Model Selection under Covariate Shift

Masashi Sugiyama

Tokyo Institute of Technology, Tokyo, Japan

Klaus-Robert Müller

Fraunhofer FIRST, Berlin, Germany

University of Potsdam, Potsdam, Germany

Standard Regression Problem

- Learning target function: $f(\boldsymbol{x})$
- Training examples:

$$\{(\boldsymbol{x}_i, y_i) \mid y_i = f(\boldsymbol{x}_i) + \epsilon_i\}_{i=1}^{n}$$

- Test input: $\{\boldsymbol{t}_i \mid \boldsymbol{t}_i \overset{i.i.d.}{\sim} p_t(\boldsymbol{x})\}_{i=1}^{m}$
- Goal: Obtain approximation $\hat{f}(\boldsymbol{x})$ that minimizes expected error for test inputs (or generalization error)

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

Training Input Distribution

- Common assumption:

  Training input $\{\boldsymbol{x}_i\}_{i=1}^{n}$ follows the same distribution as test input:

$$\boldsymbol{x}_i \overset{i.i.d.}{\sim} p_t(\boldsymbol{x})$$

- Here, we suppose distributions are different.

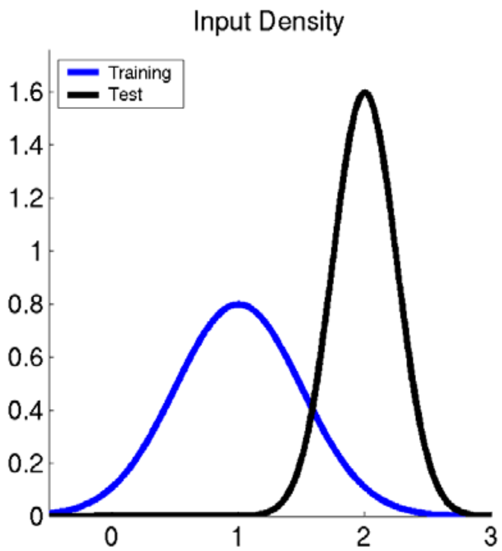
$$\boldsymbol{x}_i \overset{i.i.d.}{\sim} p_x(\boldsymbol{x}) \qquad p_x(\boldsymbol{x}) \neq p_t(\boldsymbol{x})$$

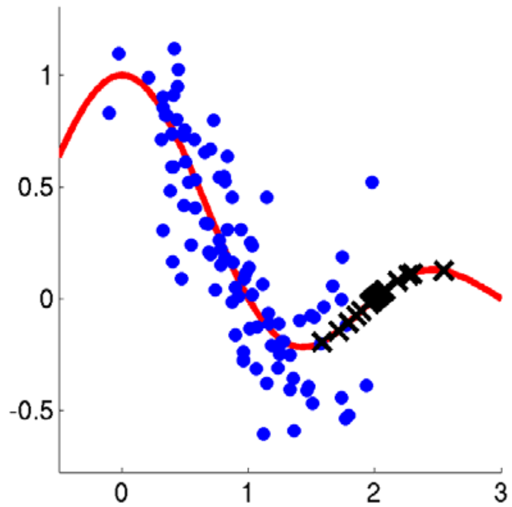
$$\boldsymbol{t}_i \overset{i.i.d.}{\sim} p_t(\boldsymbol{x})$$

$\Rightarrow$ Covariate shift

Covariate Shift

- Is covariate shift important to investigate?
- Yes! It often happens in reality.
  - Interpolation / extrapolation
  - Active learning (experimental design)
  - Classification from imbalanced data



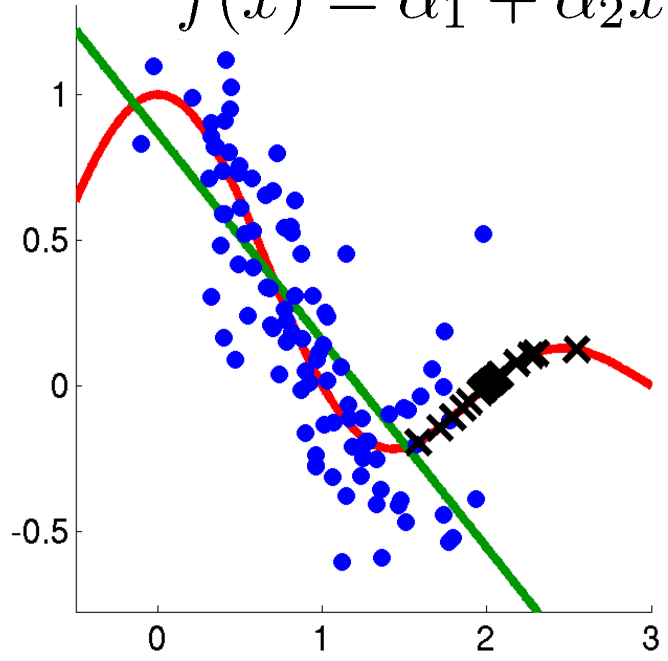


Ordinary Least Squares under Covariate Shift

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

- Asymptotically unbiased if model is correct.
- Asymptotically biased for misspecified models.
- Need to reduce bias.

$$\hat{f}(x) = \alpha_1 + \alpha_2 x$$



Weighted Least Squares for Covariate Shift

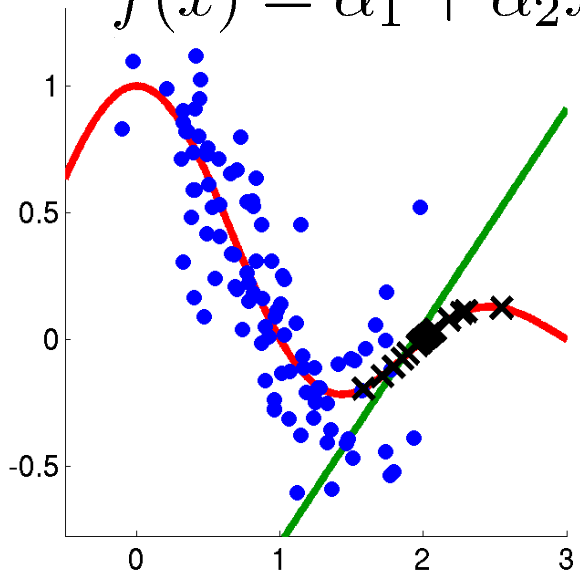
(Shimodaira, 2000)

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

$p_x(\boldsymbol{x}), p_t(\boldsymbol{x})$:Assumed known and strictly positive

- Asymptotically unbiased for misspecified models.
- Can have large variance.
- Need to reduce variance.

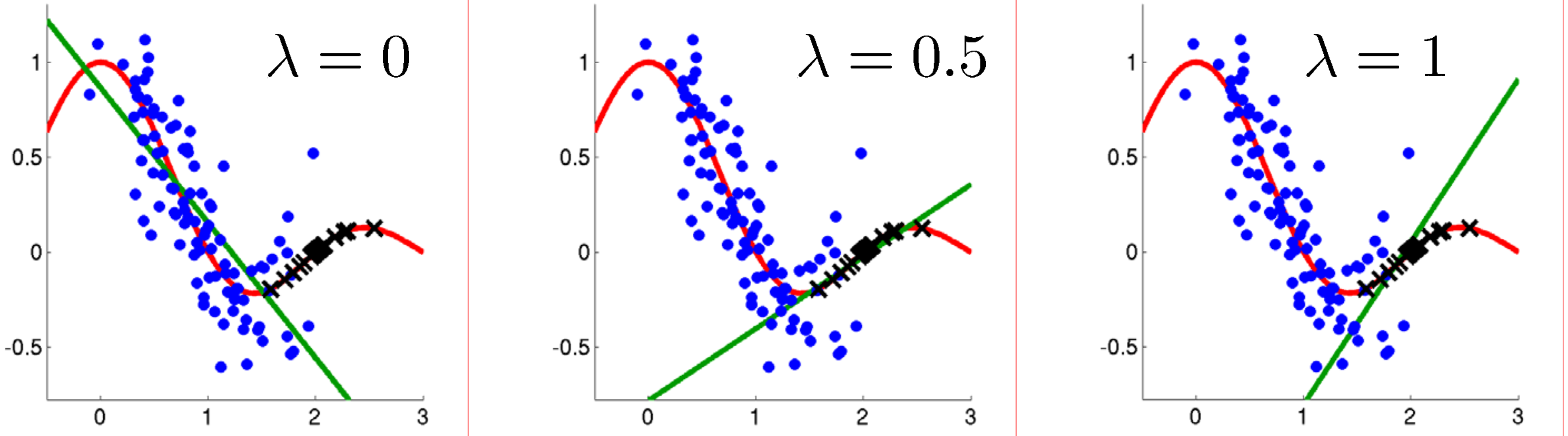
$\hat{f}(x) = \alpha_1 + \alpha_2 x$



λ -Weighted Least Squares

(Shimodaira, 2000)

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



Large bias
Small variance

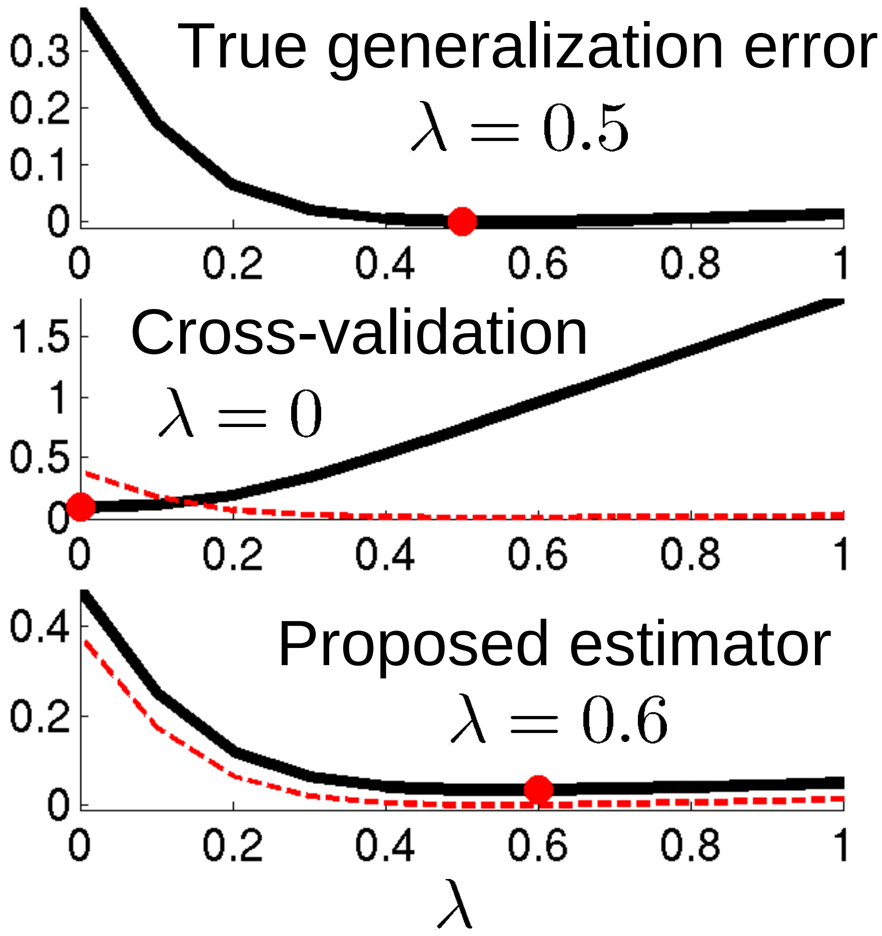
(Intermediate)

Small bias
Large variance

λ should be chosen appropriately!
(Model Selection)

Generalization Error Estimation under Covariate Shift

- λ is determined so that (estimated) generalization error is minimized.
- However, standard methods such as cross-validation is heavily biased.
- Goal: Derive better estimator



Setting

- I.i.d. noise with mean 0 and variance σ^2
- Linear regression model:

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

- λ-weighted least squares:

$$\min_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{n} \left(\frac{p_t(x_i)}{p_x(x_i)} \right)^{\lambda} \left(\hat{f}(x_i) - y_i \right)^2 \right]$$

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

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

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

$$\boldsymbol{D} = \mathrm{diag}\left(\frac{p_t(\boldsymbol{x}_i)}{p_x(\boldsymbol{x}_i)} \right)$$

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

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

Decomposition of Generalization Error

$$
\begin{aligned}
J &= \int \left(\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2 p_t(\boldsymbol{x}) d\boldsymbol{x} \\
&= \|\hat{f} - f\|^2_{L_2(p_t)} \\
&= \|\hat{f}\|^2_{L_2(p_t)} - 2\langle \hat{f}, f \rangle_{L_2(p_t)} + \|f\|^2_{L_2(p_t)}
\end{aligned}
$$

Accessible Estimated Constant (ignored)

- We estimate $Z \equiv \langle \hat{f}, f \rangle_{L_2(p_t)}$

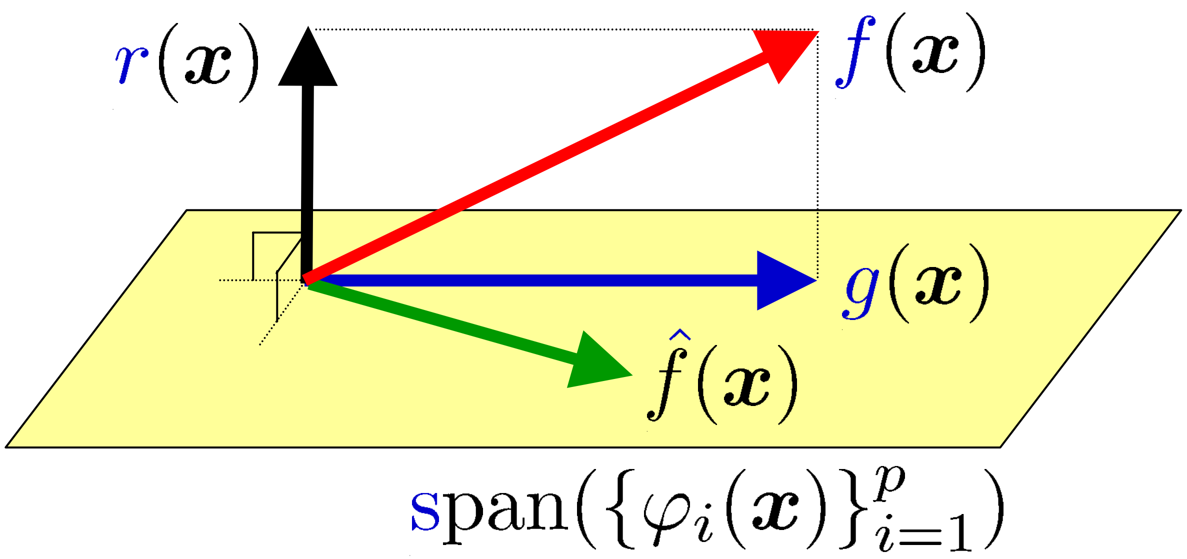
Orthogonal Decomposition of Learning Target Function

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

$$\langle \varphi_i, r \rangle_{L_2(p_t)} = 0$$

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

$\boldsymbol{\alpha}^*$:Optimal parameter



$$\begin{aligned} Z &= \langle \hat{f}, f \rangle_{L_2(p_t)} \\ &= \langle \hat{f}, g \rangle_{L_2(p_t)} \\ &= \langle \boldsymbol{U}\hat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}^* \rangle \end{aligned}$$

$$\boldsymbol{U}_{i,j} = \langle \varphi_i, \varphi_j \rangle_{L_2(p_t)}$$

Unbiased Estimation of $\mathbb{E}_\epsilon Z$

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

- Suppose we have
  - $\boldsymbol{L}_u$, which gives linear unbiased estimator of $\boldsymbol{\alpha}^*$

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

  - σ_u^2 :Unbiased estimator of noise variance

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

- Then we have an unbiased estimator of $\mathbb{E}_\epsilon Z$:

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

- But $\boldsymbol{L}_u, \sigma_u^2$ are not always available.

Use approximations instead

Approximations of $\boldsymbol{L}_u, \sigma_u^2$

- $\widehat{\boldsymbol{L}}_u = (\boldsymbol{X}^\top \boldsymbol{D} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{D}$
- $\widehat{\sigma_u^2} = \dfrac{\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y}\|^2}{n-p}$ $\quad \boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$

$\mathbb{E}_{\boldsymbol{\epsilon}} \boldsymbol{L}_u \boldsymbol{y} = \boldsymbol{\alpha}^*$ $\quad \mathbb{E}_{\boldsymbol{\epsilon}} \sigma_u^2 = \sigma^2$

- If model is correct,

$$\mathbb{E}_{\boldsymbol{\epsilon}} \widehat{\boldsymbol{L}}_u \boldsymbol{y} = \boldsymbol{\alpha}^* \qquad \mathbb{E}_{\boldsymbol{\epsilon}} \widehat{\sigma_u^2} = \sigma^2$$

- If model is misspecified,

$$\mathbb{E}_{\boldsymbol{\epsilon}} \widehat{\boldsymbol{L}}_u \boldsymbol{y} \to \boldsymbol{\alpha}^* \qquad \mathbb{E}_{\boldsymbol{\epsilon}} \widehat{\sigma_u^2} \not\to \sigma^2 \qquad (n \to \infty)$$

New 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: $B_{\boldsymbol{\epsilon}} = \mathbb{E}_{\boldsymbol{\epsilon}}[\hat{J} - J] + C$ $\quad C = \|f\|_{L_2(p_t)}^2$

- If model is correct,

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

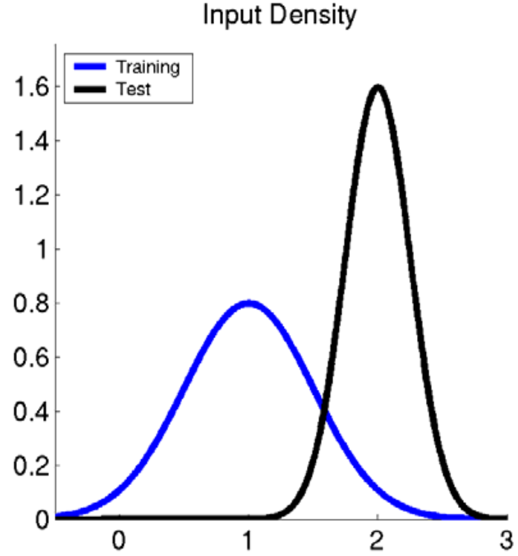
- If model is almost correct,

$$B_{\boldsymbol{\epsilon}} = \mathcal{O}(\delta) \qquad \delta = \max\{r(\boldsymbol{x}_i)\}$$

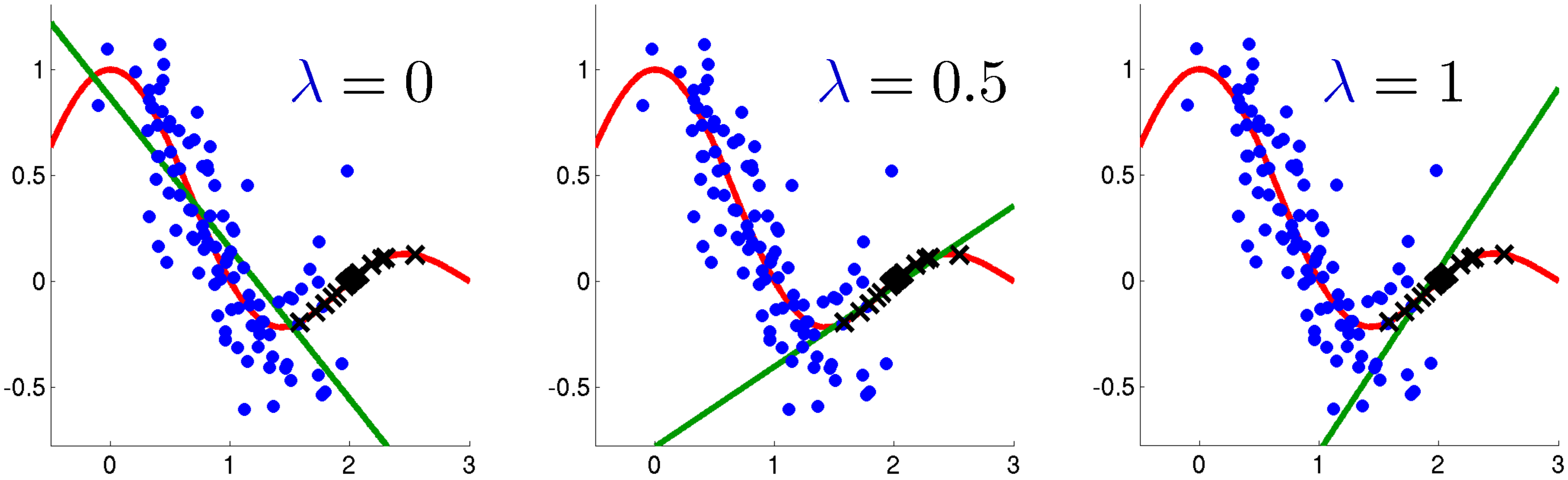
- If model is misspecified,

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

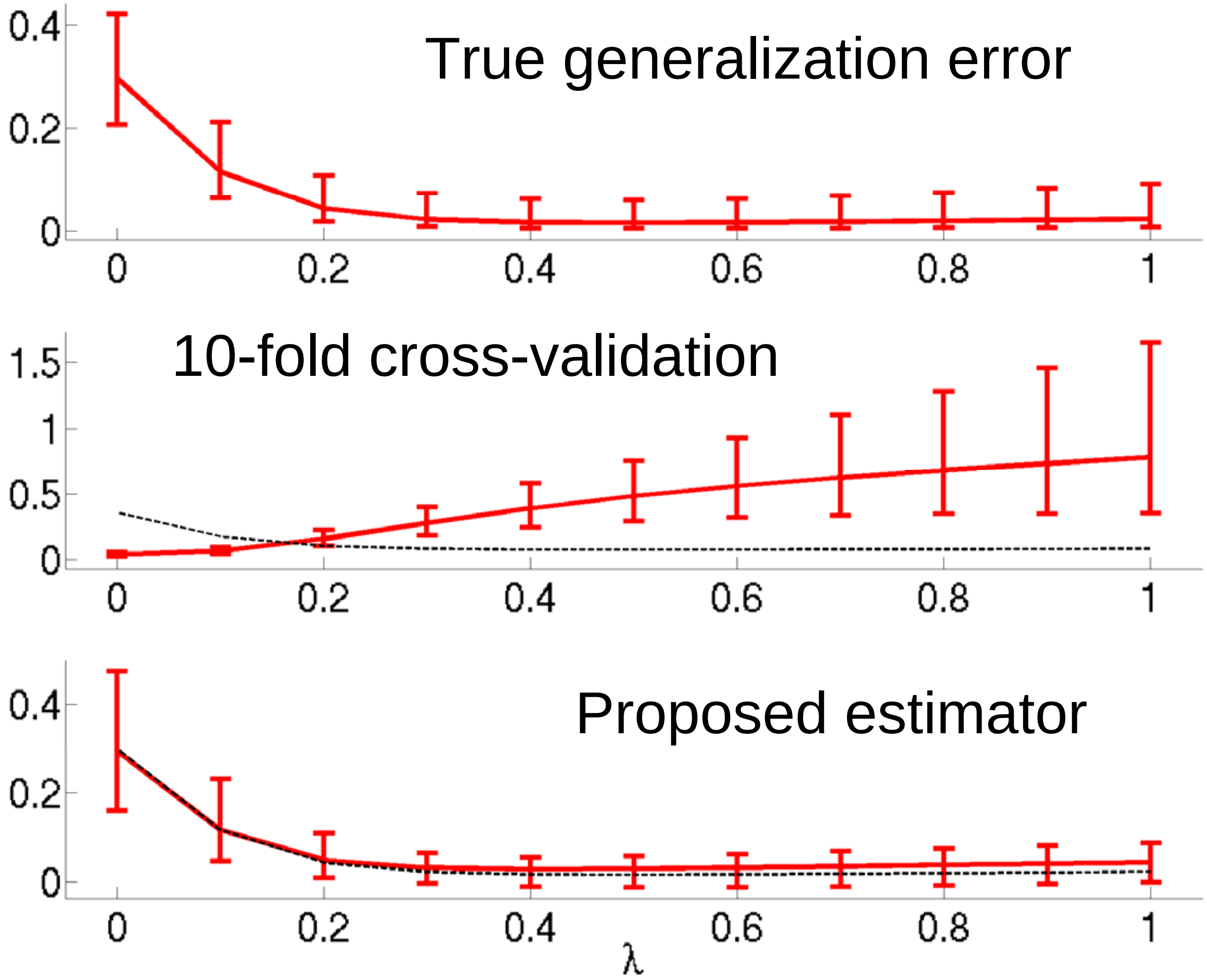
Simulation (Toy)



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

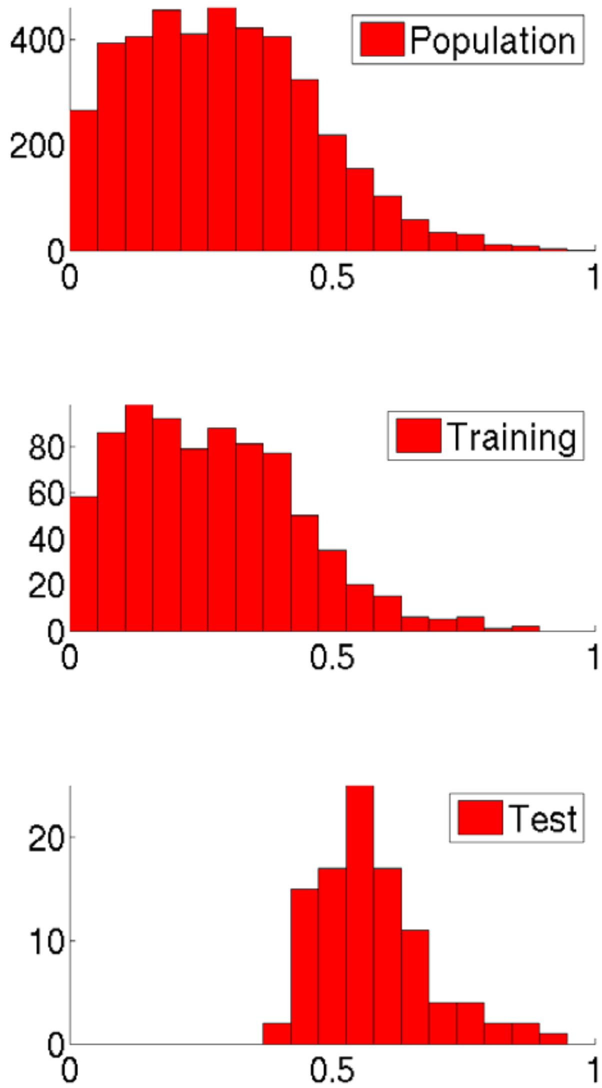


Results


True generalization error
10-fold cross-validation
Proposed estimator
0.4
0.2
0
1.5
1
0.5
0
0.4
0.2
0
0.2
0.4
0.6
0.8
1
λ

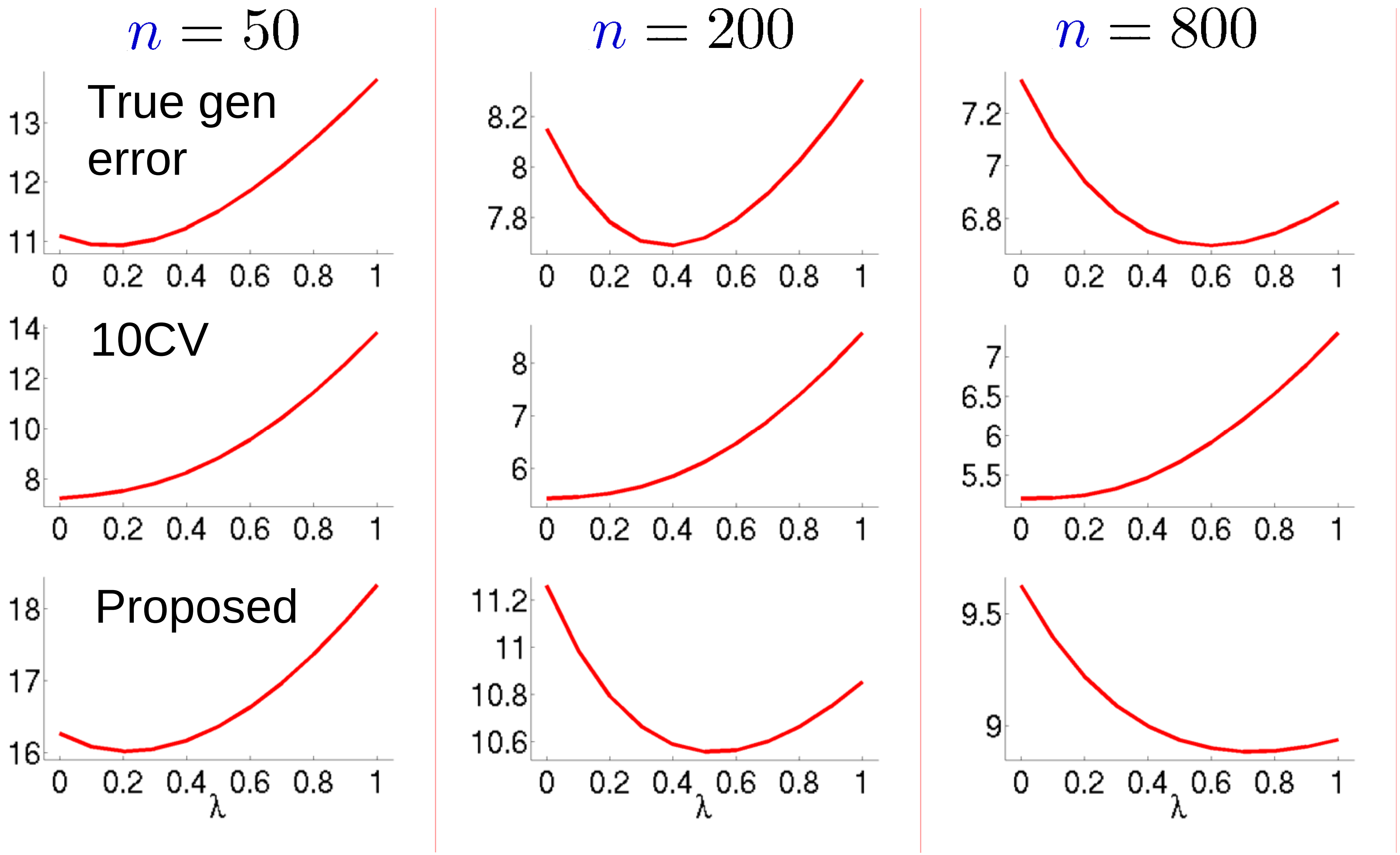
Simulation (Abalone from DELVE)

- Estimate the age of abalones from 7 physical measurements.
- We add bias to 4th attribute (weight of abalones)
- Training and test input densities are estimated by standard kernel density estimator.
- $$\hat{f}(\boldsymbol{x}) = \alpha_1 + \sum_{i=1}^{7} \alpha_{i+1} x^{(i)}$$



Generalization Error Estimation

Mean over 300 trials



Test Error After Model Selection

Extrapolation in 4th attribute

n	50	200	800
OPT	9.86±4.27	7.40±1.77	6.54±1.34
Proposed	11.67±5.74	7.95±2.15	6.77±1.40
10CV	10.88±5.05	8.06±1.91	7.24±1.37

T-test (5%)

Extrapolation in 6th attribute

n	50	200	800
OPT	9.04±4.04	6.76±1.68	6.05±1.25
Proposed	10.67±6.19	7.31±2.24	6.20±1.33
10CV	10.15±4.95	7.42±1.81	6.68±1.25

Conclusions

- Covariate shift: Training and test input distributions are different
- Ordinary LS: Biased
- Weighted LS: Unbiased but large variance.
- λ -WLS: Model selection needed.
- Cross-validation: Biased
- Proposed generalization error estimator:
  - Exactly unbiased (correct models)
  - Asymptotically unbiased (misspecified models)