ACML2009 Nov. 3, 2009

Density Ratio Estimation: A New Versatile Tool for Machine Learning

Department of Computer Science
Tokyo Institute of Technology
Masashi Sugiyama
sugi@cs.titech.ac.jp
http://sugiyama-www.cs.titech.ac.jp/~sugi

Overview of My Talk (1)

- Consider the ratio of two probability densities.

$$w(\boldsymbol{x}) = \frac{p'(\boldsymbol{x})}{p(\boldsymbol{x})}$$

- If the ratio is known, various machine learning problems can be solved!
  - Non-stationarity adaptation, domain adaptation, multi-task learning, outlier detection, change detection in time series, feature selection, dimensionality reduction, independent component analysis, conditional density estimation, classification, two-sample test

Overview of My Talk (2)

Vapnik said: When solving a problem of interest, one should not solve a more general problem as an intermediate step

Knowing densities

$$p(x), p'(x)$$

Knowing ratio

$$w(\boldsymbol{x}) = \frac{p'(\boldsymbol{x})}{p(\boldsymbol{x})}$$

- Estimating density ratio is substantially easier than estimating densities!
- Various direct density-ratio estimation methods have been developed recently.

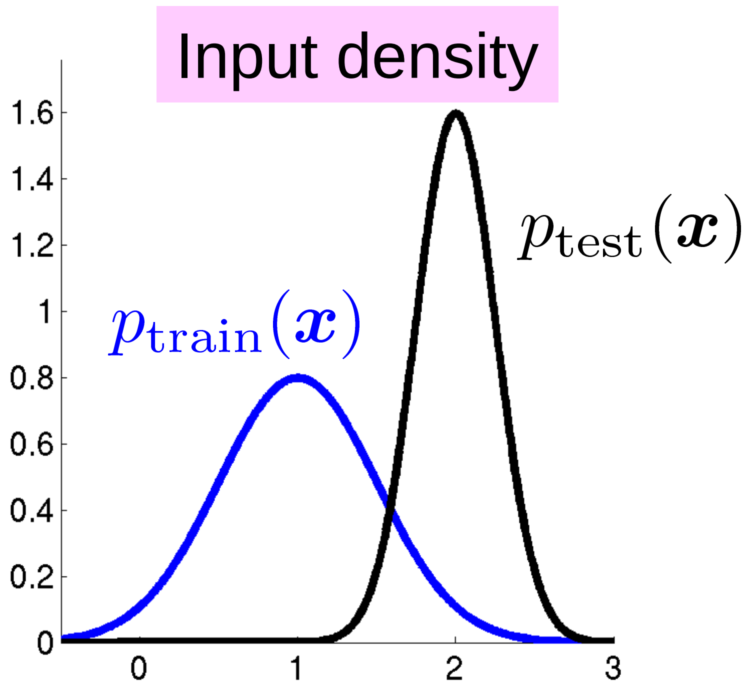
Organization of My Talk

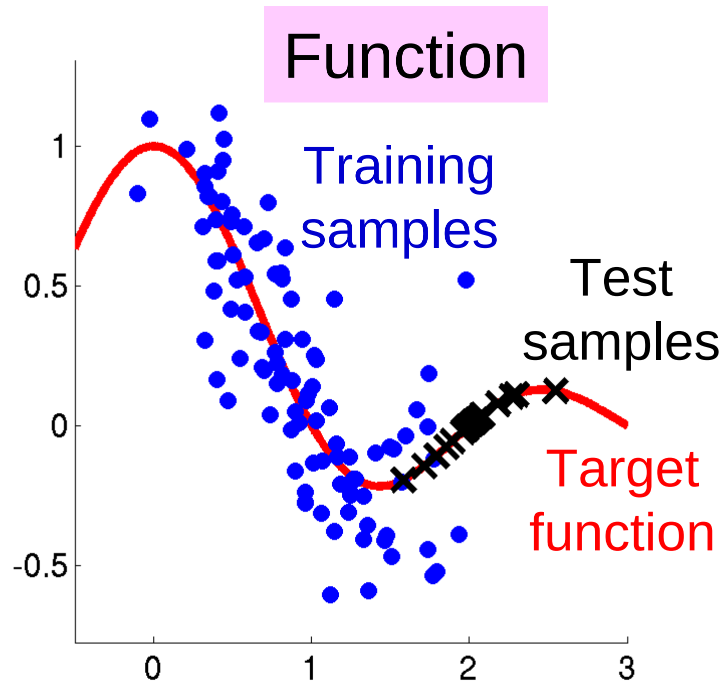
1. Applications of Density Ratios:
   - Non-stationarity adaptation, domain adaptation, and multi-task learning
   - Outlier detection and change-point detection in time series
   - Feature selection, dimensionality reduction, and independent component analysis
   - Conditional density estimation
2. Density Ratio Estimation Methods

$$\frac{p'(\boldsymbol{x})}{p(\boldsymbol{x})}$$

Non-stationarity Adaptation

- Covariate shift: training/test input distributions are different, but function remains unchanged
  - Questionnaire data analysis, robot control learning of brain-signal, speech, language, bio…





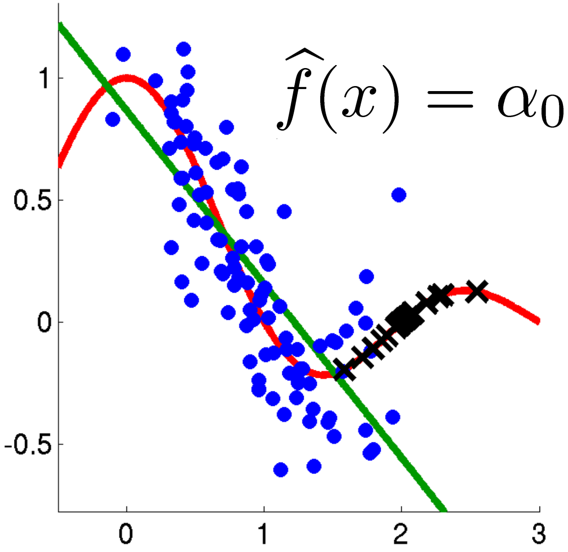
Adaptation Using Density Ratios

Shimodaira (JSPI2000), Sugiyama & Müller (ICANN2005, Stat&Deci2005)

- Ordinary least-squares is not consistent.

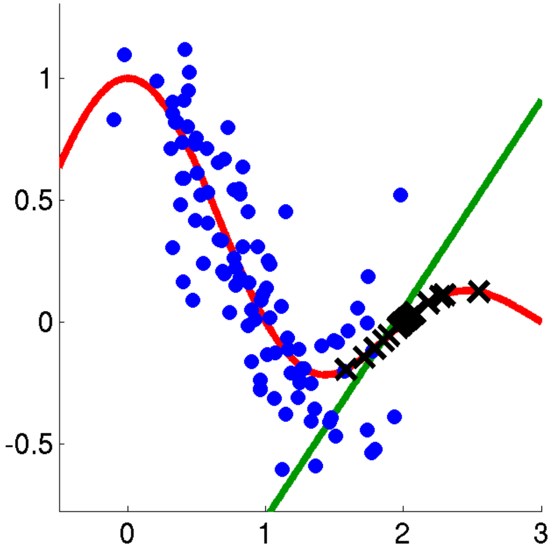
$$\min_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{n} \left(\hat{f}(\boldsymbol{x}_i) - y_i \right)^2 \right]$$

$$\widehat{f}(x) = \alpha_0 + \alpha_1 x$$



- Density-ratio weighted least-squares is consistent.

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

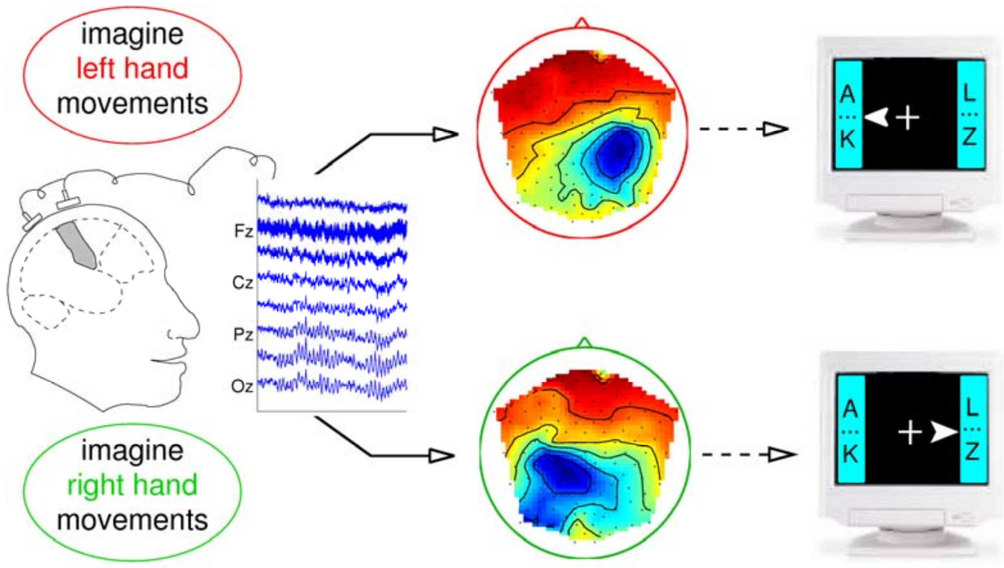


Applicable to any likelihood-based methods!

Brain-Computer Interface

Sugiyama, Krauledat & Müller (DAGM2006, JMLR2007)

- Goal: Control computers by brain signals
- Input: EEG, output: left/right
- Learn classification rules from data.
- Different mental conditions cause distribution difference in training/test phases.



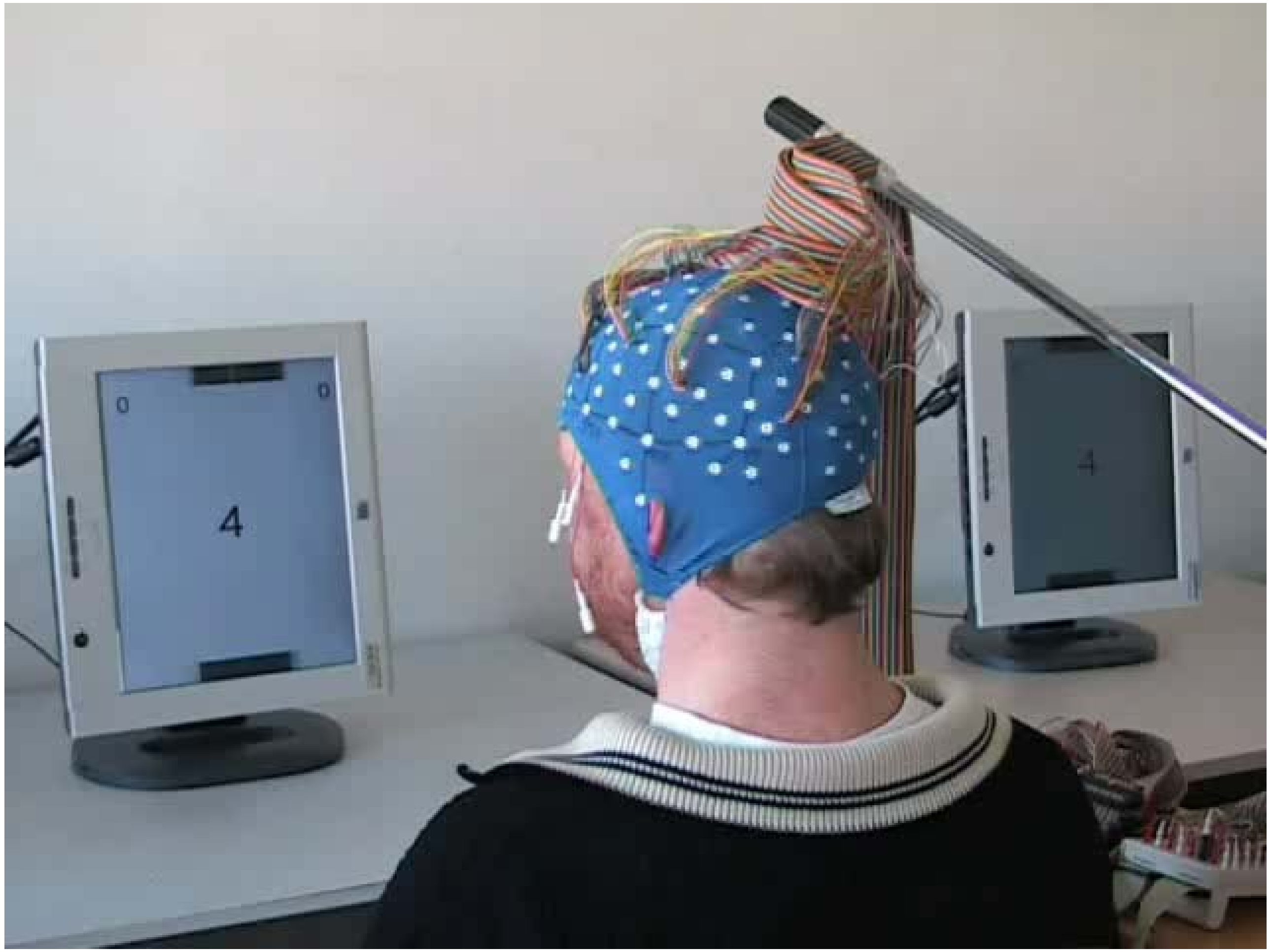
Training Phase

- Following the instruction on the screen, imagine left/right-hand movement.

Movie by Technical University of Berlin

Test Phase

- Control the pad by imagining hand movement.



- Density-ratio weighted linear discriminant analysis
- Density-ratio weighted cross-validation

Speaker Identification

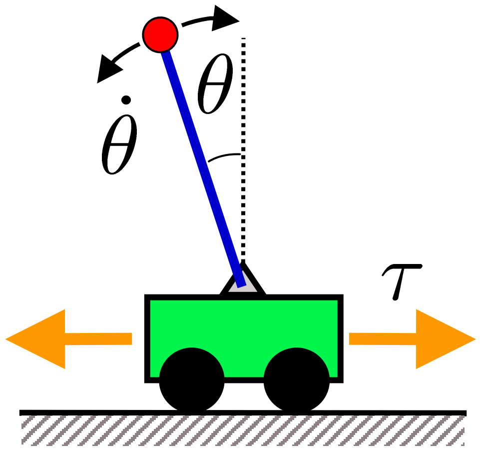
Yamada, Sugiyama & Matsui (ICASSP2009, Signal Processing 2009)

- Goal: Identify speakers from speeches
- Speech signals are not stationary.
  - Microphone / room conditions
  - Speaker’s emotion
- Performance improvement by
  - Density-ratio weighted logistic regression
  - Density-ratio weighted cross-validation

	Existing (1.5s)	Proposed (1.5s)	Existing (4.5s)	Proposed (4.5s)
3 months later	13.9 %	**13.2 %**	7.7 %	**7.4 %**
6 months later	18.0 %	**16.1 %**	7.3 %	**6.3 %**
9 months later	8.3 %	**8.0 %**	0.3 %	**0.1 %**

Robot Control

Hachiya, Akiyama, Sugiyama & Peters (AAAI2008, Neural Networks 2009)
Hachiya, Peters & Sugiyama (ECML2009)



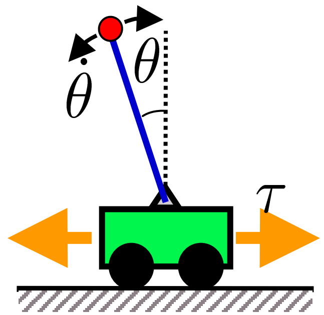
- Inverted pendulum
  - State s :angle, angular velocity
  - Action a :left/right acceleration

- Goal: Acquire a control policy of the car so that the pendulum is swung up and kept.

$$\pi(a|s)$$

Reinforcement Learning

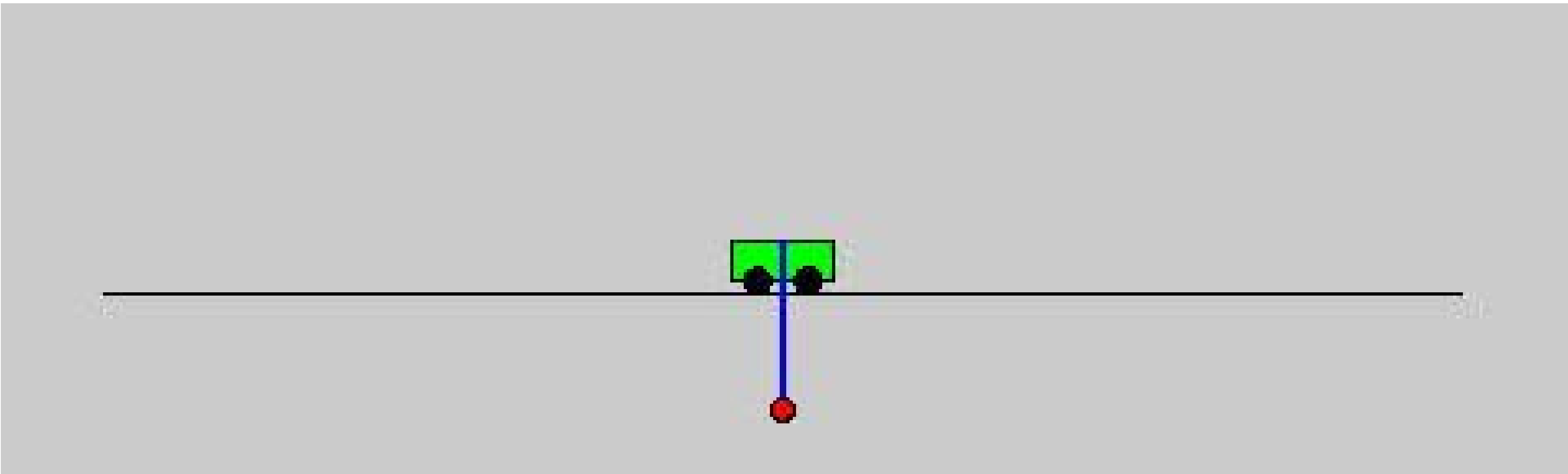
- Framework for learning the control policy $\pi(a|s)$ with maximum rewards



- Rewards: "upper is better"

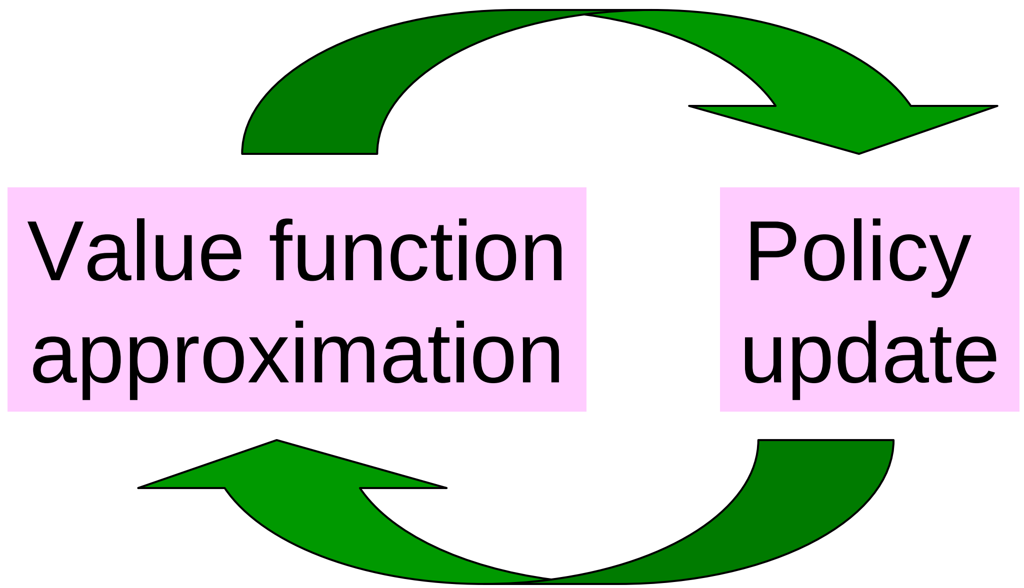
$$\cos\theta$$

  - Density-ratio weighted linear regression
  - Density-ratio weighted cross-validation



Covariate Shift in Reinforcement Learning

- Value function $Q^{\pi}(s,a)$:Sum of future rewards when taking action a at state s and following π afterwards
- Policy iteration:



- Policy update causes distribution change of (s,a).

Word Partitioning

Tsuboi, Kashima, Hido, Bickel & Sugiyama (JIP2009)

- Training data: Conversation corpus
  - (Ex.) こんな／失敗／は／ご／愛敬／だ／よ／ .
- Test data: Medical manuals
  - (Ex.) 細胞膜には受容体があり、これによって細胞を識別することができます.

	Existing	Proposed	with test labels
F-value	92.30	94.46	94.43

- Performance improvement by
  - Density-ratio weighted conditional random field
  - Density-ratio weighted cross-validation

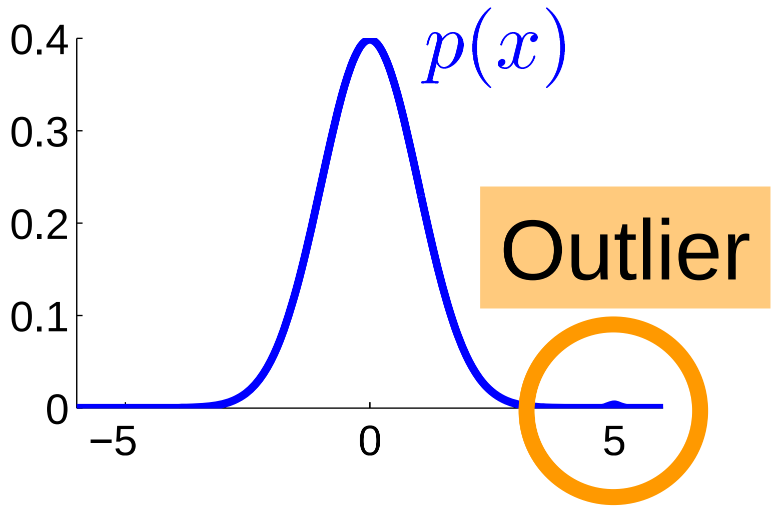
Organization of My Talk

1. Applications of Density Ratios:
   - Non-stationarity adaptation, domain adaptation, and multi-task learning
   - Outlier detection and change-point detection in time series
   - Feature selection, dimensionality reduction, and independent component analysis
   - Conditional density estimation
2. Density Ratio Estimation Methods

$$\frac{p'(\boldsymbol{x})}{p(\boldsymbol{x})}$$

Outlier Detection

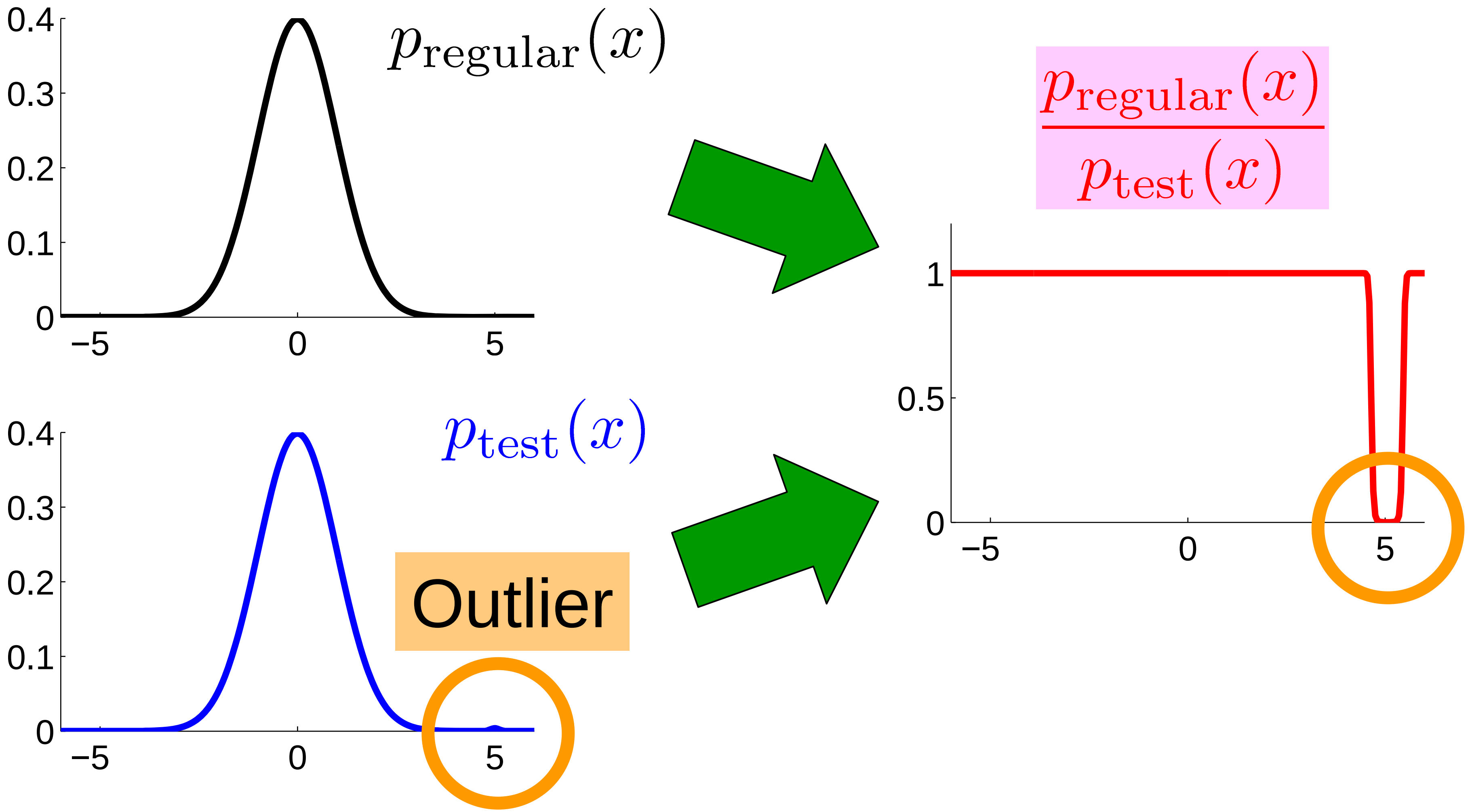
- Goal: Find "irregular" samples in dataset
  - Inferior products in assembly lines
  - Intrusions in computer networks
  - New topics in blogs
- We regard samples with low probability density as outliers.



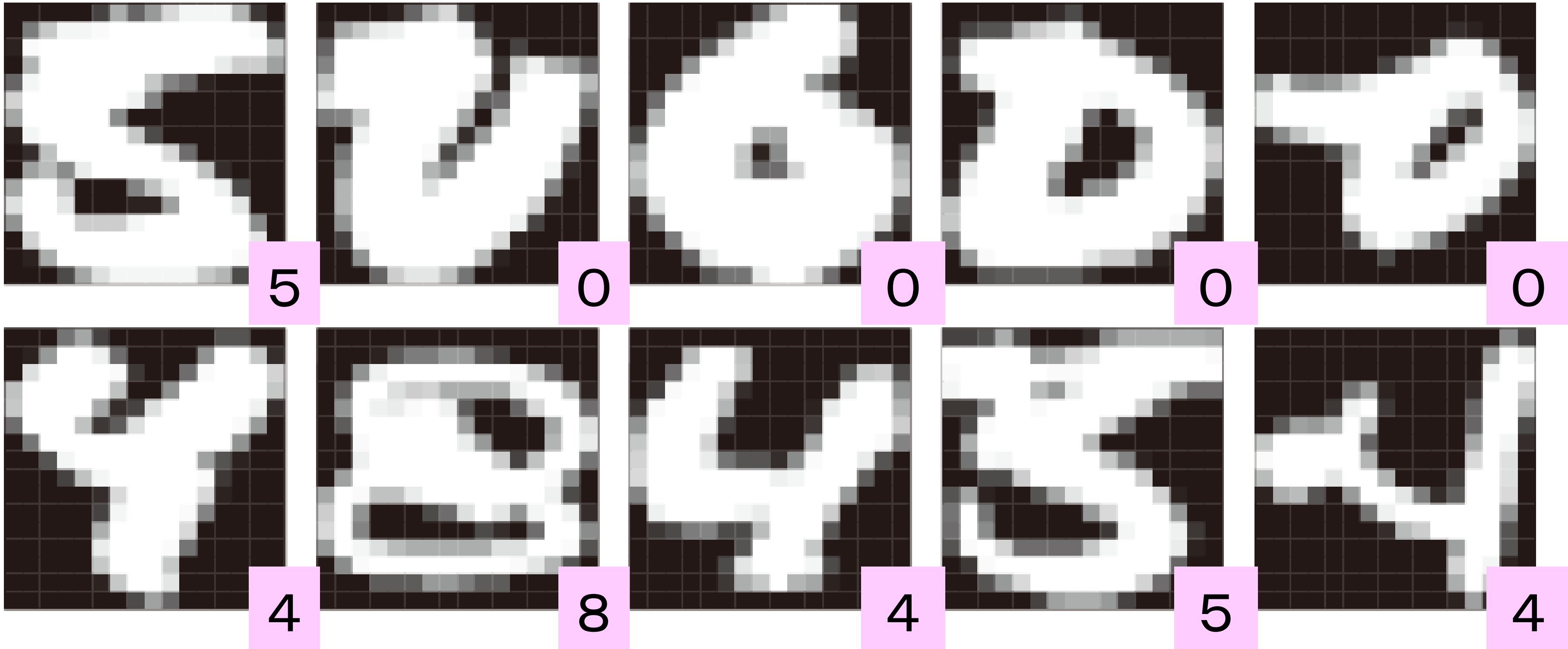
Outlier Detection with Density-Ratio

Hido, Tsuboi, Kashima, Sugiyama & Kanamori (ICDM2008)
Smola, Song & Teo (AISTATS2009)

- Outliers tend to have low density-ratio.



USPS Hand-written Digits



- USPS test data contain unclear and mislabeled samples!

Fault Diagnosis of Hard-disk Drive

- Self-Monitoring And Reporting Technology (SMART)

	Density Ratio	One-class SVM	LOF	
			NN=5	NN=30
AUC	0.881	0.843	0.847	0.924

- LOF works well if #NN is chosen appropriately; but there is no model selection method!
- Cross-validation is available for Density Ratio.

One-class SVM:

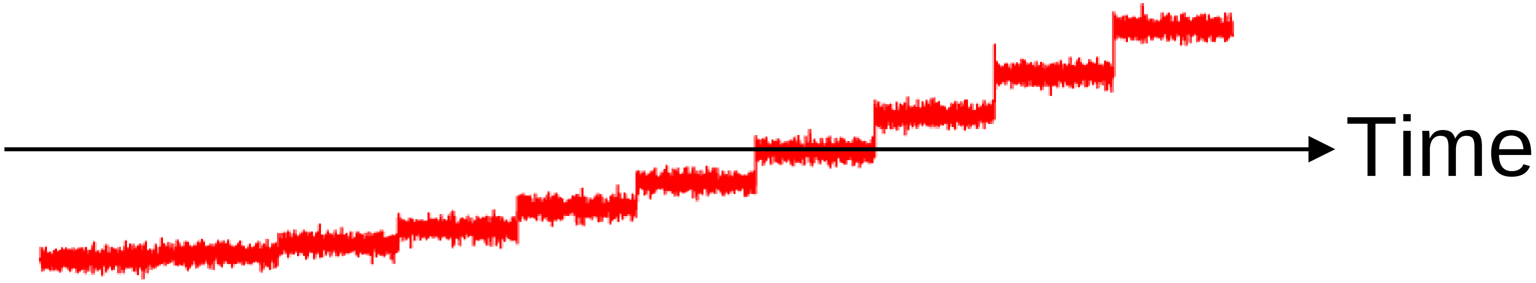
Schölkopf, Platt, Shawe-Taylor, Smola & Williamson (NeCo2001)

LOF: Local outlier factor Breunig, Kriegel, Ng & Sander (SIGMOD2000)

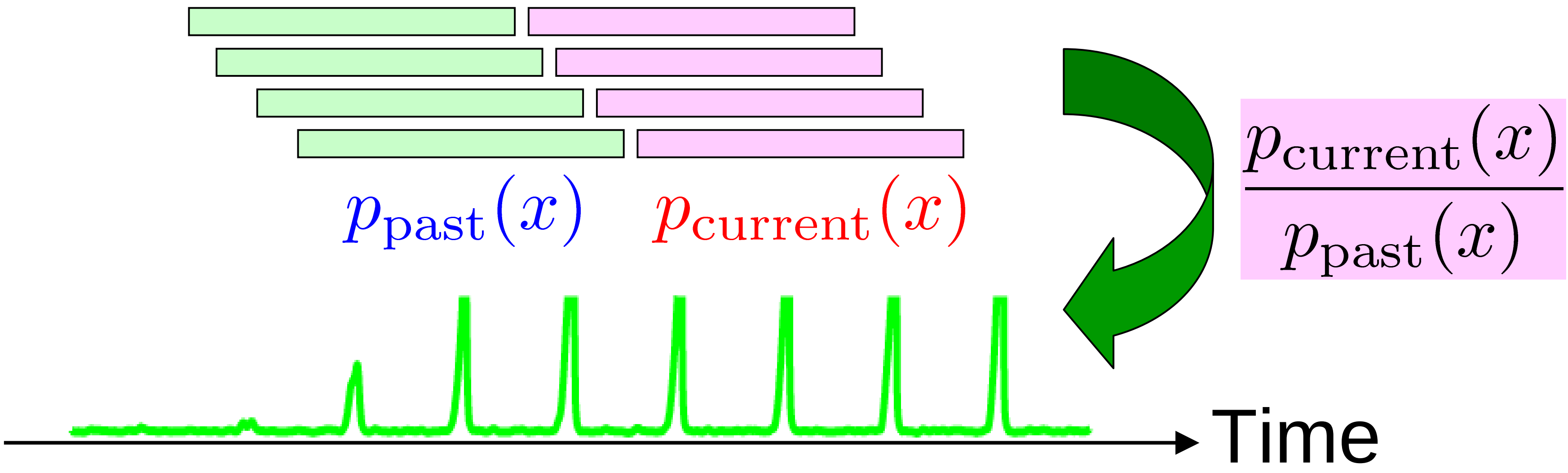
Beyond Outlier Detection

Kawahara & Sugiyama (SDM2009)

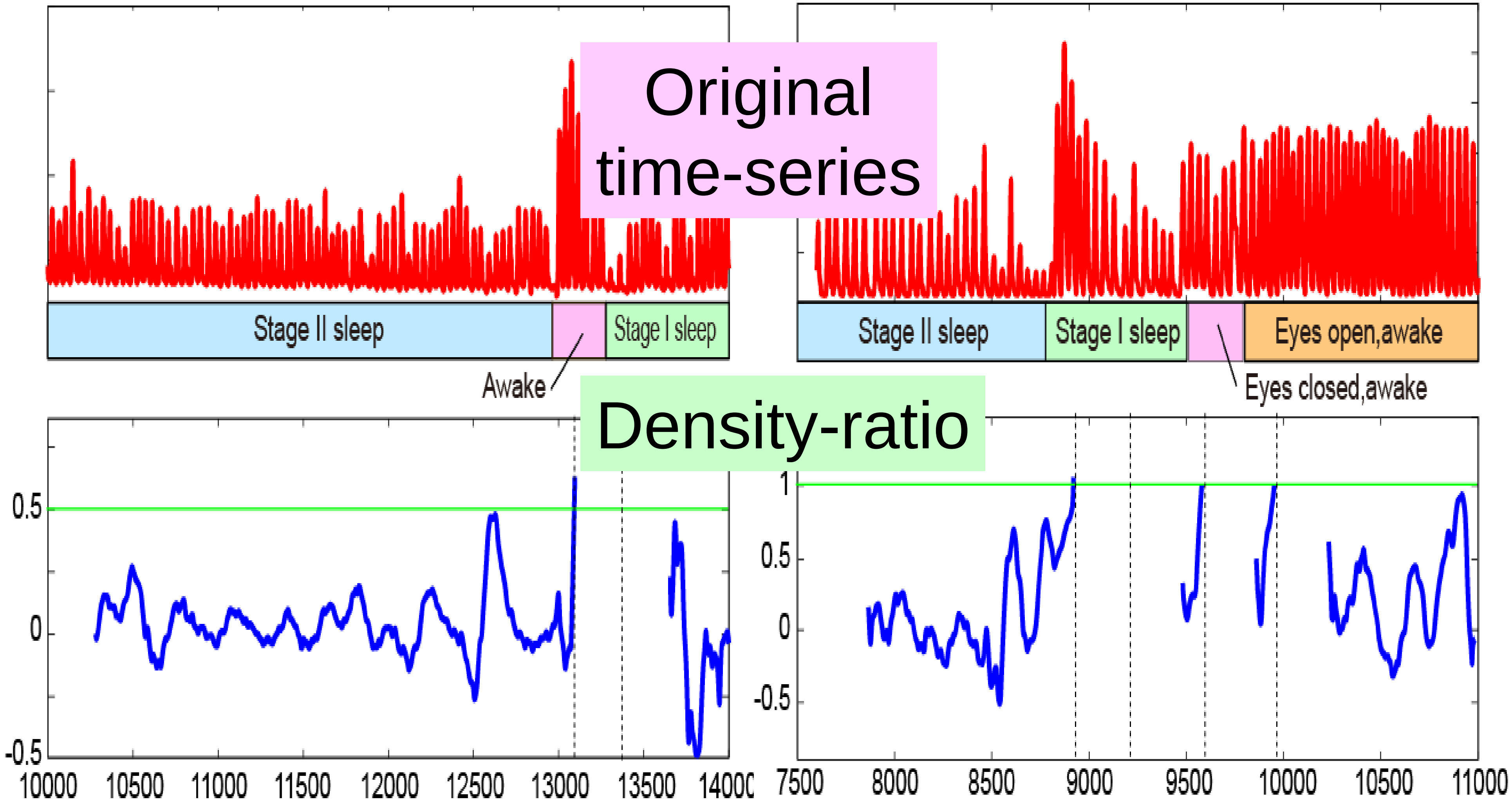
- Change detection in time series:



- Compute density-ratio in "sliding-window":



Change Detection from Breath



Organization of My Talk

1. Applications of Density Ratios:
   - Non-stationarity adaptation, domain adaptation, and multi-task learning
   - Outlier detection and change-point detection in time series
   - Feature selection, dimensionality reduction, and independent component analysis
   - Conditional density estimation
2. Density Ratio Estimation Methods

$$\frac{p'(\boldsymbol{x})}{p(\boldsymbol{x})}$$

Mutual Information Estimation

- Mutual information (MI):

$$\mathrm{MI} := \int p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy$$

- MI as an independence measure:

$\mathrm{MI} = 0$ $\Longleftrightarrow$ x and y are statistically independent

- MI can be computed using density ratio:

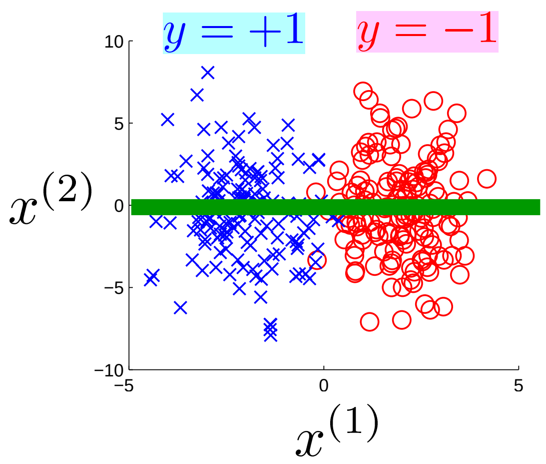
$$\frac{p(x,y)}{p(x)p(y)}$$

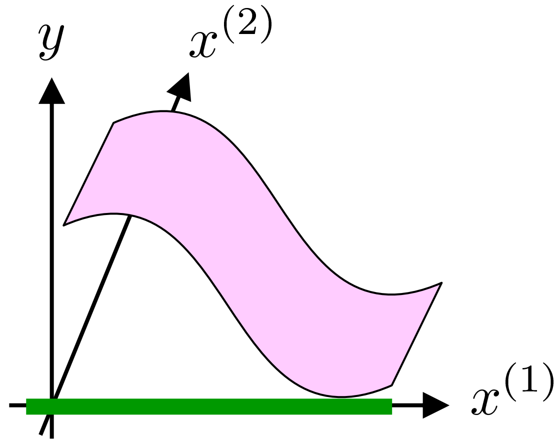
Suzuki, Sugiyama, & Tanaka (ISIT2009)
Nguyen, Wainwright & Jordan (IEEE-IT2009)

MI-Based Feature Selection

Suzuki, Sugiyama, Sese & Kanamori (FSDM2008, BMC Bioinformatics 2009)

- Goal: For $y = f(x^{(1)}, \ldots, x^{(d)})$, find the input variable $x^{(k)}$ which is the most responsible for explaining output value y
  - Gene selection, brain activity localization, drug discovery etc.



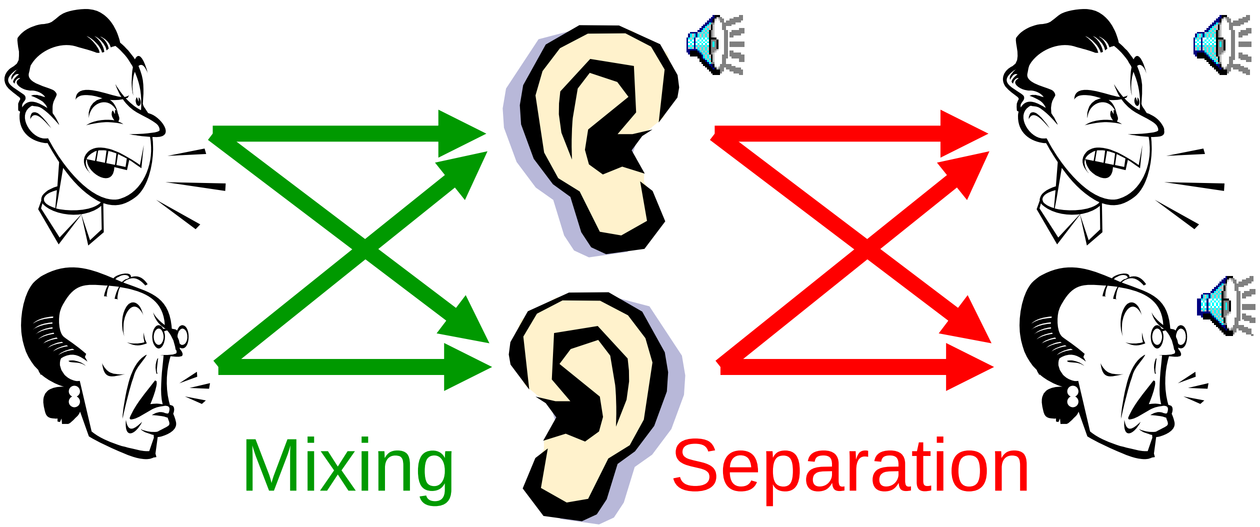


- Feature extraction is also possible.

Suzuki & Sugiyama (submitted)

MI-Based Independent Component Analysis

- Goal: Separate mixed signals into independent ones   Suzuki & Sugiyama (ICA2009)
  - Cross-validation is available for model selection (cf. no CV for kernel ICA etc.)



Organization of My Talk

1. Applications of Density Ratios:
   - Non-stationarity adaptation, domain adaptation, and multi-task learning
   - Outlier detection and change-point detection in time series
   - Feature selection, dimensionality reduction, and independent component analysis
   - Conditional density estimation
2. Density Ratio Estimation Methods

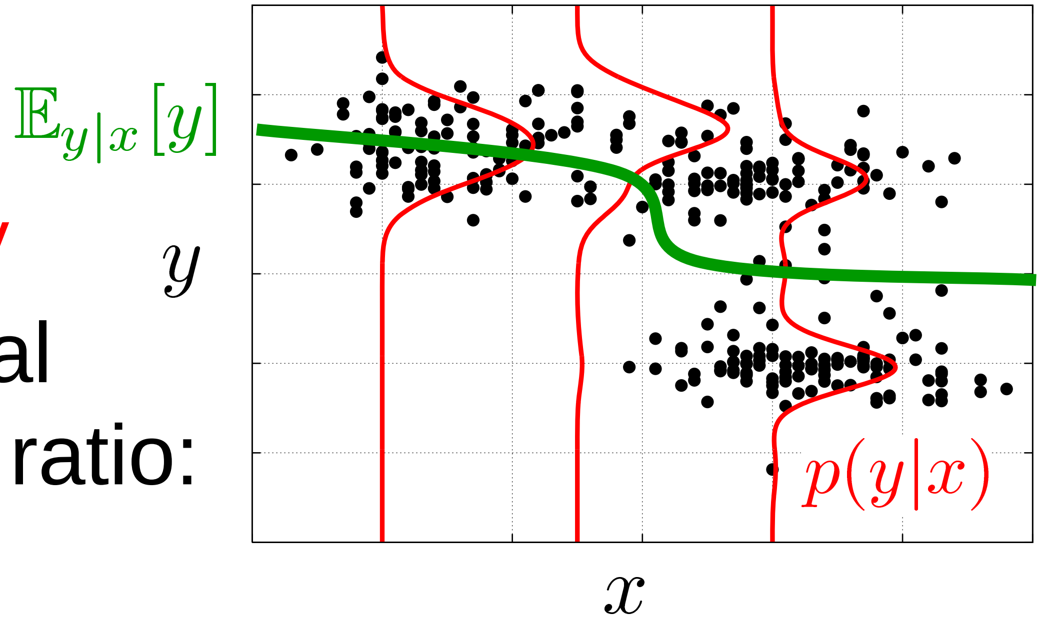
$$\frac{p'(\boldsymbol{x})}{p(\boldsymbol{x})}$$

Conditional Density Estimation

Sugiyama, Takeuchi, Suzuki, Kanamori & Hachiya (submitted)

- **Regression:** Estimating conditional mean $\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}[\boldsymbol{y}]$
- When conditional density $p(\boldsymbol{y}|\boldsymbol{x})$ is complicated, regression is not informative enough:
  - Multi-modality
  - Asymmetry
  - Hetero-scedasticity
- Estimate conditional density via density ratio:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})}$$

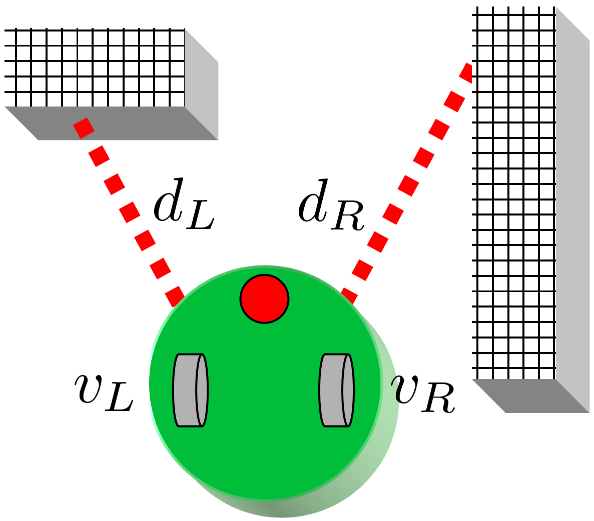


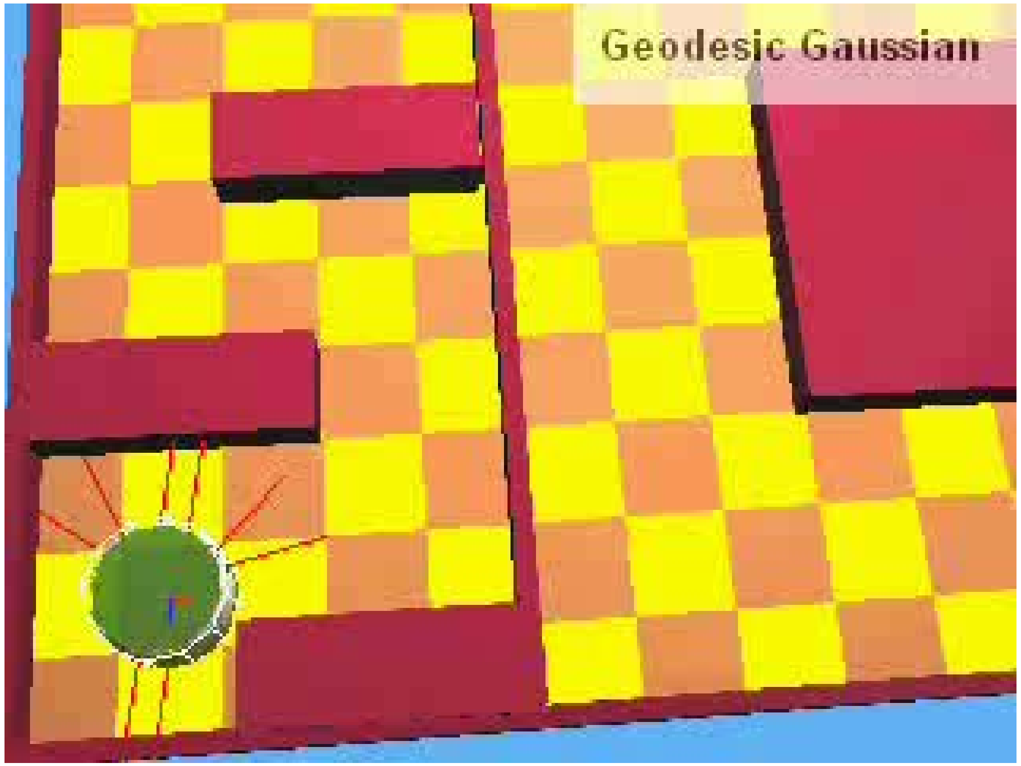
Robots' Transition Estimation

- Transition probability $p(s'|s,a)$:
  Distribution of destination state s'
  when taking action a at current state s

Kherera robot

- State: infra-red sensors
- Action: wheel speed





Organization of My Talk

1. Applications of Density Ratios
2. Density ratio estimation methods:
   A) Kullback-Leibler Importance Estimation Procedure (KLIEP)
   B) Least-Squares Importance Fitting (LSIF)
   C) Unconstrained LSIF (uLSIF)

$$\frac{p'(\boldsymbol{x})}{p(\boldsymbol{x})}$$

Density Ratio Estimation

$$w(\boldsymbol{x}) = \frac{p'(\boldsymbol{x})}{p(\boldsymbol{x})}$$

- Density ratios are shown to be versatile.
- In practice, however, the ratio should be estimated from data.

$$\{\boldsymbol{x}_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} p(\boldsymbol{x})$$

$$\{\boldsymbol{x}'_i\}_{i=1}^{n'} \overset{i.i.d.}{\sim} p'(\boldsymbol{x})$$

- Naïve approach: Estimate two densities separately and take the ratio

$$\widehat{w}(\boldsymbol{x}) = \frac{\widehat{p}'(\boldsymbol{x})}{\widehat{p}(\boldsymbol{x})}$$

Vapnik's Principle

When solving a problem, don't solve
more difficult problems as an intermediate step

Knowing densities

$$p(x), p'(x)$$

Knowing ratio

$$w(\boldsymbol{x}) = \frac{p'(\boldsymbol{x})}{p(\boldsymbol{x})}$$

- Estimating density-ratio is substantially easier than estimating densities!
- We estimate density-ratio without going through density estimation.

Organization of My Talk

1. Applications of Density Ratios
2. Density ratio estimation methods:
   - A) Kullback-Leibler Importance Estimation Procedure (KLIEP)
   - B) Least-Squares Importance Fitting (LSIF)
   - C) Unconstrained LSIF (uLSIF)

$$\frac{p'(\boldsymbol{x})}{p(\boldsymbol{x})}$$

Kullback-Leibler Importance Estimation Procedure (KLIEP)

Sugiyama, Nakajima, Kashima, von Bünau & Kawanabe (NIPS2007)
Sugiyama, Suzuki, Nakajima, Kashima, von Bünau & Kawanabe (AISM2008)

- Linear model:

$$\widehat{w}(\boldsymbol{x}) = \sum_{\ell=1}^{b} \alpha_\ell \phi_\ell(\boldsymbol{x})$$

$$= \boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x})$$

$$\alpha_\ell \geq 0$$

$$\phi_\ell(\boldsymbol{x}) \geq 0$$

(ex. Gauss kernel)

- Parameters are learned so that KL divergence from $p'(\boldsymbol{x})$ to $\widehat{p}'(\boldsymbol{x}) = \widehat{w}(\boldsymbol{x})p(\boldsymbol{x})$ is minimized:

$$\min_{\boldsymbol{\alpha}} \mathrm{KL}[p'(\boldsymbol{x}) || \widehat{p}'(\boldsymbol{x})]$$

KLIEP: Formulation

- Decomposition of KL divergence:

$$\mathrm{KL}[p'(\boldsymbol{x})||\widehat{p}'(\boldsymbol{x})] = \int p'(\boldsymbol{x}) \log \frac{p'(\boldsymbol{x})}{\widehat{w}(\boldsymbol{x})p(\boldsymbol{x})} d\boldsymbol{x}$$

$$= \underbrace{\int p'(\boldsymbol{x}) \log \frac{p'(\boldsymbol{x})}{p(\boldsymbol{x})} d\boldsymbol{x}}_{\text{Constant}} - \int p'(\boldsymbol{x}) \log \widehat{w}(\boldsymbol{x}) d\boldsymbol{x}$$

$$\min_{\boldsymbol{\alpha}} \mathrm{KL}[p'(\boldsymbol{x})||\widehat{p}'(\boldsymbol{x})] \Longleftrightarrow \max_{\boldsymbol{\alpha}} \int p'(\boldsymbol{x}) \log \widehat{w}(\boldsymbol{x}) d\boldsymbol{x}$$

- $\widehat{p}'(\boldsymbol{x}) = \widehat{w}(\boldsymbol{x})p(\boldsymbol{x})$ is probability density:

$$\int \widehat{w}(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = 1$$ (Constraint)

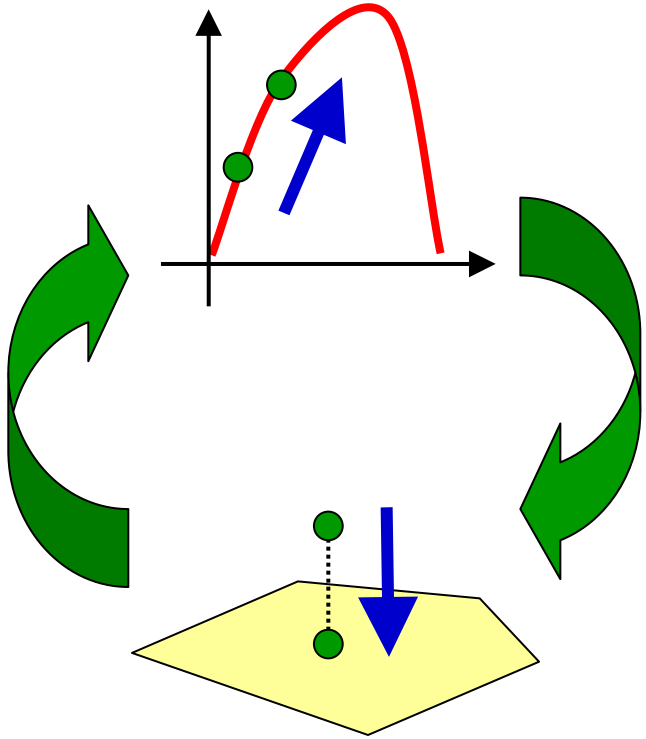
KLIEP: Algorithm

- Approximate expectation by sample average:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n'} \log(\boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x}'_i)) \qquad \text{subject to } \sum_{i=1}^{n} \boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) = n \text{ and } \boldsymbol{\alpha} \geq \mathbf{0}$$

- This is convex optimization, so repeating
  - Gradient ascent
  - Constraint satisfaction

  converges to global solution.
- Global solution is sparse!



KLIEP: Theoretical Properties

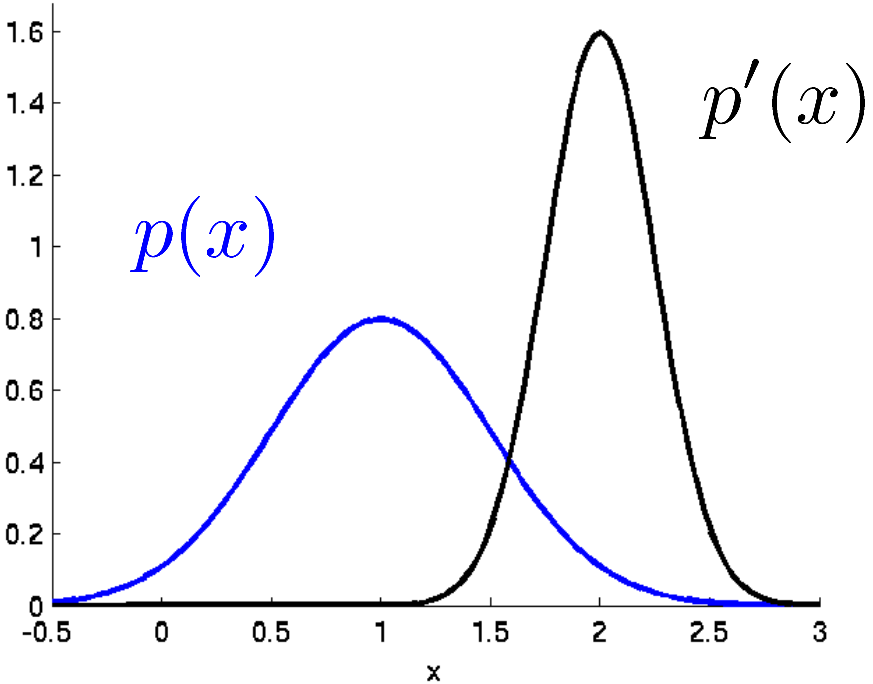
Sugiyama, Suzuki, Nakajima, Kashima, von Bünau & Kawanabe (AISM2008)
Nguyen, Wainwright & Jordan (NIPS2007)

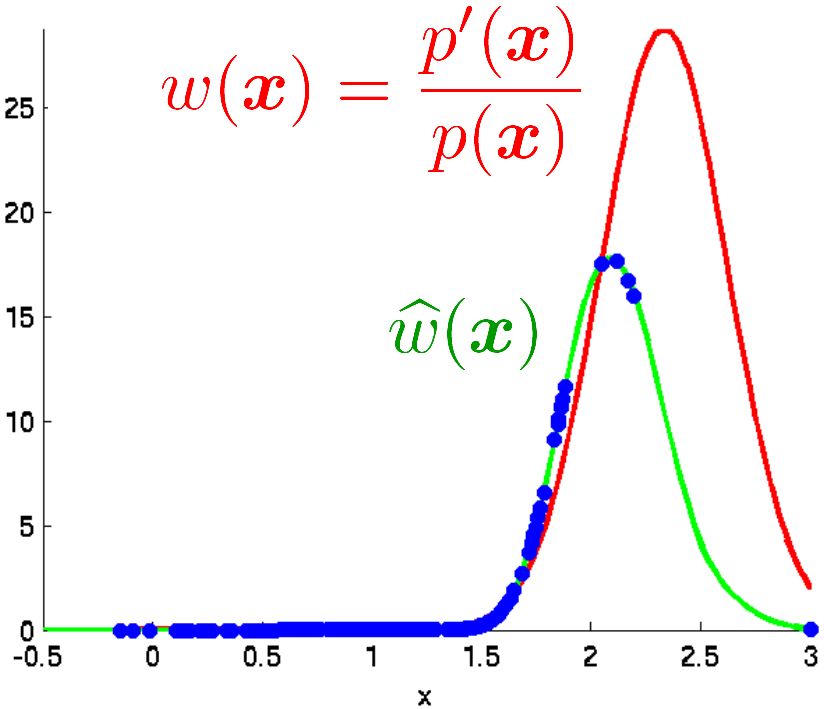
- Parametric case: $$\widehat{w}(\boldsymbol{x}) = \sum_{\ell=1}^{b} \alpha_\ell \phi_\ell(\boldsymbol{x})$$
  - Learned parameter converge to the optimal value with order $1/\sqrt{\overline{n}}$.

$$\overline{n} = \min(n, n')$$

- Non-parametric case: $$\widehat{w}(\boldsymbol{x}) = \sum_{\ell=1}^{n'} \alpha_\ell K(\boldsymbol{x}, \boldsymbol{x}_\ell)$$
  - Learned function converges to the optimal function with order slightly slower than $1/\sqrt{\overline{n}}$ (depending on complexity of function class).

KLIEP: Example



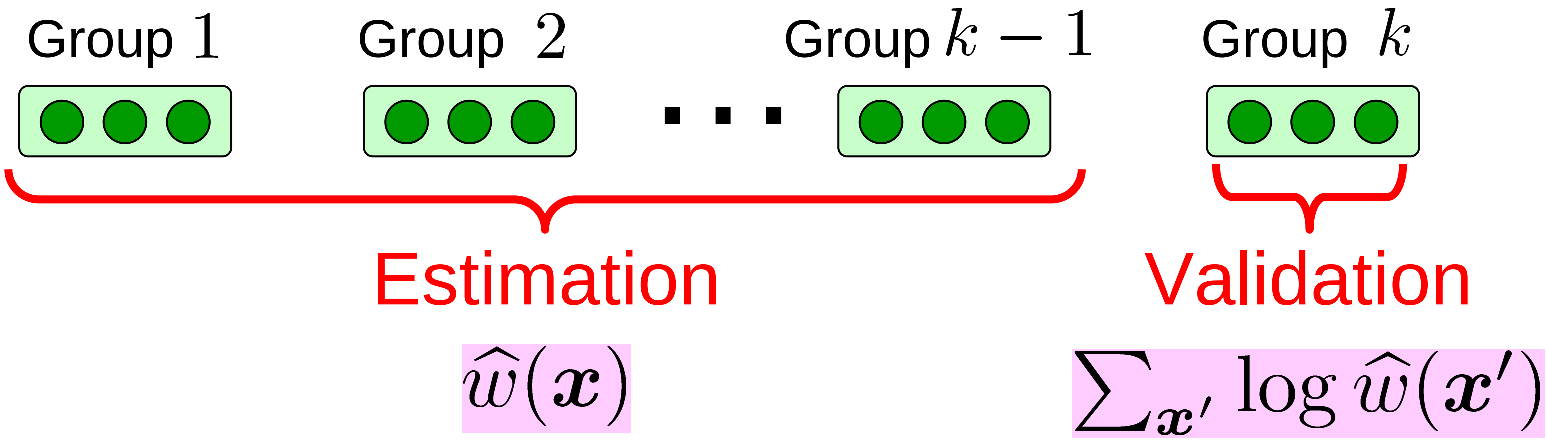


$$\widehat{w}(\boldsymbol{x}) = \sum_{\ell=1}^{n'} \alpha_\ell K(\boldsymbol{x}, \boldsymbol{x}'_\ell)$$

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma^2}\right)$$

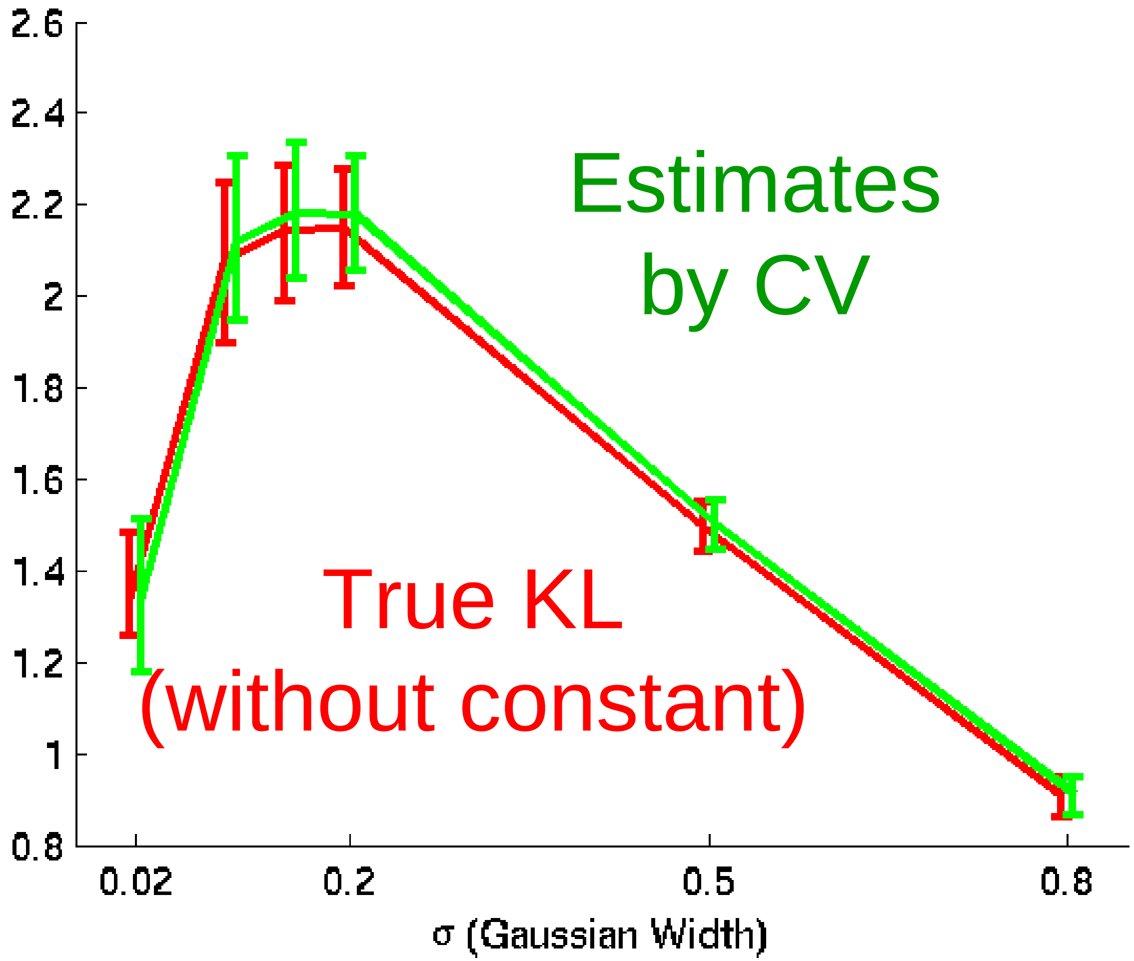
KLIEP: Model Selection

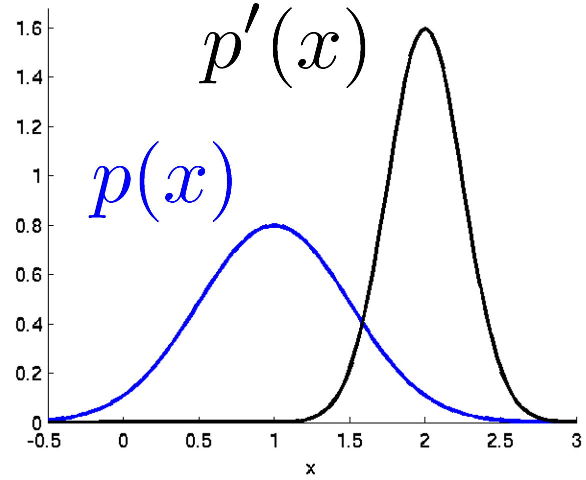
- Choice of Gaussian width is crucial.
- Cross-validation (CV):
  - Divide numerator samples for estimation and evaluation purposes.

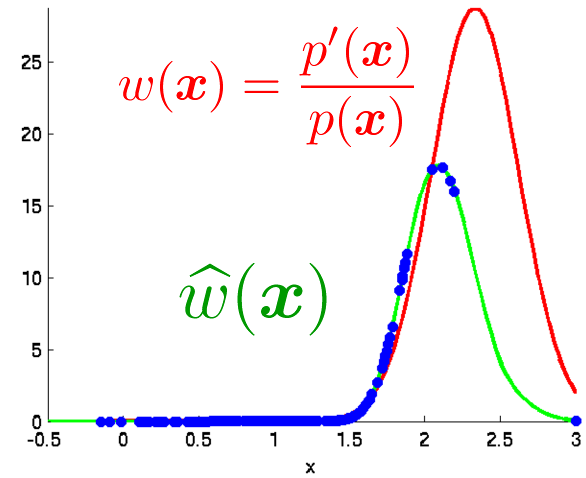


  - Repeat this for all combinations
- CV gives an unbiased estimate of KL.

KLIEP: Example of CV



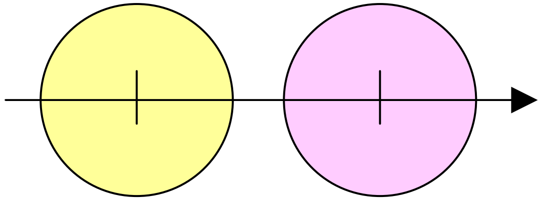




- CV is very accurate!

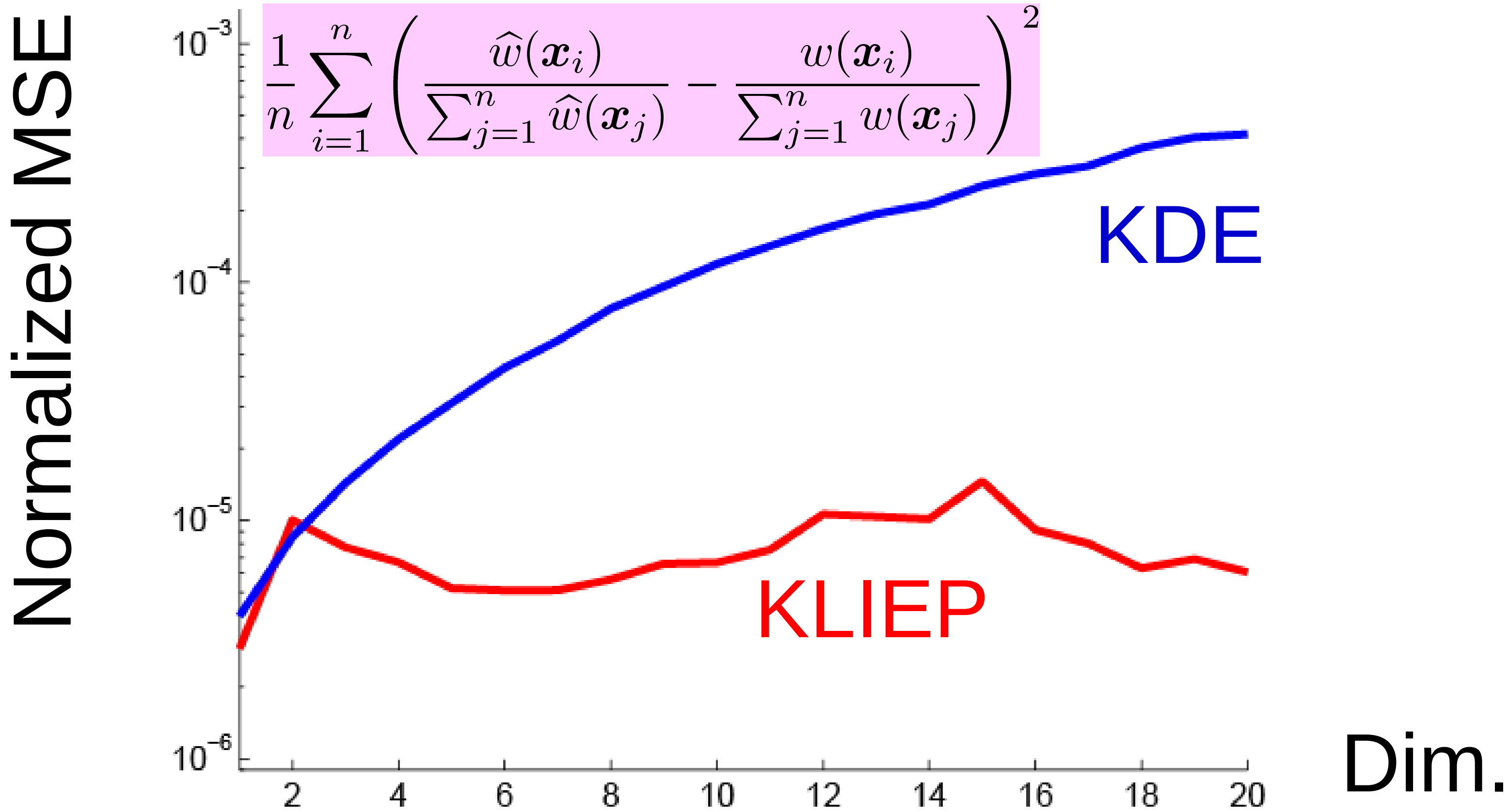
Experiments

- Setup: d-dim. Gaussian with covariance identity and
  - Denominator: mean (0,0,0,…,0)
  - Numerator: mean (1,0,0,…,0)



- Kernel density estimation (KDE):
  - Estimate two densities separately and take ratio.
  - Gaussian with is chosen by CV.
- KLIEP:
  - Estimate density-ratio directly.
  - Gaussian with is chosen by CV.

Accuracy as a Function of Input Dimensionality



- KDE: "curse of dimensionality"
- KLIEP: works well

KLIEP: Summary

- Works well in high-dimensions.
- Sparse global solution is available.
- CV by model selection is possible.
- Domains of denominator/numerator could be different (conditional density estimation).
- KL is consistent with mutual information.
- Applicable to various models such as log-linear models and Gaussian mixture models.

Tsuboi, Kashima, Hido, Bickel & Sugiyama (SDM2008, JIP2009)
Yamada & Sugiyama (IEICE2009)

Organization of My Talk

1. Applications of Density Ratios
2. Density ratio estimation methods:
   - A) Kullback-Leibler Importance Estimation Procedure (KLIEP)
   - B) Least-Squares Importance Fitting (LSIF)
   - C) Unconstrained LSIF (uLSIF)

$$\frac{p'(\boldsymbol{x})}{p(\boldsymbol{x})}$$

Least-Squares Importance Fitting (LSIF)

Kanamori, Hido & Sugiyama (NIPS2008, JMLR2009)

- Linear model:

$$\widehat{w}(\boldsymbol{x}) = \sum_{\ell=1}^{b} \alpha_\ell \phi_\ell(\boldsymbol{x})$$
$$= \boldsymbol{\alpha}^\top \boldsymbol{\phi}(\boldsymbol{x})$$

$$\alpha_\ell \geq 0$$
$$\phi_\ell(\boldsymbol{x}) \geq 0$$

(ex. Gauss kernel)

- Squared-loss:

$$J_0(\boldsymbol{\alpha}) = \frac{1}{2} \int \left(\widehat{w}(\boldsymbol{x}) - w(\boldsymbol{x}) \right)^2 p(\boldsymbol{x}) d\boldsymbol{x}$$

$$w(\boldsymbol{x}) = \frac{p'(\boldsymbol{x})}{p(\boldsymbol{x})}$$

LSIF: Formulation

- Decomposition of squared-loss:

$$\begin{aligned} J_0(\boldsymbol{\alpha}) &= \frac{1}{2}\int \left(\widehat{w}(\boldsymbol{x}) - w(\boldsymbol{x})\right)^2 p(\boldsymbol{x})d\boldsymbol{x} \\ &= \frac{1}{2}\int \left(\widehat{w}(\boldsymbol{x})\right)^2 p(\boldsymbol{x})d\boldsymbol{x} - \int \widehat{w}(\boldsymbol{x})p'(\boldsymbol{x})d\boldsymbol{x} \\ &\quad \underbrace{+\frac{1}{2}\int \left(w(\boldsymbol{x})\right)^2 p(\boldsymbol{x})d\boldsymbol{x}}_{\text{constant}} \end{aligned}$$

- Constraint: $\boldsymbol{\alpha} \geq \mathbf{0}$

$$w(\boldsymbol{x}) = \frac{p'(\boldsymbol{x})}{p(\boldsymbol{x})}$$

LSIF: Algorithm

- Approximate expectation by sample average and include a regularizer, we have:

$$\min_{\boldsymbol{\alpha}} \left[\frac{1}{2} \boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}} \boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \lambda \boldsymbol{\alpha}^\top \mathbf{1} \right] \quad \text{subject to } \boldsymbol{\alpha} \geq \mathbf{0}$$

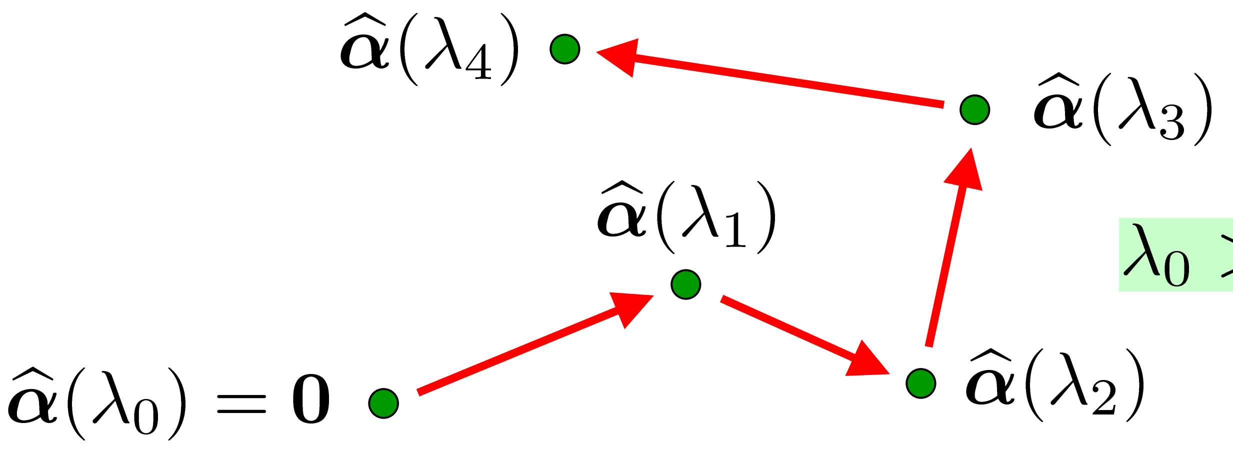
$$\widehat{\boldsymbol{H}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\phi}(\boldsymbol{x}_i)^\top \boldsymbol{\phi}(\boldsymbol{x}_i) \qquad \widehat{\boldsymbol{h}} = \frac{1}{n'} \sum_{j=1}^{n'} \boldsymbol{\phi}(\boldsymbol{x}'_j)$$

- This is a convex quadratic program (QP), so the global solution can be efficiently computed by standard optimization software.
- The optimal solution is sparse!

LSIF: Regularization Path Tracking

$$\min_{\boldsymbol{\alpha}} \left[\frac{1}{2} \boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}} \boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \lambda \boldsymbol{\alpha}^\top \mathbf{1} \right] \quad \text{subject to } \boldsymbol{\alpha} \geq \mathbf{0}$$

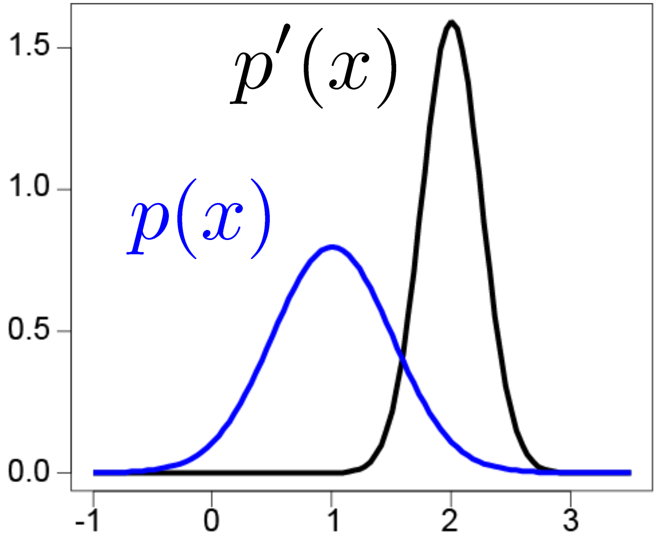
- Solution is piece-wise linear with respect to the regularization parameter λ .



$\lambda_0 > \lambda_1 > \cdots > 0$

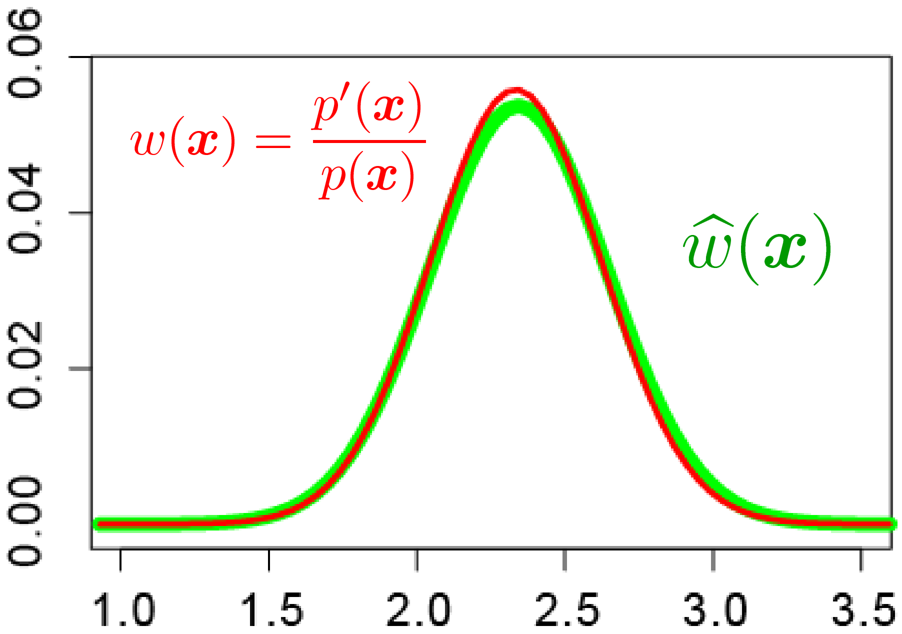
- Solutions for all λ can be computed efficiently without QP solvers!

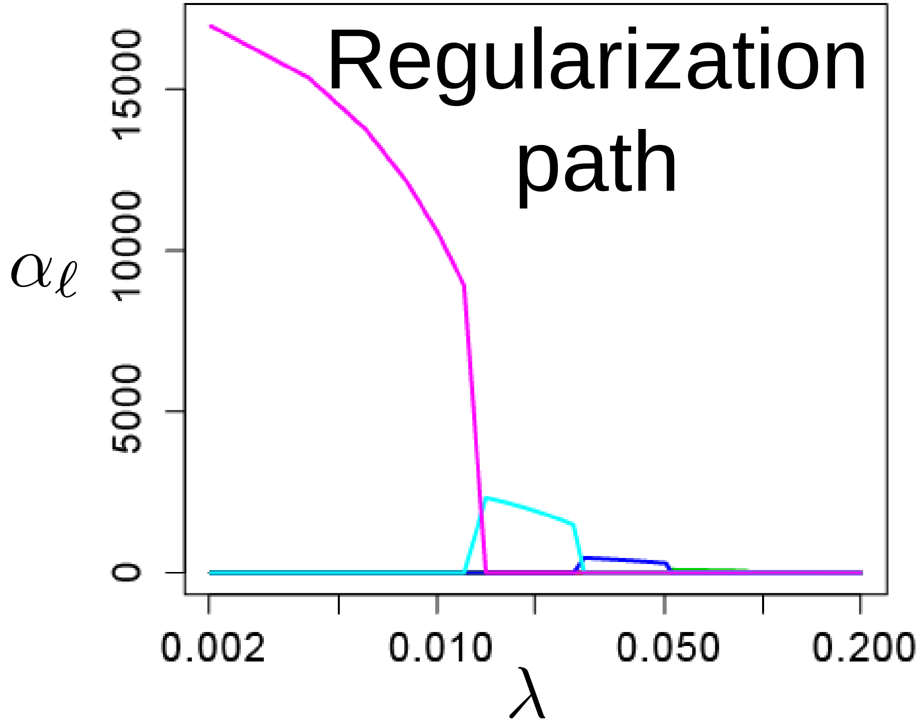
LSIF: Examples



- Regularization parameter and Gaussian width are chosen by CV

$$\widehat{w}(\boldsymbol{x}) = \sum_{\ell=1}^{n'} \alpha_\ell \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'_\ell\|^2}{2\sigma^2}\right)$$





LSIF: Summary

- Squared-loss is often preferred to KL divergence in conditional density estimation.
- Regularization path algorithm is computationally very efficient.
- However, it is numerically rather unstable.

Organization of My Talk

1. Applications of Density Ratios
2. Density ratio estimation methods:
   - A) Kullback-Leibler Importance Estimation Procedure (KLIEP)
   - B) Least-Squares Importance Fitting (LSIF)
   - C) Unconstrained LSIF (uLSIF)

$$\frac{p'(\boldsymbol{x})}{p(\boldsymbol{x})}$$

Unconstrained LSIF (uLSIF)

Kanamori, Hido & Sugiyama (NIPS2008, JMLR2009)

$$\min_{\boldsymbol{\alpha}} \left[\frac{1}{2} \boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}} \boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \lambda \boldsymbol{\alpha}^\top \mathbf{1} \right] \quad \text{subject to } \boldsymbol{\alpha} \geq \mathbf{0}$$

- Slightly modify LSIF:
  - Ignore non-negativity
  - Use a quadratic regularizer

$$\widetilde{\boldsymbol{\alpha}} = \operatorname*{argmin}_{\boldsymbol{\alpha}} \left[\frac{1}{2} \boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}} \boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \frac{\lambda}{2} \boldsymbol{\alpha}^\top \boldsymbol{\alpha} \right]$$

$$\widehat{\boldsymbol{H}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\phi}(\boldsymbol{x}_i)^\top \boldsymbol{\phi}(\boldsymbol{x}_i) \qquad \widehat{\boldsymbol{h}} = \frac{1}{n'} \sum_{j=1}^{n'} \boldsymbol{\phi}(\boldsymbol{x}'_j)$$

uLSIF: Algorithm

$$\widetilde{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \left[\frac{1}{2} \boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}} \boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \frac{\lambda}{2} \boldsymbol{\alpha}^\top \boldsymbol{\alpha} \right]$$

- Solution $\widetilde{\alpha}$ can be computed analytically!

$$\widetilde{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{h}}$$

- Ignored non-negativity constraint is imposed as post-processing:

$$\widehat{\boldsymbol{\alpha}} = \max(\mathbf{0}, \widetilde{\boldsymbol{\alpha}})$$

uLSIF : Model Selection

- Leave-one-out CV (LOOCV):

Sample 1 Sample 2 ... Sample $n-1$ Sample n

Estimation Validation

$\widehat{w}(\boldsymbol{x})$ $\frac{1}{2}\widehat{w}(\boldsymbol{x}')^2 - \widehat{w}(\boldsymbol{x}')$

- LOOCV generally requires n repetitions.
- However, for uLSIF, it is analytic!
  (Sherman-Woodbury-Morrison formula)
- Computation time including model selection is dramatically improved!

Density-Ratio Estimation Methods

Method	Density estimation	Domains of denom/nume	Model selection	Computation time
KDE	Involved	Could differ	Possible	Very fast
KMM	Free	Same	Not possible	Slow
LogReg	Free	Same	Possible	Slow
KLIEP	Free	Could differ	Possible	Slow
LSIF	Free	Could differ	Possible	Rather fast
uLSIF	Free	Could differ	Possible	Fast

- Kernel density estimation (KDE)
- Kernel mean matching (KMM)

  Huang, Smola, Gretton, Borgwardt & Schölkopf (NIPS2006)
- Logistic regression based method (LogReg)

  Qin (Biometrica1998), Cheng & Chu (Bernoulli2004)
  Bickel, Brückner & Scheffer (ICML2007)

Conclusions

- Many ML tasks can be formulated as the problem of estimating density ratios.
  - Non-stationarity adaptation, domain adaptation, multi-task learning, outlier detection, change detection in time series, feature selection, dimensionality reduction, independent component analysis, conditional density estimation, classification, two-sample test
- Directly estimating density ratios without going through density estimation is the key.
  - KMM, LogReg, KLIEP, LSIF, and uLSIF.

Books

- Quiñonero-Candela, Sugiyama, Schwaighofer & Lawrence (Eds.), Dataset Shift in Machine Learning, MIT Press, 2009.

- Sugiyama, von Bünau, Kawanabe & Müller, Covariate Shift Adaptation in Machine Learning, MIT Press (coming soon!)
- Sugiyama, Suzuki & Kanamori, Density Ratio Estimation in Machine Learning, Cambridge University Press (coming soon!)

The World of Density Ratios

Real-world applications:
Brain-computer interface, Robot control, Speech recognition
Image recognition, Natural language processing, Bioinformatics

Machine learning algorithms:
Importance sampling (domain adaptation, multi-task learning)
Statistical test (two-sample test, outlier/change detection)
Conditional density estimation (visualization, transition estimation)
Mutual information estimation (feature selection/extraction, ICA)

Density ratio estimation:
Fundamental algorithms (KMM, LogReg, KLIEP, LSIF, uLSIF)
Large-scale, High-dimensionality, Stabilization, Robustification

Theoretical analysis:
Convergence, Information criteria, Numerical stability