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

XVIth Quark Confinement and the Hadron Spectrum Conference
Cairns Convention Centre, Cairns, Queensland, Australia
2024. 8. 22 (Thu.)

Yung-Kyun Noh

Hanyang University & Korea Institute for Advanced Study

Joint work with Dr. Cheongjae Jang

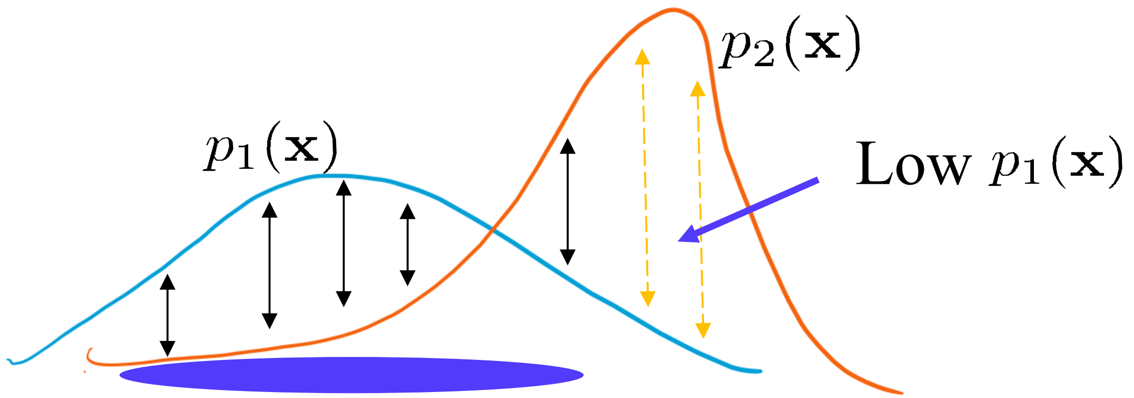
f-divergences

f-divergences

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

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

f-divergences



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

f-divergences

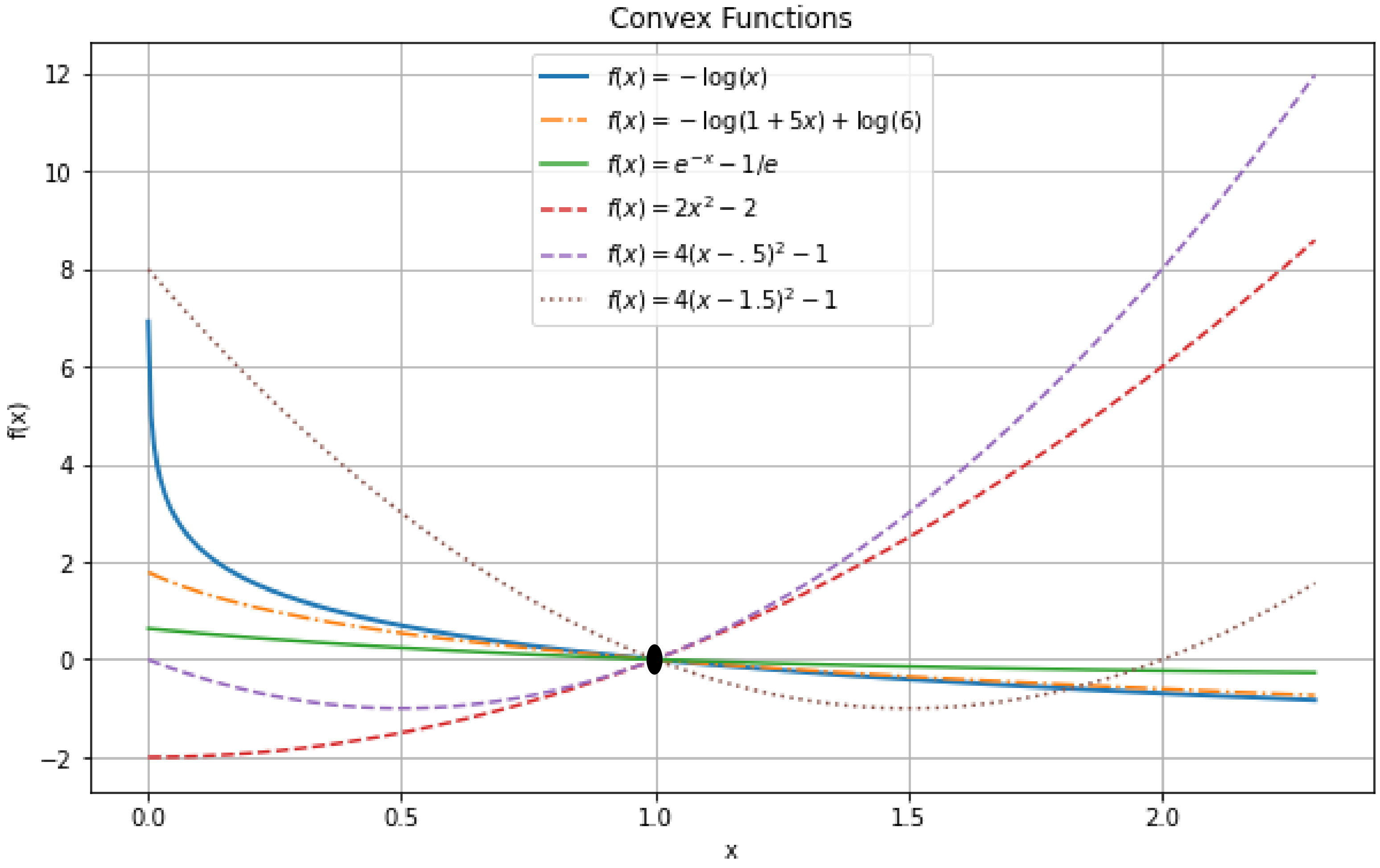
$f(.)$: convex

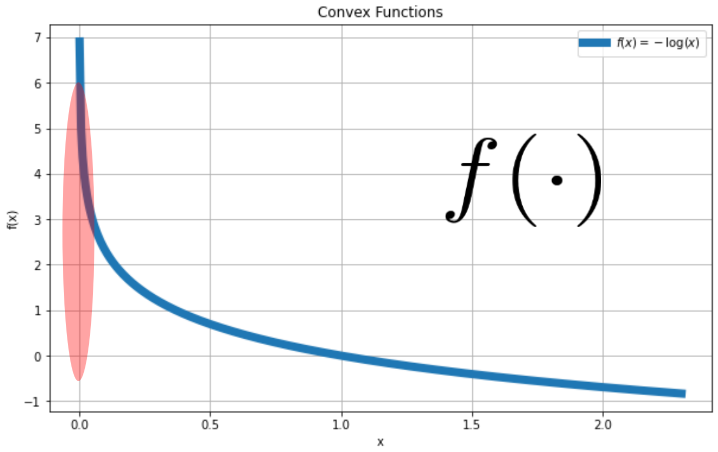
& $f(1) = 0$

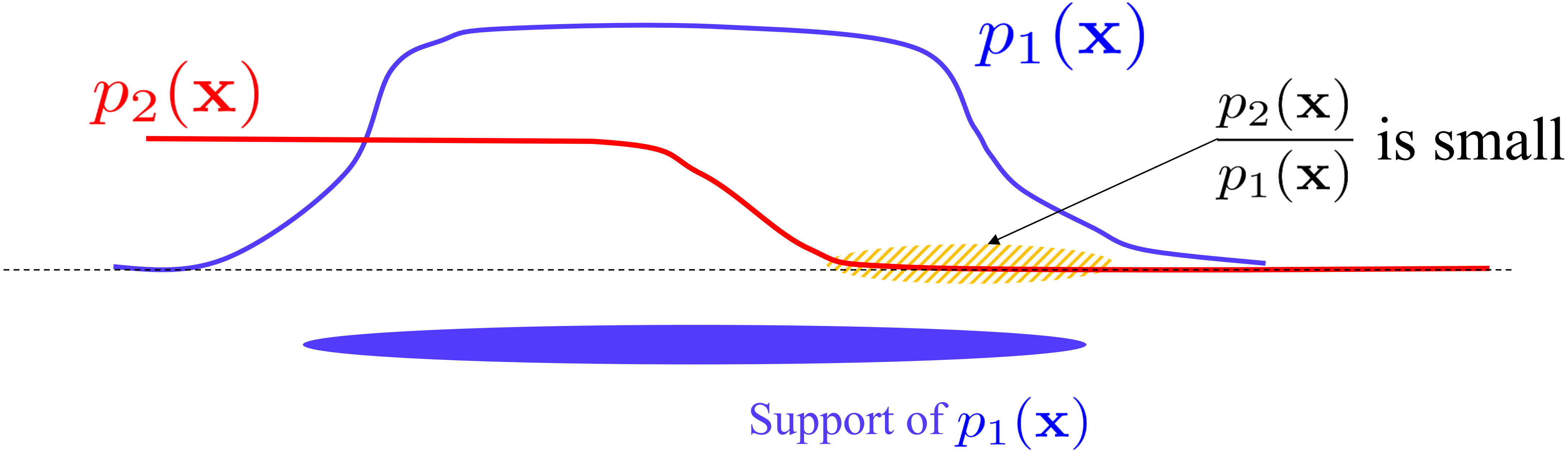
⟺

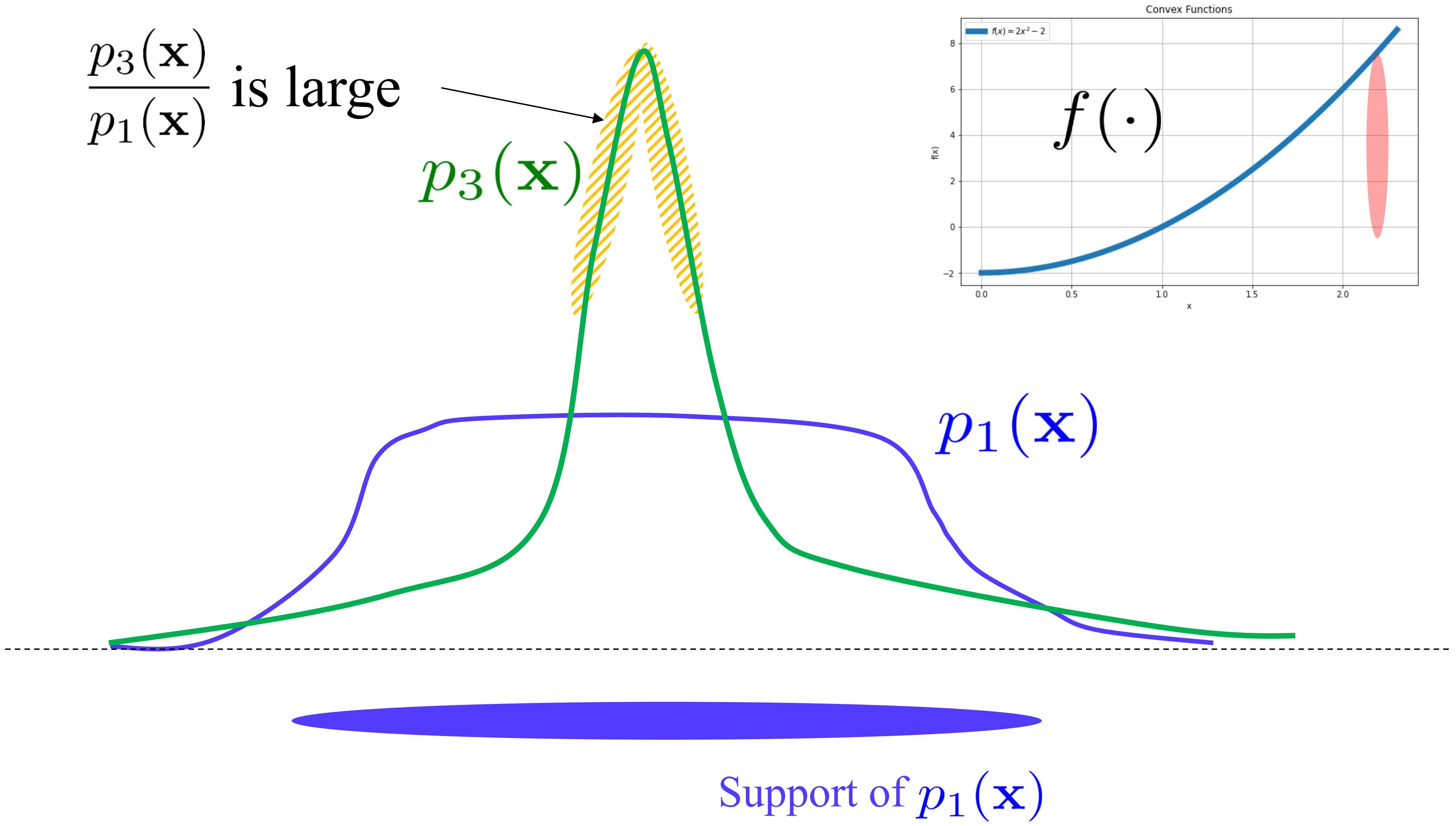
$D_f(p_1(\mathbf{x}), p_2(\mathbf{x})) = 0$
and is minimized when
$p_1(\mathbf{x}) = p_2(\mathbf{x})$ for all $\mathbf{x}$

Candidates of *f*-functions



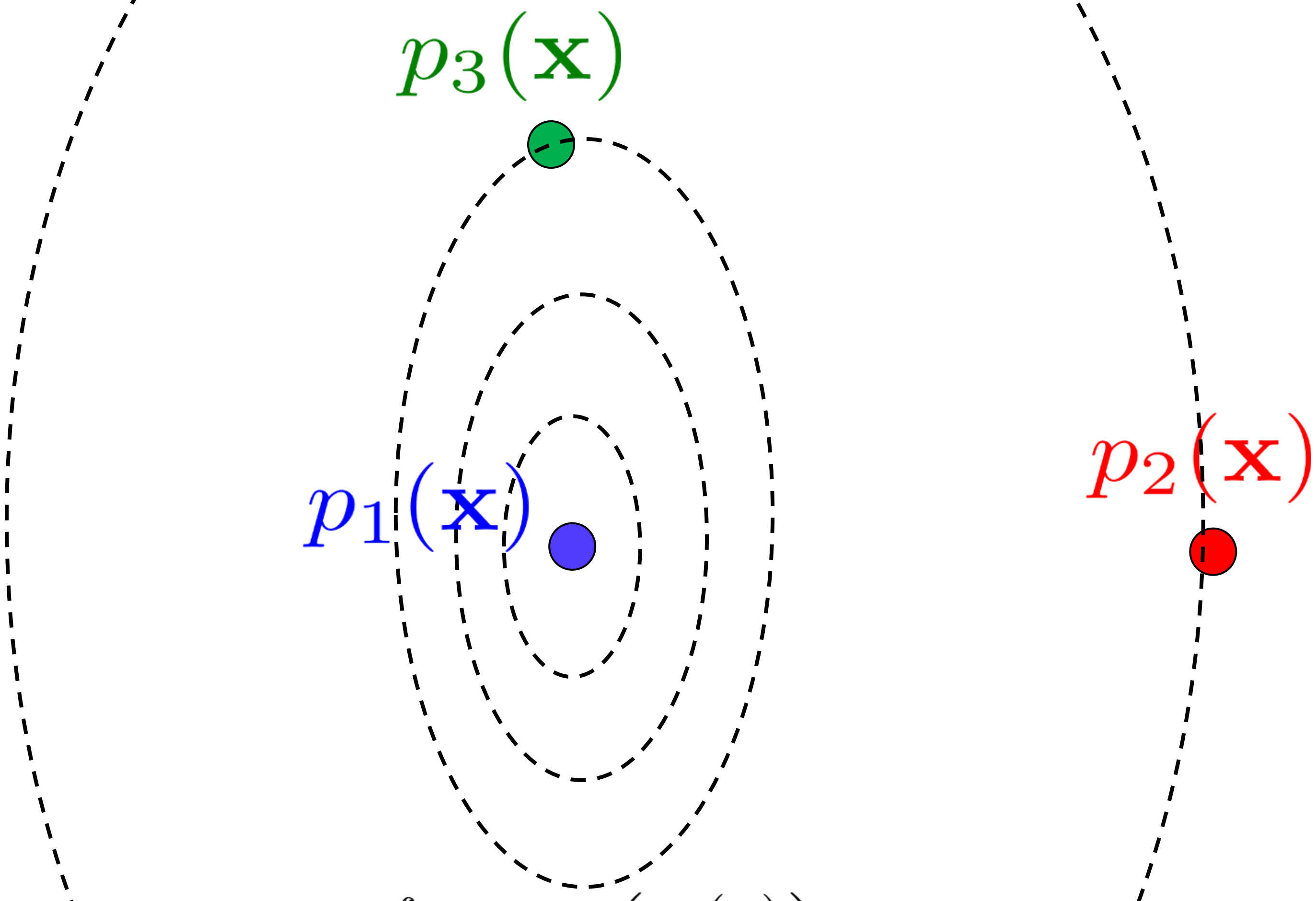

Convex Functions
f(x) = −log(x)
f(x)
x
7
6
5
4
3
2
1
0
−1
0.0
0.5
1.0
1.5
2.0
$f(\cdot)$

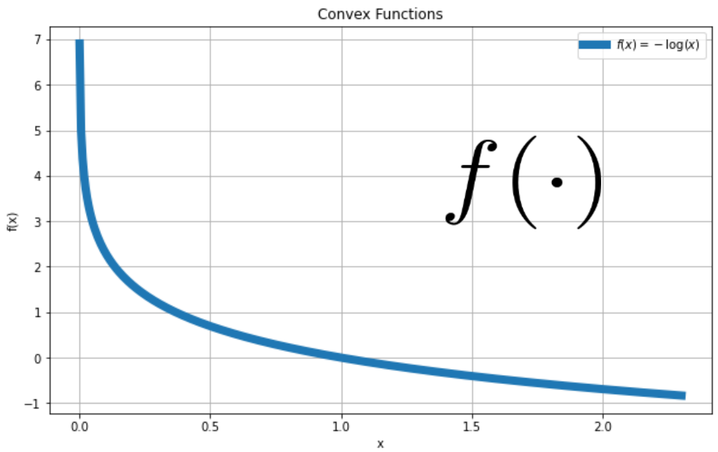

$p_1(\mathbf{x})$
$p_2(\mathbf{x})$
$\frac{p_2(\mathbf{x})}{p_1(\mathbf{x})}$ is small
Support of $p_1(\mathbf{x})$



$\frac{p_3(\mathbf{x})}{p_1(\mathbf{x})}$ is large
$p_3(\mathbf{x})$
$p_1(\mathbf{x})$
Support of $p_1(\mathbf{x})$
Convex Functions
$f(x) = 2x^2 - 2$
$f(\cdot)$
f(x)
x

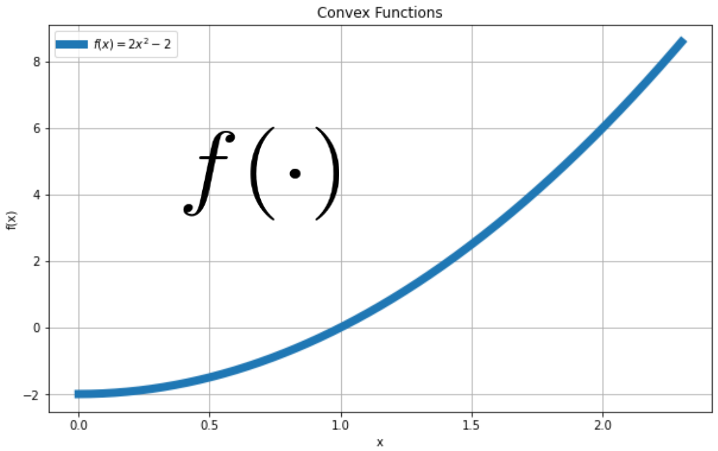
Equi-Divergence contour

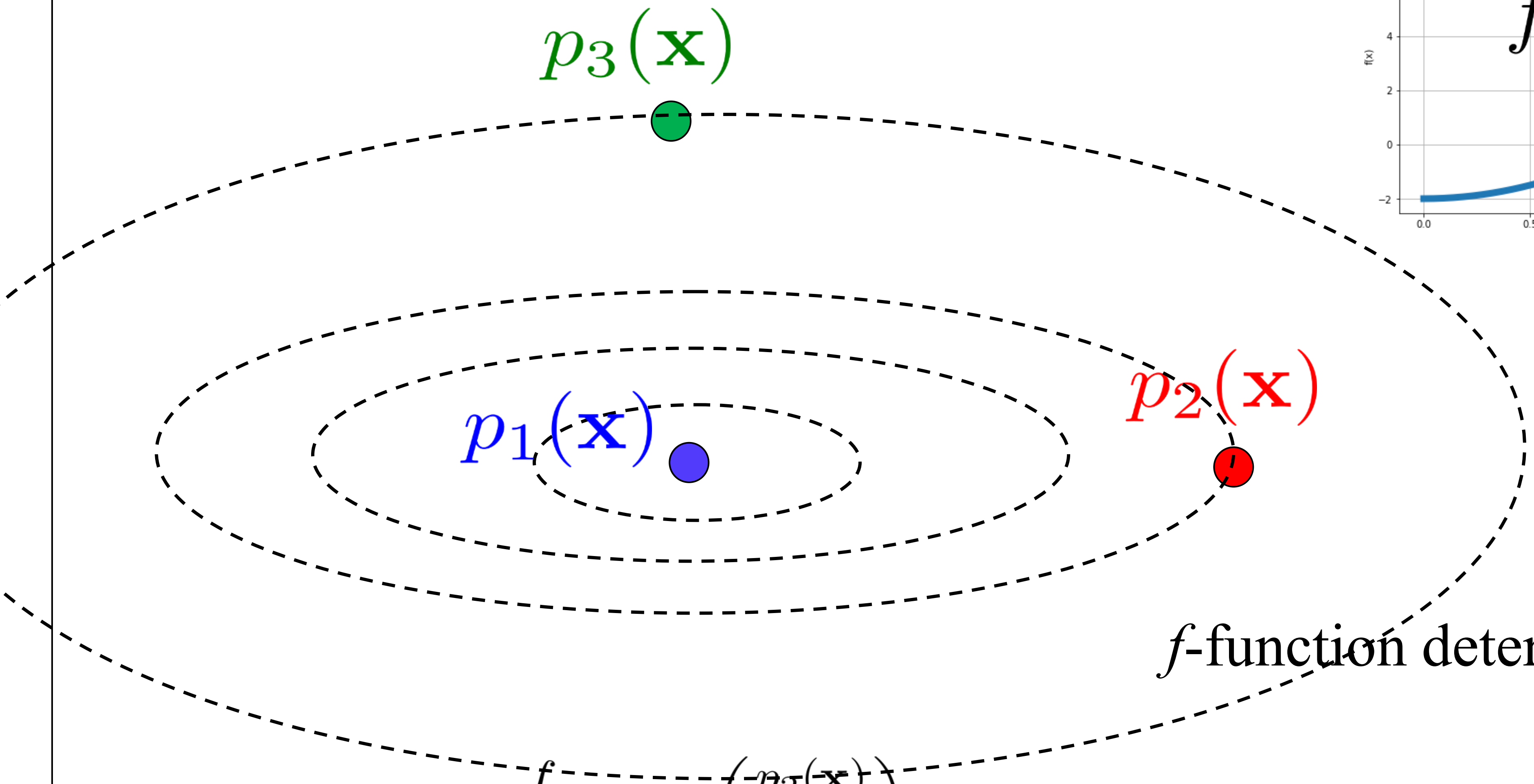




$$D_f(p_1(\mathbf{x}), p_i(\mathbf{x})) = \int p_1(\mathbf{x}) f\left(\frac{p_2(\mathbf{x})}{p_1(\mathbf{x})}\right) d\mathbf{x}$$

Equi-Divergence contour





f-function determines the metric

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

Our Research with *f*-divergences

- Construction of nonparametric estimators (using nearest neighbors)
  - [Ryu et al. (2022) *IEEE TIT*]

- Addressing finite sampling bias of estimation
  - [Noh et al. (2010) *NeurIPS*, Noh et al. (2017) *NeurIPS*, Noh et al. (2018) *IEEE TPAMI*, Noh et al. (2018) *Neural Computation*, Yoon et al. (2023) *NeurIPS*]

- Application to eliminating harmful variables – ongoing work

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*

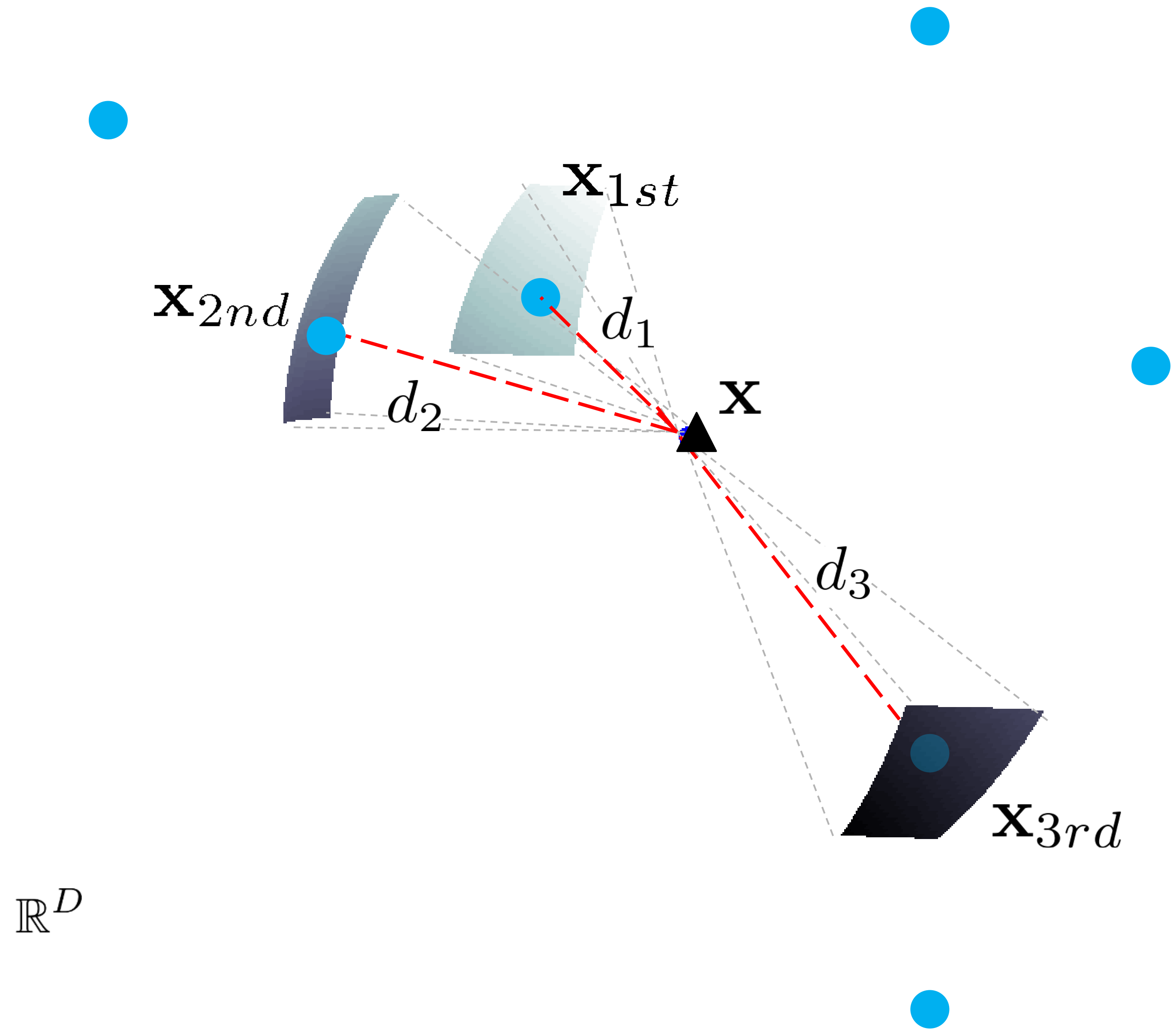
***Abstract*—A new approach to L_2-consistent estimation of a general density functional using k-nearest neighbor distances is proposed, where the functional under consideration is in the form of the expectation of some function f of the densities at each point. The estimator is designed to be asymptotically 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.**

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,

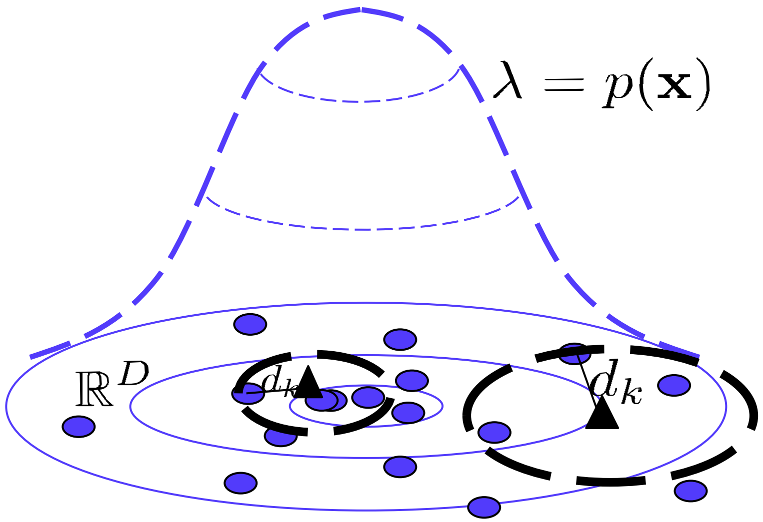
$$\lim_{m\to\infty} \mathbb{E}\big[\big(\hat{T}_f^m(\mathbf{X}_{1:m}) - T_f(p)\big)^2\big] = 0.$$

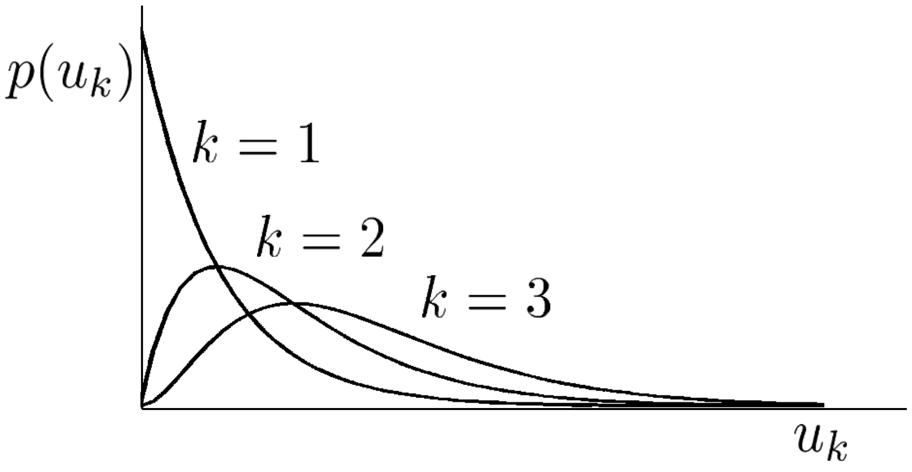
More generally, let $f\colon \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ and consider a divergence functional

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


$\mathbf{x}_{1st}$
$\mathbf{x}_{2nd}$
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*

$N \to \infty,$

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

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

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

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

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

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 Laplace Transforms

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

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

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

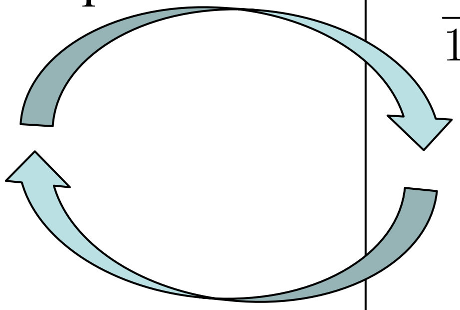
$$\mathbb{E}_{u_1,u_2}\phi(u_1, u_2) = -\log \frac{p_2}{p_1}$$

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

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

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

$D_f(p_1(\mathbf{x}), p_2(\mathbf{x}))$	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	$-\log t$
$1 - \int \sqrt{p_1 p_2}\, d\mathbf{x}$	$1 - \frac{1}{\Gamma(1.5)\Gamma(2.5)}\sqrt{\frac{v_1^{(2)}}{u_2^{(2)}}}$	$1 - \sqrt{t}$
$1 - \int \frac{p_1 p_2}{p_1 + p_2}\, d\mathbf{x}$	$\mathbb{I}(u_1^{(1)} < u_2^{(1)})$	$\frac{1}{1+t}$
⋮	⋮	⋮



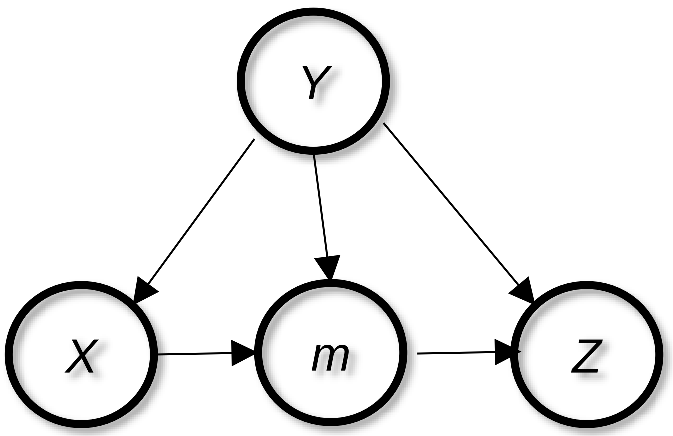
Extension to Engaging Problems

- We use information-theory to construct complex, non-trivial problems.

- For example, by using estimators of the Kullback-Leibler (KL) divergence, we can formulate information-theoretic objective functions, such as conditional mutual information:

$$\begin{aligned} I(r;m|y) &= I(r,y;m) - I(y;m) \\ &= KL(p(r,m,y)||p(r,y)p(m)) - KL(p(m,y)||p(m)p(y)) \end{aligned}$$

Application - We know *m* should not be a relevant feature

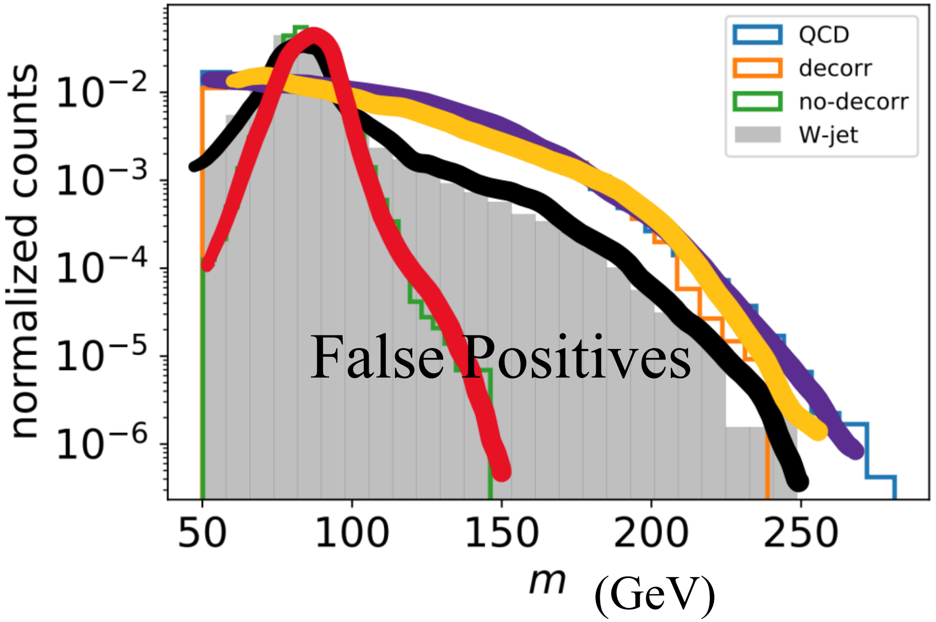


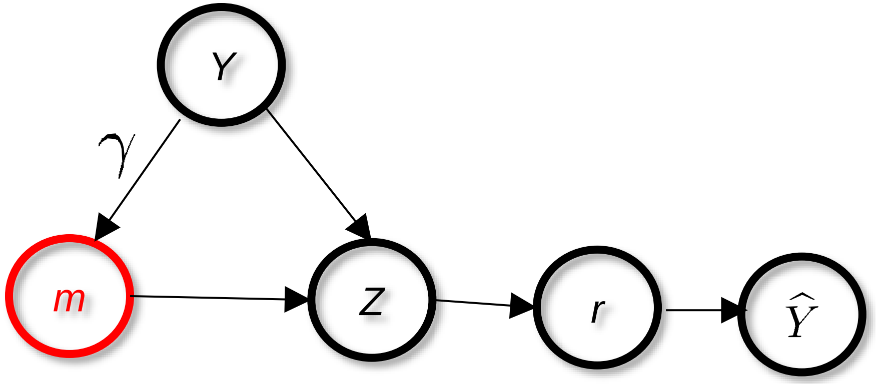
Y: target to predict
X: feature variables used to predict Y
m: Contains information about Y, but a variable that should "not" be used. (harmful variable)
Z: feature variables containing its own information about Y, but corrupted by m

We do not want to include m into the set of relevant features for various reasons:

- moral reasons
- prohibited by law
- Real data will not have the effect from m.
- We want to eliminate the effect of one variable (e.g. medicine)

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*

Experimented by Do-Hyun Song

Summary

- *f*-divergence is a fundamental information-theoretic measure that can be used for making many interesting machine learning problems.
- Since *f*-divergence is defined with underlying densities, we need estimators.
- The flow of harmful information can be blocked using these estimators to construct reliable machine learning models.

Yung-Kyun Noh
nohyung@hanyang.ac.kr