

Constructing Novel Nonparametric Estimators for f -divergences and Its Applications to High-energy Physics

Yang Seungtaik Auditorium (E9 building), KAIST
2023. 11. 15 (Wed.)

Yung-Kyun Noh (노영균)

*Hanyang University &
Korea Institute for Advanced Study (KIAS)*

Many contents in the slides are based on the discussion with **Dr. Cheongjae Jang** (Hanyang University) and **Dr. Sangwoong Yoon** (KIAS).

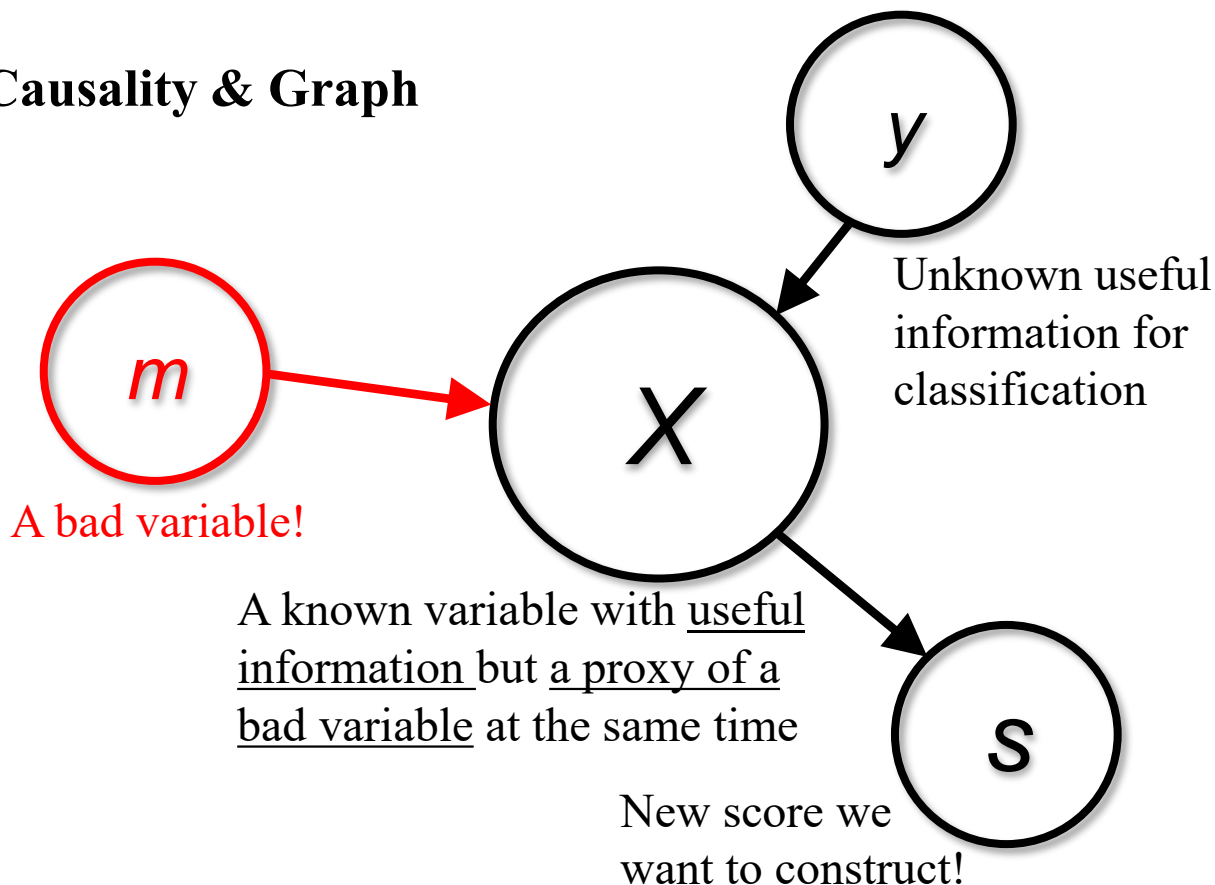


Contents

- Eliminating the flow of bad information in Bayesian Networks
- Bias of nonparametric estimators
- Reduction of the bias

Estimation of Information Contents and Decorrelation

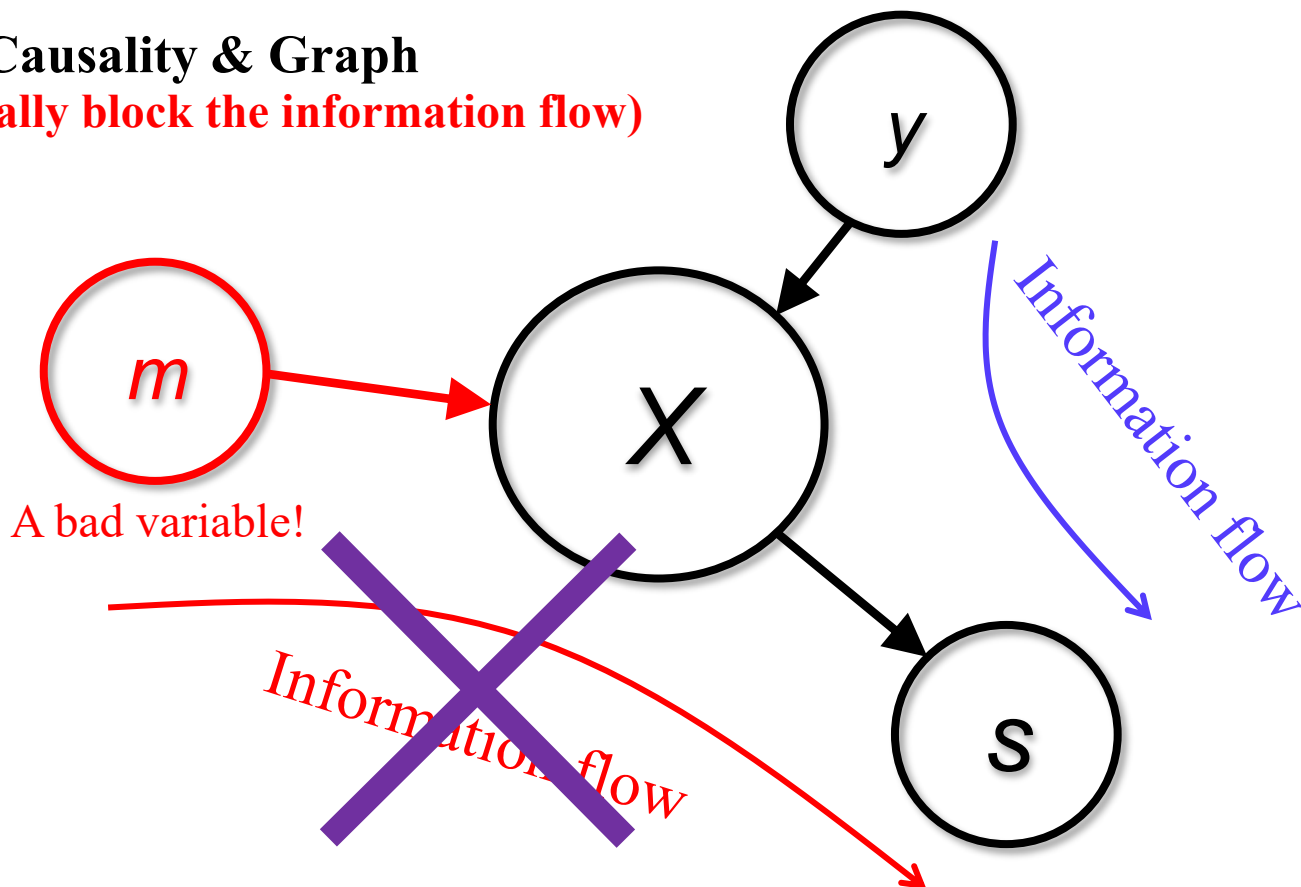
Causality & Graph



Estimation of Information Contents and Decorrelation

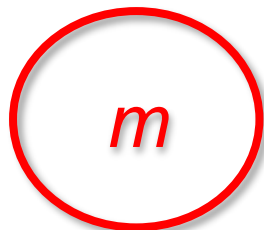
Causality & Graph

(Intentionally block the information flow)

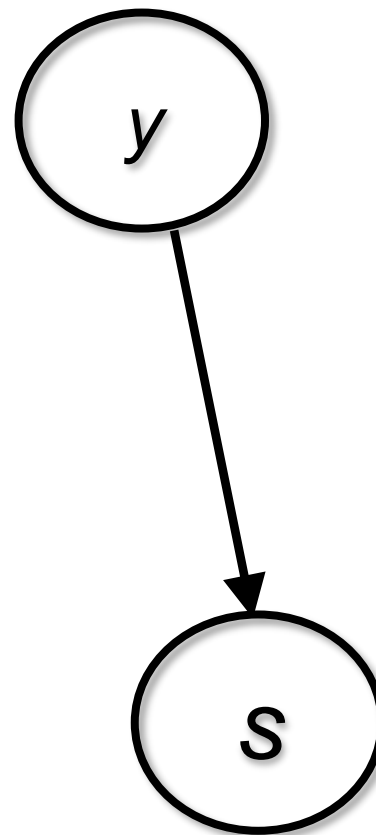


Estimation of Information Contents and Decorrelation

Causality & Graph
(Now after marginalizing X)



$$m \perp\!\!\!\perp s$$



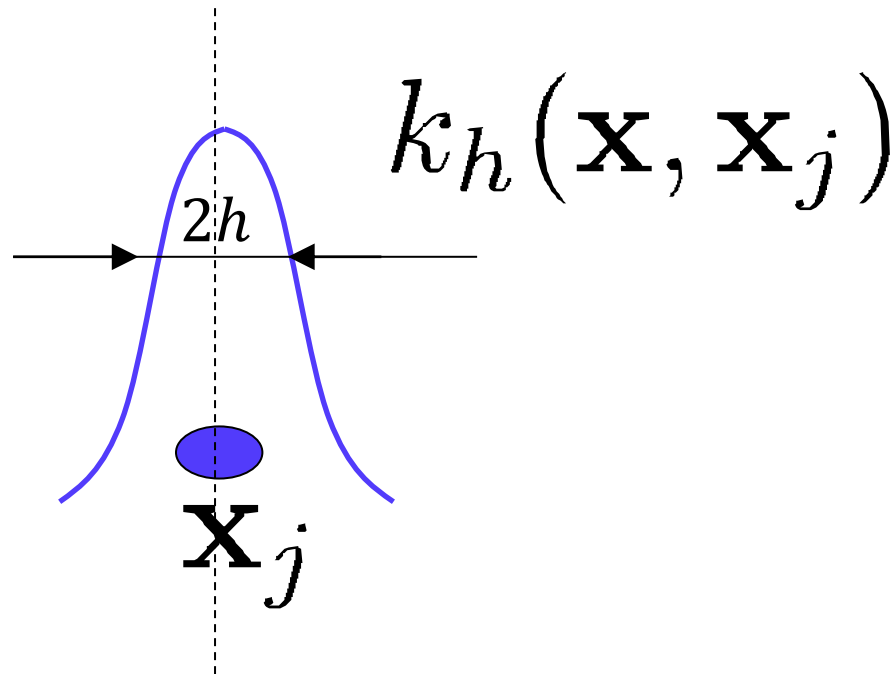
Why is This Useful?

- We do not want to use gender or ethnicity information including their proxy for classification because it is prohibited by law!
- In hospital H1, drug D1 is used for disease α . A classifier is trained using data from H1. We want to use the classifier for the patients in hospital H2, which uses drug D2 (instead of D1) for the same disease. We want D1 as well as its effect on other variables to be **excluded** in the classifier for generalization in hospital H2.
- Data are not sufficient. We decided to use the simulated data. There are some variables that we arbitrarily determined because we do not know the true distribution for those variables. We need to make sure that our classifier does not learn the patterns of those variables that we arbitrarily set.

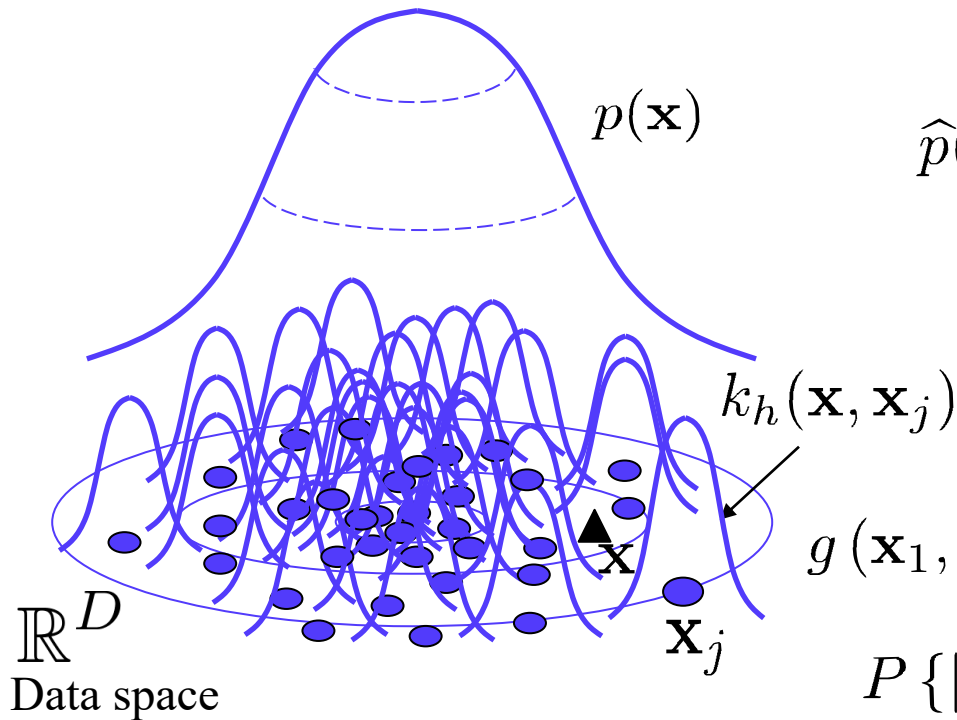
Information-theoretic Measure Estimators

- Nonparametric estimators
 - Plug-in estimators (Kernel estimator)
 - Nearest neighbor estimator
- Hilbert Schmidt Independence Criterion (Maximum mean discrepancy), distance correlation
 - Zero values imply independences.

Kernel Density Estimation (KDE)



Kernel Density Estimation (KDE)



$$\hat{p}(\mathbf{x}) = \frac{1}{N} \sum_{j=1}^N k_h(\mathbf{x}, \mathbf{x}_j)$$

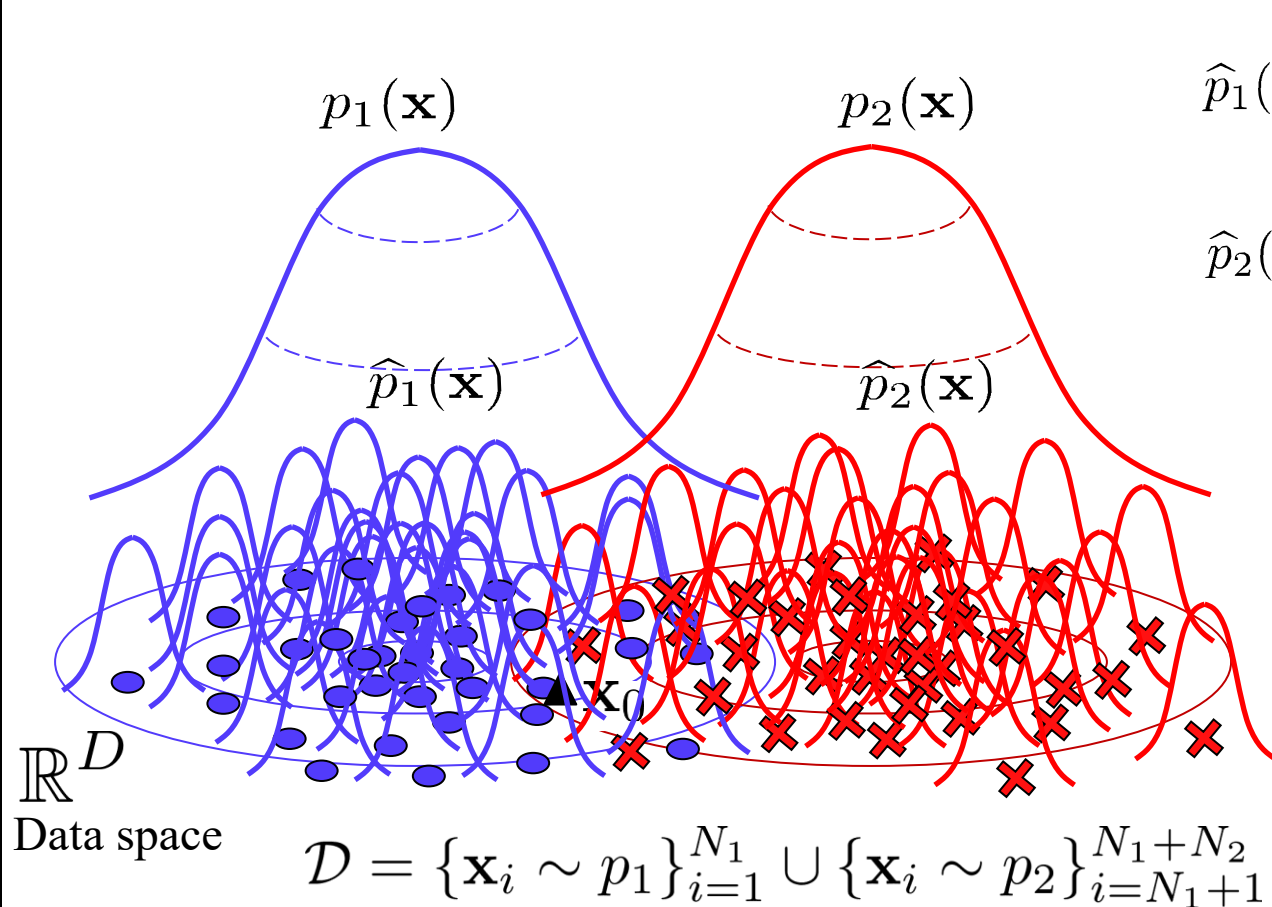
$$g(\mathbf{x}_1, \dots, \mathbf{x}_N) = \int |p(\mathbf{x}) - \hat{p}(\mathbf{x})| d\mathbf{x}$$

$$P\{|g - \mathbb{E}g| \geq \epsilon\} \leq 2e^{-N\epsilon^2/2}$$

Efron-Stein inequality bound

$$\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N \sim p(\mathbf{x})$$

KDE for Ratio Estimation



$$\hat{p}_1(\mathbf{x}) = \frac{1}{N_1} \sum_{j=1}^{N_1} k_h(\mathbf{x}, \mathbf{x}_j)$$

$$\hat{p}_2(\mathbf{x}) = \frac{1}{N_2} \sum_{j=N_1+1}^{N_1+N_2} k_h(\mathbf{x}, \mathbf{x}_j)$$

$$\frac{\hat{p}_1(\mathbf{x})}{\hat{p}_2(\mathbf{x})} \rightarrow \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})}$$

Nearest Neighbor Density Functional Estimation From Inverse Laplace Transform

J. Jon Ryu[✉], *Student Member, IEEE*, Shouvik Ganguly[✉], *Member, IEEE*, Young-Han Kim[✉], *Fellow, IEEE*,
Yung-Kyun Noh[✉], *Member, IEEE*, and Daniel D. Lee, *Fellow, IEEE*

Abstract—A new approach to L_2 -consistent estimation of a general density functional using k -nearest neighbor distances is proposed, where the functional under consideration is in the form of the expectation of some function f of the densities at each point. The estimator is designed to be asymptotically unbiased, using the convergence of the normalized volume of a k -nearest neighbor ball to a Gamma distribution in the large-sample limit, and naturally involves the inverse Laplace transform of a scaled version of the function f . Some instantiations of the proposed estimator recover existing k -nearest neighbor based estimators of Shannon and Rényi entropies and Kullback–Leibler and Rényi divergences, and discover new consistent estimators for many other functionals such as logarithmic entropies and divergences. The L_2 -consistency of the proposed estimator is established for a broad class of densities for general functionals, and the convergence rate in mean squared error is established as a function of the sample size for smooth, bounded densities.

Index Terms—Density functional estimation, information measure, nearest neighbor, inverse Laplace transform.

I. INTRODUCTION

THIS paper studies the problem of estimating an entropy functional of the form

where $f: \mathbb{R}_+ \rightarrow \mathbb{R}$ is a given function and p is a probability density over \mathbb{R}^d . Table I lists examples of f and the corresponding functional T_f . The goal is to estimate $T_f(p)$ based on independent and identically distributed (i.i.d.) samples $\mathbf{X}_{1:m} = (\mathbf{X}_1, \dots, \mathbf{X}_m)$ from p by forming an estimator $\hat{T}_f^m(\mathbf{X}_{1:m})$ that converges to $T_f(p)$ in L_2 as the sample size m grows to infinity, that is,

$$\lim_{m \rightarrow \infty} \mathbb{E}[(\hat{T}_f^m(\mathbf{X}_{1:m}) - T_f(p))^2] = 0.$$

More generally, let $f: \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}$ and consider a divergence functional

$$T_f(p, q) := \mathbb{E}_{\mathbf{X} \sim p}[f(p(\mathbf{X}), q(\mathbf{X}))] = \int f(p(\mathbf{x}), q(\mathbf{x}))p(\mathbf{x}) \, d\mathbf{x}$$

of a pair of probability densities p and q over \mathbb{R}^d . Table II lists examples of f and the corresponding T_f . In this case, the main problem is to construct an estimator $\hat{T}_f^{m,n}(\mathbf{X}_{1:m}, \mathbf{Y}_{1:n})$ based on i.i.d. samples $\mathbf{X}_{1:m}$ from p and $\mathbf{Y}_{1:n}$ from q , independent of each other, such that

Construction of the f -divergence Estimator in A Systematic Way

$$D_f(p_1(\mathbf{x}), p_2(\mathbf{x})) = \int p_1(\mathbf{x}) f\left(\frac{p_2(\mathbf{x})}{p_1(\mathbf{x})}\right) d\mathbf{x}$$

$$\widehat{D}_f = \frac{1}{N} \sum_{\mathbf{x}_i \sim p_1(\mathbf{x})} \phi(\mathbf{x}_i)$$

Hint: let $\mathbb{E}_{d_{k_1}^{(1)}, d_{k_2}^{(2)}}[\phi(\mathbf{x})] = f\left(\frac{p_2(\mathbf{x})}{p_1(\mathbf{x})}\right)$

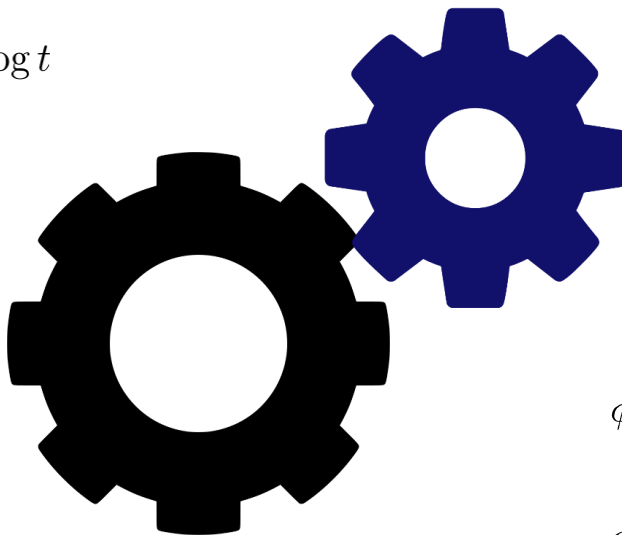
Systematic Methods of Constructing Estimators

$$\phi(u(\mathbf{x}_i), v(\mathbf{x}_i)) = \frac{(k-1)!(l-1)!}{u^{k-1}v^{l-1}} \mathcal{L}_{(u,v)}^{-1} \left[\frac{f(s,t)}{s^k t^l} \right]$$

$$f(s,t) = -\log s + \log t$$

$$f(s,t) = 1 - \sqrt{\frac{s}{t}}$$

$$f(s,t) = \frac{t}{s+t}$$



$$\phi(u,v) = \log u - \log v$$

$$\phi(u,v) = 1 - \frac{1}{\Gamma(1.5)\Gamma(2.5)} \sqrt{\frac{v}{u}} \quad (k=2)$$

$$\phi(u,v) = \mathbb{I}(u > v)$$

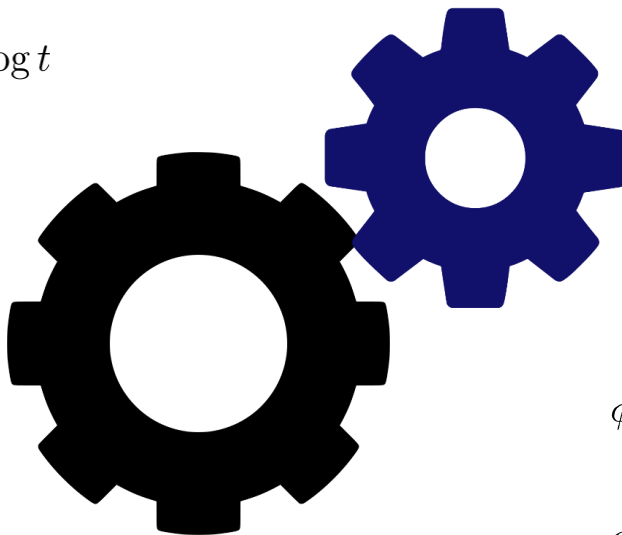
Systematic Methods of Constructing Estimators

$$\phi(u(\mathbf{x}_i), v(\mathbf{x}_i)) = \frac{(k-1)!(l-1)!}{u^{k-1}v^{l-1}} \mathcal{L}_{(u,v)}^{-1} \left[\frac{f(s,t)}{s^k t^l} \right]$$

$$f(s,t) = -\log s + \log t$$

$$f(s,t) = 1 - \sqrt{\frac{s}{t}}$$

$$f(s,t) = \frac{t}{s+t}$$



$$\phi(u,v) = \log u - \log v$$

$$\phi(u,v) = 1 - \frac{1}{\Gamma(1.5)\Gamma(2.5)} \sqrt{\frac{v}{u}} \quad (k=2)$$

$$\phi(u,v) = \mathbb{I}(u > v)$$

Optimization for Training

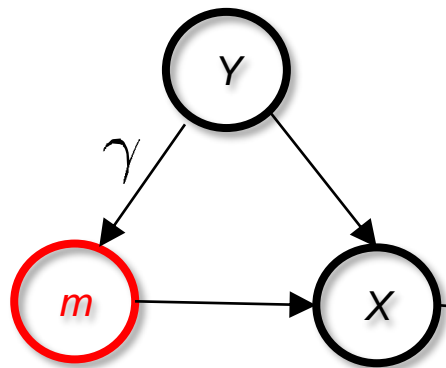
Data: $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$

$$\theta^* = \arg \min_{\theta} \underbrace{\frac{1}{N} \sum_{i=1}^N l(\hat{y}(s(\mathbf{x}_i; \theta)), y_i)}_{\text{Minimize the expected loss } \mathbb{E}[l(\hat{y}, y)]} + \underbrace{\lambda \hat{I}_{\theta}(s; m|y)}_{\text{Maximize the decorrelation } -\hat{I}_{\theta}(s; m|y)}$$

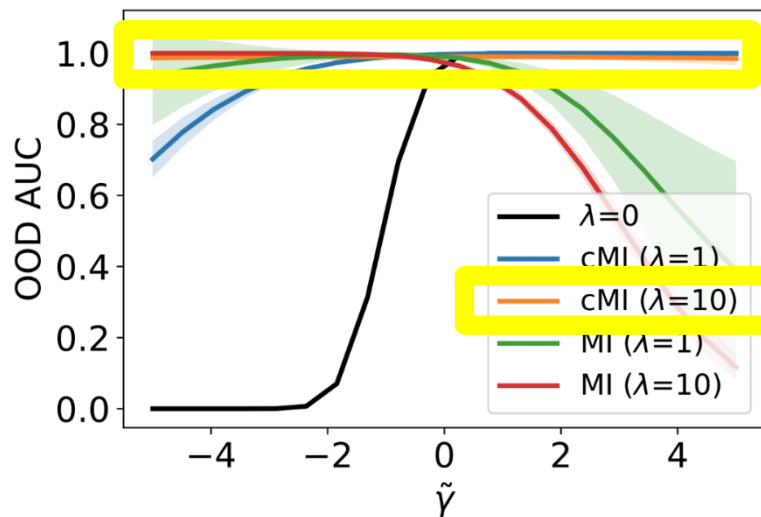
Minimize the expected loss $\mathbb{E}[l(\hat{y}, y)]$

Maximize the decorrelation $-\hat{I}_{\theta}(s; m|y)$

Use tradeoff constant λ



When γ is different in training and testing



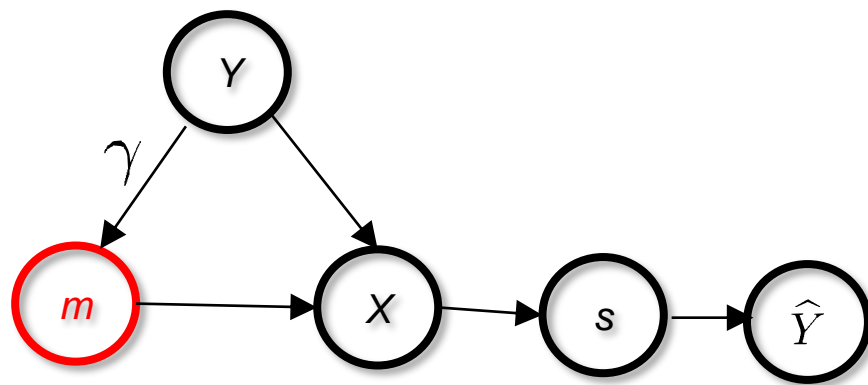
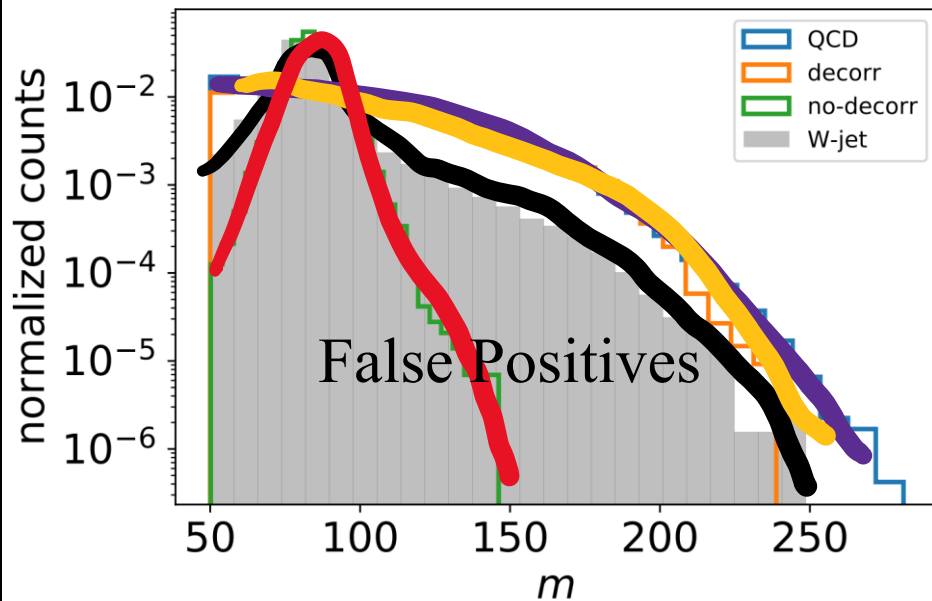
cMI:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N l(\hat{y}(s(\mathbf{x}_i; \theta)), y_i) + \lambda \hat{I}_{\theta}(s; m|y)$$

MI:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N l(\hat{y}(s(\mathbf{x}_i; \theta)), y_i) + \lambda \hat{I}_{\theta}(s; m)$$

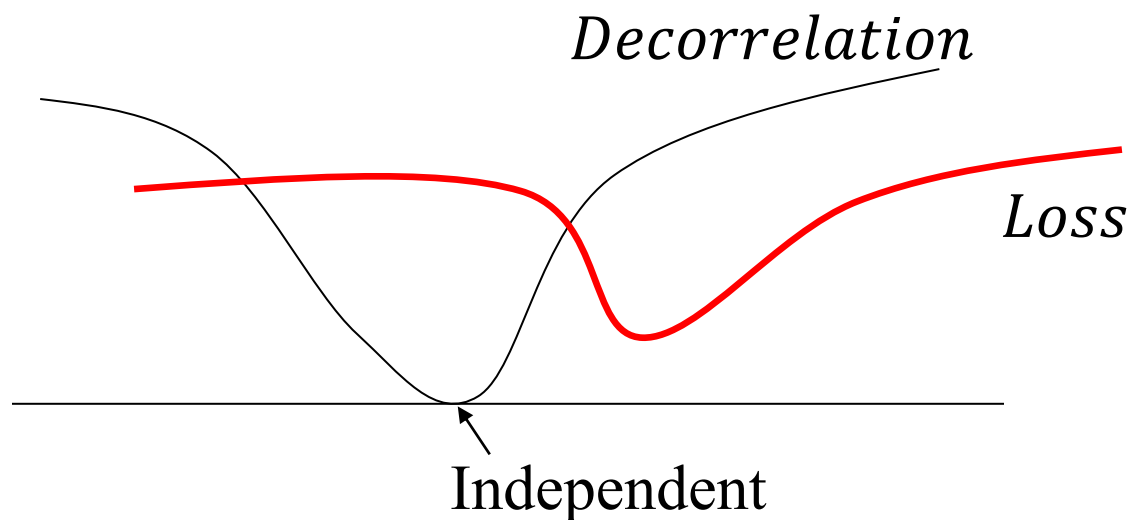
Reconstruction of W-jet Decorrelation Experiment



Reconstruction of decorrelation experiment in
Kasieczka, G., Shih, D. (2020) Robust Jet Classifiers through
Distance Correlation, *Phys. Rev. Lett. Vol. 125, Iss. 12 — 18*

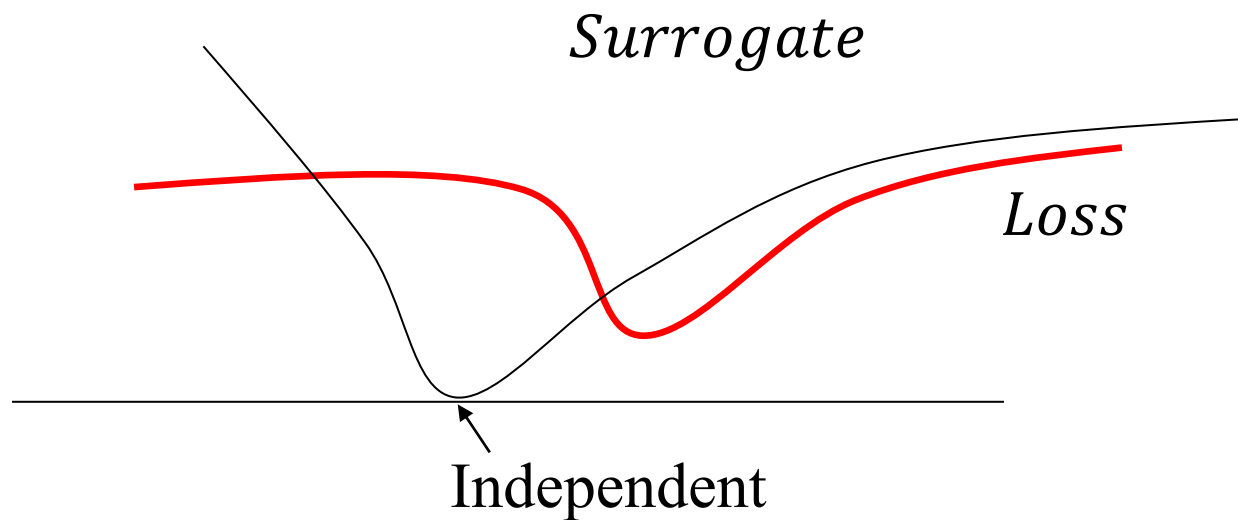
Problems with Surrogate Decorrelation Objective

Loss + Decorrelation

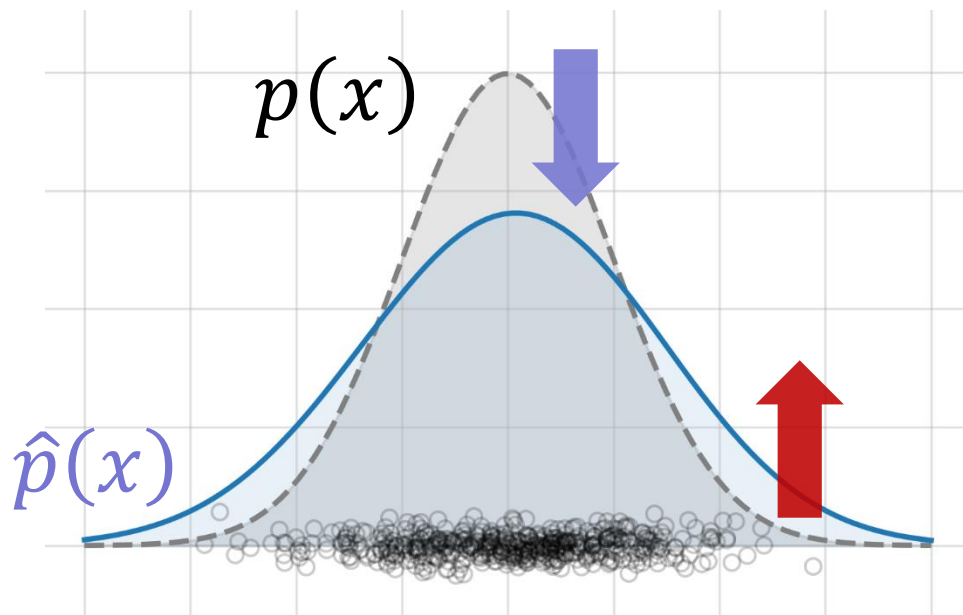


Problems with Surrogate Decorrelation Objective

Loss + Surrogate



Bias of KDE

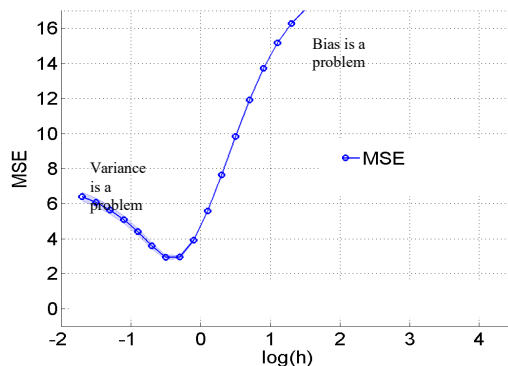


Underestimation of Peaks

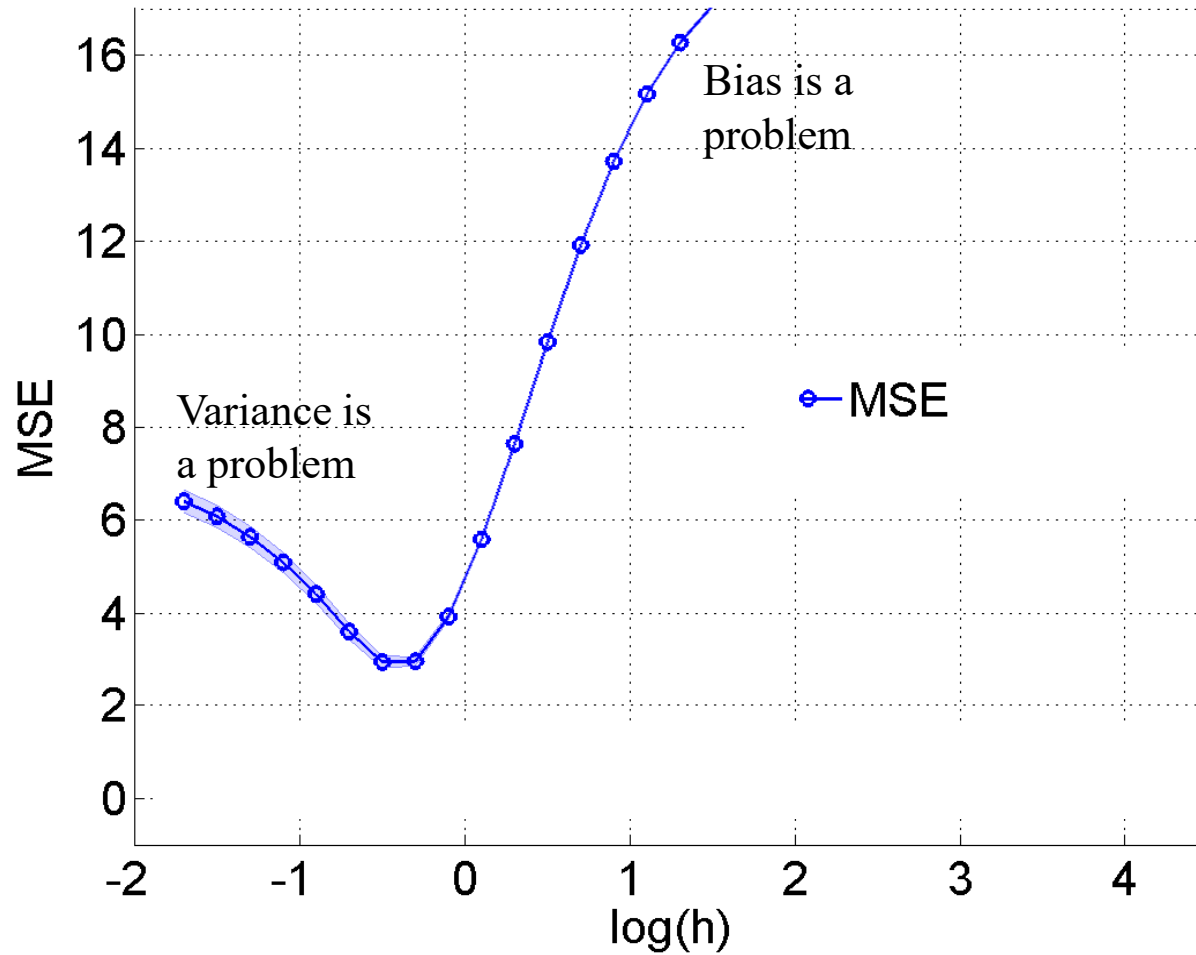
Overestimation of Tails

Peter Hall (1992)

- How to alleviate bias of KDE?
 - Undersmoothing
 - Choose bandwidth h as small as possible
 - Explicit bias correction
 - Bias of symmetric kernels depends on the second derivative of density function
 - Calculate leading order bias and subtract the bias



Peter Hall (1992) Effect of Bias Estimation on Coverage Accuracy of Bootstrap Confidence Intervals for a Probability Density, *Annals of Statistics*

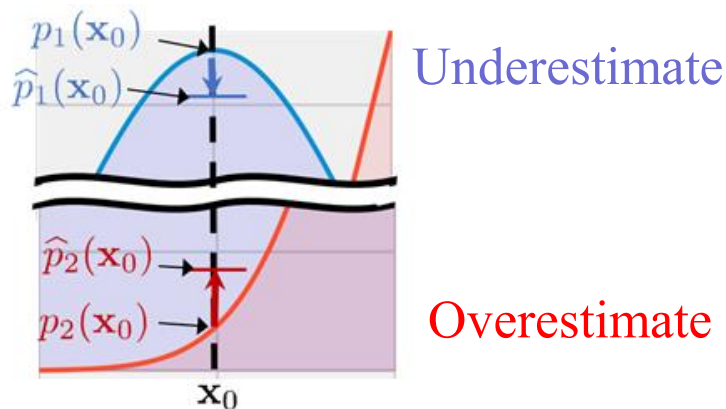
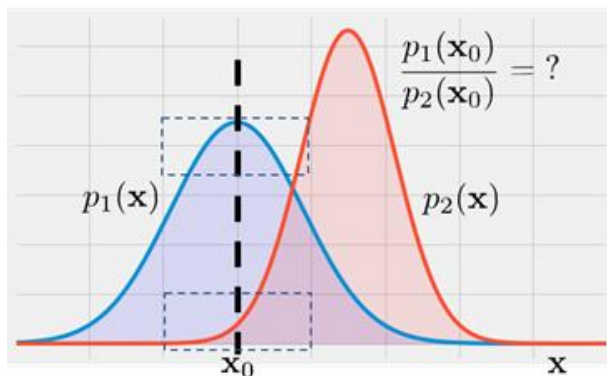


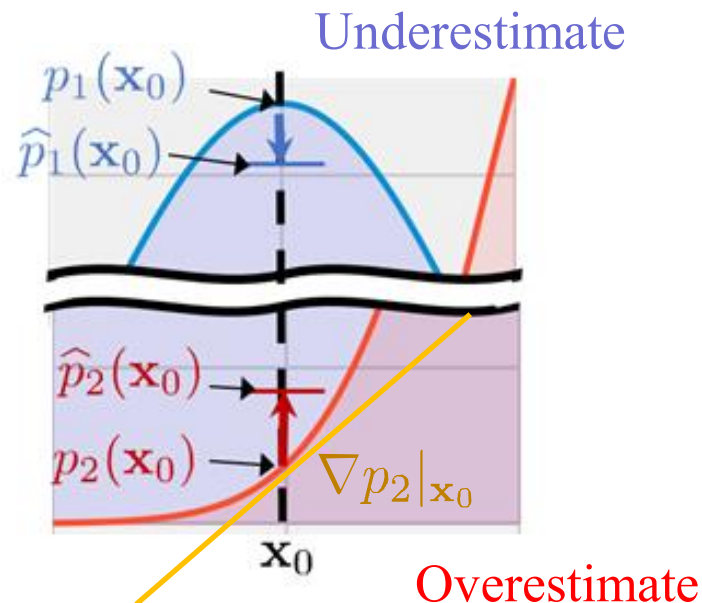
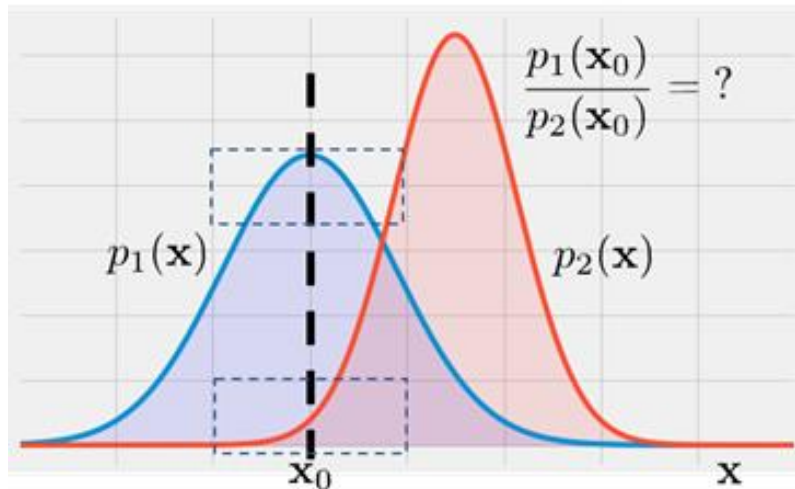
Bias in the “Ratio” Estimators

$$\hat{p}_1(\mathbf{x}) = \frac{1}{N_1} \sum_{j=1}^{N_1} k_h(\mathbf{x}, \mathbf{x}_j), \quad \hat{p}_2(\mathbf{x}) = \frac{1}{N_2} \sum_{j=N_1+1}^{N_1+N_2} k_h(\mathbf{x}, \mathbf{x}_j)$$

Interested in using $\frac{p_1(\mathbf{x})}{p_2(\mathbf{x})}$

Ex) For posterior: $P(y = 1|\mathbf{x}) = \frac{p_1(\mathbf{x})}{p_1(\mathbf{x}) + p_2(\mathbf{x})} = \frac{\frac{p_1(\mathbf{x})}{p_2(\mathbf{x})}}{\frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} + 1}$





$$\hat{p}_1(\mathbf{x}) = \frac{1}{N_1} \sum_{j=1}^{N_1} \alpha(\mathbf{x}_j) k_h(\mathbf{x}, \mathbf{x}_j)$$

$$\hat{p}_2(\mathbf{x}) = \frac{1}{N_2} \sum_{j=N_1+1}^{N_1+N_2} \alpha(\mathbf{x}_j) k_h(\mathbf{x}, \mathbf{x}_j)$$

Variational Weighting for Kernel Density Ratios

Sangwoong Yoon

Korea Institute for Advanced Study
swyoon@kias.re.kr

Frank C. Park

Seoul National University / Saige Research
fcp@snu.ac.kr

Gunsu Yun

POSTECH
gunsu@postech.ac.kr

Iljung Kim

Hanyang University
iljung0810@hanyang.ac.kr

Yung-Kyun Noh

Hanyang University / Korea Institute for Advanced Study
nohyung@hanyang.ac.kr

Abstract

Kernel density estimation (KDE) is integral to a range of generative and discriminative tasks in machine learning. Drawing upon tools from the multidimensional

NeurIPS 2023

Perturbation and Bias

$$\hat{p}_1(\mathbf{x}) = \frac{1}{N_1} \sum_{j=1}^{N_1} \underline{\alpha(\mathbf{x}_j)} k_h(\mathbf{x}, \mathbf{x}_j) \quad \hat{p}_2(\mathbf{x}) = \frac{1}{N_2} \sum_{j=N_1+1}^{N_1+N_2} \underline{\alpha(\mathbf{x}_j)} k_h(\mathbf{x}, \mathbf{x}_j)$$

$$\begin{aligned} \mathbb{E}_{\mathcal{D}_1, \mathcal{D}_2} [f(\mathbf{x})] &\rightarrow \frac{\mathbb{E}_{\mathcal{D}_1} [\hat{p}_1(\mathbf{x})]}{\mathbb{E}_{\mathcal{D}_1} [\hat{p}_1(\mathbf{x})] + \gamma \mathbb{E}_{\mathcal{D}_1} [\hat{p}_1(\mathbf{x})]} \quad \leftarrow \text{Perturb this equation} \\ &= f(\mathbf{x}) + \frac{h^2}{2} P(y=1|\mathbf{x}) P(y=2|\mathbf{x}) B_{\alpha; p_1, p_2}(\mathbf{x}) + \mathcal{O}(h^4) \end{aligned}$$

$$\begin{aligned} B_{\alpha; p_1, p_2}(\mathbf{x}) &= \nabla^\top \log \alpha|_{\mathbf{x}} \left(\underbrace{\frac{\nabla p_1|_{\mathbf{x}}}{p_1(\mathbf{x})} - \frac{\nabla p_2|_{\mathbf{x}}}{p_2(\mathbf{x})}}_{\vec{h}(\mathbf{x})} \right) + \frac{1}{2} \left[\underbrace{\frac{\nabla^2 p_1|_{\mathbf{x}}}{p_1(\mathbf{x})} - \frac{\nabla^2 p_2|_{\mathbf{x}}}{p_2(\mathbf{x})}}_{g(\mathbf{x})} \right] \\ &\equiv \nabla^\top \log \alpha|_{\mathbf{x}} \vec{h}(\mathbf{x}) + g(\mathbf{x}) \end{aligned}$$

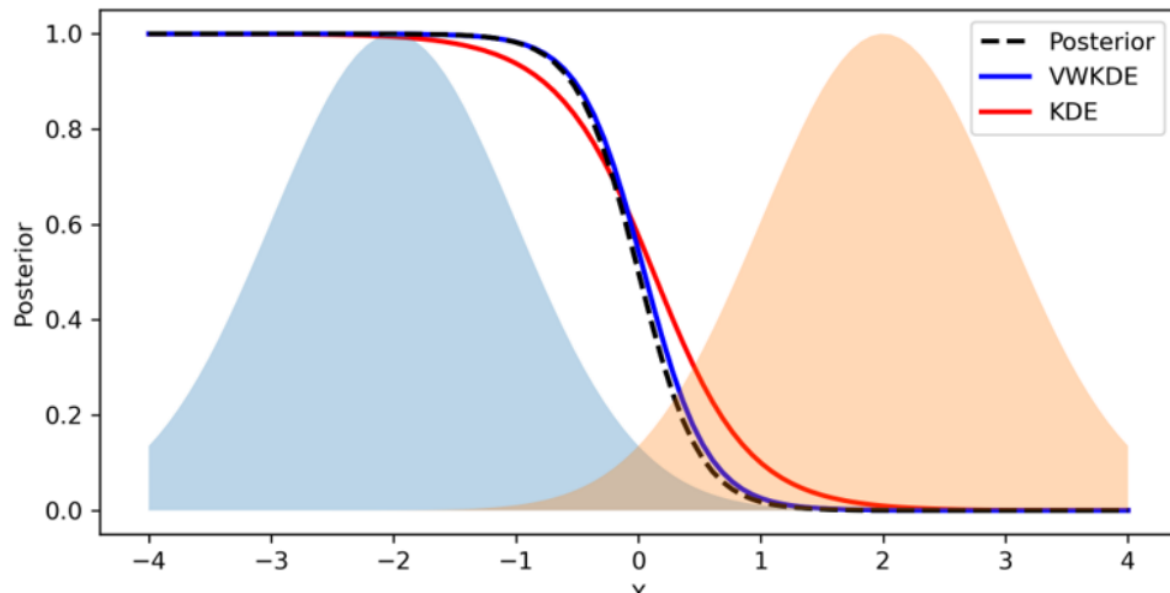
NeurIPS 2023

Calculus of Variation

$$\arg \min_{\alpha(\mathbf{x})} \int \underbrace{\left((\nabla \log \alpha|_{\mathbf{x}})^\top \vec{h}(\mathbf{x}) + g(\mathbf{x}) \right)^2}_{\text{(Pointwise Bias)}^2} r(\mathbf{x}) d\mathbf{x}$$

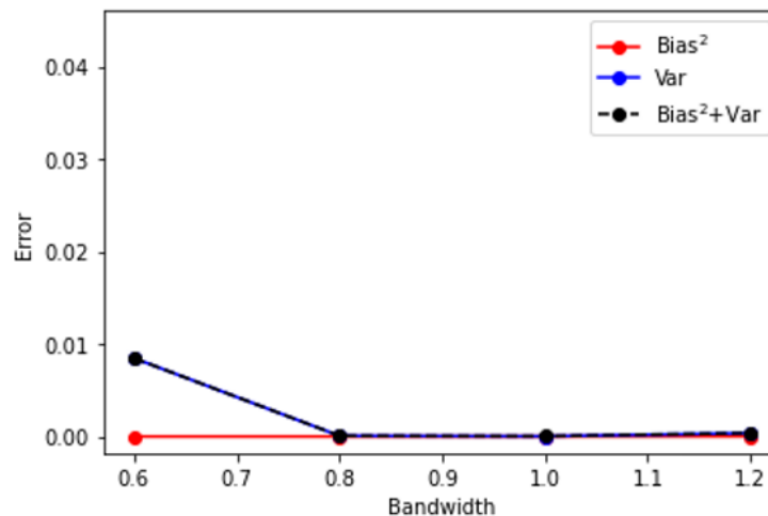
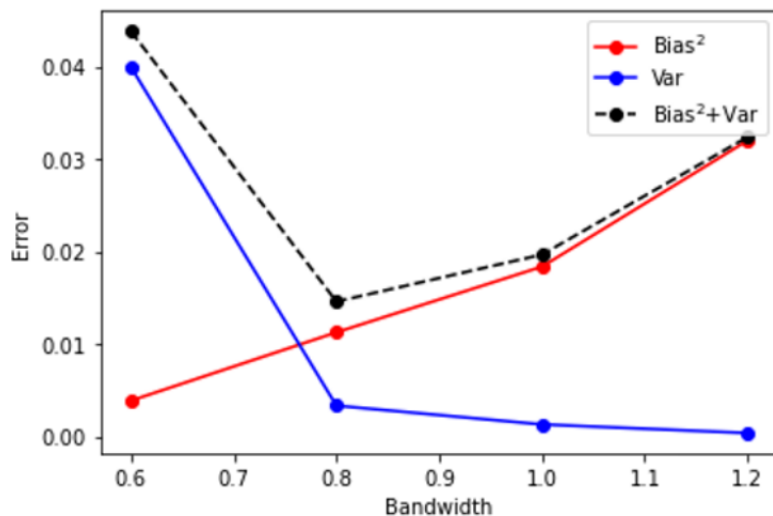
$$\longrightarrow \nabla \cdot \left[r((\nabla \log \alpha)^\top \vec{h} + g) \vec{h} \right] = 0$$

Two 20-Dimensional Gaussians



NeurIPS 2023

Two 20-Dimensional Gaussians



Kullback Divergence Estimation

$$KL(p_1||p_2) = \mathbb{E}_{\mathbf{x} \sim p_1} \left[\log \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} \right]$$

$$\widehat{KL}(p_1||p_2) = \frac{1}{N_1} \sum_{i=1}^{N_1} \log \frac{\widehat{p}_1(\mathbf{x}_i)}{\widehat{p}_2(\mathbf{x}_i)}$$

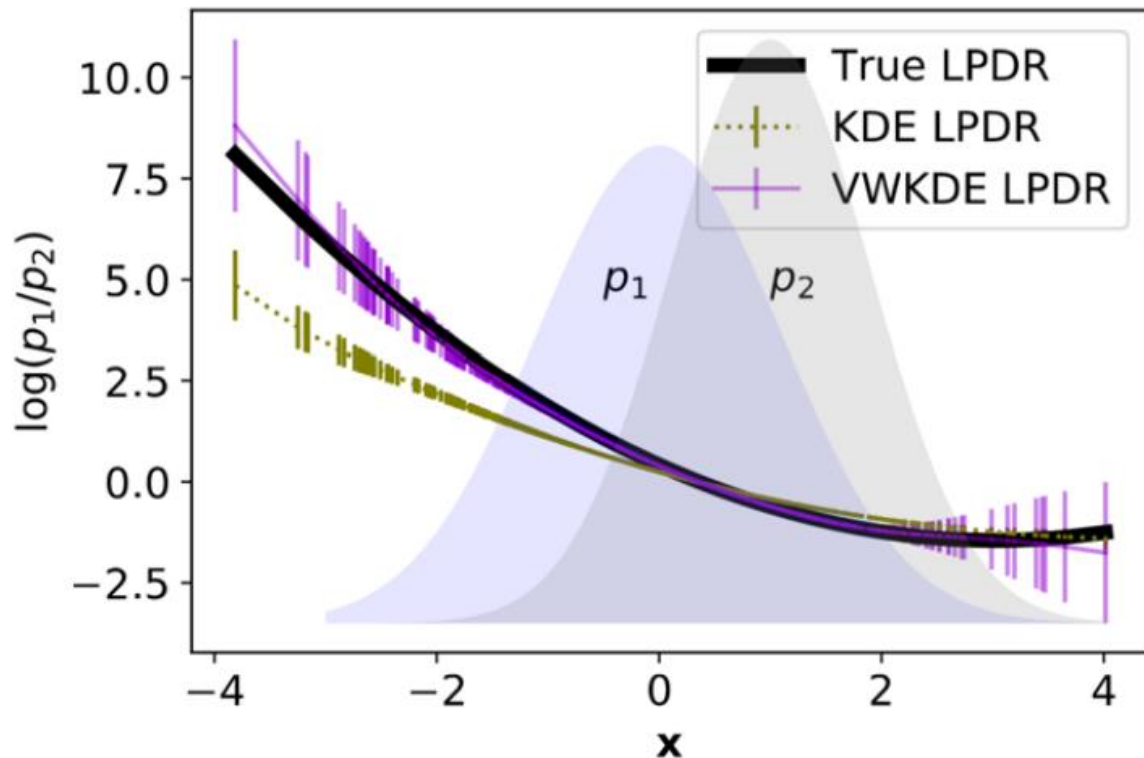
$$\text{Bias}(\mathbf{x}) = \frac{h^2}{2} \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} B_{\alpha;p_1,p_2}(\mathbf{x})$$

Previously used $B_{\alpha;p_1,p_2}(\mathbf{x})$

$$B_{\alpha;p_1,p_2}(\mathbf{x}) = \nabla^\top \log \alpha|_{\mathbf{x}} \left(\frac{\nabla p_1|_{\mathbf{x}}}{p_1(\mathbf{x})} - \frac{\nabla p_2|_{\mathbf{x}}}{p_2(\mathbf{x})} \right) + \frac{1}{2} \left[\frac{\nabla^2 p_1|_{\mathbf{x}}}{p_1(\mathbf{x})} - \frac{\nabla^2 p_2|_{\mathbf{x}}}{p_2(\mathbf{x})} \right]$$

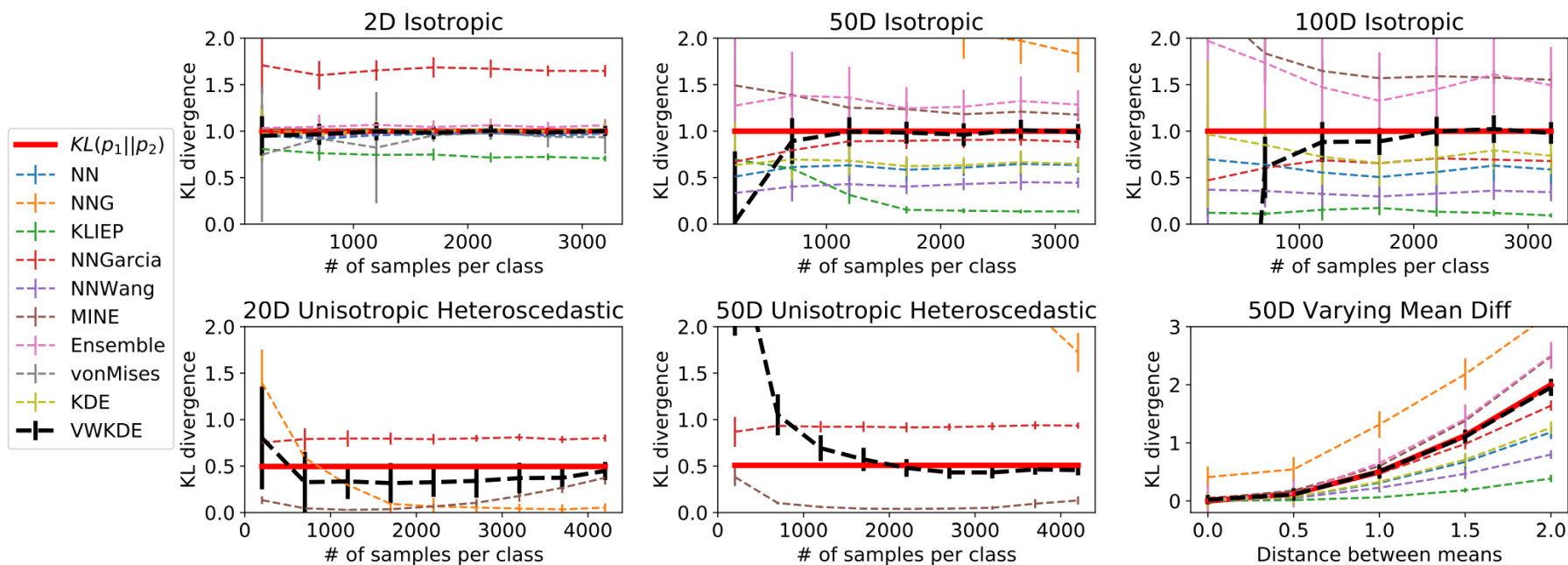
NeurIPS 2023

Log Probability Density Ratio (p_1/p_2)



NeurIPS 2023

KL-Divergence Estimation



- Diffusion-Convection Equation for Bias

Direction of Convection

Concentrations of two fluids

Convection Diffusion

$$\text{Bias} \propto \left[\nabla^\top \left(\log \alpha + \frac{1}{2} \log p_1 \right) \nabla \log p_1 + \frac{1}{2} \nabla^2 \log p_1 \right] - \left[\nabla^\top \left(\log \alpha + \frac{1}{2} \log p_2 \right) \nabla \log p_2 + \frac{1}{2} \nabla^2 \log p_2 \right]$$

Figure credit: JOOINN

Diffusion-Convection for Bias

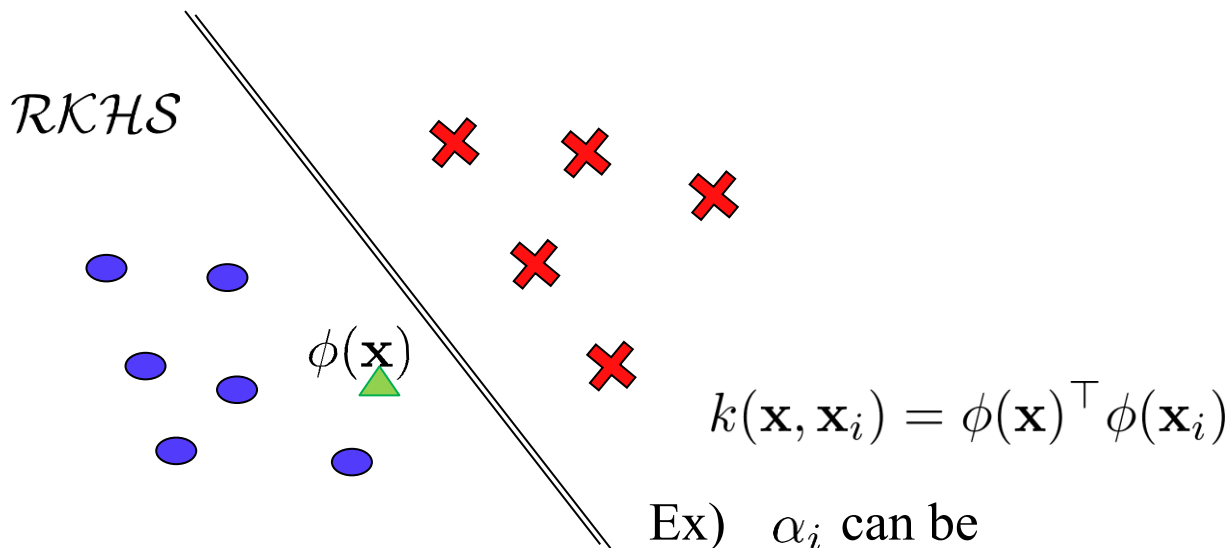
$$\begin{aligned} \text{Bias} \quad \propto \quad & \left[\nabla^\top \left(\log \alpha + \frac{1}{2} \log p_1 \right) \nabla \log p_1 + \frac{1}{2} \nabla^2 \log p_1 \right] \\ & - \left[\nabla^\top \left(\log \alpha + \frac{1}{2} \log p_2 \right) \nabla \log p_2 + \frac{1}{2} \nabla^2 \log p_2 \right] \end{aligned}$$

RKHS Representation of P.D. Kernels

$$\sum_{i=1}^{N_1+N_2} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) \geq 0$$

$$y_i = 1 \quad \text{if } \mathbf{x}_i \sim p_1(\mathbf{x})$$

$$y_i = -1 \quad \text{if } \mathbf{x}_i \sim p_2(\mathbf{x})$$



Ex) α_i can be optimized to produce large margin in the feature space.

Correspondences

Probability density
estimation interpretation

$$\hat{p}_1(\mathbf{x}) - \hat{p}_2(\mathbf{x}) = \sum_{i=1}^{N_1+N_2} y_i k_h(\mathbf{x}, \mathbf{x}_i)$$

$$\begin{aligned} y_i &= 1 && \text{if } \mathbf{x}_i \sim p_1(\mathbf{x}) \\ y_i &= -1 && \text{if } \mathbf{x}_i \sim p_2(\mathbf{x}) \end{aligned}$$

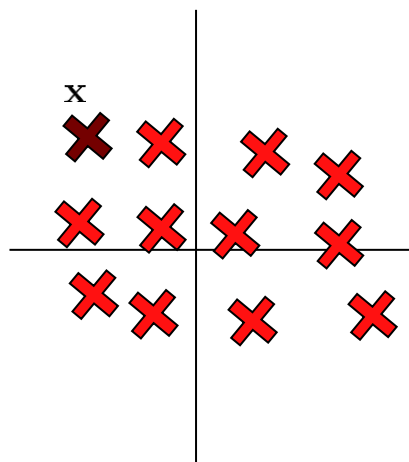


$$\sum_{i=1}^{N_1+N_2} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) = \sum_{i=1}^{N_1+N_2} \alpha_i y_i \phi(\mathbf{x})^\top \phi(\mathbf{x}_i)$$

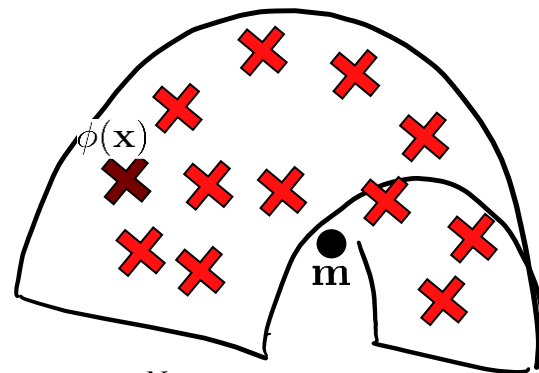
Geometric interpretation
in RKHS

RKHS Geometric Interpretation of KDE

- *RKHS* geometry



$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$$

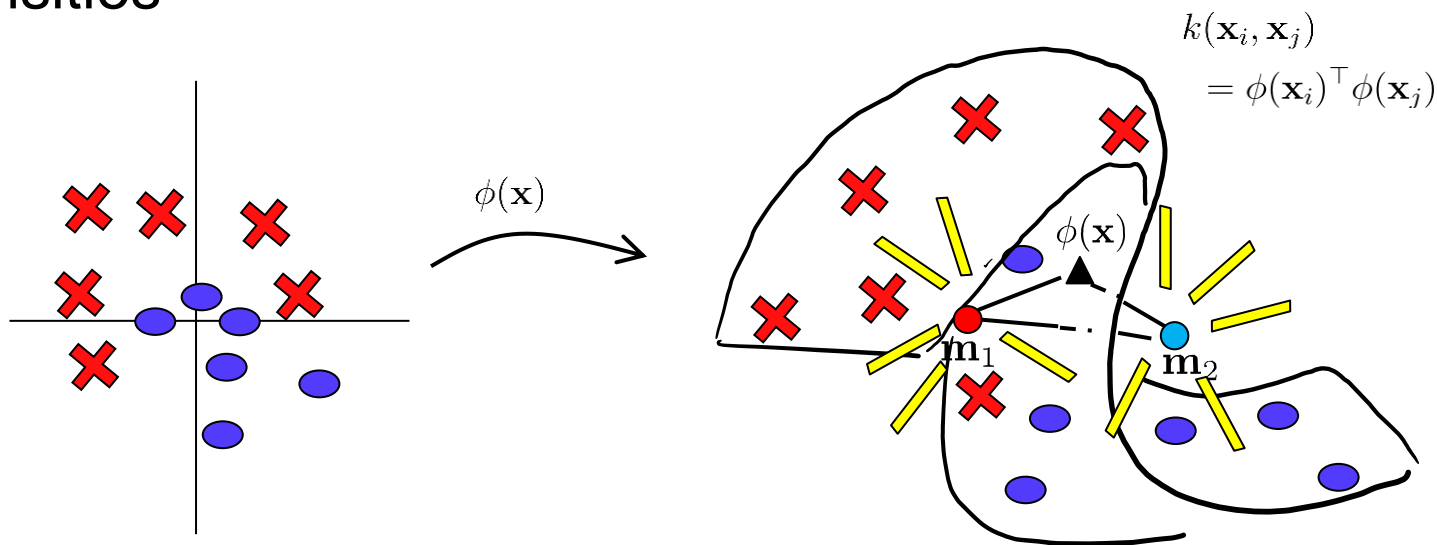


$$\begin{aligned}\hat{p}(\mathbf{x}) &= \frac{1}{N} \sum_{j=1}^N k(\mathbf{x}, \mathbf{x}_j) \\ &= \phi(\mathbf{x})^\top \left(\frac{1}{N} \sum_{i=1}^N \phi(\mathbf{x}_i) \right)\end{aligned}$$

$$\begin{aligned}&= \phi(\mathbf{x})^\top \mathbf{m} \quad \mathbf{m} = \frac{1}{N} \sum_{i=1}^N \phi(\mathbf{x}_i)\end{aligned}$$

RKHS Geometric Interpretation of KDE

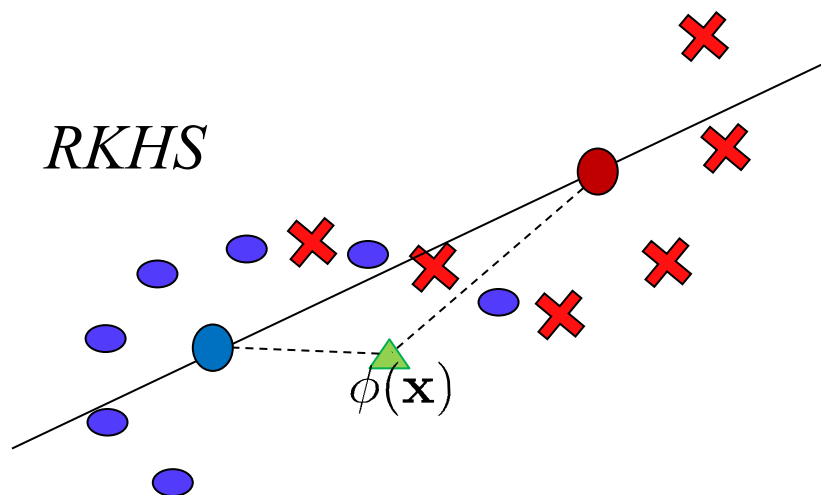
- RKHS* representation of comparing two probability densities



$$\begin{aligned}
 \hat{p}_1(\mathbf{x}) - \hat{p}_2(\mathbf{x}) &= \phi(\mathbf{x})^\top \left\{ \frac{1}{N_1} \sum_{i=1}^{N_1} \phi(\mathbf{x}_i) - \frac{1}{N_2} \sum_{i=N_1+1}^{N_1+N_2} \phi(\mathbf{x}_i) \right\} \\
 &= \phi(\mathbf{x})^\top \{ \mathbf{m}_1 - \mathbf{m}_2 \} \quad \mathbf{m}_1 = \frac{1}{N_1} \sum_{i=1}^{N_1} \phi(\mathbf{x}_i), \quad \mathbf{m}_2 = \frac{1}{N_2} \sum_{i=N_1+1}^{N_1+N_2} \phi(\mathbf{x}_i)
 \end{aligned}$$

RKHS Representation

$$\hat{p}_1(\mathbf{x}) - \hat{p}_2(\mathbf{x}) = \sum_{i=1}^{N_1+N_2} y_i k_h(\mathbf{x}, \mathbf{x}_i) \geq 0$$

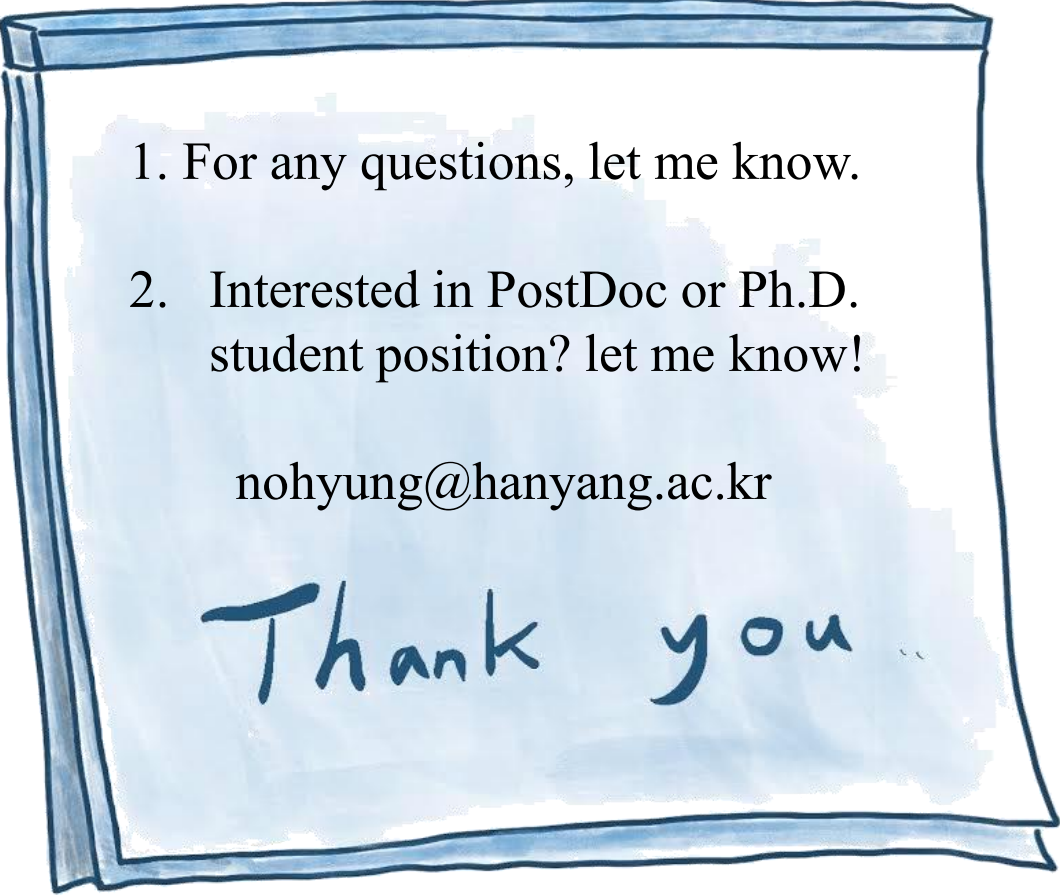


$$y_i = 1 \quad \text{if } \mathbf{x}_i \sim p_1(\mathbf{x})$$

$$y_i = -1 \quad \text{if } \mathbf{x}_i \sim p_2(\mathbf{x})$$

Summary

- 1. Decorrelation for eliminating undesired information (for our safe classifier)
- 2. Construction of novel nonparametric estimators for f -divergences
- 3. Elimination of high-dimensional bias in nonparametric estimators

- 
1. For any questions, let me know.
 2. Interested in PostDoc or Ph.D. student position? let me know!

nohyung@hanyang.ac.kr

Thank you ..