Importance-Weighted Cross-Validation for Covariate Shift

Masashi Sugiyama[1], Benjamin Blankertz[2], Matthias Krauledat[2,3], Guido Dornhege[2], Klaus-Robert Müller[3,2]

[1] Tokyo Institute of Technology, Tokyo, Japan
[2] Fraunhofer FIRST.IDA, Berlin, Germany
[3] Technical University Berlin, Berlin, Germany

Common Assumption in Supervised Learning

- Goal: from given training samples, predict output of unseen test samples
- To do so, we always assume

Training and test samples are drawn from the same distribution

$$P_{train}(\boldsymbol{x}, y) = P_{test}(\boldsymbol{x}, y)$$

- Is this assumption really true?

Not Always True!

- Less women in face dataset than reality.
- More criticisms in survey sampling than reality.
- Tend to collect easy-to-gather samples for training.
- Sample generation mechanism varies over time.

The Yale Face Database B

Brain activity data

Covariate Shift

- However, no chance for generalization if training and test samples have nothing in common.

$$P_{train}(\boldsymbol{x}, y) \neq P_{test}(\boldsymbol{x}, y)$$

- Covariate shift:
  - Input distribution changes

$$P_{train}(\boldsymbol{x}) \neq P_{test}(\boldsymbol{x})$$

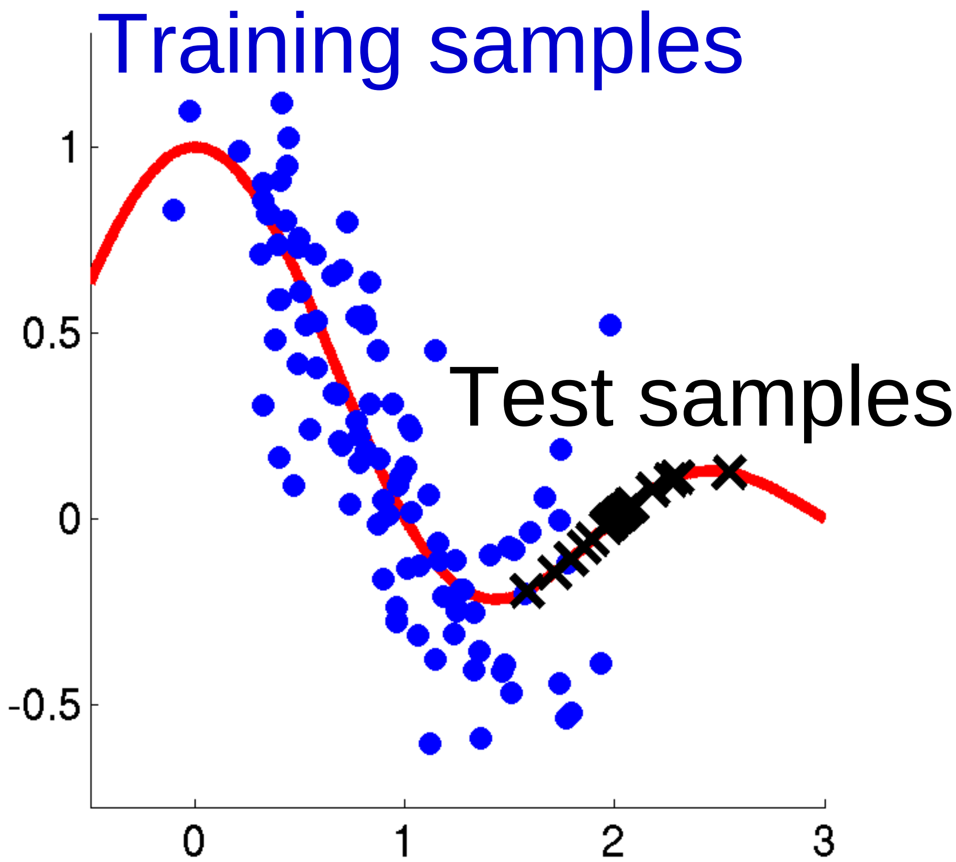
  - Functional relation remains unchanged

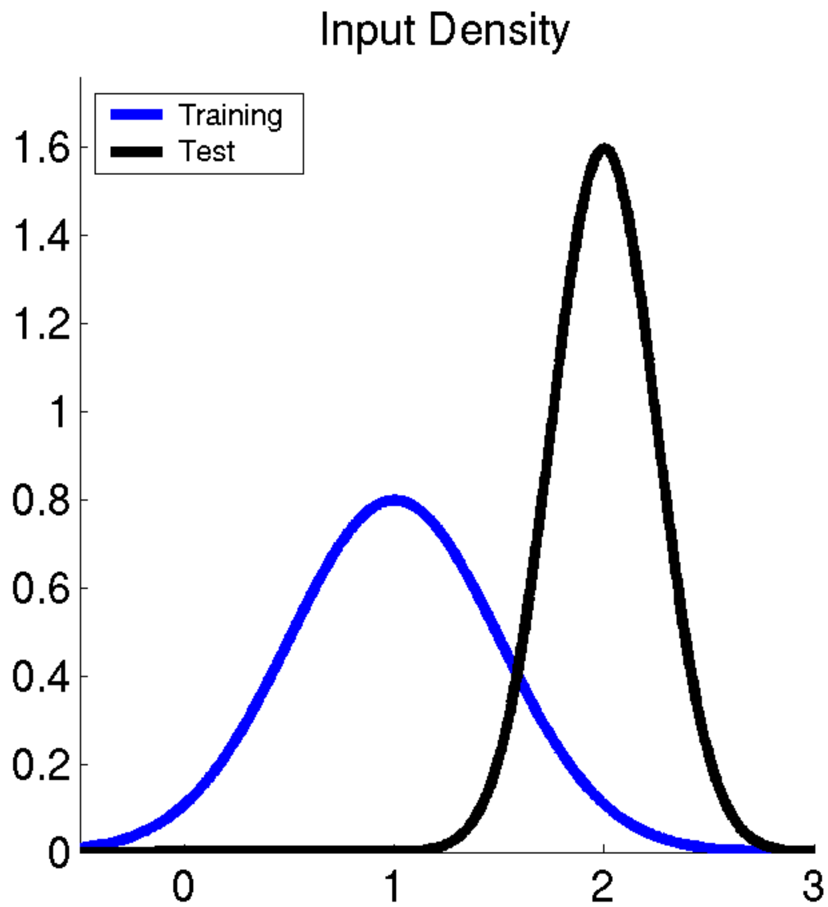
$$P_{train}(y|\boldsymbol{x}) = P_{test}(y|\boldsymbol{x})$$

Examples of Covariate Shift

(Weak) extrapolation:

Predict output values outside training region

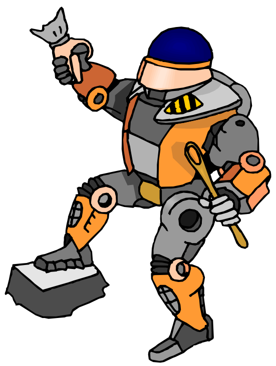




Examples (cont.)

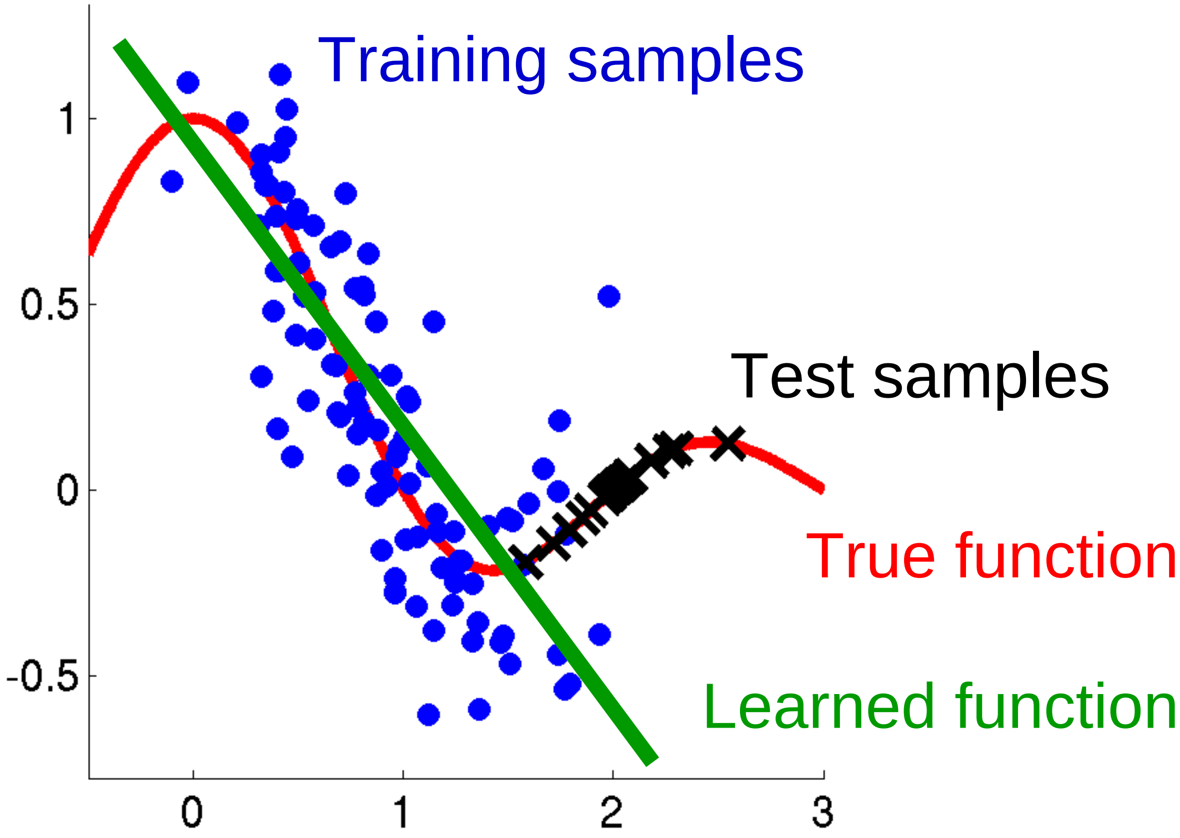
- Possible applications:
  - Non-stationarity compensation in brain-computer interface
  - Online system adaptation in robot motor control
  - Correcting sample selection bias in survey sampling
  - Active learning (experimental design)

Sugiyama (JMLR2006)



Covariate Shift

- To illustrate the effect of covariate shift, let's focus on linear extrapolation

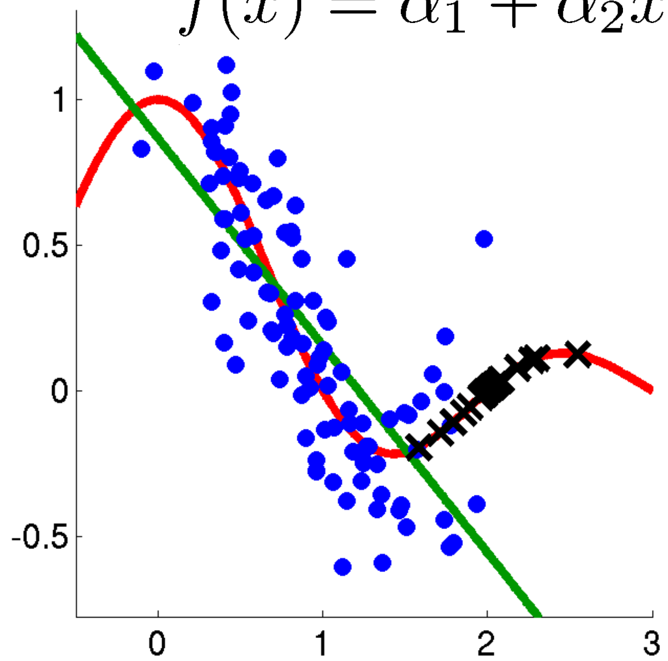


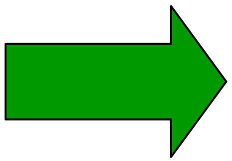
Ordinary Least-Squares

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

- If model is correct:
  OLS minimizes bias asymptotically
- If model is misspecified:
  OLS does not minimize bias even asymptotically.

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





We don't have correct model in practice, so we need to reduce bias!

Law of Large Numbers

- Sample average converges to the population mean:

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

$$\frac{1}{n}\sum_{i=1}^{n} A(\boldsymbol{x}_i) \longrightarrow \int A(\boldsymbol{x}) p_{train}(\boldsymbol{x}) d\boldsymbol{x}$$

- We want to estimate the expectation over test input points only using training input points $\{\boldsymbol{x}_i\}_{i=1}^{n}$.

$$\int A(\boldsymbol{t}) p_{test}(\boldsymbol{t}) d\boldsymbol{t} \qquad \boldsymbol{t} \sim p_{test}(\boldsymbol{x})$$

Key Trick: Importance-Weighted Average

- Importance: Ratio of test and training input densities

$$\frac{p_{test}(\boldsymbol{x})}{p_{train}(\boldsymbol{x})}$$

- Importance-weighted average:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{p_{test}(\boldsymbol{x}_i)}{p_{train}(\boldsymbol{x}_i)}A(\boldsymbol{x}_i) \longrightarrow \int \frac{p_{test}(\boldsymbol{x})}{p_{train}(\boldsymbol{x})}A(\boldsymbol{x})p_{train}(\boldsymbol{x})d\boldsymbol{x}$$

$$\boldsymbol{x}_i \overset{i.i.d.}{\sim} p_{train}(\boldsymbol{x}) \qquad = \int A(\boldsymbol{x})p_{test}(\boldsymbol{x})d\boldsymbol{x}$$

$$\boldsymbol{t} \sim p_{test}(\boldsymbol{x})$$

(cf. importance sampling)

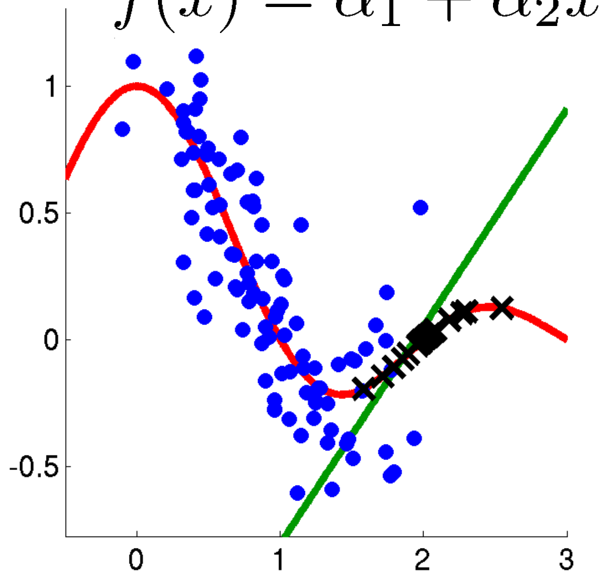
Importance-Weighted LS

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

$p_{train}(\boldsymbol{x}), p_{test}(\boldsymbol{x})$:Assumed known and strictly positive

- Even for misspedified models, IWLS minimizes bias asymptotically.

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

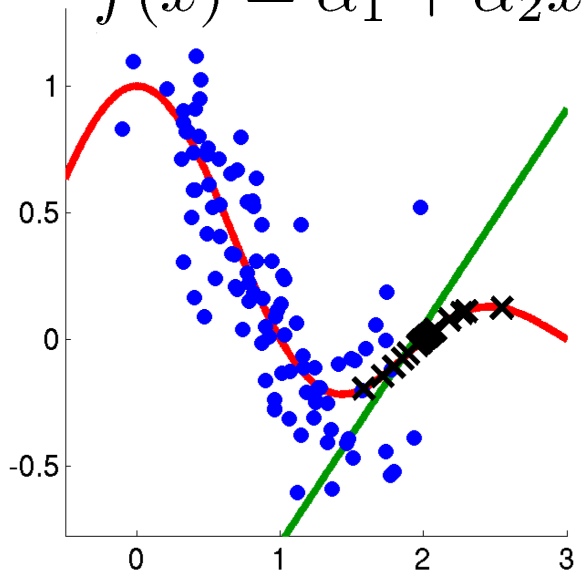


Importance-Weighted LS (cont.)

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

- However, variance of IWLS is larger than OLS (cf. BLUE)

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



⇒ We want to reduce variance

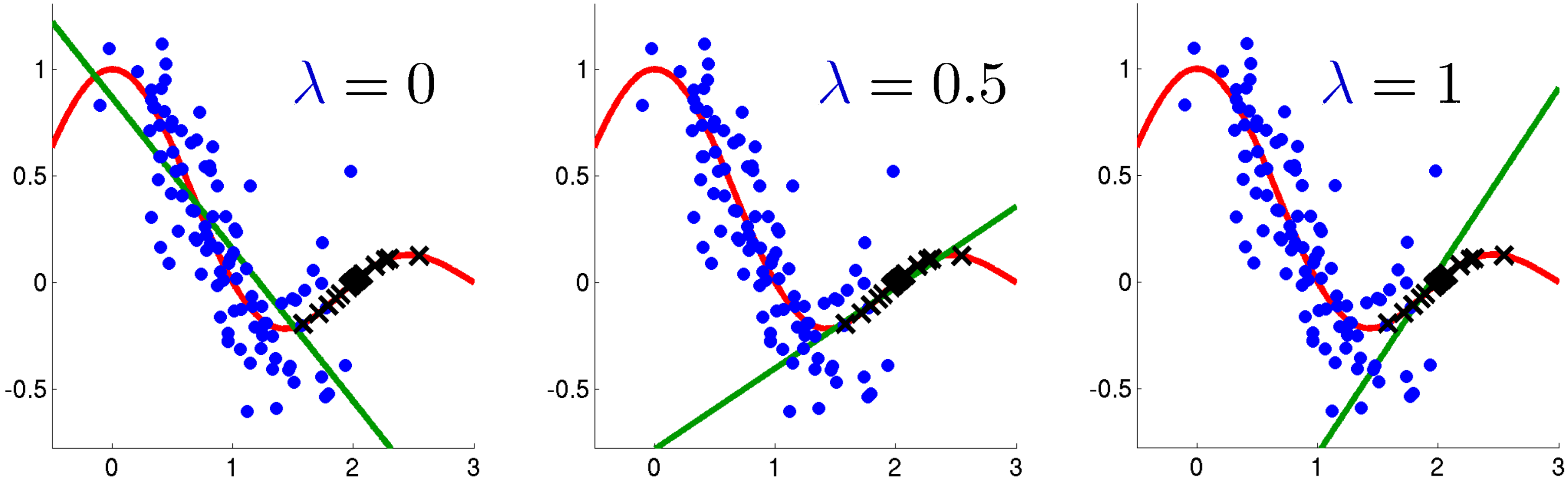
We reduce variance by adding small bias to IWLS (e.g., changing weight, regularization)

Adaptive IWLS

(Shimodaira, 2000)

$$\min_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{n} \left(\frac{p_{test}(\boldsymbol{x}_i)}{p_{train}(\boldsymbol{x}_i)} \right)^{\lambda} \left(\widehat{f}(\boldsymbol{x}_i) - y_i \right)^2 \right]$$

$$0 \leq \lambda \leq 1$$



Large bias
Small variance

Intermediate

Small bias
Large variance

Model Selection

$$\min_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{n} \left(\frac{p_{test}(\boldsymbol{x}_i)}{p_{train}(\boldsymbol{x}_i)} \right)^{\lambda} \left(\widehat{f}(\boldsymbol{x}_i) - y_i \right)^2 \right]$$

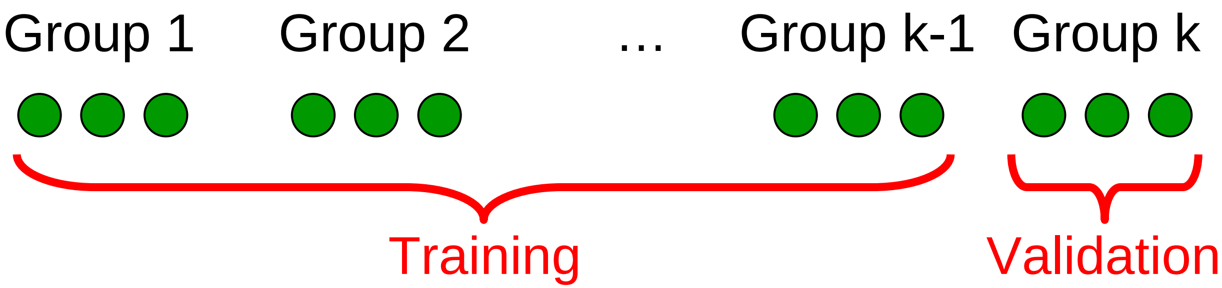
- We want to determine λ so that generalization error (bias+var) is minimized.

$$G = \int \left(\widehat{f}(\boldsymbol{x}) - f(\boldsymbol{x}) \right)^2 p_{test}(\boldsymbol{x}) d\boldsymbol{x}$$

- However, gen. error is inaccessible.
- We use a gen. error estimator instead.

Cross-Validation

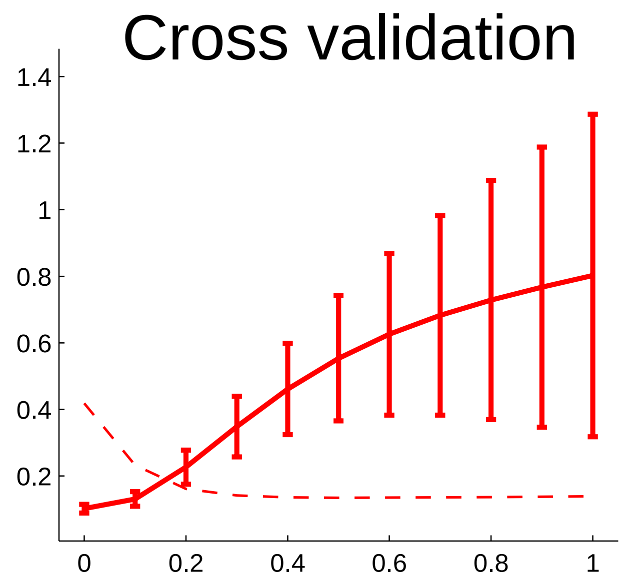
- A standard method for gen. error estimation
  - Divide training samples into k groups.
  - Train a learning machine with $k-1$ groups.
  - Validate the trained machine using the rest.
  - Repeat this for all combinations and output the mean validation error.



$$\widehat{f}(\boldsymbol{x}) \qquad \left(\widehat{f}(\boldsymbol{x}_t) - y_t\right)^2$$

CV under Covariate Shift

- CV is almost unbiased without covariate shift.
- However, it is heavily biased under covariate shift.

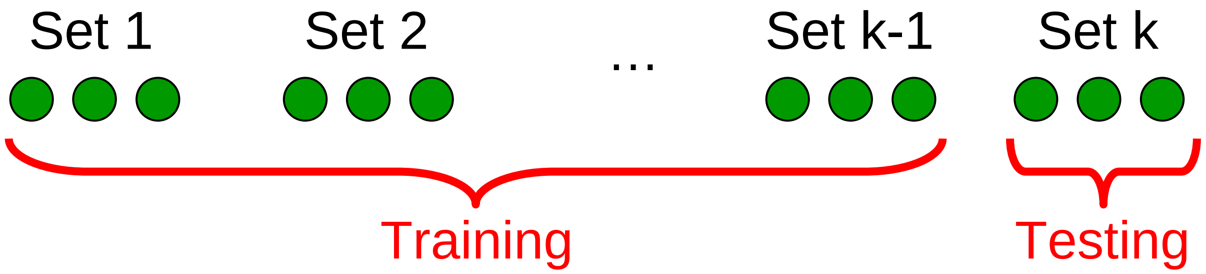


Goal of This Talk

- We propose a better generalization error estimator under covariate shift!

Importance-Weighted CV (IWCV)

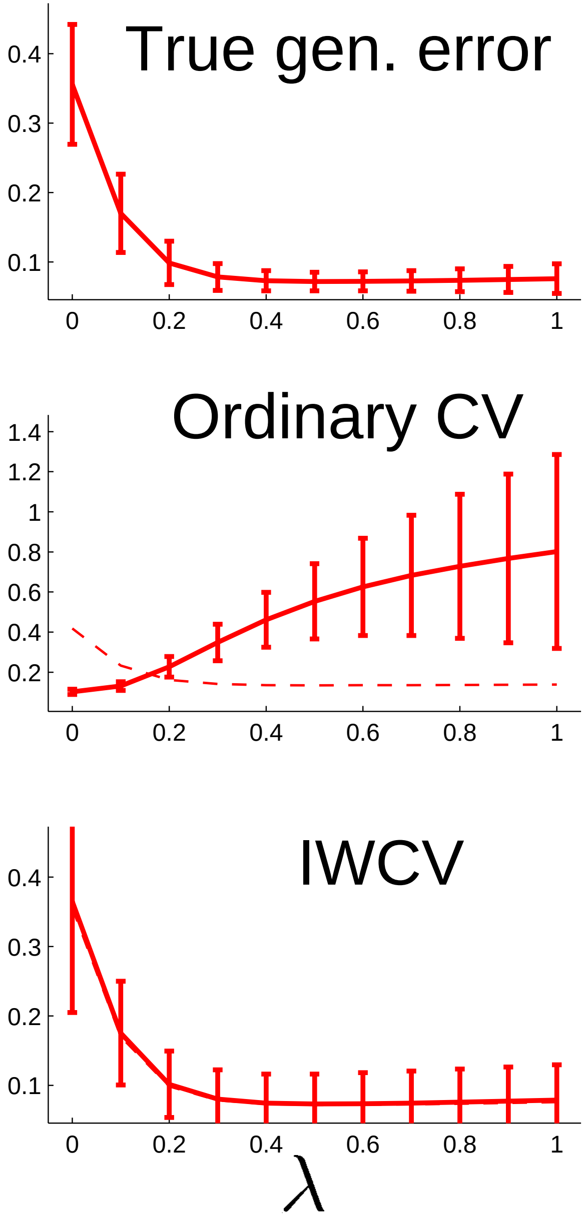
- When testing the classifier in CV process, we also importance-weight the test error.



$$\widehat{f}(\boldsymbol{x}) \qquad \frac{p_{test}(\boldsymbol{x}_t)}{p_{train}(\boldsymbol{x}_t)}\left(\widehat{f}(\boldsymbol{x}_t) - y_t\right)^2$$

IWCV gives almost unbiased estimates of gen. error even under covariate shift

Example of IWCV



Obtained generalization error

Ordinary CV	0.356(0.086)
IWCV	0.077(0.020)

Mean(Std.)

- IWCV is nicely unbiased
- Model selection by IWCV outperforms CV!

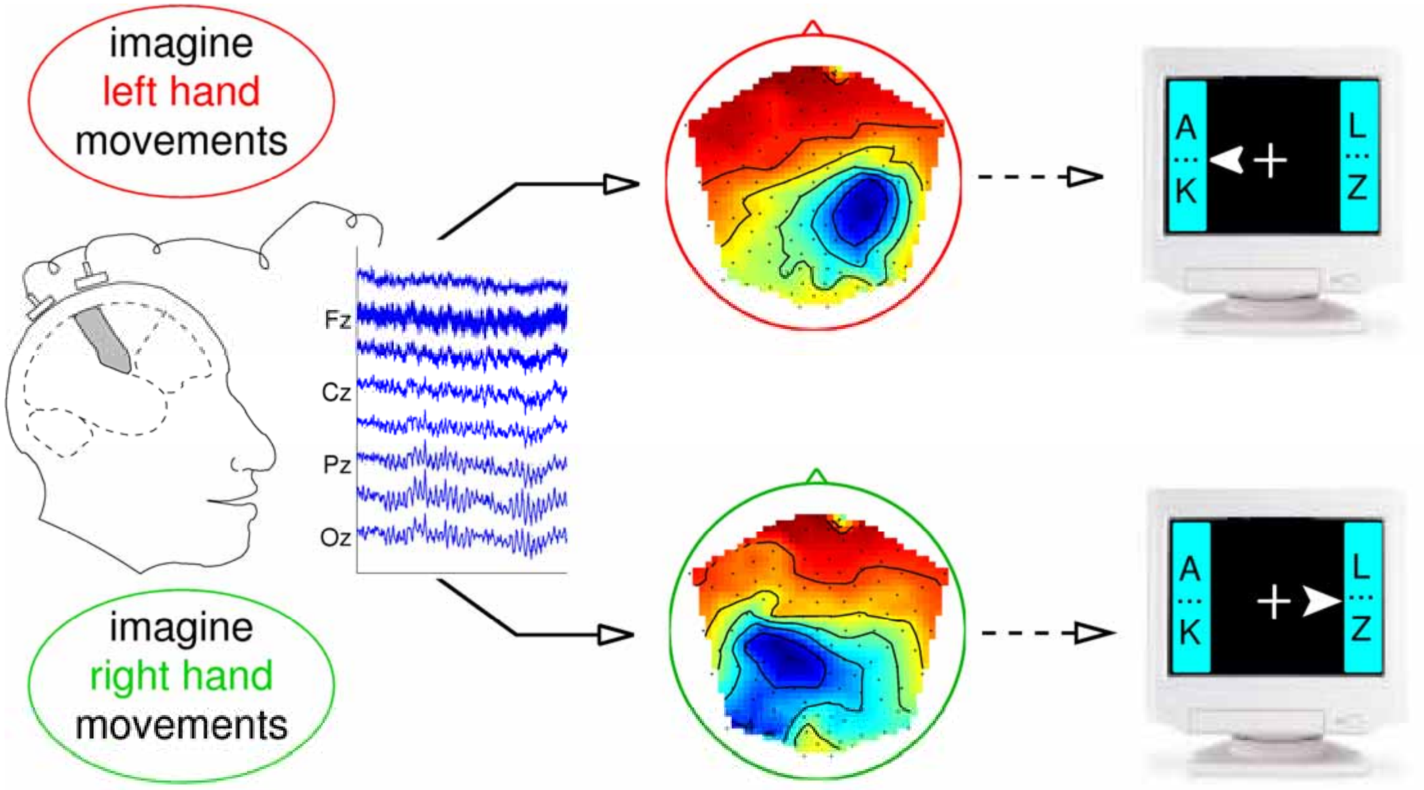
Relation to Existing Methods

IWAIC (Shimodaira, JSPI 2000) IWSIC (Sugiyama & Müller, Stat. & Deci. 2005)

	IWAIC	IWSIC	IWCV
Unbiasedness	Asymptotic	Asymptotic & Finite	Finite sample
Loss	Smooth	Squared	Arbitrary
Model	Regular	Linear	Arbitrary
Parameter learning	Smooth	Linear	Arbitrary
Computation	Fast	Fast	Slow

IWCV is the first method that is applicable to classification with covariate shift!

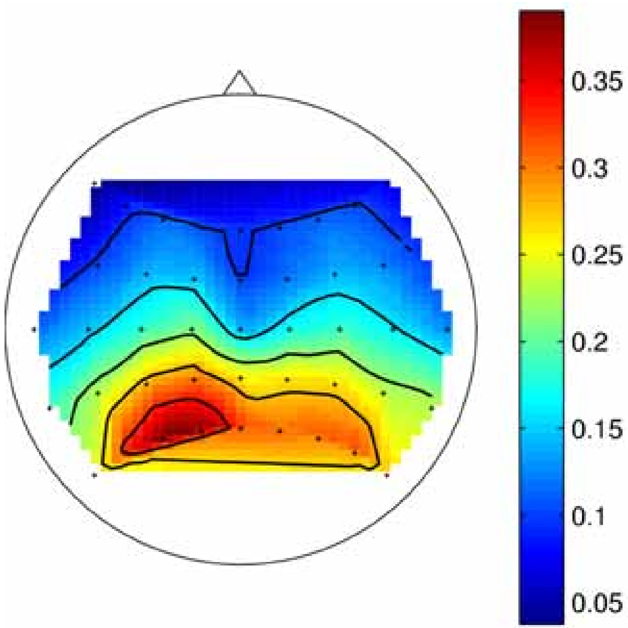
Application: Brain-Computer Interface



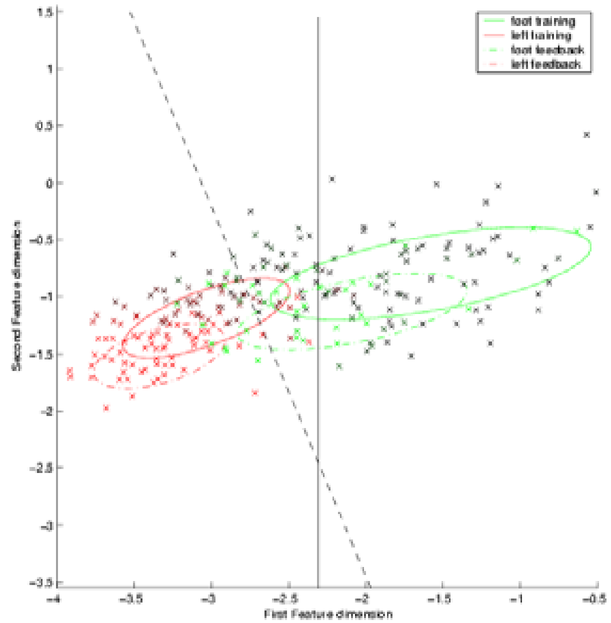
Brain activity in different mental states is transformed into control signals

Non-Stationarity in EEG Features

- Different mental conditions (attention, sleepiness etc.) between training and test phases may change the EEG signals.



Bandpower differences between training and test phases



Features extracted from brain activity during training and test phases

Adaptive Importance-Weighted Linear Discriminant Analysis

- Standard classification method in BCI: LDA (after appropriate feature extraction)
- We use its variant: AIWLDA

$$\min_{\theta_0,\boldsymbol{\theta}}\left[\sum_{i=1}^{n}\left(\frac{p_{test}(\boldsymbol{x}_i)}{p_{train}(\boldsymbol{x}_i)}\right)^{\lambda}\left(\theta_0+\boldsymbol{\theta}^{\top}\boldsymbol{x}_i-y_i\right)^2\right]$$

$$0\leq\lambda\leq 1$$

- $\lambda=0$: Ordinary LDA (standard method)
- $\lambda=1$: IWLDA (consistent)
- λ is tuned by proposed IWCV

BCI Results

Sub-ject	Trial	Ordinary LDA	AIWLDA +10IWCV
1	1	9.3 %	10.0 %
	2	8.8 %	8.8 %
	3	4.3 %	4.3 %
2	1	40.0 %	40.0 %
	2	39.3 %	38.7 %
	3	25.5 %	25.5 %
3	1	36.9 %	34.4 %
	2	21.3 %	19.3 %
	3	22.5 %	17.5 %
4	1	21.3 %	21.3 %
	2	2.4 %	2.4 %
	3	6.4 %	6.4 %
5	1	21.3 %	21.3 %
	2	15.3 %	14.0 %

- Proposed method outperforms existing one in 5 cases!

BCI Results

Sub-ject	Trial	Ordinary LDA	AIWLDA +10IWCV	KL
1	1	9.3 %	10.0 %	0.76
	2	8.8 %	8.8 %	1.11
	3	4.3 %	4.3 %	0.69
2	1	40.0 %	40.0 %	0.97
	2	39.3 %	38.7 %	1.05
	3	25.5 %	25.5 %	0.43
3	1	36.9 %	34.4 %	2.63
	2	21.3 %	19.3 %	2.88
	3	22.5 %	17.5 %	1.25
4	1	21.3 %	21.3 %	9.23
	2	2.4 %	2.4 %	5.58
	3	6.4 %	6.4 %	1.83
5	1	21.3 %	21.3 %	0.79
	2	15.3 %	14.0 %	2.01

KL divergence from training to test input distributions

- When KL is large, IWCV is better.
- When KL is small, no difference.
- Non-stationarity in EEG could be successfully modeled by covariate shift!

Conclusions

- Covariate shift: input distribution varies but functional relation remains unchanged.
- Importance weight plays a central role in compensating covariate shift.
- IW cross-validation: unbiased and general
- IWCV improves the performance of BCI.

- Class-prior change: a variant of IWCV works
- Latent distribution shift:

Storkey & Sugiyama (to be presented at NIPS2006)