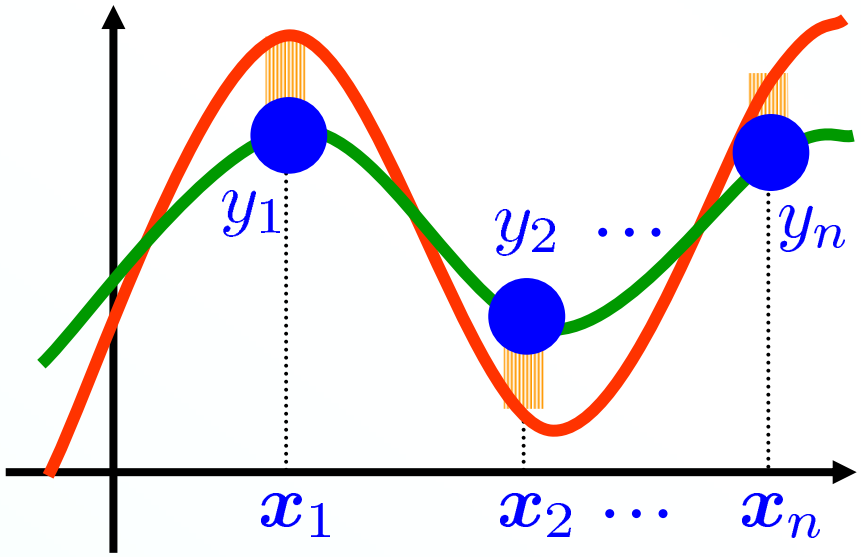
Estimating the Error at Given Test Input Points for Linear Regression

Masashi Sugiyama

Fraunhofer FIRST-IDA, Berlin, Germany

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

- Linear regression model

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

α_i :Parameters

$\{\varphi_i(\boldsymbol{x})\}_{i=1}^{p}$:Fixed basis functions

- Ridge estimation

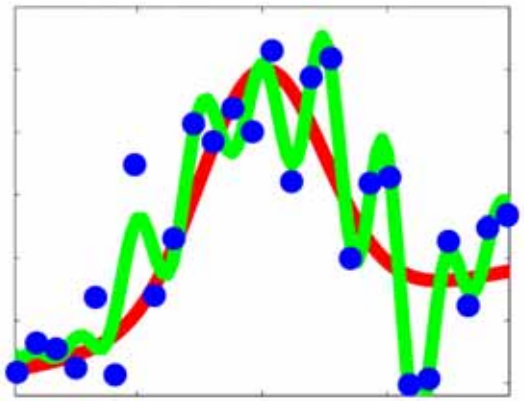
$$\min_{\{\alpha_i\}} \left[\sum_{i=1}^{n} \left(\hat{f}(\boldsymbol{x}_i) - y_i \right)^2 + \lambda \sum_{i=1}^{p} \alpha_i^2 \right]$$

λ :Ridge parameter (model parameter)

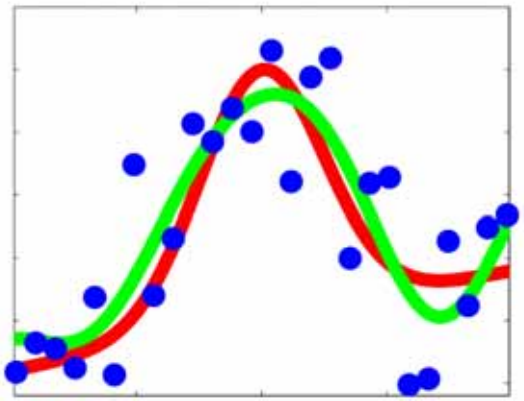
Model Selection

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

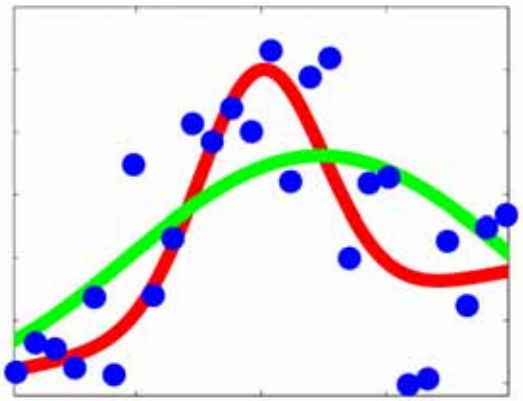
Learned function $\hat{f}(\boldsymbol{x})$



λ is too small



λ is appropriate

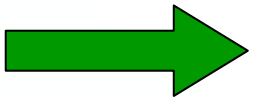


λ is too large

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

Generalization Error

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



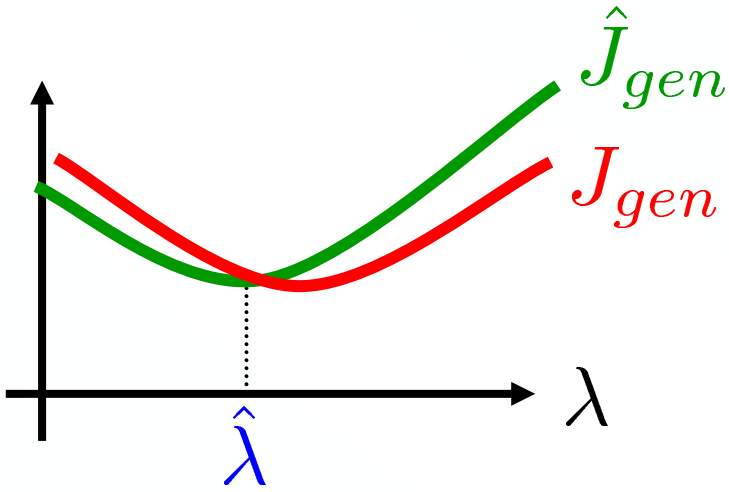
Generalization error, e.g.,

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

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

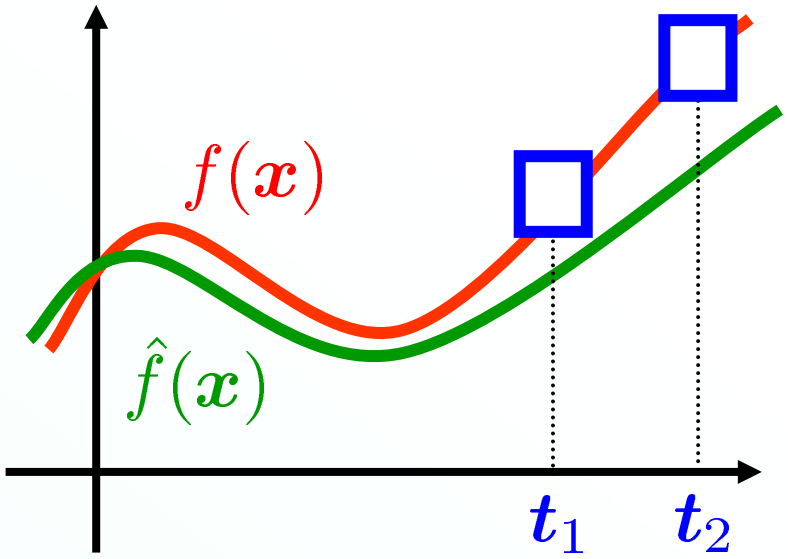
$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmin}}\, \hat{J}_{gen}(\lambda)$$

$p(\boldsymbol{x})$:Probability density of test input points



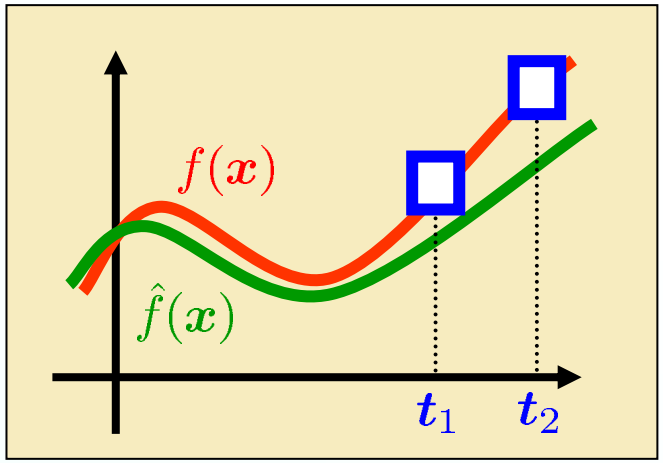
Transductive Inference

- Test input points are specified in advance.
- We do not have to estimate the entire function $f(\boldsymbol{x})$, but just estimate the values of the function at the test input points $\{\boldsymbol{t}_i\}_{i=1}^{n_t}$.

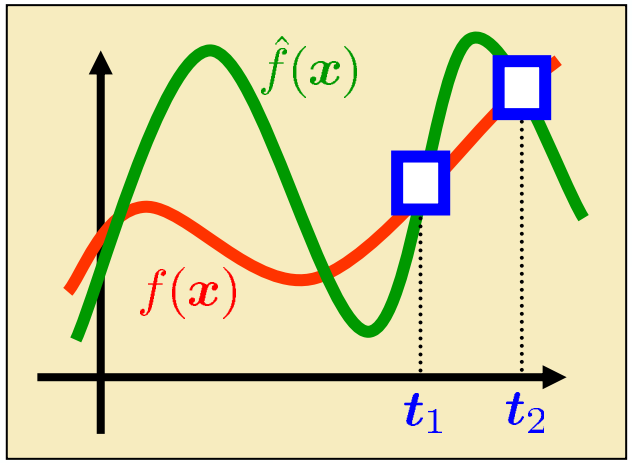


Model Selection for Transductive Inference

- Test error at given test input points is different from the generalization error.
- Model should be chosen so that the test error only at $\{\boldsymbol{t}_i\}_{i=1}^{n_t}$ is minimized.



Small generalization error
Large test error



Large generalization error
Small test error

Goal of Our Research

- We want to estimate the test error at the given test input points!

$$J_{test} = \mathrm{E} \sum_{i=1}^{n_t} \left(\hat{f}(\boldsymbol{t}_i) - f(\boldsymbol{t}_i) \right)^2$$

E :Expectation over noise

Setting

- Linear regression model

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

α_i :Parameters

$\{\varphi_i(\boldsymbol{x})\}_{i=1}^{p}$:Fixed basis functions

- Linear estimation

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

$\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_p)^\top$

$\boldsymbol{X}$:A matrix

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

- Realizability

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

α_i^* :Unknown true parameters

Bias / Variance Decomposition

$$J_{test} = \mathrm{E} \sum_{i=1}^{n_t} \left(\hat{f}(\boldsymbol{t}_i) - f(\boldsymbol{t}_i) \right)^2$$

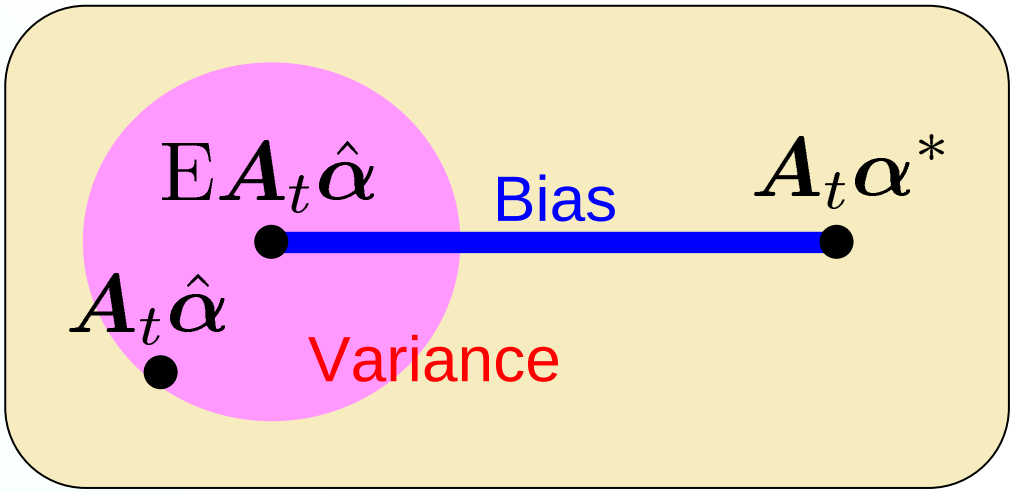
$$= \mathrm{E} \| \boldsymbol{A}_t \hat{\boldsymbol{\alpha}} - \boldsymbol{A}_t \boldsymbol{\alpha}^* \|^2$$

$$= \underbrace{\| \mathrm{E} \boldsymbol{A}_t \hat{\boldsymbol{\alpha}} - \boldsymbol{A}_t \boldsymbol{\alpha}^* \|^2}_{\text{Bias}} + \underbrace{\mathrm{E} \| \boldsymbol{A}_t \hat{\boldsymbol{\alpha}} - \mathrm{E} \boldsymbol{A}_t \hat{\boldsymbol{\alpha}} \|^2}_{\text{Variance}}$$

$$[\boldsymbol{A}_t]_{i,j} = \varphi_j(\boldsymbol{t}_i)$$

$$\boldsymbol{A}_t \hat{\boldsymbol{\alpha}} = (\hat{f}(\boldsymbol{t}_1), \hat{f}(\boldsymbol{t}_2), \ldots, \hat{f}(\boldsymbol{t}_{n_t}))^\top$$

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



Tricks for Estimating Bias

Sugiyama & Ogawa (Neural Comp., 2001)
Sugiyama & Müller (JMLR, 2002)

$$Bias = \|\mathrm{E}\boldsymbol{A}_t\hat{\boldsymbol{\alpha}} - \boldsymbol{A}_t\boldsymbol{\alpha}^*\|^2$$

- True parameter $\boldsymbol{\alpha}^*$ is unknown.
- We utilize an unbiased estimator of the true parameter for estimating the bias.

$$\hat{\boldsymbol{\alpha}}_u = \boldsymbol{A}^\dagger \boldsymbol{y}$$

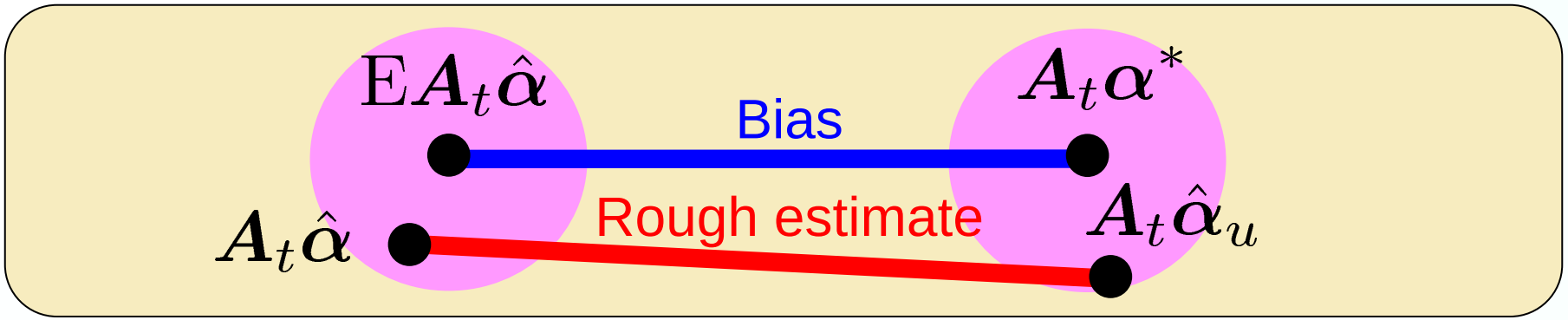
$$\mathrm{E}\hat{\boldsymbol{\alpha}}_u = \boldsymbol{\alpha}^*$$

$\boldsymbol{A}$:Design matrix $\quad$ $\dagger$:Generalized inverse

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

$$\boldsymbol{A}\hat{\boldsymbol{\alpha}} = (\hat{f}(\boldsymbol{x}_1), \hat{f}(\boldsymbol{x}_2), \ldots, \hat{f}(\boldsymbol{x}_n))^\top$$

Unbiased Estimator of Bias



$$Bias = \|\mathrm{E}\boldsymbol{A}_t\hat{\boldsymbol{\alpha}} - \boldsymbol{A}_t\boldsymbol{\alpha}^*\|^2$$

$$\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^\top$$
$$\boldsymbol{z} = (f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots, f(\boldsymbol{x}_n))^\top$$

$$= \|\boldsymbol{A}_t\hat{\boldsymbol{\alpha}} - \boldsymbol{A}_t\hat{\boldsymbol{\alpha}}_u\|^2 - 2\langle \boldsymbol{A}_t\boldsymbol{X}\boldsymbol{z}, \boldsymbol{A}_t\boldsymbol{A}^\dagger\boldsymbol{\epsilon}\rangle - \|\boldsymbol{A}_t(\boldsymbol{X} - \boldsymbol{A}^\dagger)\boldsymbol{\epsilon}\|^2$$

E E

$$\widehat{Bias} = \|\boldsymbol{A}_t\hat{\boldsymbol{\alpha}} - \boldsymbol{A}_t\hat{\boldsymbol{\alpha}}_u\|^2 - 0 - \sigma^2\mathrm{tr}\left(\boldsymbol{A}_t(\boldsymbol{X} - \boldsymbol{A}^\dagger)[\boldsymbol{A}_t(\boldsymbol{X} - \boldsymbol{A}^\dagger)]^\top\right)$$

$$\mathrm{E}\,\widehat{Bias} = Bias$$

Unbiased Estimator of Variance

- $$Var = \mathrm{E}\|\boldsymbol{A}_t\hat{\boldsymbol{\alpha}} - \mathrm{E}\boldsymbol{A}_t\hat{\boldsymbol{\alpha}}\|^2$$
$$= \sigma^2 \mathrm{tr}\left(\boldsymbol{A}_t\boldsymbol{X}(\boldsymbol{A}_t\boldsymbol{X})^\top\right)$$

  σ^2 :Noise variance

- An unbiased estimator of noise variance:

$$\hat{\sigma}^2 = \frac{\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{A}^\dagger\boldsymbol{y}\|^2}{n-p}$$

- $\widehat{Var} = \hat{\sigma}^2 \mathrm{tr}\left(\boldsymbol{A}_t\boldsymbol{X}(\boldsymbol{A}_t\boldsymbol{X})^\top\right)$

$$\mathrm{E}\,\widehat{Var} = Var$$

Unbiased Estimator of Test Error

- Adding bias and variance estimators, we have an unbiased estimator of test error.

$$\hat{J}_{test} = \widehat{Bias} + \widehat{Var}$$

- For simplicity, we ignore constant terms

$$\hat{J}_t = \|\boldsymbol{A}_t\boldsymbol{X}\boldsymbol{y}\|^2 - 2\langle \boldsymbol{A}_t\boldsymbol{X}\boldsymbol{y}, \boldsymbol{A}_t\boldsymbol{A}^\dagger\boldsymbol{y}\rangle + 2\hat{\sigma}^2 \operatorname{tr}\left(\boldsymbol{A}_t\boldsymbol{X}(\boldsymbol{A}_t\boldsymbol{A}^\dagger)^\top\right)$$

$$\mathrm{E}\hat{J}_t = J_t$$

$$J_t = \mathrm{E}\sum_{i=1}^{n_t}\left(\hat{f}(\boldsymbol{t}_i) - f(\boldsymbol{t}_i)\right)^2 - \sum_{i=1}^{n_t} f(\boldsymbol{t}_i)^2$$

Unrealizable Cases

- So far, we assumed that the model includes the underlying function.

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

α_i^* :Unknown true parameters

- We can prove that even when the above assumption is not rigorously fulfilled, $\hat{J}_t$ is still almost unbiased.

$$\mathrm{E}\hat{J}_t \approx J_t$$

Simulation: Toy Data Sets

- Basis functions: 10 Gaussian functions centered at equally located points in $[-\pi, \pi]$.
- Target function: sinc-like function (realizable).
- Training examples $\{(x_i, y_i)\}_{i=1}^{n}$:

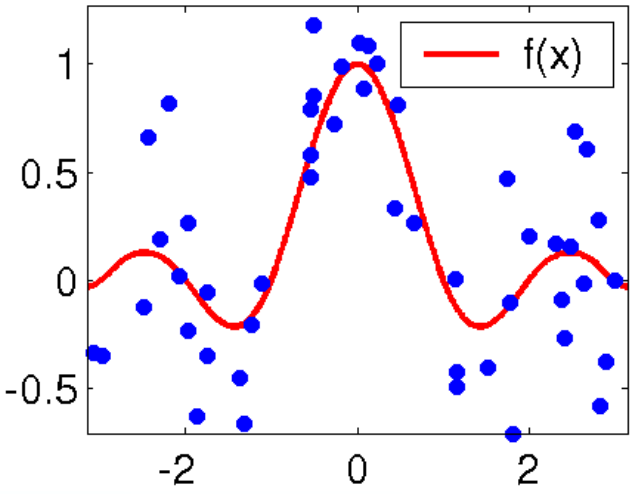
$$x_i \overset{i.i.d.}{\sim} U(-\pi, \pi)$$

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

- Test input points $\{t_i\}_{i=1}^{50}$:

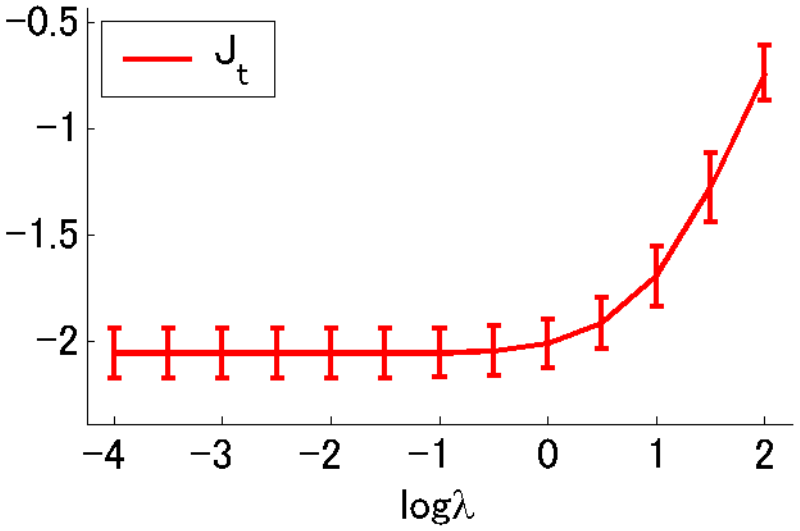
$$t_i \overset{i.i.d.}{\sim} U(-\pi, \pi)$$

- Ridge estimation is used for learning.

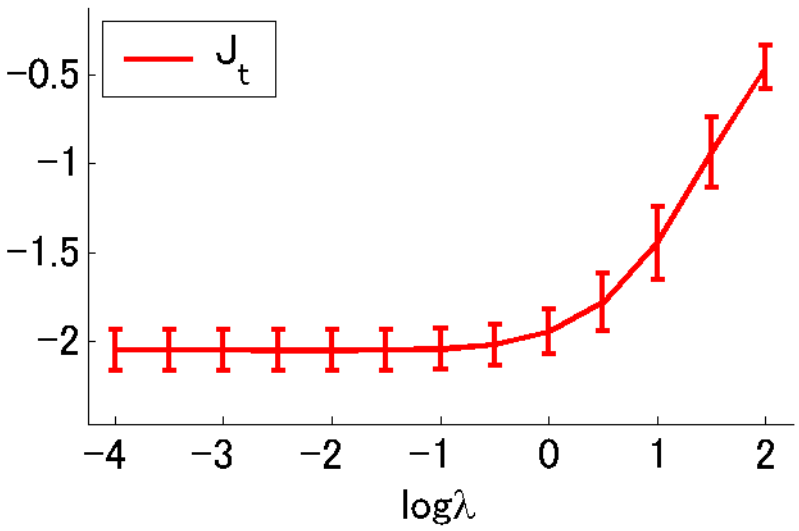


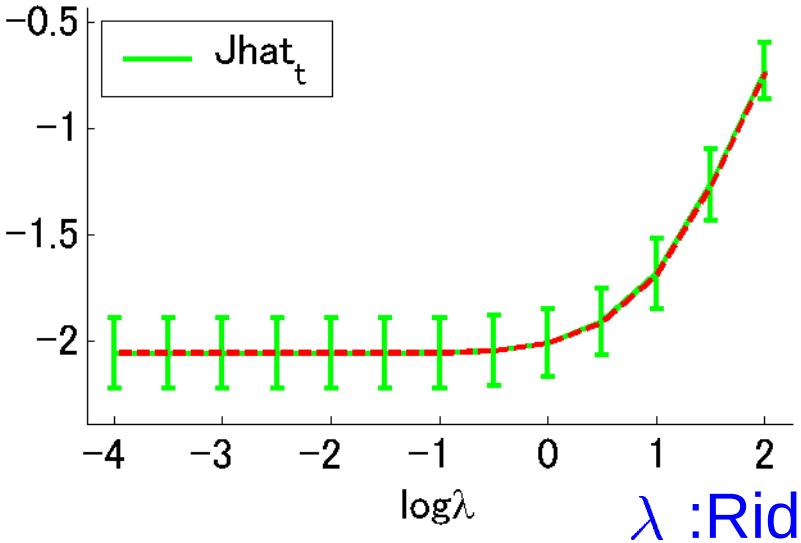
Results (1)

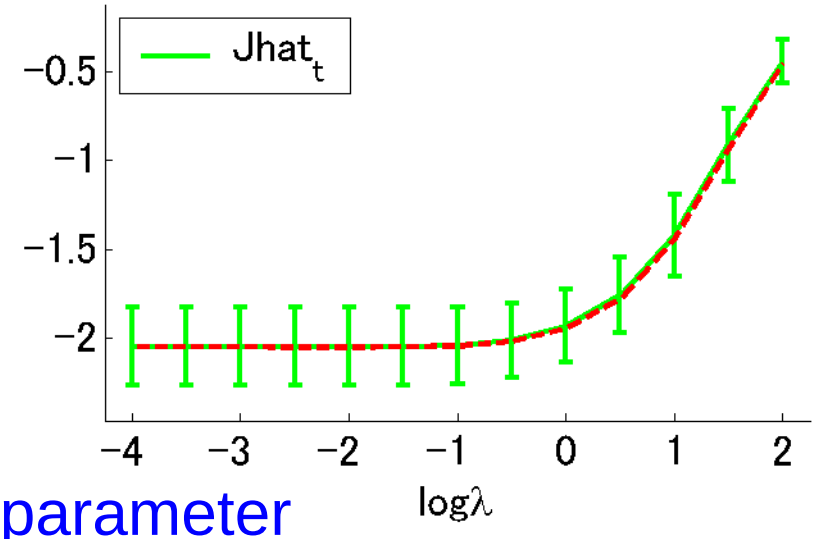
$(n, \sigma^2) = (100, 0.01)$



$(n, \sigma^2) = (50, 0.01)$



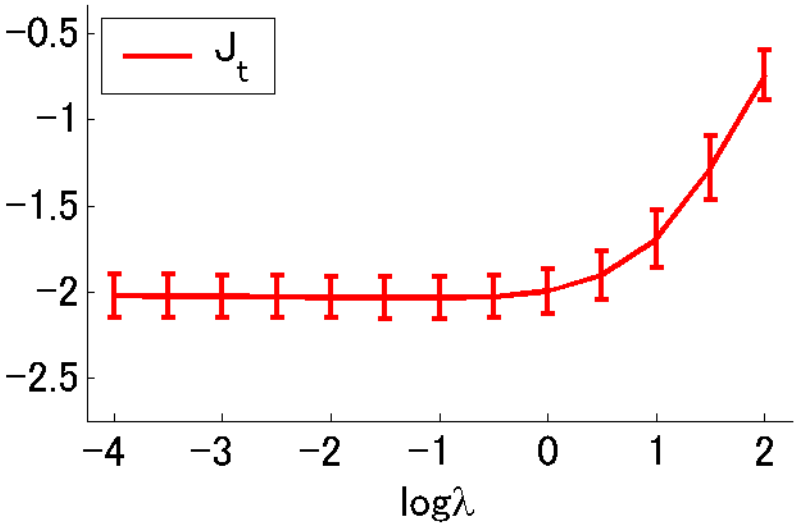




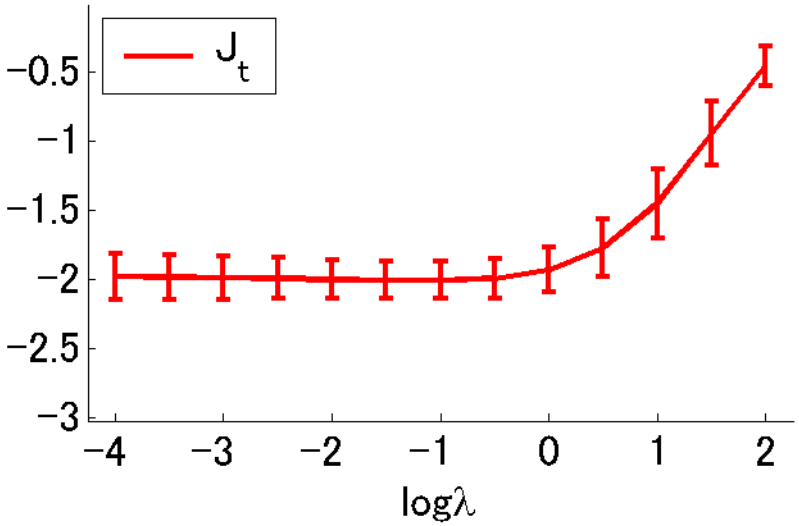
λ :Ridge parameter

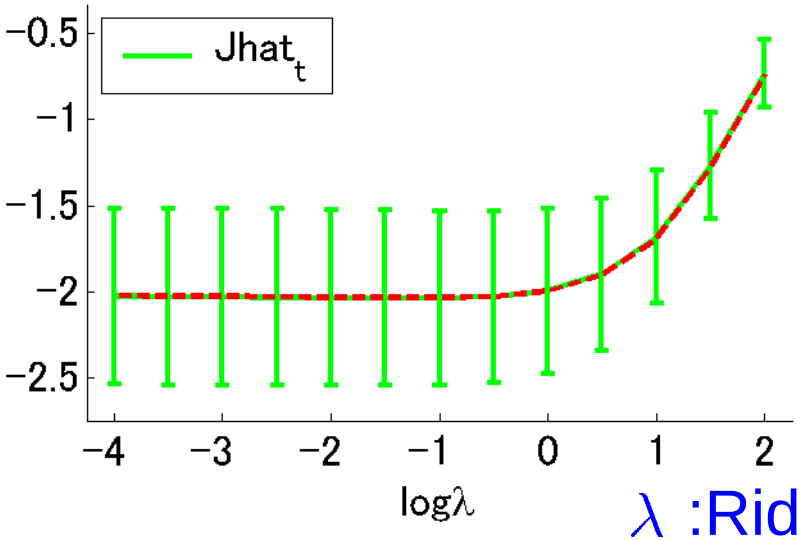
Results (2)

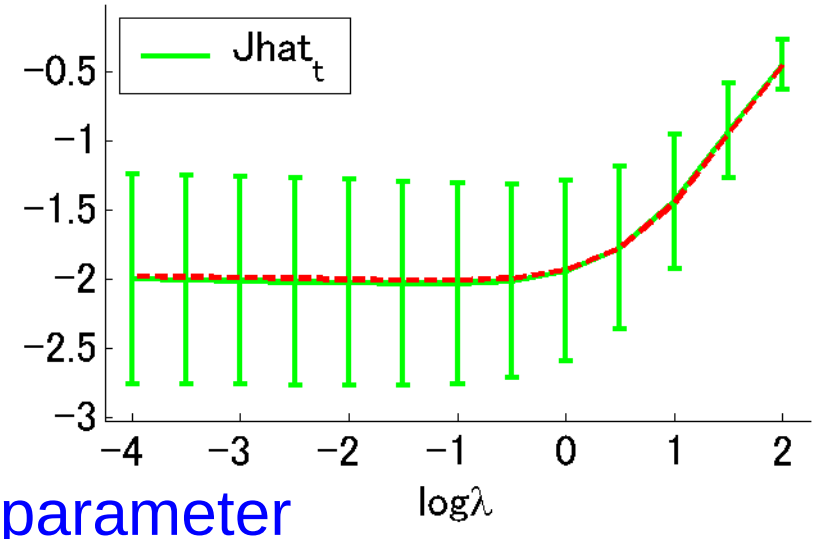
$(n, \sigma^2) = (100, 0.09)$



$(n, \sigma^2) = (50, 0.09)$







λ :Ridge parameter

Simulation: DELVE Data Sets

- Training set: 100 randomly selected samples.
- Test set: 50 randomly selected samples.
- Basis functions: Gaussian function centered at first 50 training input points.
- Ridge estimation is used for learning.
- Ridge parameter is selected by the proposed method, leave-one-out cross-validation, or an empirical Bayesian method.

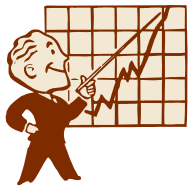
Normalized Test Errors

Mean (Standard deviation)

Data set	Proposed method	LOO cross-validation	Empirical Bayes
Boston	1.17 (0.54)	1.26 (0.58)	1.39 (0.59)
Bank-8fm	1.07 (0.29)	1.11 (0.32)	1.09 (0.31)
Bank-8nm	1.09 (0.51)	1.12 (0.56)	1.18 (0.60)
Kin-8fm	1.06 (0.32)	1.17 (0.36)	1.68 (0.48)
Kin-8nm	1.11 (0.27)	1.09 (0.24)	1.15 (0.24)

Red: Best and others with no significant difference by 99% t-test

Proposed method can be successfully applied to transductive model selection!



Conclusions

- Model selection is usually carried out so that estimated generalization error is minimized.
- When test input points are specified in advance (transductive inference), it is natural to choose a model so that the test error only at the test input points is minimized.
- We derived an unbiased estimator of the test error at given test input points.
- Simulation showed the proposed method works well in practical situations.