Nonparametric *f*-Divergence Estimation and its Application to Eliminating Harmful Variables

KIAS Center for AI and Natural Sciences
2024 Fall Workshop
Nest Hotel, Incheon, Rep. of Korea
2024. 11. 6 (Wed.)

Yung-Kyun Noh

Hanyang University &

Korea Institute for Advanced Study

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

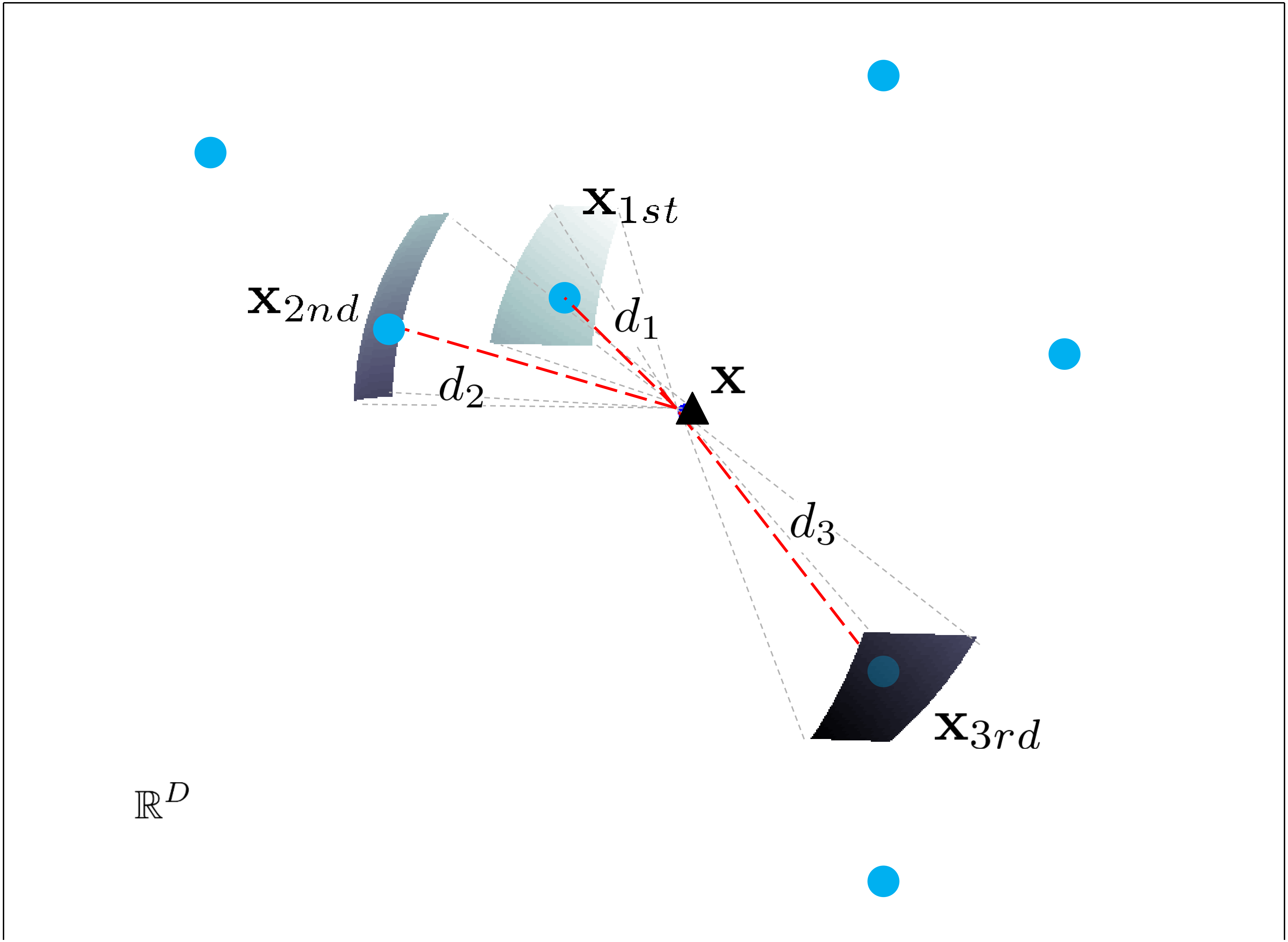
where $f\colon \mathbb{R}_+ \to \mathbb{R}$ is a given function and p is a probability density over $\mathbb{R}^d$. Table I lists examples of f and the corresponding functional T_f. The goal is to estimate $T_f(p)$ based on independent and identically distributed (i.i.d.) samples $\mathbf{X}_{1:m} = (\mathbf{X}_1, \ldots, \mathbf{X}_m)$ 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,

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

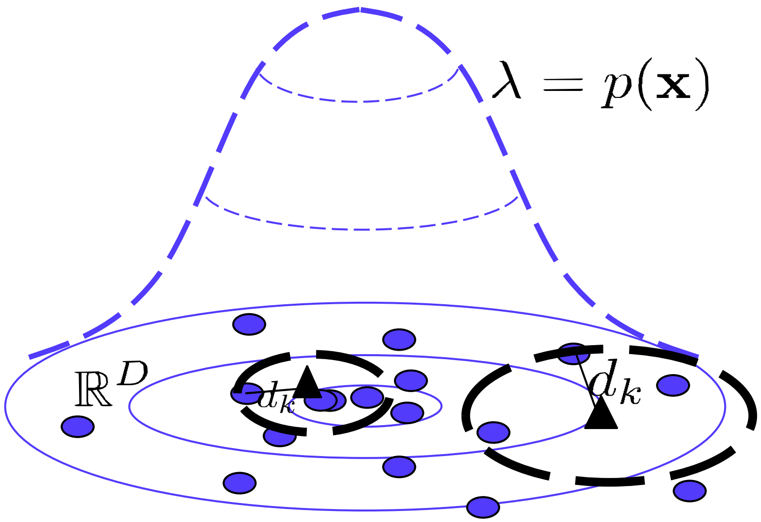
More generally, let $f\colon \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ and consider a divergence functional

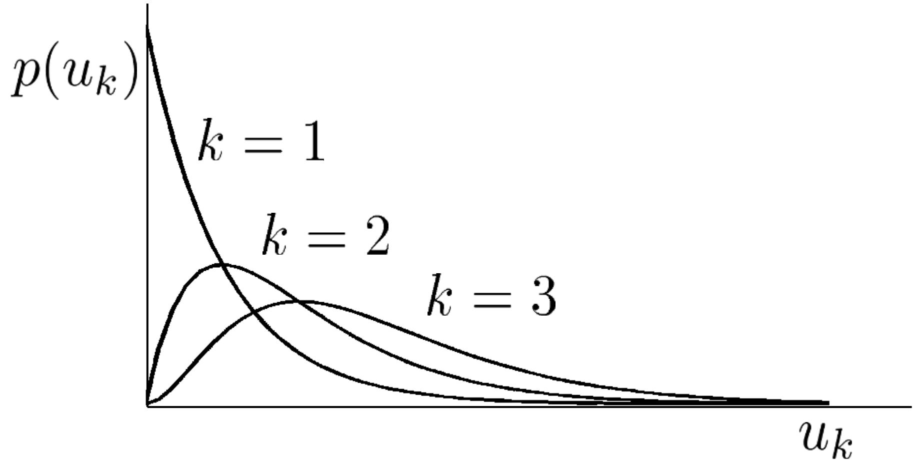
$$T_f(p,q) := \mathbb{E}_{\mathbf{X}\sim p}[f(p(\mathbf{X}), q(\mathbf{X}))] = \int f(p(\mathbf{x}), q(\mathbf{x}))p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

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


$\mathbf{x}_{1st}$
$\mathbf{x}_{2nd}$
d_1
d_2
$\mathbf{x}$
d_3
$\mathbf{x}_{3rd}$
$\mathbb{R}^D$

Density Function for Nearest Neighbor Distances





Gamma (Erlang) function of order *k*

$N \rightarrow \infty,$

$$p(u^{(k)}|\lambda) = \frac{\lambda^k}{\Gamma(k)} \exp\left(-\lambda u^{(k)}\right) (u^{(k)})^{k-1}$$

$$(\lambda = p(\mathbf{x}))$$

Volume of sphere

$$u^{(k)} = N\gamma d_k^D, \quad \gamma = \frac{\pi^{\frac{D}{2}}}{\Gamma(\frac{D}{2}+1)}$$

Karl W. Pettis et al. (1979) TPAMI
Hertz, P. (1909) Mathematische Annalen

Construction of the Estimator

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

$$\widehat{D_f}(p_1(\mathbf{x}), p_2(\mathbf{x})) = \frac{1}{N} \sum_{\mathbf{x}_i \sim p_1(\mathbf{x})} \phi\left(u_1^{(k_1)}(\mathbf{x}_i), u_2^{(k_2)}(\mathbf{x}_i)\right)$$

classes

Let $\mathbb{E}_{u_1^{(k_1)}, u_2^{(k_2)}}[\phi(\mathbf{x})] = f\left(\frac{p_2(\mathbf{x})}{p_1(\mathbf{x})}\right)$

Example – How to Build an Estimator

- Kullback-Leibler Estimator

$$D_{\mathrm{KL}}(p_1(\mathbf{x}), p_2(\mathbf{x})) = -\int p_1(\mathbf{x}) \log\left(\frac{p_2(\mathbf{x})}{p_1(\mathbf{x})}\right) d\mathbf{x}$$

$$\mathbb{E}_{u_1^{(k)}, u_2^{(k)}}[\phi] =$$

$$\int_0^\infty \int_0^\infty \frac{p_1^k}{\Gamma(k)} \exp(-p_1 u_1^{(k)}) {u_1^{(k)}}^{k-1} \frac{p_2^k}{\Gamma(k)} \exp(-p_2 u_2^{(k)}) {u_2^{(k)}}^{k-1} \underline{\phi(u_1^{(k)}, u_2^{(k)})} du_1^{(k)} du_2^{(k)}$$

$$= \frac{p_1^k p_2^k}{\Gamma(k)^2} \mathcal{L}_{p_1}\left[\mathcal{L}_{p_2}\left[\underline{\phi(u_1^{(k)}, u_2^{(k)})} {u_1^{(k)}}^{k-1} {u_2^{(k)}}^{k-1}\right]\right] = -\log\left(\frac{p_2}{p_1}\right)$$

Laplace transform: $\mathcal{L}_s[f(t)] = \int_0^\infty f(t) \exp(-st) dt$

Laplace Transform

$$u_1 = u_1^{(k_1)}, u_2 = u_2^{(k_2)}$$

$$\mathcal{L}_{p_1}\left[\mathcal{L}_{p_2}\left[\phi(u_1, u_2){u_1}^{k_1-1}{u_2}^{k_2-1}\right]\right] = -\frac{\Gamma(k_1)\Gamma(k_2)}{p_1^{k_1} p_2^{k_2}} \log\left(\frac{p_2}{p_1}\right)$$

- Perform the inverse Laplace transform of $-\frac{\Gamma(k_1)\Gamma(k_2)}{p_1^{k_1} p_2^{k_2}} \log\left(\frac{p_2}{p_1}\right)$ with respect to p_1 and p_2, then multiply $\frac{1}{u_1^{k_1-1} u_2^{k_2-1}}$ to obtain $\phi(u_1, u_2)$.

- Use the following two Laplace Transforms

$$\mathcal{L}_s[t^n \log t] = \Gamma(n+1) s^{-(n+1)}(\psi(n+1) - \log s), \quad n > -1$$

$$\mathcal{L}_s[t^n] = \Gamma(n+1) s^{-(n+1)}, \quad n > -1$$

$$\phi(u_1, u_2) = \log u_1 - \log u_2 - \psi(k_1) + \psi(k_2)$$

$$\mathbb{E}_{u_1,u_2}\phi(u_1, u_2) = -\log \frac{p_2}{p_1}$$

- Convergence?
  - It is practically working to check whether the variance (expectation of the square) diverges or not.

$$Var\left[\phi(u_1, u_2)^2\right] =$$

$$\mathbb{E}_{u_1,u_2}\left[\phi(u_1, u_2)^2\right] - \mathbb{E}_{u_1,u_2}\left[\phi(u_1, u_2)\right]^2 < \infty$$

$$\mathbb{E}_{u_1,u_2}\left[\phi(u_1,u_2)^2\right] < \infty$$

$$\Leftrightarrow$$

$$\mathcal{L}_{p_1}\mathcal{L}_{p_2}\left[\phi(u_1,u_2)^2 u_1^{k_1-1} u_2^{k_2-1}\right] < \infty$$

$$\Leftrightarrow$$

$$\mathcal{L}_{p_1}\left[u_1^{k_1-1}(\log u_1)^2\right] = (-1)^{k_1-1}\frac{d^{k_1-1}}{dp_1^{k_1-1}}\frac{1}{p_1}\left((\log p_1 + C)^2 + \frac{1}{6}\pi^2\right) < \infty$$

$$\mathcal{L}_{p_2}\left[u_2^{k_2-1}(\log u_2)^2\right] < \infty$$

Kozachenko-Leonenko estimator

$$\widehat{D_f}(p_1(\mathbf{x}), p_2(\mathbf{x})) = \frac{1}{N} \sum_{\mathbf{x}_i \sim p_1(\mathbf{x})} \phi(u_1^{(k_1)}(\mathbf{x}_i), u_2^{(k_2)}(\mathbf{x}_i))$$

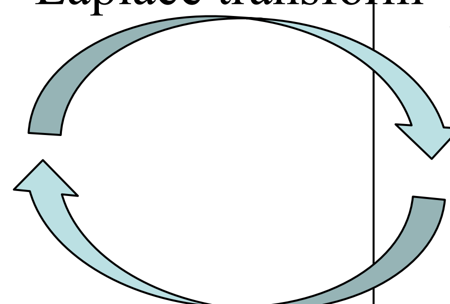
$$\phi(u_1^{(k_1)}, u_2^{(k_2)}) = \log u_1^{(k_1)} - \log u_2^{(k_2)} - \psi(k_1) + \psi(k_2)$$

L. F. Kozachenko and N. N. Leonenko (1987) Problemy Peredachi Informatsii
N. Leonenko, L. Pronzato, &V. Savani, (2008) Annals of Statistics
B. Poczos and J. Schneider (2011) AISTATS

– For the analysis with finite *N*, see

D. Lombardi and S. Pant (2016) Phys. Rev. E
A. Kraskov, H. Stögbauer, and P. Grassberger (2004) Phys. Rev. E

$D_f(p_1(\mathbf{x}), p_2(\mathbf{x}))$	Estimator $\phi(u_1, u_2)$	$f(t)$
$\frac{1}{\alpha-1}\left(\int p_1^{(1-\alpha)} p_2^{\alpha} d\mathbf{x} - 1\right)$ $(\alpha \neq 1)$	$\frac{1}{\alpha-1}\left(\frac{\Gamma(k_1)\Gamma(k_2)}{\Gamma(\alpha+k_1)\Gamma(k_2-\alpha)}\left(\frac{u_1}{u_2}\right)^{\alpha} - \frac{\Gamma(k_1)\Gamma(k_2)}{\Gamma(k_1+1)\Gamma(k_2-1)}\frac{u_1}{u_2}\right)$	$\frac{t^{\alpha} - t}{\alpha-1}$
$-\int p_1 \log\left(\frac{p_2}{p_1}\right) d\mathbf{x}$	$\log u_1^{(k_1)} - \log u_2^{(k_2)} - \psi(k_1) + \psi(k_2)$ $\psi(.)$: digamma	$-\log t$
$1 - \int \sqrt{p_1 p_2}\, d\mathbf{x}$	$1 - \frac{1}{\Gamma(1.5)\Gamma(2.5)}\sqrt{\frac{v_1^{(2)}}{u_2^{(2)}}}$	$1 - \sqrt{t}$
$1 - \int \frac{p_1 p_2}{p_1 + p_2} d\mathbf{x}$	$\mathbb{I}(u_1^{(1)} < u_2^{(1)})$	$\frac{1}{1+t}$
⋮	⋮	⋮



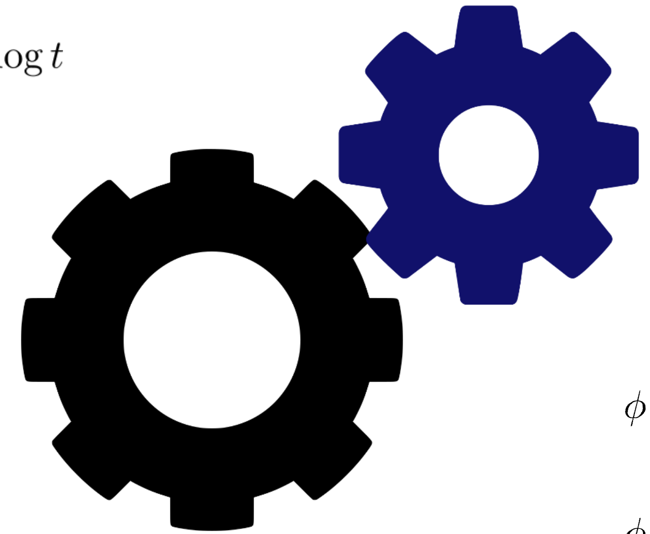
Systematic Methods of Constructing Estimators

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

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

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

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



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

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

$$(k=2)$$

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

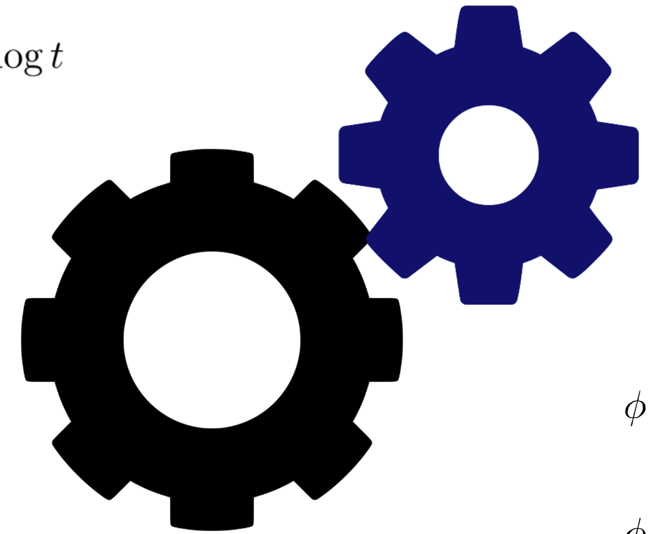
Systematic Methods of Constructing Estimators

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

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

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

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



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

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

$$(k=2)$$

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

Estimation of Bhattacharyya Coefficient

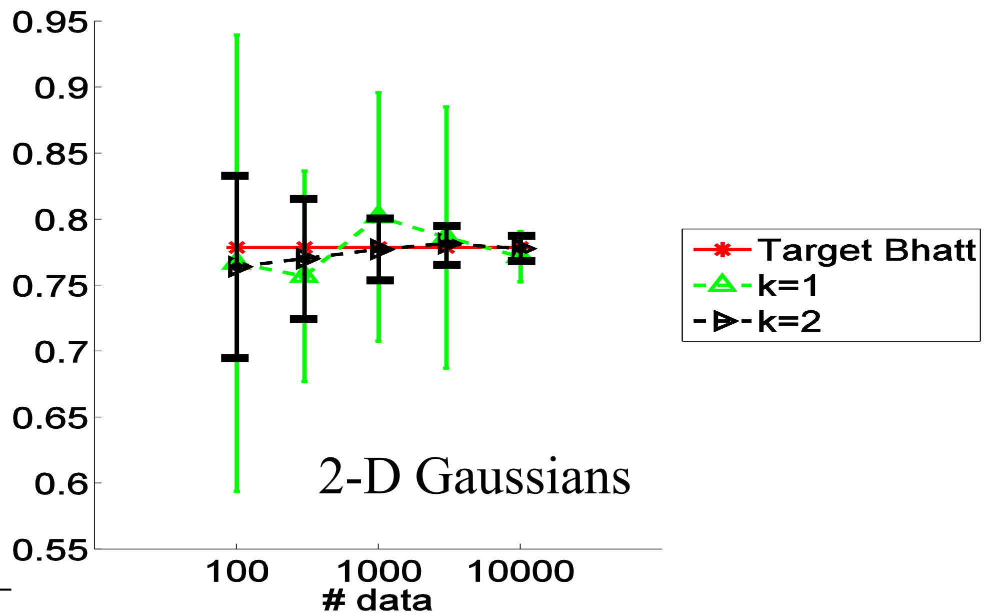
$$D_{\text{Batt}}(p_1, p_2) = \int \sqrt{p_1 p_2}\, d\mathbf{x}$$

$k_1=k_2=1$ ➔ $$\phi(u_1, u_2) = \frac{1}{\Gamma(1.5)\Gamma(0.5)}\sqrt{\frac{u_1}{u_2}}$$

Condition for k_2 is not satisfied

$k_1=k_2=2$ ➔ $$\phi(u_1, u_2) = \frac{1}{\Gamma(2.5)\Gamma(1.5)}\sqrt{\frac{u_1}{u_2}}$$

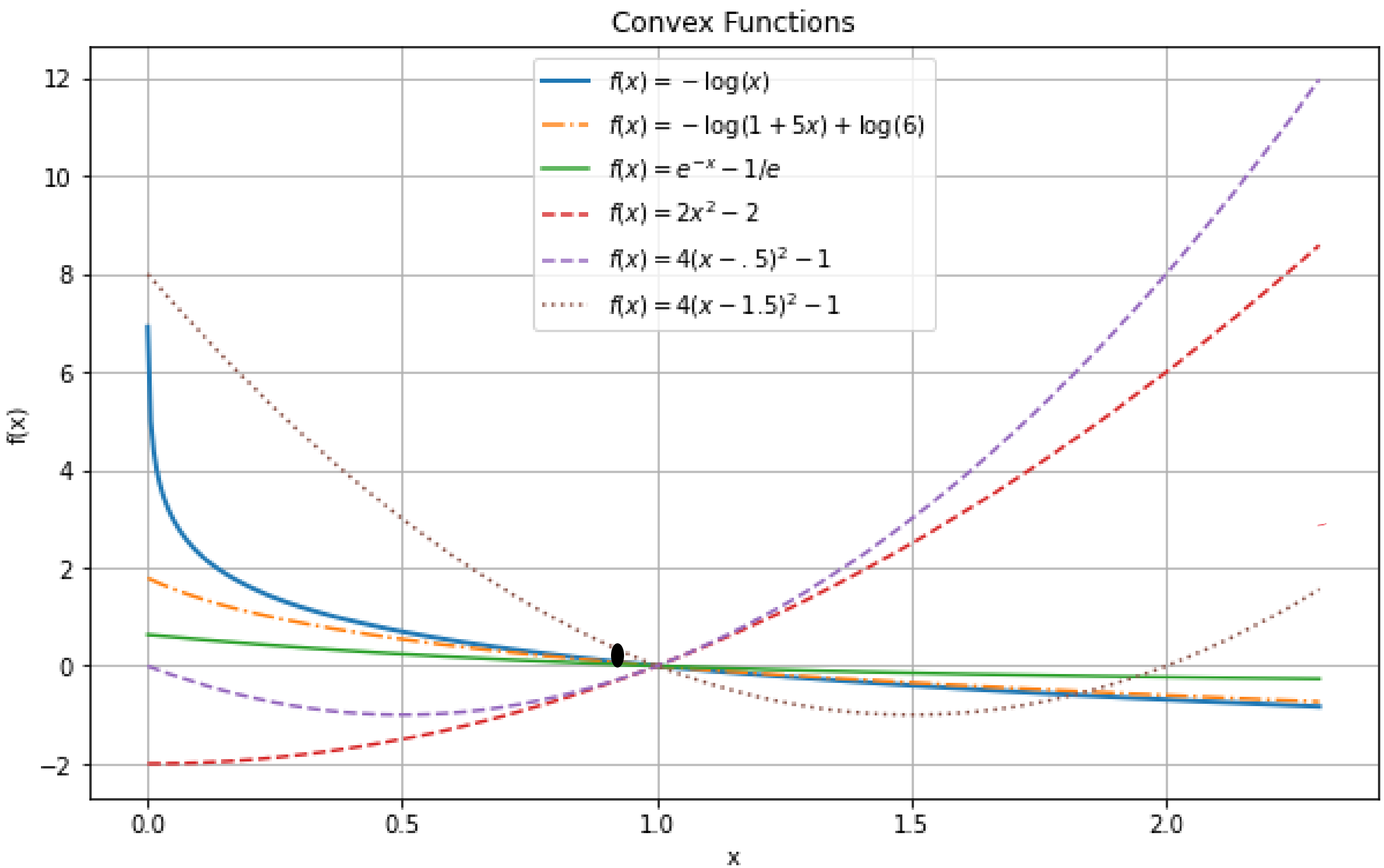
- For two 2-D Gaussian data:

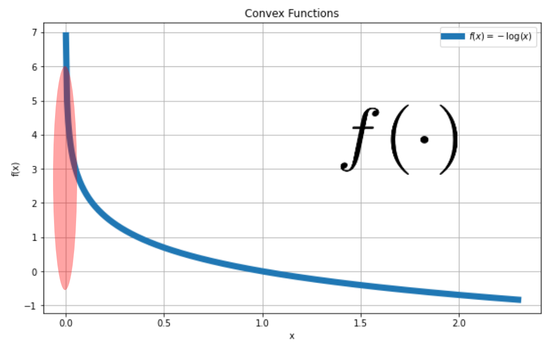


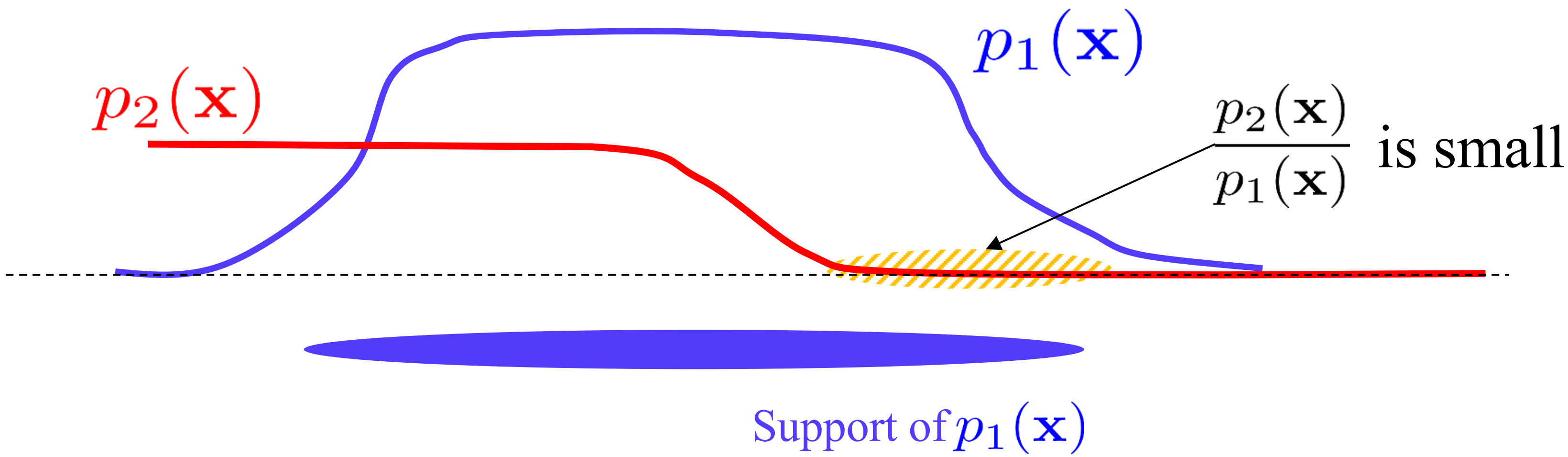
f-divergences

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

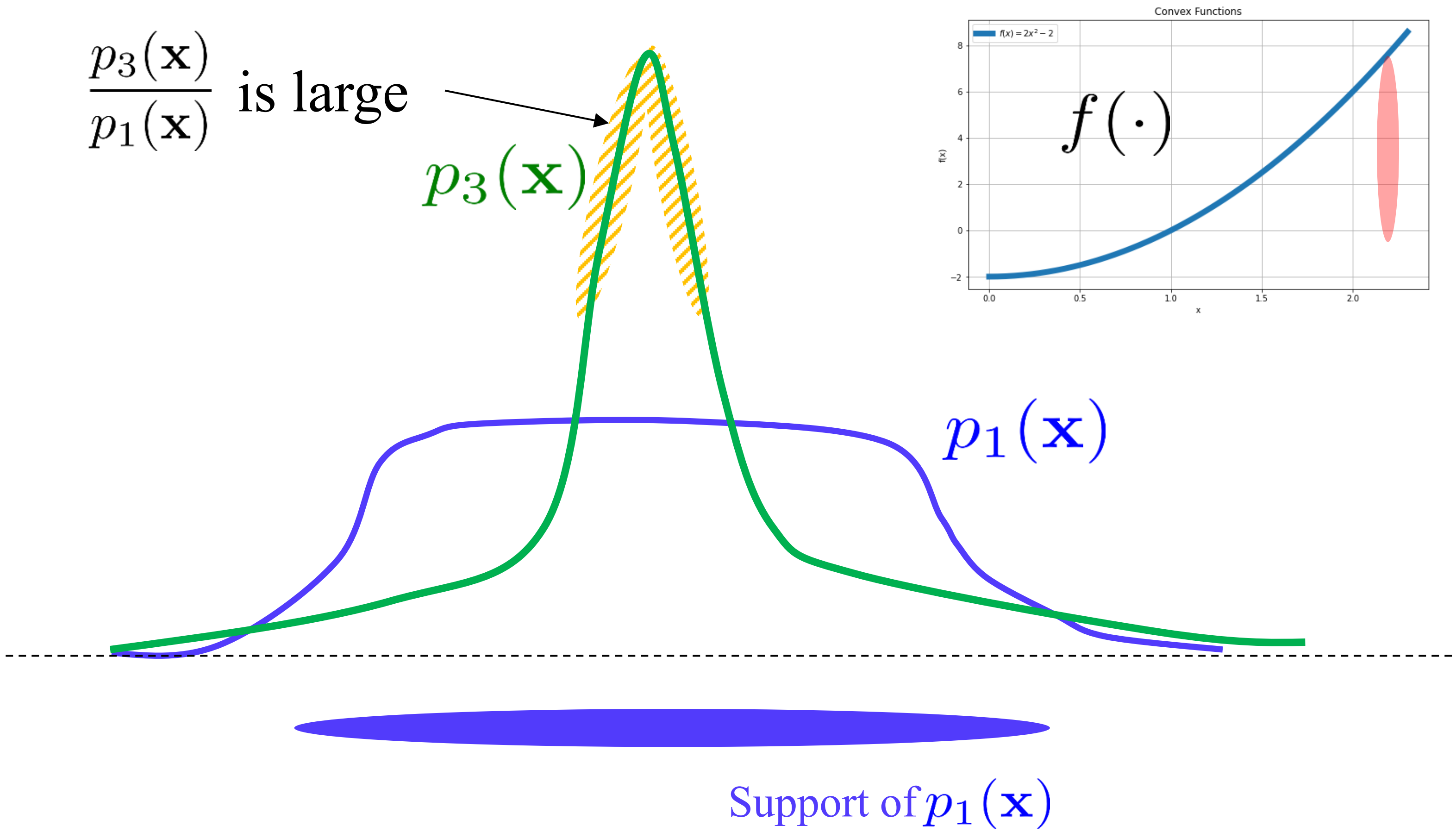
Candidates of *f*-functions





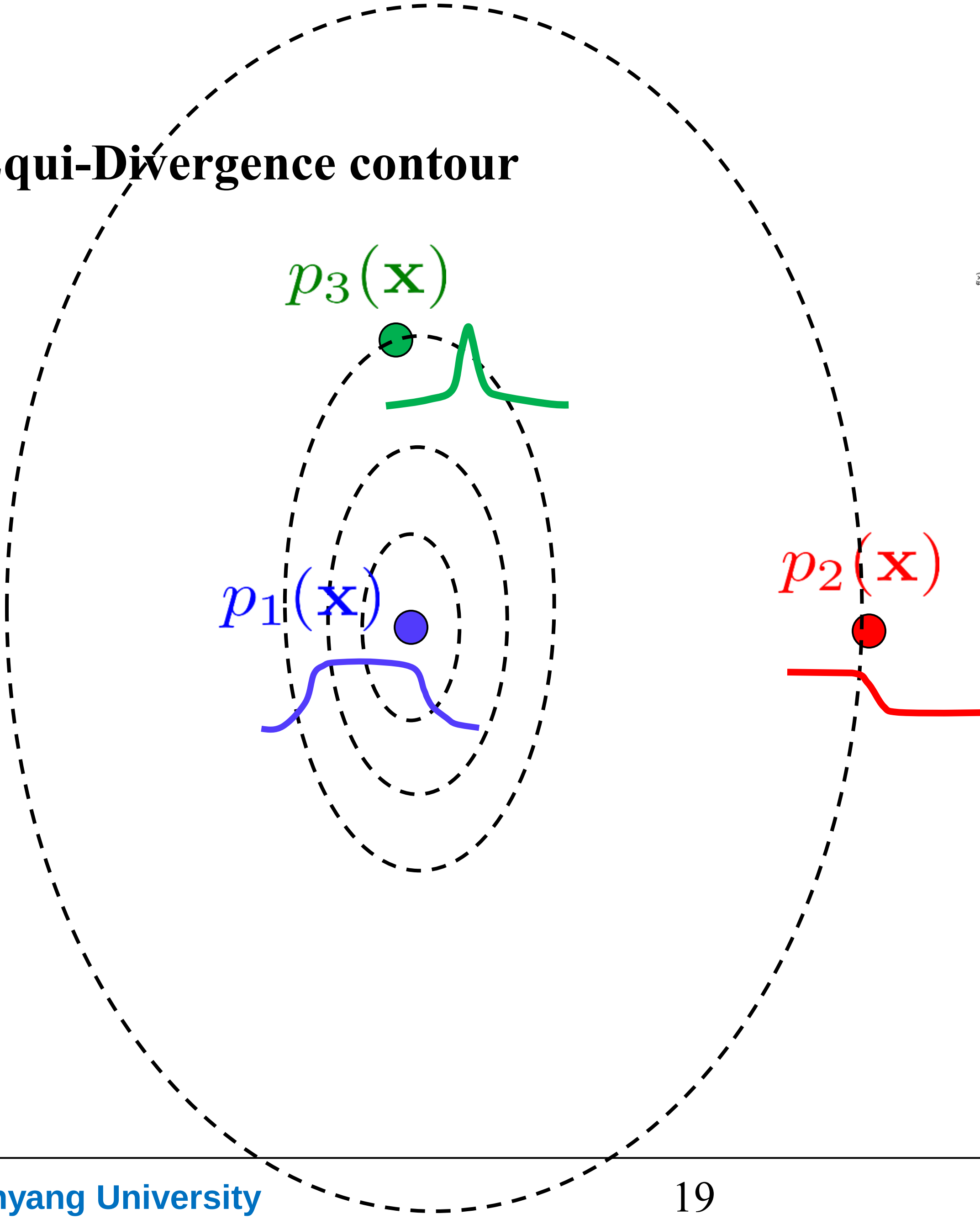


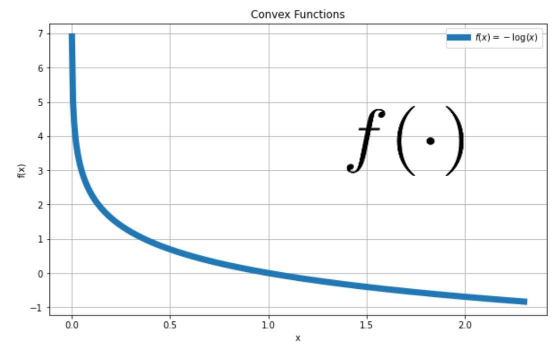
$$D_f(p_1(\mathbf{x}), p_2(\mathbf{x})) = \int p_1(\mathbf{x}) f\left(\frac{p_2(\mathbf{x})}{p_1(\mathbf{x})}\right) d\mathbf{x}$$



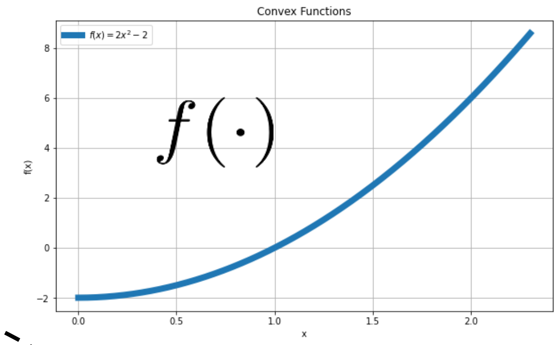
$$D_f(p_1(\mathbf{x}), p_3(\mathbf{x})) = \int p_1(\mathbf{x}) f\left(\frac{p_3(\mathbf{x})}{p_1(\mathbf{x})}\right) d\mathbf{x}$$

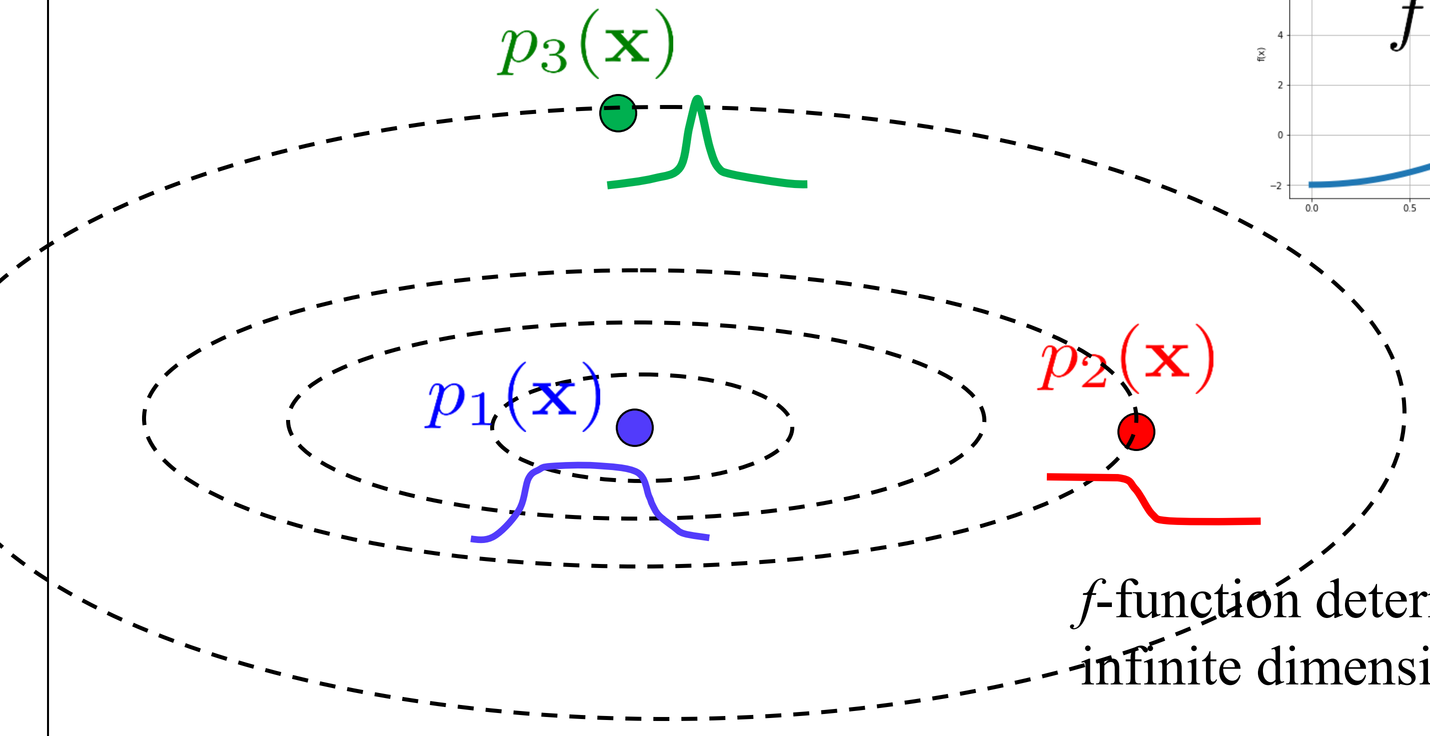
Equi-Divergence contour





Equi-Divergence contour





f-function determines the infinite dimensional metric

Similar to Loss, but Not the Same

- Invariant to the coordinate transformation once the dimensionality is conserved.

$$\mathbf{z} = T(\mathbf{x}) \ , \quad \mathbf{z}, \mathbf{x} \in \mathbb{R}^D$$

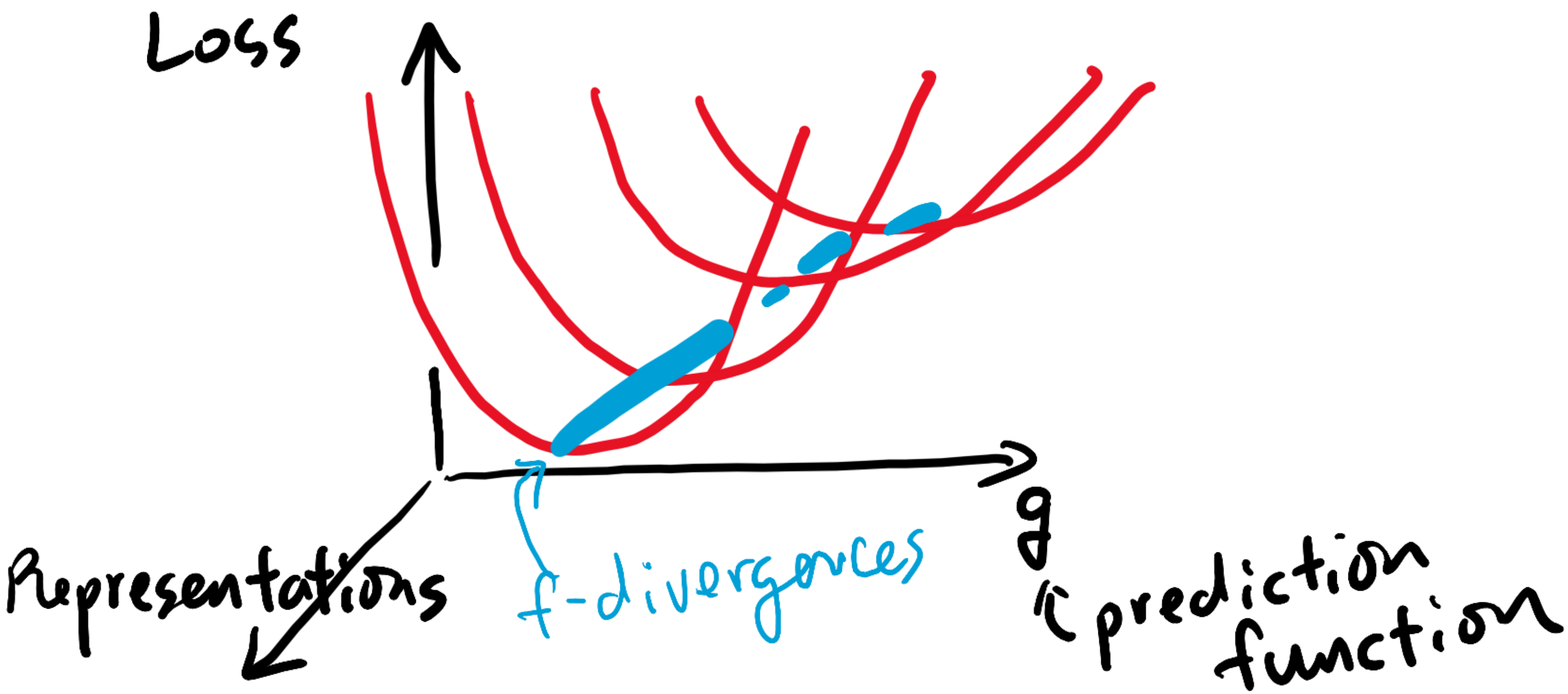
$$\int p_1(\mathbf{x}) f\left(\frac{p_2(\mathbf{x})}{p_1(\mathbf{x})}\right) d\mathbf{x} = \int p_1(\mathbf{z}) \cdot \cancel{j(\mathbf{x})} f\left(\frac{p_2(\mathbf{z}) \cdot \cancel{j(\mathbf{x})}}{p_1(\mathbf{z}) \cdot \cancel{j(\mathbf{x})}}\right) \frac{d\mathbf{z}}{\cancel{j(\mathbf{x})}}$$

$$= \int p_1(\mathbf{z}) f\left(\frac{p_2(\mathbf{z})}{p_1(\mathbf{z})}\right) d\mathbf{z} \qquad j(\mathbf{x}) = \left| \frac{dT}{d\mathbf{x}} \Big|_{\mathbf{x}} \right|$$

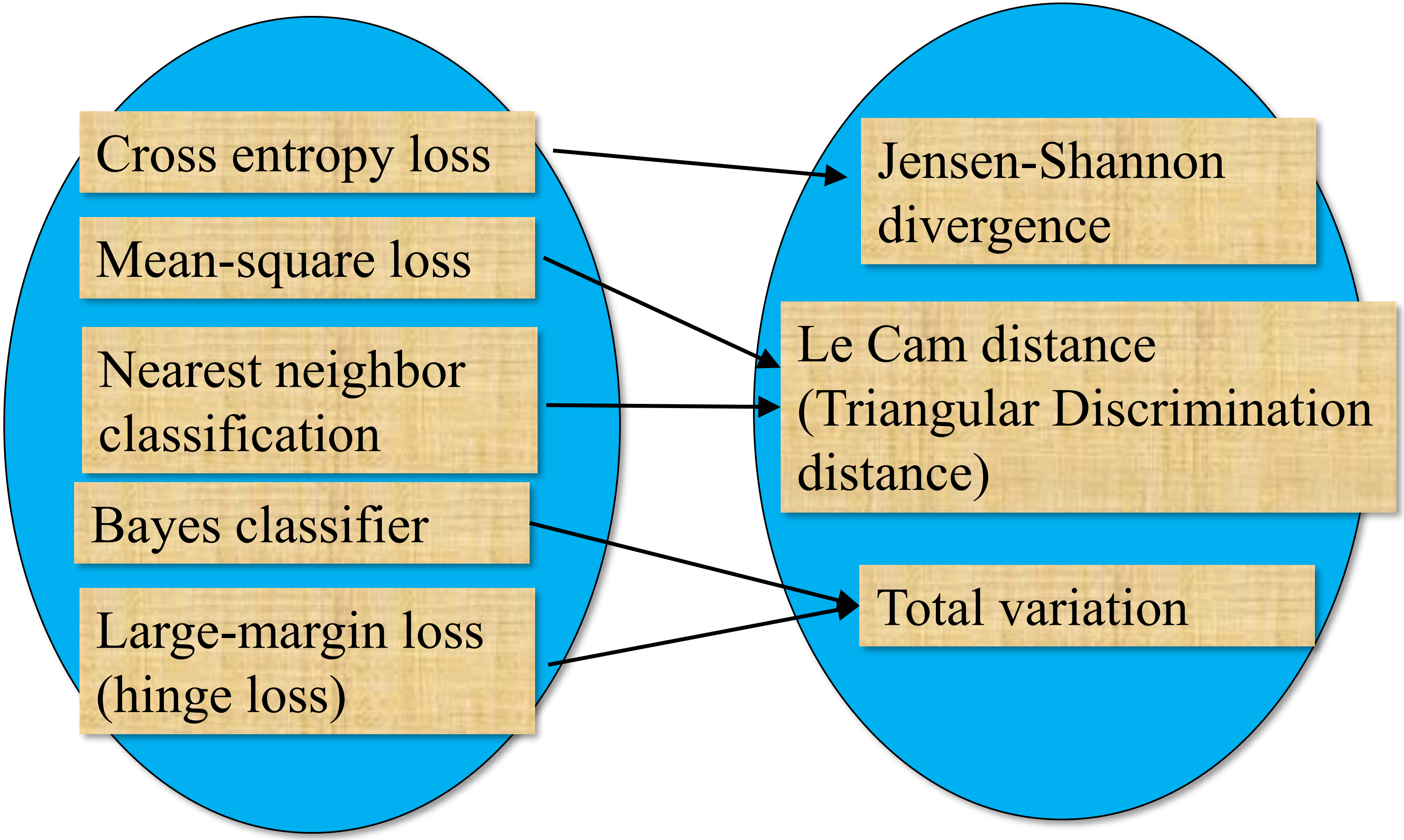
- When considering the separation property of densities after eliminating all properties obtained through coordinate transformation (← in contrast to Loss), it captures the information-based differences between underlying densities, independent of the coordinate choice.

Loss and *f*-divergences

- *f*-divergence: Set of minimum values obtained when the optimal prediction function is chosen



Loss functions – *f*-divergences



X.Nguyen, M.J. Wainwright, M.I. Jordan (2009) On surrogate loss functions and *f*-divergences, *Annals of Statistics*

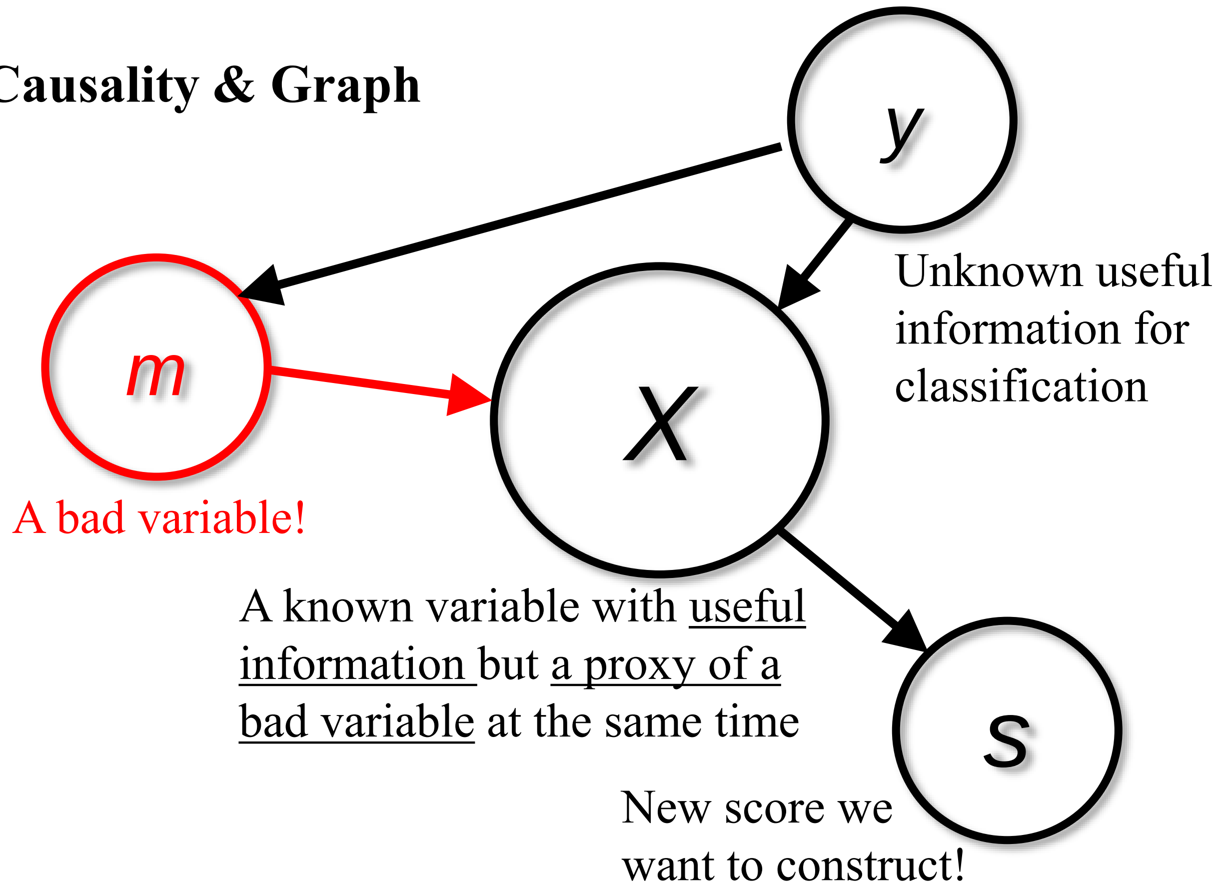
Blocking Information Flow

- 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 (seed variable) 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.

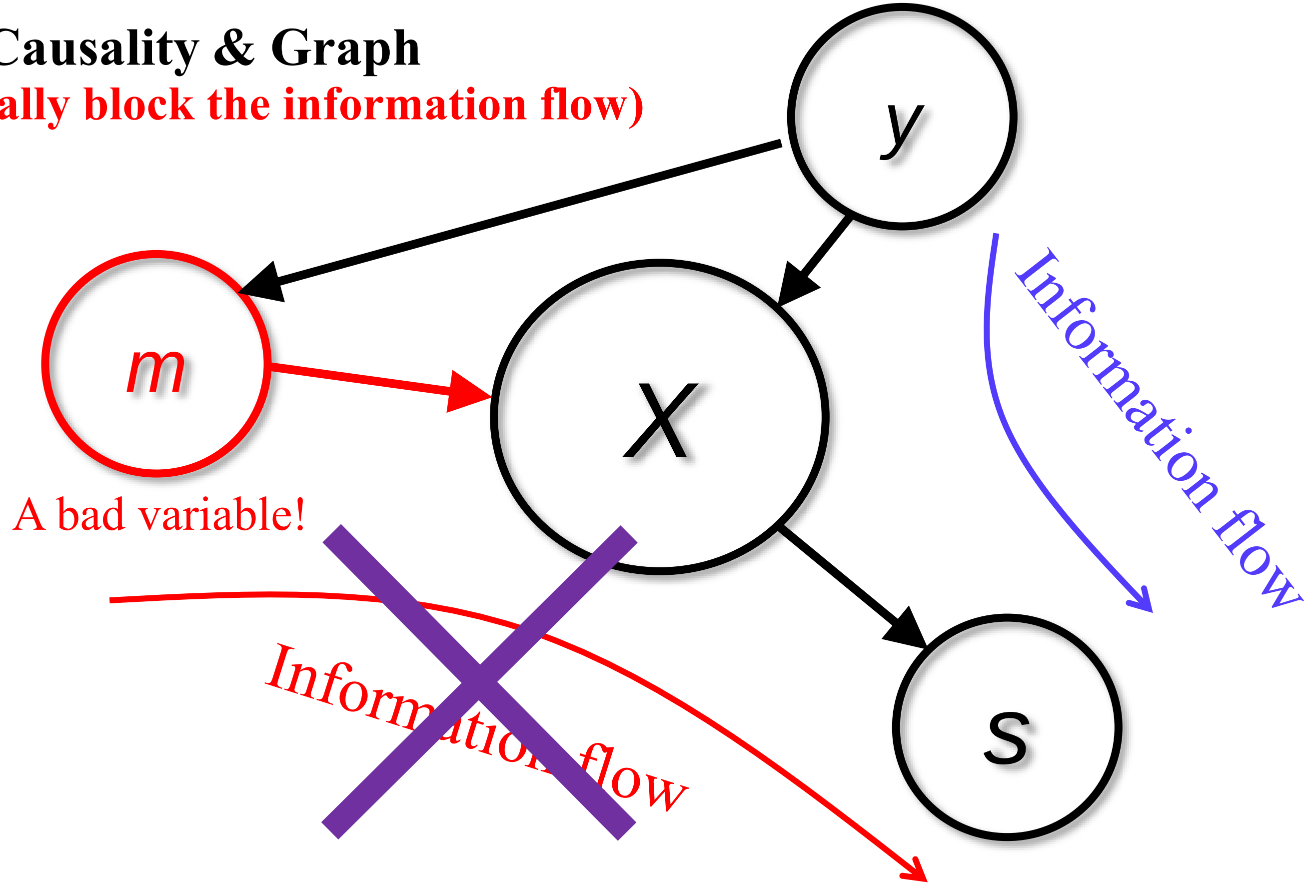
Causality & Graph



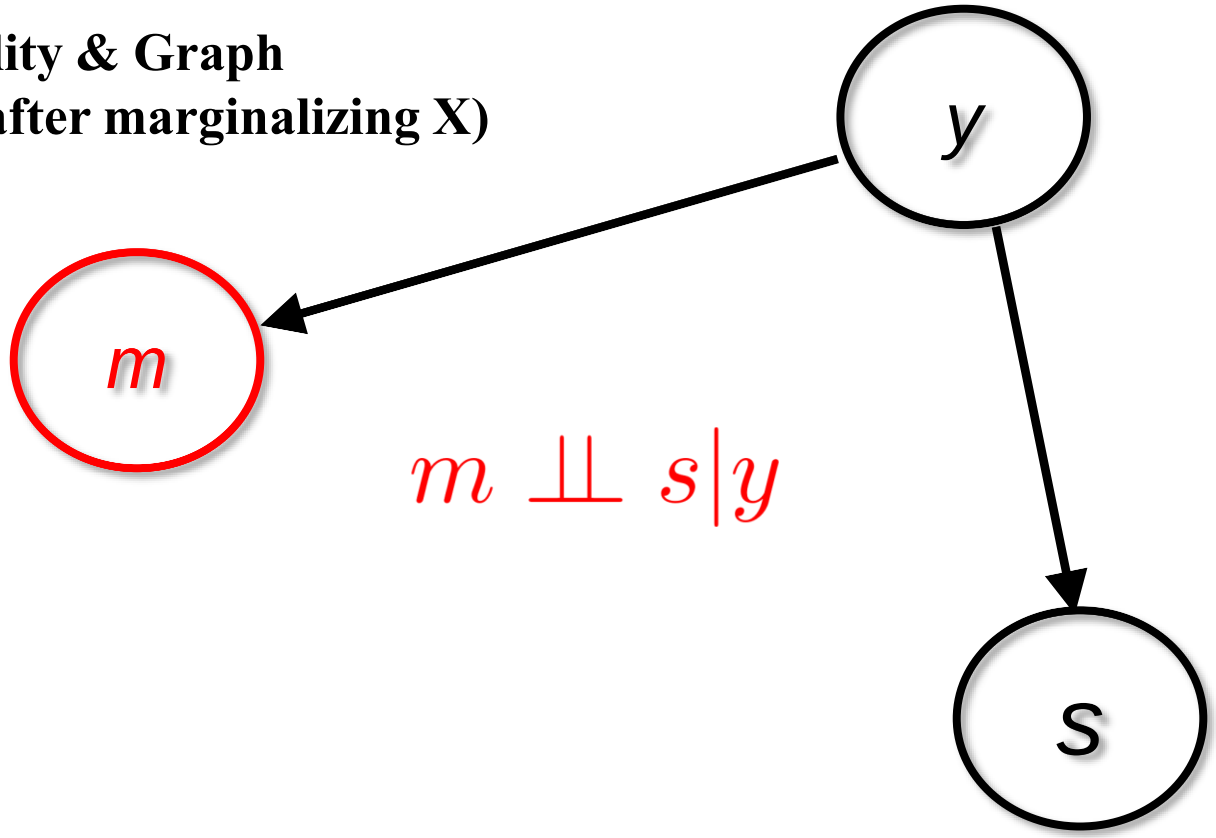
Estimation of Information Contents and Decorrelation

Causality & Graph

(Intentionally block the information flow)



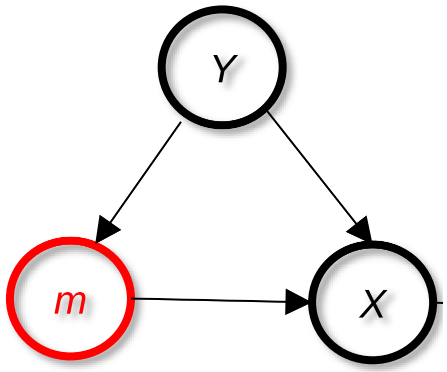
Causality & Graph
(Now after marginalizing X)



$$m \perp\!\!\!\perp s \mid y$$

Optimization for Training

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

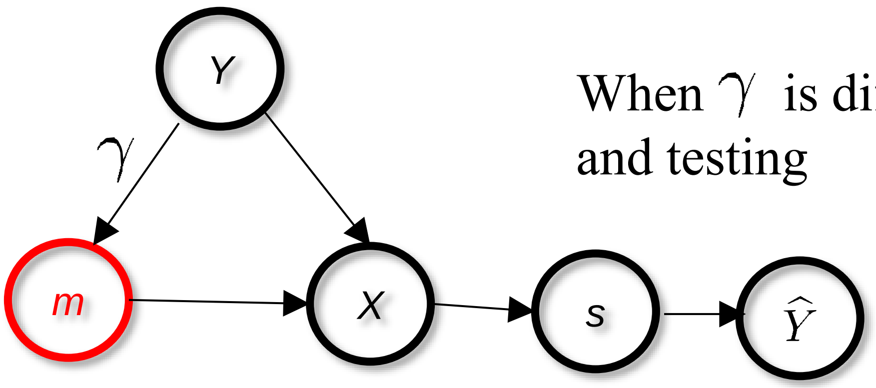


$$\theta^* = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} l(\widehat{y}(s(\mathbf{x}_i;\theta)), y_i) + \lambda \hat{I}_\theta(s; m|y)$$

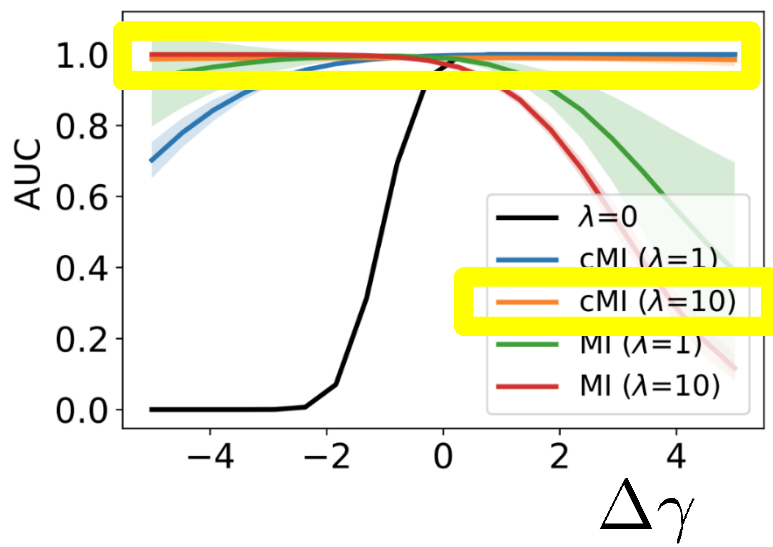
Minimize the expected loss $\mathbb{E}[l(\widehat{y}, y)]$

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

Use tradeoff constant λ



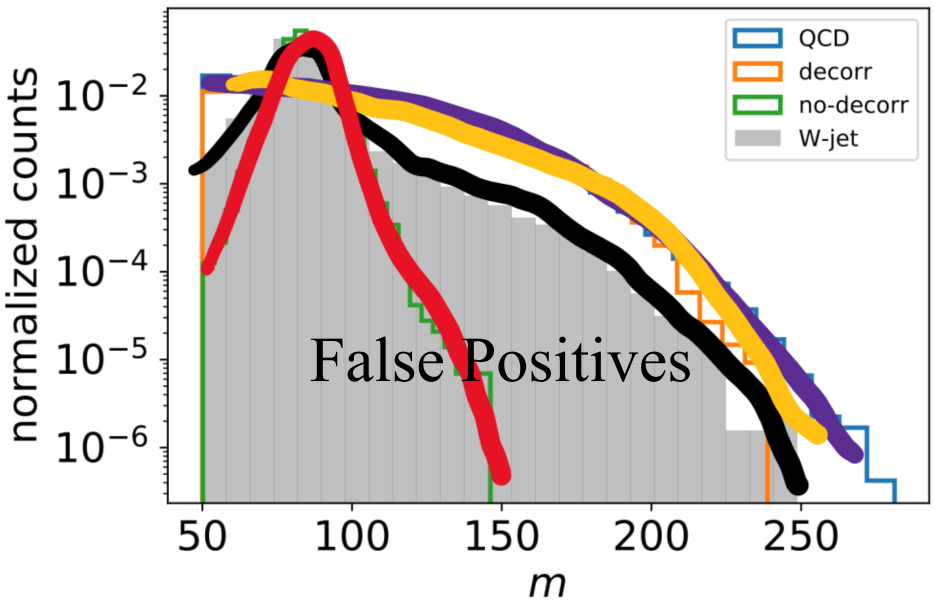
When γ is different in training and testing

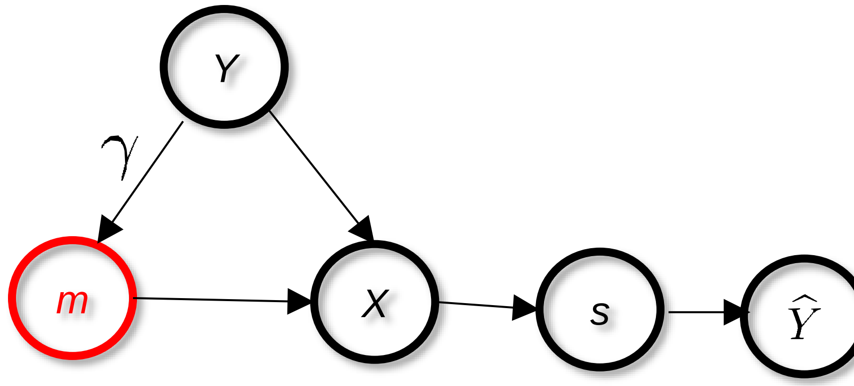


Decorrelation:

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

Reconstruction of W-jet Decorrelation Experiment

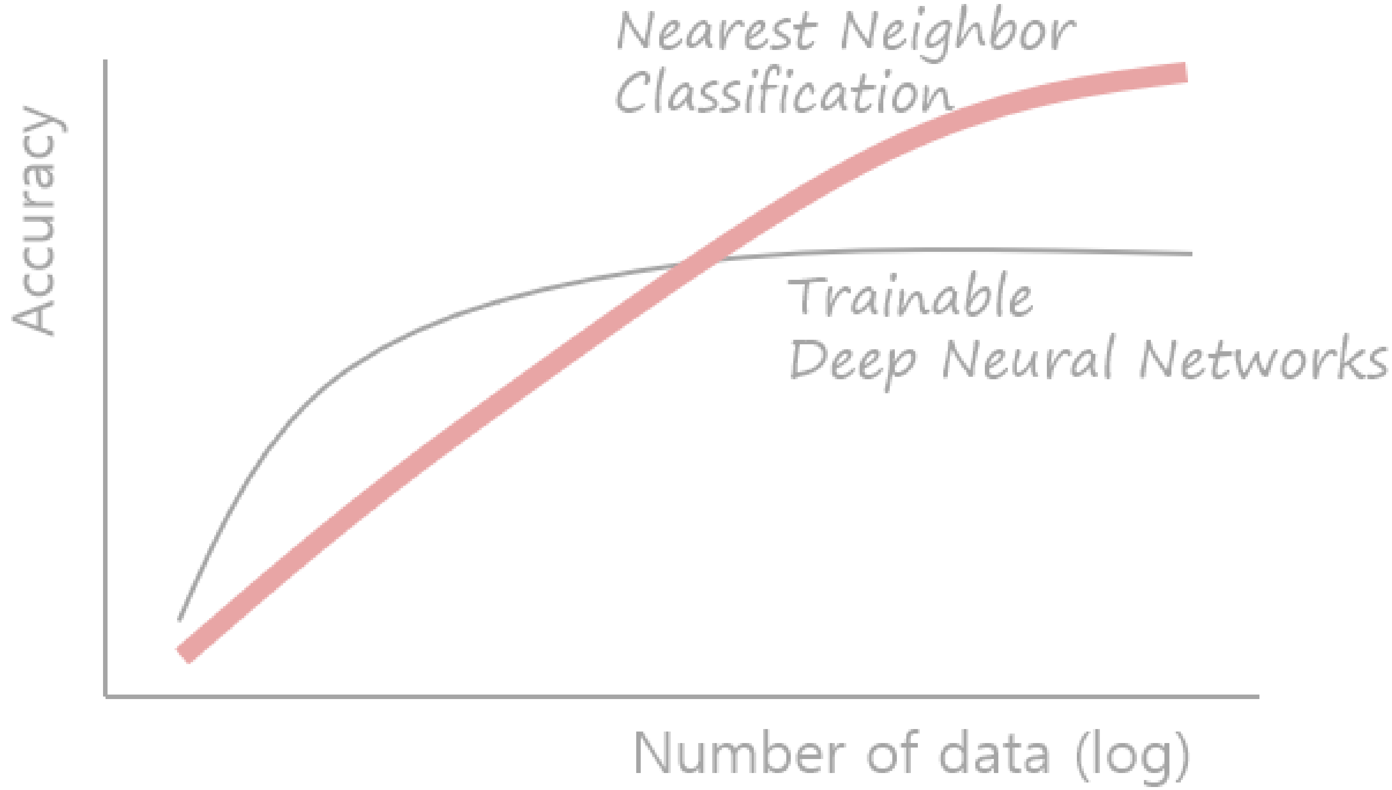




Reconstruction of decorrelation experiment in Kasieczka, G., Shih, D. (2020) Robust Jet Classifiers through Distance Correlation, *Phys. Rev. Lett. Vol. 125, Iss. 12 — 18*

Summary

- Estimating *f*-divergence using nearest neighbor information
- Finding loss-aware representations based on the intended loss function
- Blocking Information flow by purifying the variables

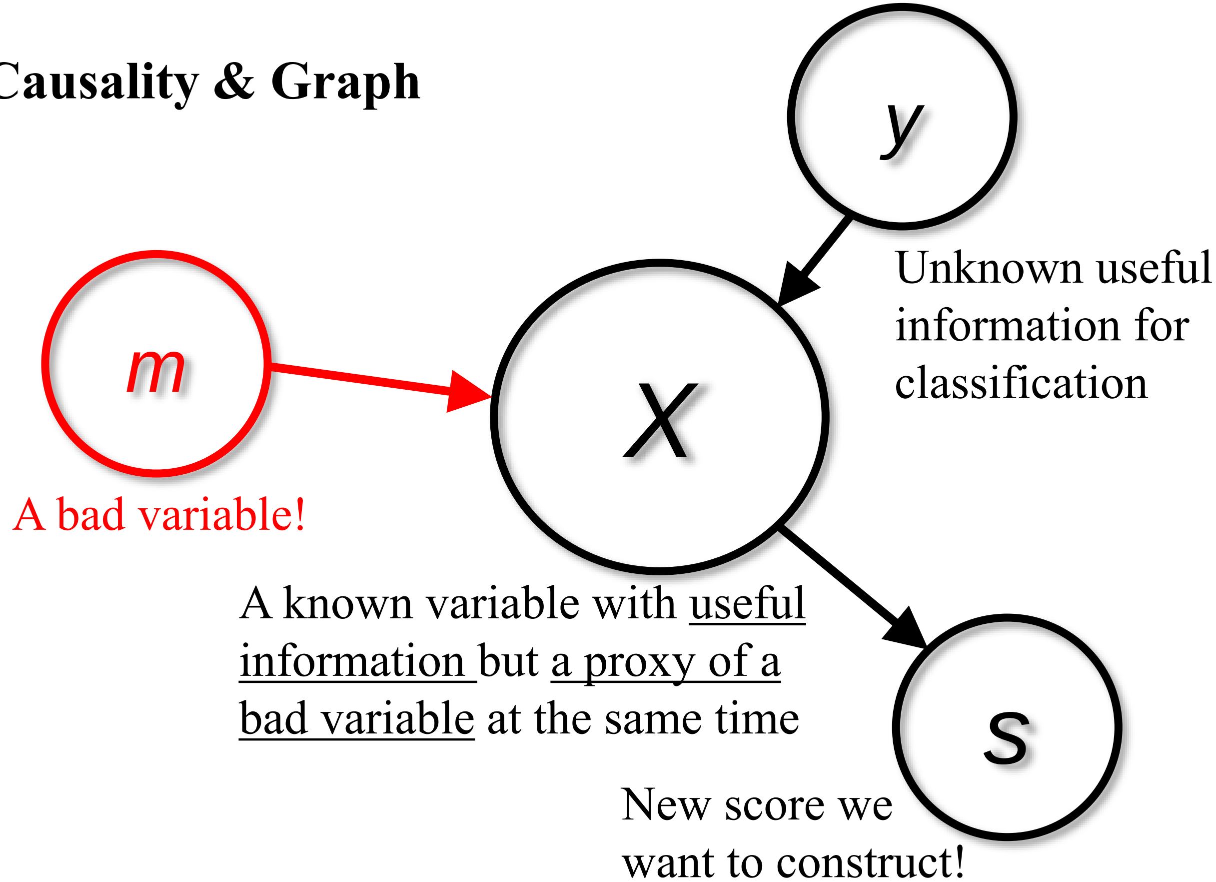


THANK YOU

Yung-Kyun Noh
nohyung@hanyang.ac.kr

Estimation of Information Contents and Decorrelation

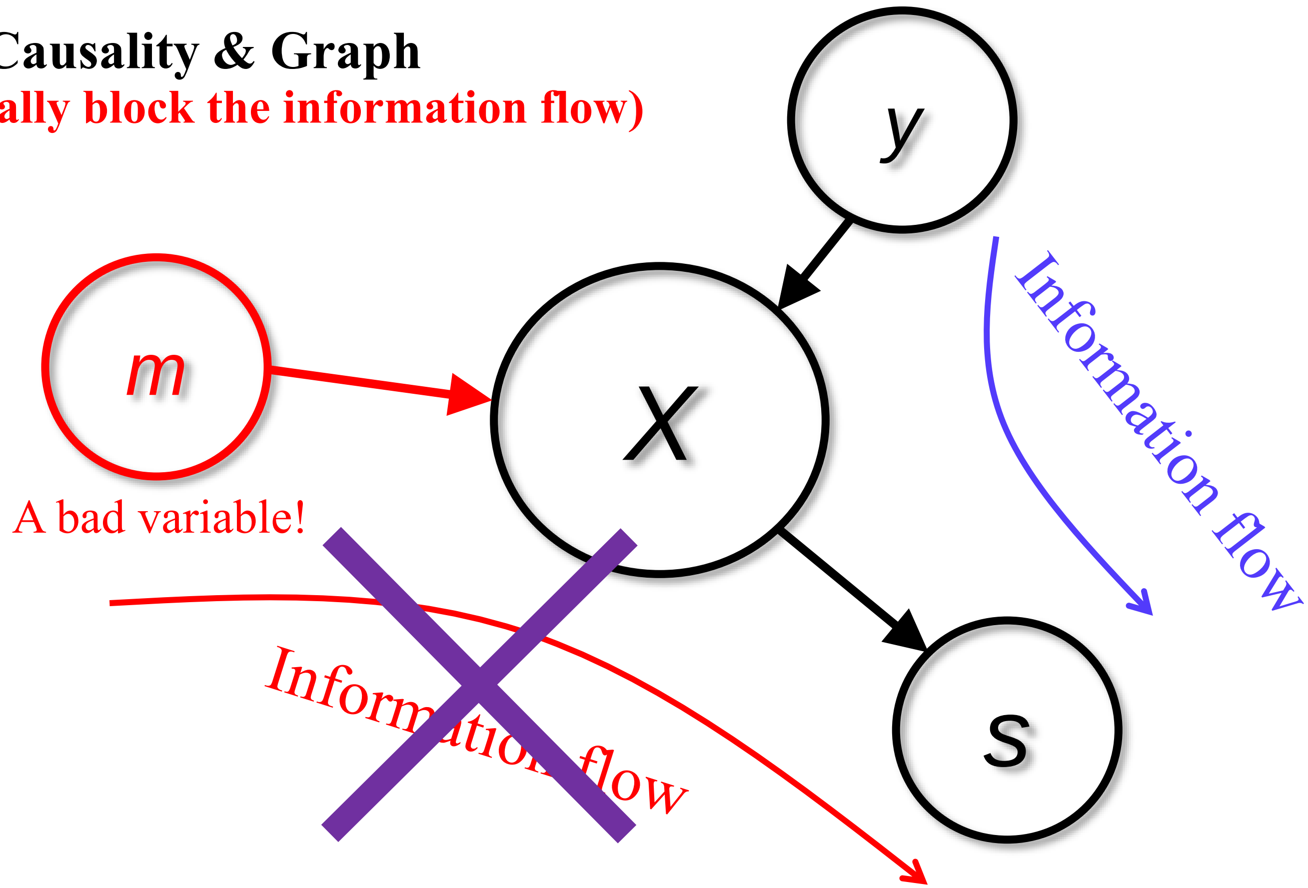
Causality & Graph



Estimation of Information Contents and Decorrelation

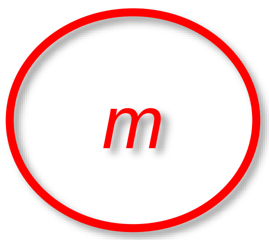
Causality & Graph

(Intentionally block the information flow)



Estimation of Information Contents and Decorrelation

Causality & Graph
(Now after marginalizing X)



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