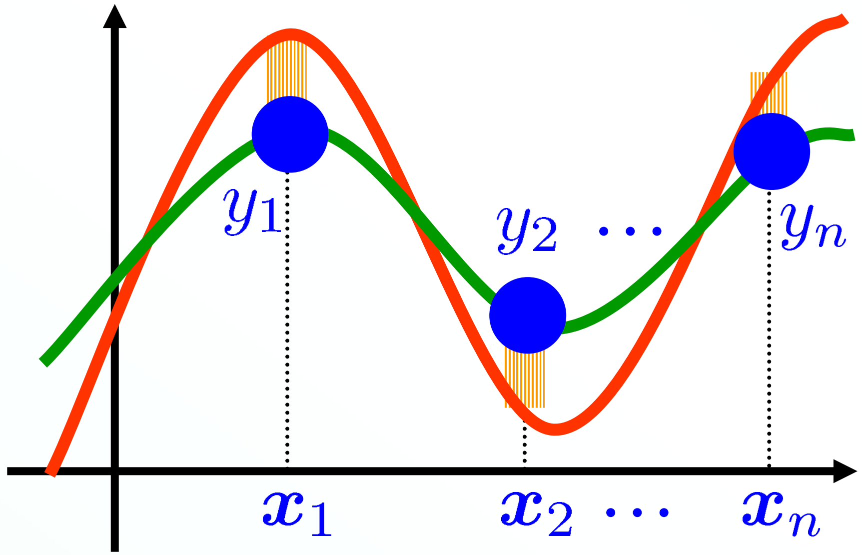
On the Influence of Input Noise on a Generalization Error Estimator

Fraunhofer Institut Rechnerarchitektur und Softwaretechnik

Masashi Sugiyama [1,2]
Yuta Okabe [2]
Hidemitsu Ogawa [2]

[1] Fraunhofer FIRST-IDA, Berlin, Germany
[2] Tokyo Institute of Technology, Tokyo, Japan

Regression Problem



$f(\boldsymbol{x})$:Underlying function

$\hat{f}(\boldsymbol{x})$:Learned function

$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$:Training examples

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

(noise)

$$\epsilon_i \overset{i.i.d.}{\sim} \text{mean } 0, \ \text{variance } \sigma^2$$

From $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, obtain a good approximation $\hat{f}(\boldsymbol{x})$ to $f(\boldsymbol{x})$

Typical Method of Learning

- Kernel regression model

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

α_i :Parameters to be learned

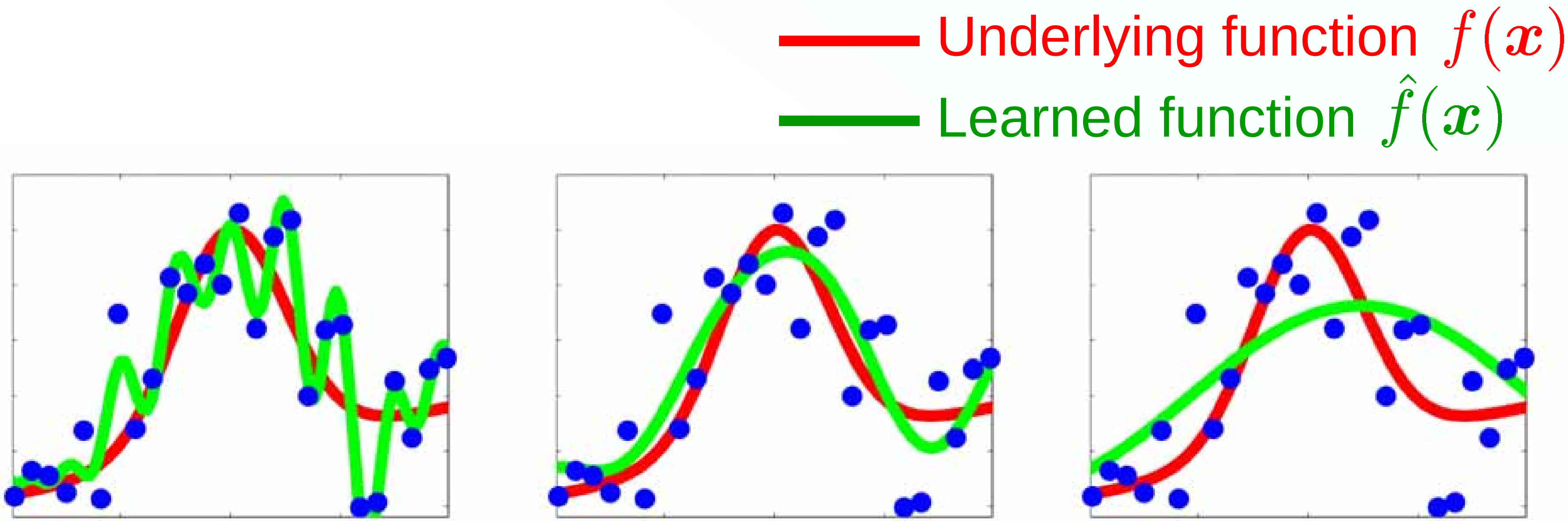
$K(\boldsymbol{x}, \boldsymbol{x}')$:Kernel function (e.g., Gaussian)

- Ridge estimation

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

λ :Ridge parameter (model parameter)

Model Selection



λ is too small　　λ is appropriate　　λ is too large

Choice of the model is crucial
for obtaining good learned function $\hat{f}(x)$!

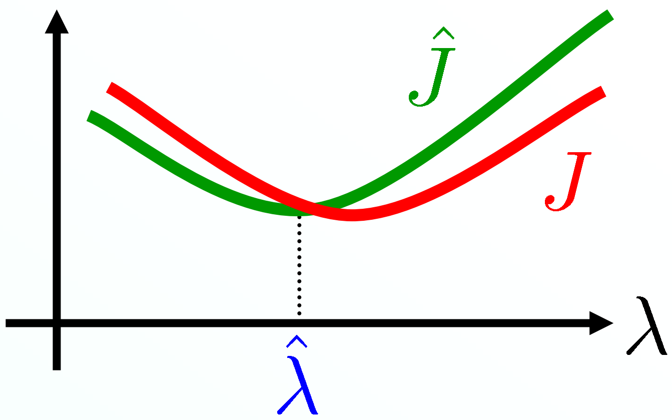
Generalization Error

For model selection, we need a criterion that measures 'closeness' between $\hat{f}(\boldsymbol{x})$ and $f(\boldsymbol{x})$:

→ Generalization error

Determine the model λ so that an estimator $\hat{J}$ of the unknown generalization error J is minimized.

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmin}}\, \hat{J}(\lambda)$$



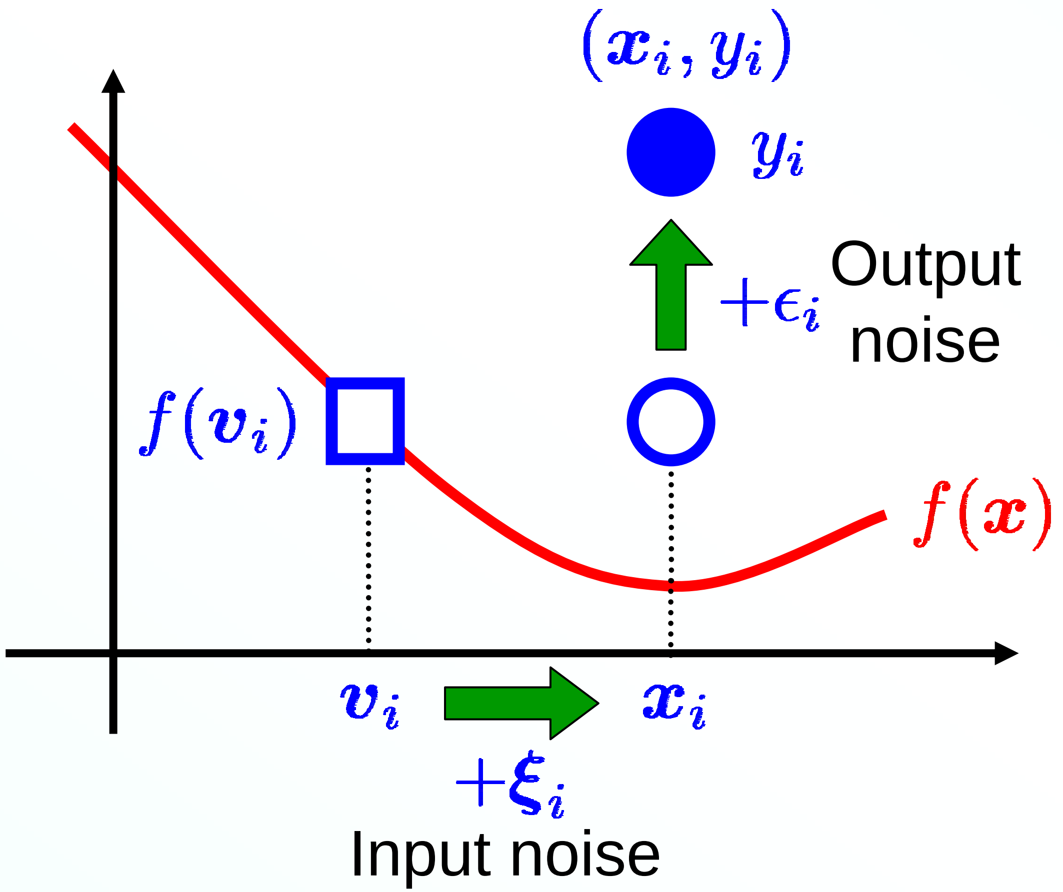
Noise in Input Points

- Previous research mainly deals with the cases where noise is included only in output values.
- However, noise is sometimes included also in input points, e.g.,
  - Input points are measured:

    Signal/image recognition, robot motor control, and bioinformatic data analysis.
  - Input points are estimated:

    Time series prediction of multiple-step ahead.

Noise in Input Points (cont.)



- We want to measure output values $f(x_i)$ at x_i
- But measurement is actually done at unknown v_i
- Output noise ϵ_i is then added

$$x_i = v_i + \xi_i$$
$$y_i = f(v_i) + \epsilon_i$$

Aim of Our Research

- So far, it seems that model selection in the presence of input noise has not been well studied yet.
- We investigate the effect of input noise on a generalization error estimator called the subspace information criterion (SIC).

Sugiyama & Ogawa (Neural Computation, 2001)
Sugiyama & Müller (JMLR, 2002)

Generalization Error in RKHS

- $\mathcal{H}$: A reproducing kernel Hilbert space
- We assume $f, \hat{f} \in \mathcal{H}$
- We shall measure the generalization error by

$$\boxed{J = \mathrm{E}_{\boldsymbol{\epsilon}} \|\hat{f} - f\|^2 - \|f\|^2}$$

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

$\|\cdot\|$:Norm

Setting

Kernel regression model

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

α_i :Parameters to be learned

$K(\boldsymbol{x}, \boldsymbol{x}')$:Kernel function (e.g., Gaussian)

Linear estimation

$$\boldsymbol{\alpha} = \boldsymbol{X}\boldsymbol{y}$$

$\boldsymbol{X}$:Learning matrix

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

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

Subspace Information Criterion

Sugiyama & Ogawa (Neural Computation, 2001)
Sugiyama & Müller (JMLR, 2002)

$$SIC = \langle \boldsymbol{K}\boldsymbol{X}\boldsymbol{y}, \boldsymbol{X}\boldsymbol{y}\rangle - 2\langle \boldsymbol{K}\boldsymbol{X}\boldsymbol{y}, \boldsymbol{K}^{\dagger}\boldsymbol{y}\rangle + 2\sigma^2 \mathrm{tr}(\boldsymbol{K}^{\dagger}\boldsymbol{K}\boldsymbol{X})$$

$\boldsymbol{K}_{i,j} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$

$\boldsymbol{K}^{\dagger}$:Pseudo inverse of $\boldsymbol{K}$

$\langle \cdot, \cdot \rangle$:Inner product

- In the absence of input noise, SIC is an unbiased estimator of J :

$$\mathrm{E}_{\boldsymbol{\epsilon}} SIC = J \qquad J = \mathrm{E}_{\boldsymbol{\epsilon}}\|\hat{f} - f\|^2 - \|f\|^2$$

- We investigate how the unbiasedness of SIC is affected by input noise.

Unbiasedness of SIC in the Presence of Input Noise

- In the presence of input noise,

$$\mathrm{E}_{\boldsymbol{\epsilon}} SIC = J + \Delta J$$

$$\Delta J = \langle \boldsymbol{K}^{\dagger} \boldsymbol{K} \boldsymbol{X} \boldsymbol{z}, \boldsymbol{z}_{\boldsymbol{x}} - \boldsymbol{z} \rangle$$

$$\boldsymbol{z} = (f(\boldsymbol{v}_1), f(\boldsymbol{v}_2), \ldots, f(\boldsymbol{v}_n))^{\top}$$

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

$\boldsymbol{v}_i$:Noiseless input points

$\boldsymbol{x}_i$:Noisy input points

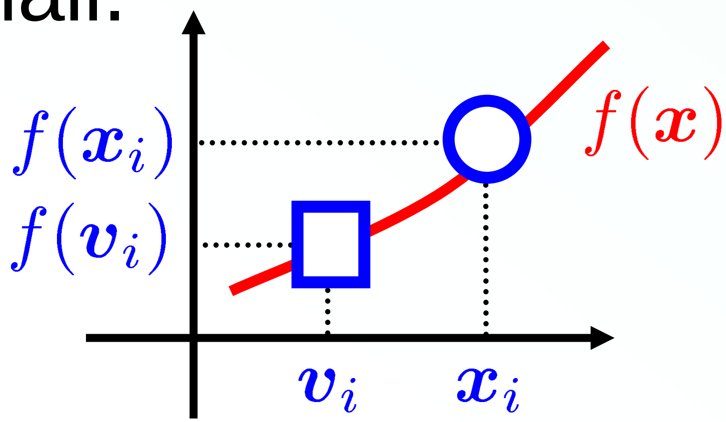
Unbiasedness of SIC does not generally hold in the presence of input noise.

Effect of Small Input Noise

- When $f(\boldsymbol{x})$ is continuous, small input noise varies the output value only slightly, i.e., $|f(\boldsymbol{x}_i) - f(\boldsymbol{v}_i)|$ is small.

$\boldsymbol{v}_i$:Noiseless input points

$\boldsymbol{x}_i$:Noisy input points



- Therefore, we expect that the unbiasedness of SIC is not severely affected (ΔJ is small) by small input noise.

$$\mathrm{E}_{\boldsymbol{\epsilon}} SIC = J + \Delta J$$

Effect of Small Input Noise (cont.)

- However, we can show that, for some learning matrix $\boldsymbol{X}$, it holds that

$$|\Delta J| \not\to 0 \text{ as } \|\boldsymbol{\xi}_i\| \to 0 \text{ for all } i\ .$$

$\boldsymbol{\xi}_i$:Input noise

- This implies that, for some $\boldsymbol{X}$, the unbiasedness of SIC is heavily affected even when input noise is very small.

Theorem

- Let $\|\boldsymbol{X}\|$ be the matrix norm defined by

$$\|\boldsymbol{X}\| = \sup_{\boldsymbol{z} \neq 0} \frac{\|\boldsymbol{X}\boldsymbol{z}\|}{\|\boldsymbol{z}\|}$$

- If the learning matrix $\boldsymbol{X}$ satisfies

$$\|\boldsymbol{X}\| = o(1/\delta) \qquad \delta = \|\boldsymbol{z}_{\boldsymbol{x}} - \boldsymbol{z}\|$$

then $|\Delta J| \to 0$ as $\|\boldsymbol{\xi}_i\| \to 0$ for all i .

$\boldsymbol{z} = (f(\boldsymbol{v}_1), f(\boldsymbol{v}_2), \ldots, f(\boldsymbol{v}_n))^\top$ $\qquad$ $\boldsymbol{v}_i$:Noiseless input points

$\boldsymbol{z}_{\boldsymbol{x}} = (f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots, f(\boldsymbol{x}_n))^\top$ $\qquad$ $\boldsymbol{x}_i$:Noisy input points

Ridge Estimation

- Ridge estimation

λ :Ridge parameter

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

$$\boldsymbol{X}=(\boldsymbol{K}^2+\lambda\boldsymbol{I})^{-1}\boldsymbol{K}$$

$\boldsymbol{K}_{i,j}=K(\boldsymbol{x}_i,\boldsymbol{x}_j)$

$\boldsymbol{I}$:Identity matrix

- We can prove that ridge estimation satisfies

$$\|\boldsymbol{X}\|=O(1)=o(1/\delta)$$

- Therefore, SIC with ridge estimation is robust against small input noise.

Simulation

- H :Gaussian RKHS

$$K(x,x') = \exp\left(-(x-x')^2/2\right)$$

- Learning target function $f(\boldsymbol{x})$: sinc function
- Training examples $\{(x_i, y_i)\}_{i=1}^{n}$:

$$v_i \overset{i.i.d.}{\sim} U(-\pi, \pi)$$

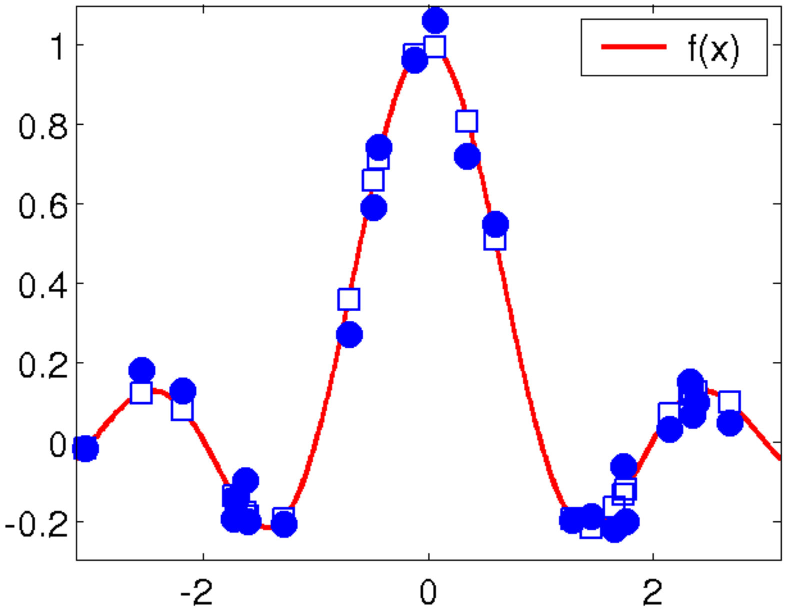
$$x_i = v_i + \xi_i, \quad \xi_i \overset{i.i.d.}{\sim} N(0, \sigma_x^2)$$

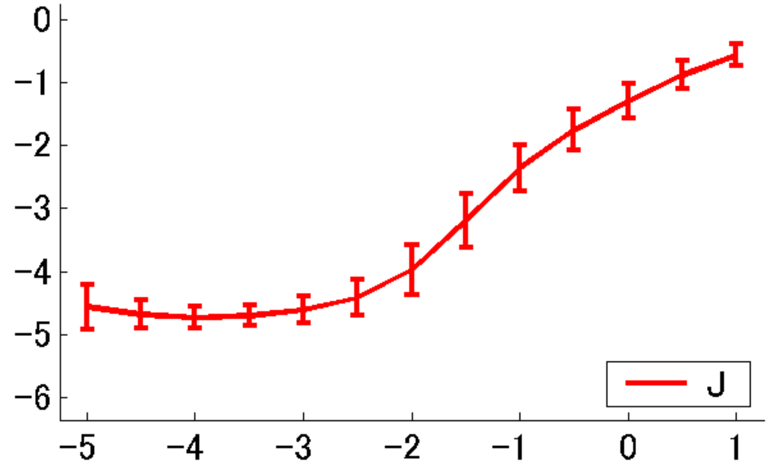
$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$$

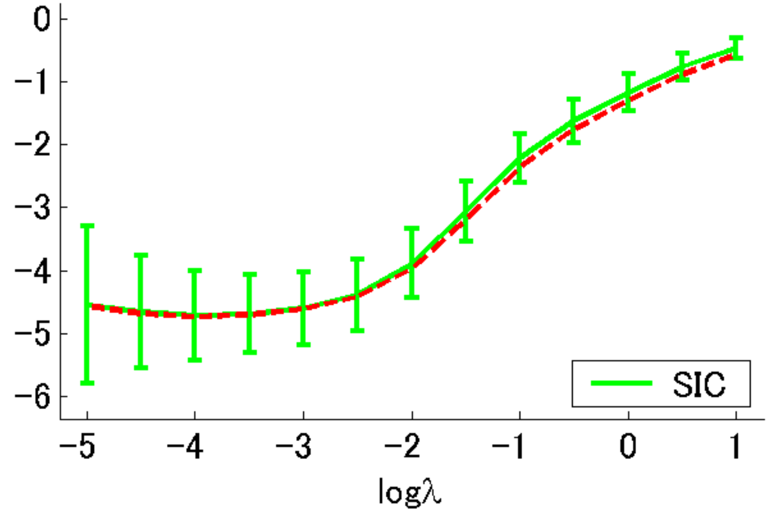
- $n = 25, \;\; \sigma = 0.05, \;\; \sigma_x = 0, 0.1, 0.2$
- Ridge estimation is used for learning.

Result (No Input Noise)

$\sigma_x = 0$







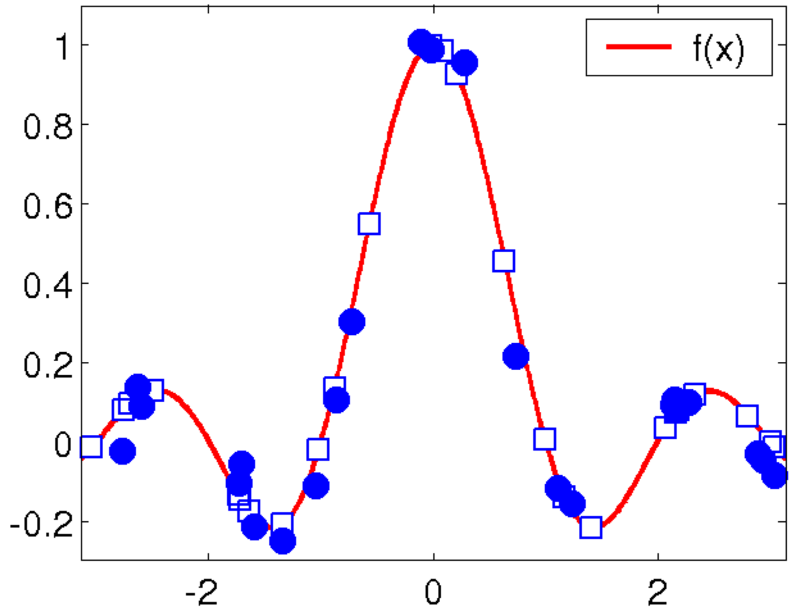
- SIC is surely unbiased without input noise

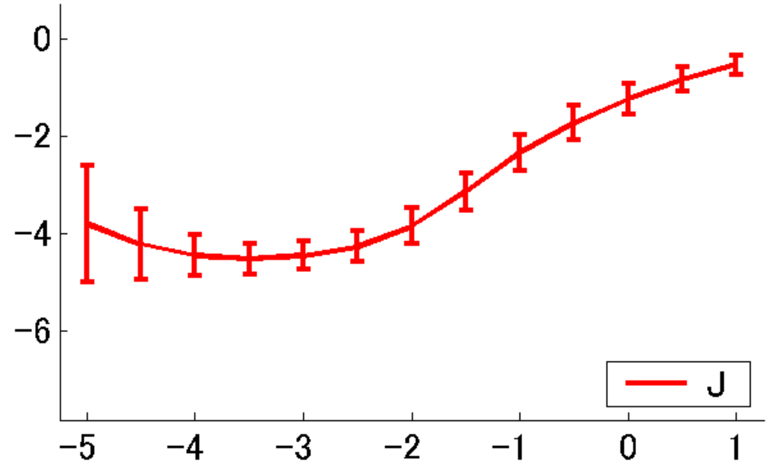
$$J = \mathrm{E}_{\epsilon} \|\hat{f} - f\|^2 - \|f\|^2$$

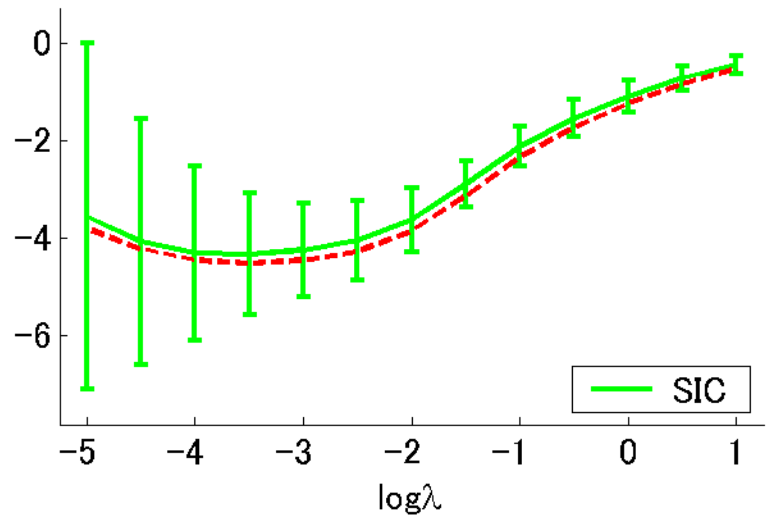
λ :Ridge parameter

Result (Small Input Noise)

$$\sigma_x = 0.1$$







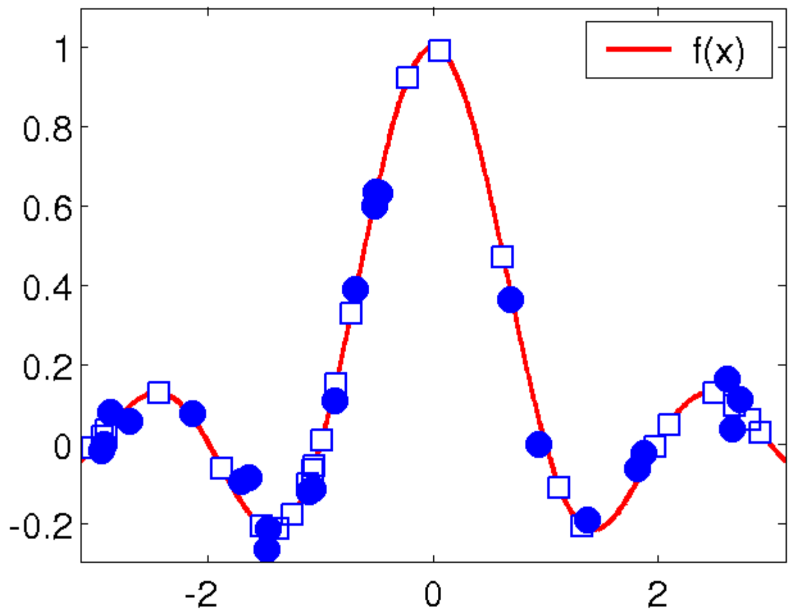
- SIC is still almost unbiased with small input noise

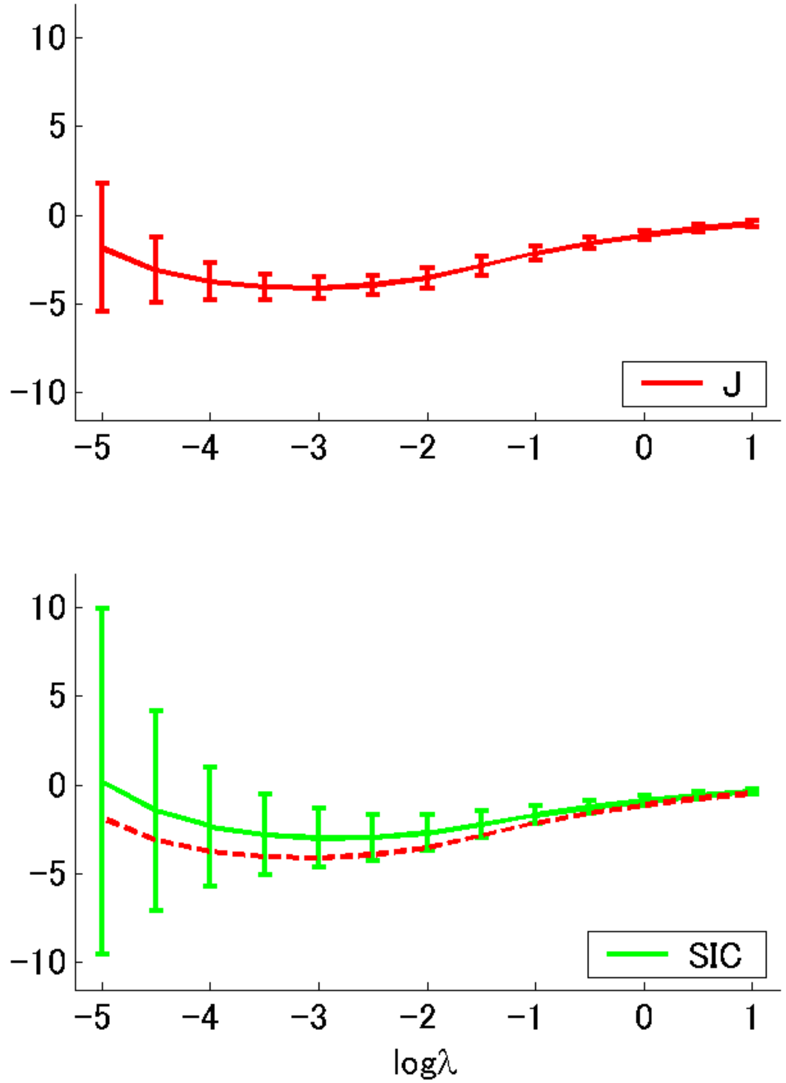
$$J = \mathrm{E}_{\epsilon} \|\hat{f} - f\|^2 - \|f\|^2$$

λ :Ridge parameter

Result (Large Input Noise)

$$\sigma_x = 0.2$$





- SIC is no longer reliable with large input noise

$$J = \mathrm{E}_{\epsilon} \|\hat{f} - f\|^2 - \|f\|^2$$

λ :Ridge parameter

Conclusions

- Effect of input noise on SIC.
- In some cases, the unbiasedness of SIC is heavily affected even by small input noise.
- A sufficient condition for unbiasedness.
- Ridge estimation satisfies this condition.
- Experiments: SIC is still almost unbiased for small input noise.
- Future work: Accurately estimate the generalization error when input noise is large.