Cauchy-Schwarz Divergence Information Bottleneck

Shujian Yu

Vrije Universiteit Amsterdam

Outline

- Information Bottleneck
- Deep Information Bottleneck
  - Cauchy-Schwarz divergence Information Bottleneck

Information Bottleneck

Information Bottleneck

- Let T be a representation of X
  - Which T is useful?
    - Disentangled
    - Interpretable

Information Bottleneck

- Let T be a representation of X
  - Which T is useful?
    - Disentangled
    - Interpretable

“cat” laying on a “laptop”

X T

Federici, Marco, et al. "Learning Robust Representations via Multi-View Information Bottleneck." ICLR, 2020.

Information Bottleneck

- Let T be a representation of X
  - Which T is useful?
    - Disentangled
    - Interpretable

"cat" laying on a "laptop"

X T

  - Tasks
    - Is there a cat? relevant: "cat"; irrelevant: "laptop"
    - How many pixels are there in the image? irrelevant: "cat" and "laptop"

Information Bottleneck

- Let T be a representation of X
  - Which T is useful?
    - Disentangled
    - Interpretable
    - Related to task Y → Useful for predicting Y

"cat" laying on a "laptop"

X T

  - Tasks
    - Is there a cat? relevant: "cat"; irrelevant: "laptop"
    - How many pixels are there in the image? irrelevant: "cat" and "laptop"

Information Bottleneck

- Let T be a representation of X
  - Which T is useful?
    - Disentangled
    - Interpretable
    - Related to task $Y \rightarrow$ Useful for predicting Y

  - How to define the optimal representation T?

Sufficient Statistics $S(X)$

$$\boxed{I(S(X);Y)} = \boxed{I(X;Y)}$$

A representation T of X is sufficient for Y if and only if $I(X;Y) = I(T;Y)$;
T contains all information regarding Y that can be obtained also from X

Tishby, Naftali, Fernando C. Pereira, and William Bialek. "The information bottleneck method." *37th Allerton Conference on Communication and Computation, 2000.*

Information Bottleneck

- Let T be a representation of X
  - Which T is useful?
    - Disentangled
    - Interpretable
    - Related to task Y → Useful for predicting Y

  - How to define the optimal representation T?

Sufficient Statistics $S(X)$

$$\boxed{I(S(X);Y)} = \boxed{I(X;Y)}$$

Minimal Sufficient Statistics $T(X)$

$$T(X) = \arg \min_{S(X):\boxed{I(S(X);Y)}=\boxed{I(X;Y)}} \boxed{I(S(X);X)}$$

T contains only relevant information regarding Y, but least information from X.

Information Bottleneck

- Let T be a representation of X
  - Which T is useful?
    - Disentangled
    - Interpretable
    - Related to task Y → Useful for predicting Y
  - How to define the optimal representation T?

Sufficient Statistics $S(X)$

$$\boxed{I(S(X);Y)} = \boxed{I(X;Y)}$$

Minimal Sufficient Statistics $T(X)$

$$T(X) = \arg \min_{S(X):\boxed{I(S(X);Y)}=\boxed{I(X;Y)}} \boxed{I(S(X);X)}$$

Sufficiency vs Minimality

Gilad-Bachrach, Ran, Amir Navot, and Naftali Tishby. "An information theoretic tradeoff between complexity and accuracy." *Learning Theory and Kernel Machines*. Springer, Berlin, Heidelberg, 2003. 595-609.

Information Bottleneck

- Let T be a representation of X
  - Which T is useful?
    - Disentangled
    - Interpretable
    - Related to task Y → Useful for predicting Y
  - How to define the optimal representation T?

Sufficient Statistics $S(X)$

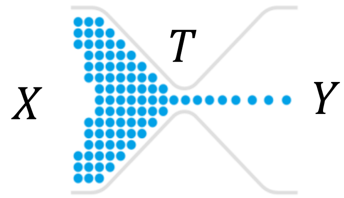
$$\boxed{I(S(X);Y)} = \boxed{I(X;Y)}$$

Minimal Sufficient Statistics $T(X)$

$$T(X) = \arg \min_{S(X):\boxed{I(S(X);Y)}=\boxed{I(X;Y)}} \boxed{I(S(X);X)}$$

Information Bottleneck as an Approximation

$$\max_{p(t|x)} \boxed{I(T;Y)} - \beta \boxed{I(T;X)}$$

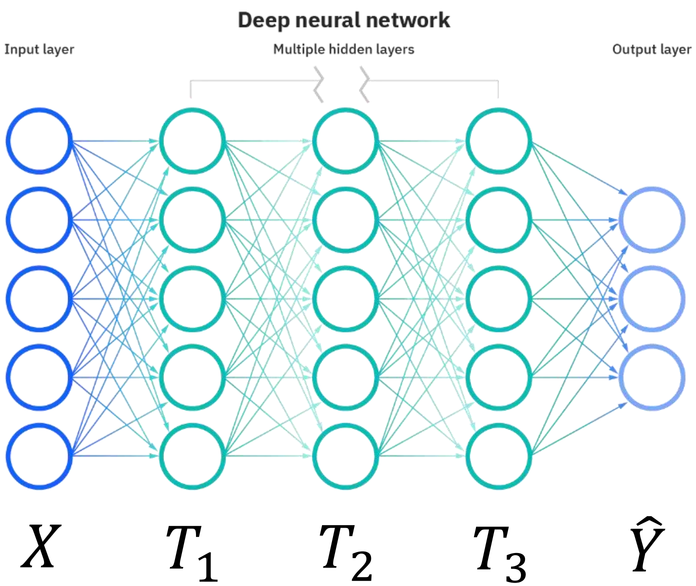


Information Bottleneck Approach in Deep Neural Networks

Information Bottleneck in DNNs

- Deep Information Bottleneck
  - Neural Network parameterization of IB

$$\max_{\theta} I(T;Y) - \beta I(T;X)$$



Information Bottleneck in DNNs

- Deep Information Bottleneck
  - Neural Network parameterization of IB
    - $\max_{\theta} I(T;Y) - \beta I(T;X)$
    - How to estimate $I(T;Y)$ and $I(T;X)$
    - How to optimize?

Information Bottleneck in DNNs

- Estimate $I(T;Y)$
  - $\max_{\theta} I(T;Y) \Leftrightarrow \min_{\theta} D_{\mathrm{KL}}(p(y|\boldsymbol{x}); q_{\theta}(\hat{y}|\boldsymbol{x}))$
  - If $q_{\theta}(\hat{y}|\boldsymbol{x}) \sim \mathcal{N}\left(h_{\theta}(\boldsymbol{x}), \sigma^2 I\right)$, we obtain $\max_{\theta} \mathbb{E}\left(\|y - h_{\theta}(\boldsymbol{x})\|_2^2\right)$
  - How to evaluate $I(T;Y)$ without any parametric assumption?

Rodriguez Galvez, Borja. "The Information Bottleneck: Connections to Other Problems, Learning and Exploration of the IB Curve." (2019).

Information Bottleneck in DNNs

- Estimate $I(T;X)$
  - An upper bound of $I(T;X)$

Variational upper bound (Variational IB)

$$I(T;X) = \mathbb{E}_{p(x,t)} \log p(t|x) - \mathbb{E}_{p(t)} \log p(t)$$
$$\leq \mathbb{E}_{p(x,t)} \log p(t|x) - \mathbb{E}_{p(t)} \log v(t) = D_{\mathrm{KL}}(p(t|x); v(t))$$

Alemi, Alexander A., et al. "Deep variational information bottleneck." ICLR, 2017.

Information Bottleneck in DNNs

- Estimate $I(T;X)$
    - An upper bound of $I(T;X)$

Non-parametric upper bound (Nonlinear IB)

$$I(T;X) \leq -\frac{1}{N}\sum_{i=1}^{N} \log \frac{1}{N}\sum_{j=1}^{N} \exp\left(-D_{\mathrm{KL}}\left(p(t|x_i);p\left(t\middle|x_j\right)\right)\right)$$

Kolchinsky, Artemy, Brendan D. Tracey, and David H. Wolpert. "Nonlinear information bottleneck." *Entropy* 21.12 (2019): 1181.

Information Bottleneck in DNNs

- Estimate $I(T;X)$
  - An upper bound of $I(T;X)$
  - How to evaluate $I(T;X)$ without using an upper bound or variational approximation?

How to carry out <u>nonlinear Information Bottleneck</u> ***without*** variational approximation or parametric distributional assumption?

original IB Lagrangian

$$\max_{p(t|x)} I(T;Y) - \beta I(T;X)$$

deep IB objective

$$\min_{p(t|x)} D_{\mathrm{KL}}(p(y|\boldsymbol{x}); q_\theta(\hat{y}|\boldsymbol{x})) + \beta I(T;X)$$

Cauchy-Schwarz divergence IB

$$\min_{p(t|x)} D_{\mathrm{CS}}(p(y|\boldsymbol{x}); q_\theta(\hat{y}|\boldsymbol{x})) + \beta I_{\mathrm{CS}}(T;X)$$

Yu, Shujian, et al., “Cauchy-Schwarz Divergence Information Bottleneck for Regression”, ICLR, 2024.

How to carry out nonlinear Information Bottleneck ***without*** variational approximation or parametric distributional assumption?

original IB Lagrangian

$$\max_{p(t|x)} I(T;Y) - \beta I(T;X)$$

deep IB objective

$$\min_{p(t|x)} D_{\mathrm{KL}}(p(y|\boldsymbol{x}); q_\theta(\hat{y}|\boldsymbol{x})) + \beta I(T;X)$$

Cauchy-Schwarz divergence IB

$$\min_{p(t|x)} \boxed{D_{\mathrm{CS}}(p(y|\boldsymbol{x}); q_\theta(\hat{y}|\boldsymbol{x}))} + \beta I_{\mathrm{CS}}(T;X)$$

Conditional CS Divergence between $p(y|\boldsymbol{x})$ and $q_\theta(\hat{y}|\boldsymbol{x})$

How to carry out nonlinear Information Bottleneck ***without*** variational approximation or parametric distributional assumption?

original IB Lagrangian

$$\max_{p(t|x)} I(T;Y) - \beta I(T;X)$$

deep IB objective

$$\min_{p(t|x)} D_{\mathrm{KL}}(p(y|\boldsymbol{x});q_\theta(\hat{y}|\boldsymbol{x})) + \beta I(T;X)$$

Cauchy-Schwarz divergence IB

$$\min_{p(t|x)} \boxed{D_{\mathrm{CS}}(p(y|\boldsymbol{x});q_\theta(\hat{y}|\boldsymbol{x}))} + \beta \boxed{I_{\mathrm{CS}}(T;X)}$$

Conditional CS Divergence between $p(y|\boldsymbol{x})$ and $q_\theta(\hat{y}|\boldsymbol{x})$

$I_{\mathrm{CS}}(T;X) = D_{\mathrm{CS}}\big(p(\boldsymbol{x},\boldsymbol{t});p(\boldsymbol{x})p(\boldsymbol{t})\big)$
CS Quadratic Mutual Information

Cauchy-Schwarz Divergence

Integral probability metrics

f-divergence

$P - Q$

$$D_{\mathcal{H}}(P; Q) = \sup_{\varphi \in \mathcal{H}} \left| \mathbb{E}_{X \sim P} \varphi(X) - \mathbb{E}_{Y \sim Q} \varphi(Y) \right|$$

(MMD)

$$D_{KL}(P; Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

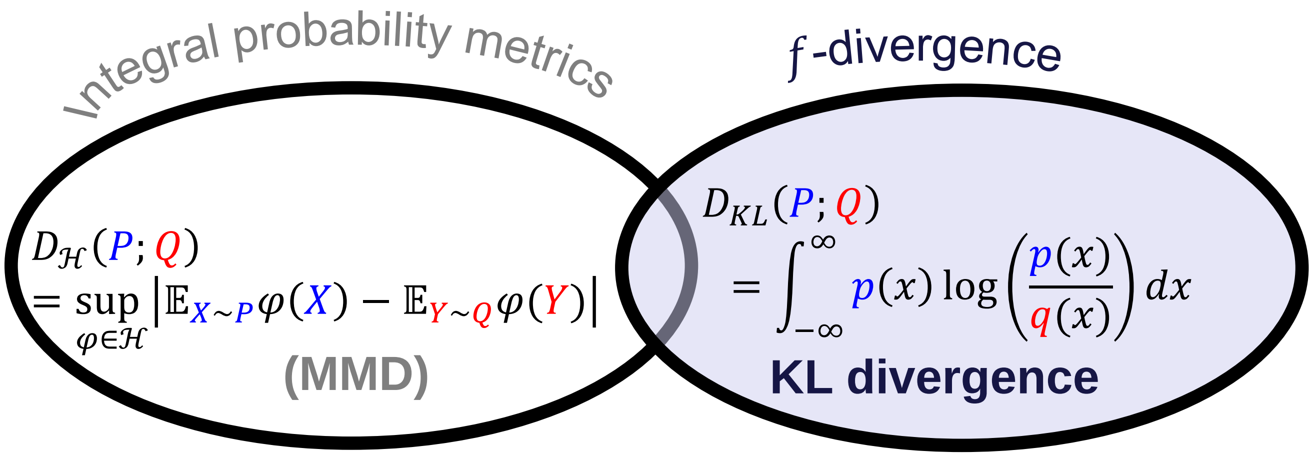
KL divergence

$\frac{P}{Q}$

$$D_f(P; Q) = \int_{-\infty}^{\infty} q(x)\, f\left(\frac{p(x)}{q(x)}\right) dx$$

Support constraint due to ratio

$P - Q$



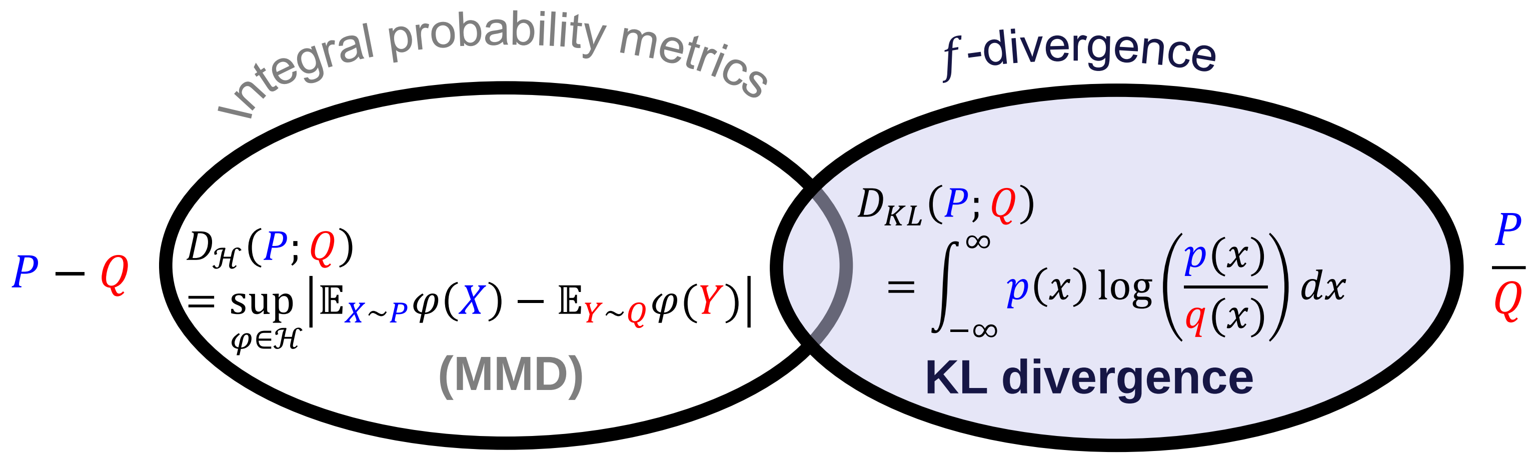
$\frac{P}{Q}$

Cauchy-Schwarz inequality
for square-integral functions

$$\left(\int p(x)q(x)dx\right)^2 \leq \int p^2(x)dx \int q^2(x)dx$$

$$\frac{\int p^2(x)dx \int q^2(x)dx}{\left(\int p(x)q(x)dx\right)^2}$$

Principe, Jose C. *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.



Cauchy-Schwarz inequality
for square-integral functions

$$\left(\int p(x)q(x)dx\right)^2 \leq \int p^2(x)dx \int q^2(x)dx$$

$$\log\left(\frac{\int p^2(x)dx \int q^2(x)dx}{\left(\int p(x)q(x)dx\right)^2}\right)$$

Principe, Jose C. *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.

Integral probability metrics

f-divergence

$P - Q$

$$D_{\mathcal{H}}(P;Q) = \sup_{\varphi \in \mathcal{H}} \left| \mathbb{E}_{X\sim P}\varphi(X) - \mathbb{E}_{Y\sim Q}\varphi(Y) \right|$$

(MMD)

$$D_{KL}(P;Q) = \int_{-\infty}^{\infty} p(x)\log\left(\frac{p(x)}{q(x)}\right)dx$$

KL divergence

$\frac{P}{Q}$

$$D_{CS}(P;Q) = \log\left(\frac{\int p^2(x)dx \int q^2(x)dx}{\left(\int p(x)q(x)dx\right)^2}\right)$$

$\int PQ$

$$\left(\int p(x)q(x)dx\right)^2 \leq \int p^2(x)dx \int q^2(x)dx$$

Sample Estimator of the CS Divergence

- Cauchy-Schwarz divergence between $p(x)$ and $q(x)$, $x \in \mathbb{R}^d$
  - $\{x_i^p\}_{i=1}^M \sim p$, $\{x_i^q\}_{i=1}^N \sim q$

$$\widehat{D}_{\mathrm{CS}}(p;q) = \log\underbrace{\left(\frac{1}{M^2}\sum_{i,j=1}^M \kappa_\sigma\left(x_i^p - x_j^p\right)\right)}_{\textit{within distr. similarity}} + \log\underbrace{\left(\frac{1}{N^2}\sum_{i,j=1}^N \kappa_\sigma\left(x_i^q - x_j^q\right)\right)}_{\textit{within distr. similarity}}$$

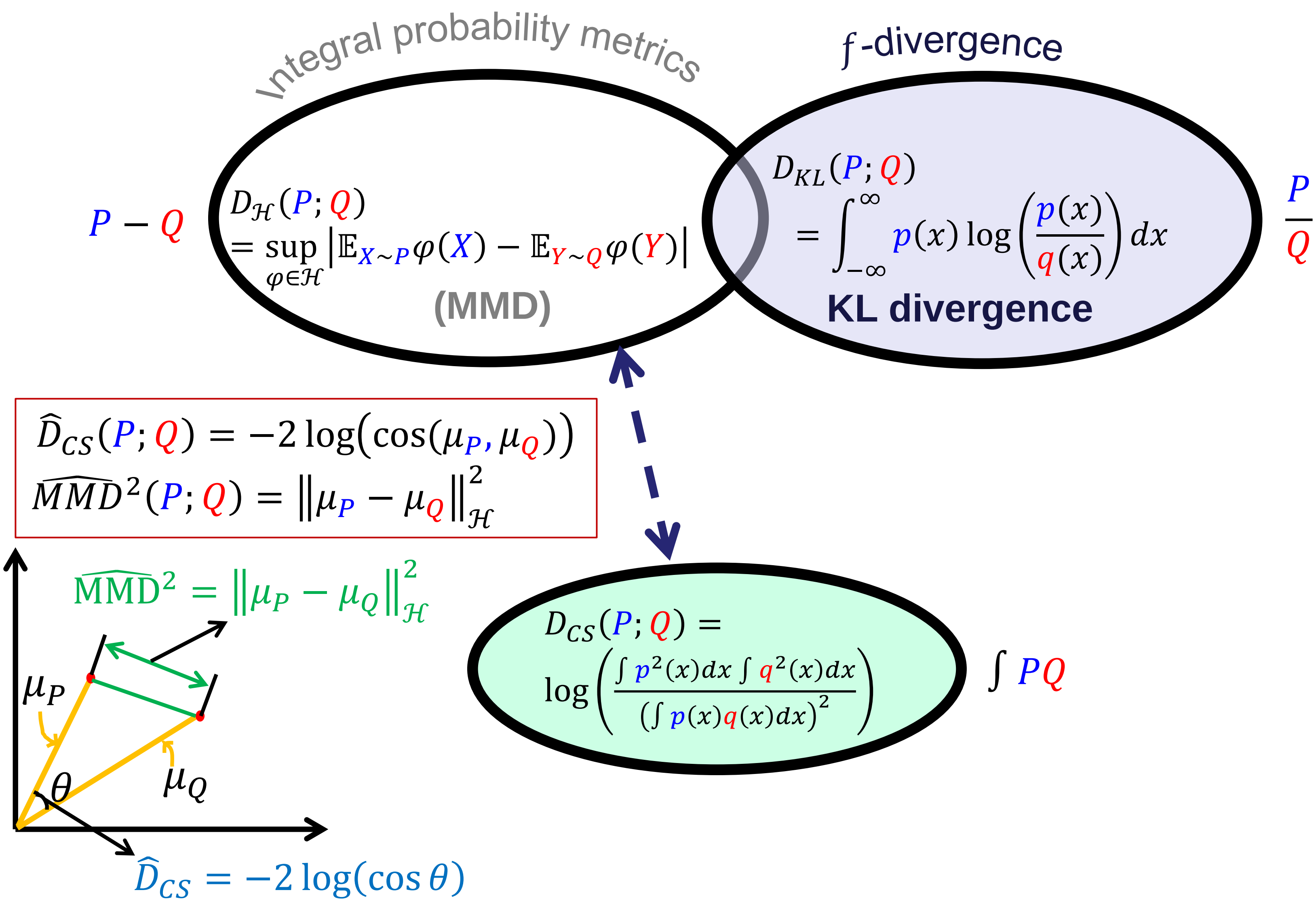
$$-2\log\underbrace{\left(\frac{1}{MN}\sum_{i=1}^M\sum_{j=1}^N \kappa_\sigma\left(x_i^p - x_j^q\right)\right)}_{\textit{cross distr. similarity}}$$

$$\widehat{MMD}^2(p;q) = \frac{1}{M^2}\sum_{i,j=1}^M \kappa_\sigma\left(x_i^p - x_j^p\right) + \frac{1}{N^2}\sum_{i,j=1}^N \kappa_\sigma\left(x_i^q - x_j^q\right)$$

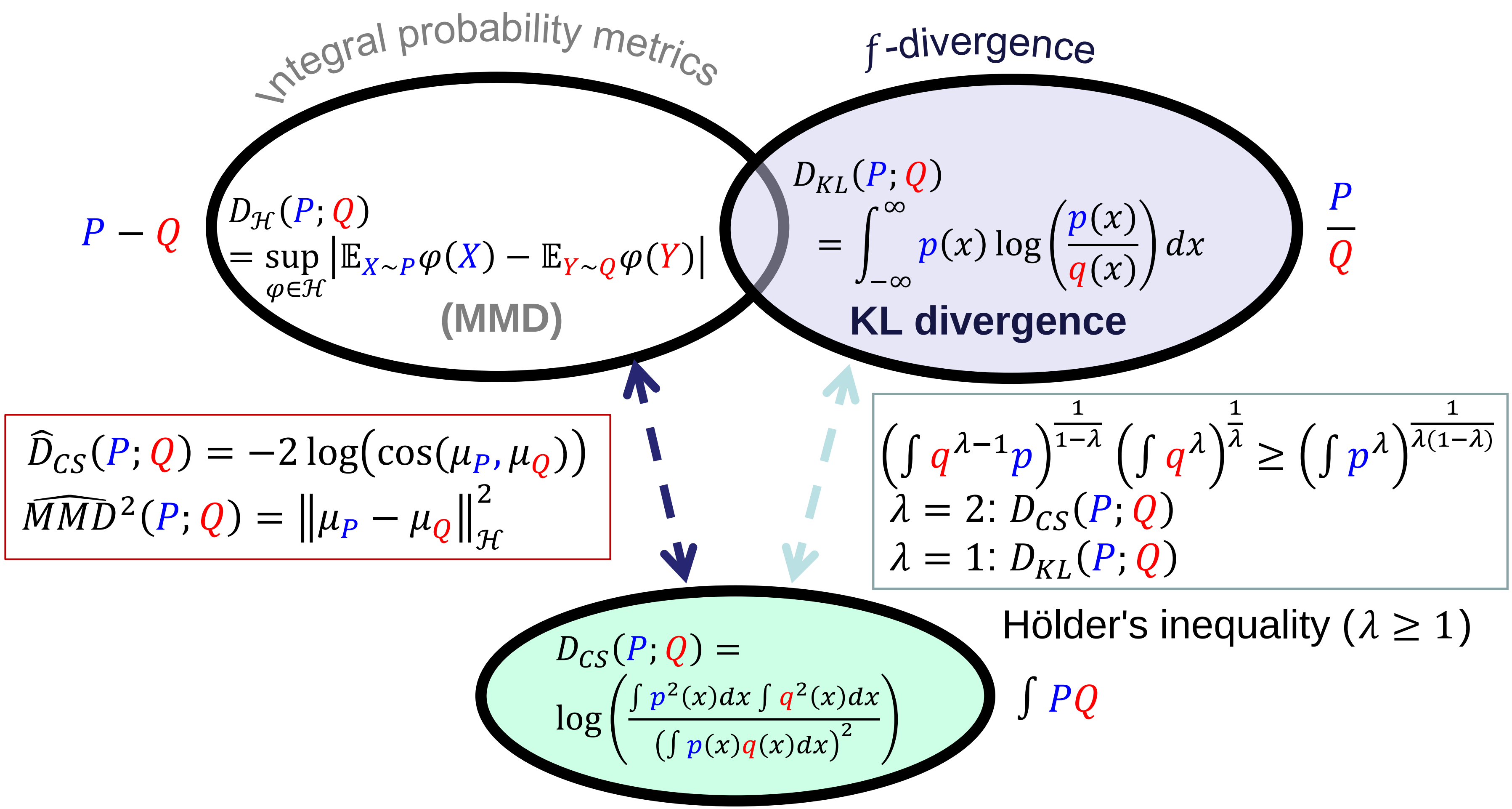
$$-\frac{2}{MN}\sum_{i=1}^M\sum_{j=1}^N \kappa_\sigma\left(x_i^p - x_j^q\right)$$

Gretton, Arthur, et al. "A kernel two-sample test." *The Journal of Machine Learning Research* 13.1 (2012): 723-773.

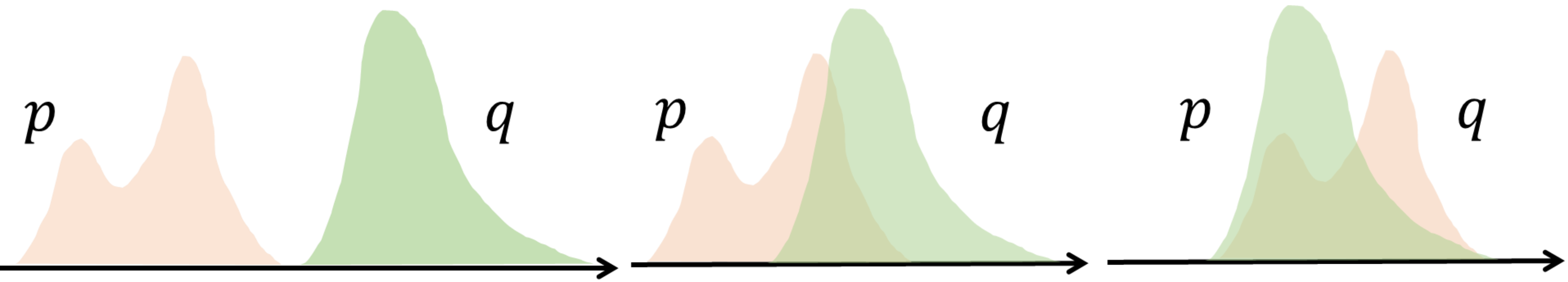
Part 2: CS-IB



$$\mu_P = \frac{1}{M}\sum_{i=1}^{M}\varphi(x_i^p) \qquad \mu_Q = \frac{1}{N}\sum_{j=1}^{N}\varphi(x_j^q)$$


Integral probability metrics
f-divergence
$P - Q$
$D_{\mathcal{H}}(P;Q) = \sup_{\varphi\in\mathcal{H}} \left|\mathbb{E}_{X\sim P}\varphi(X) - \mathbb{E}_{Y\sim Q}\varphi(Y)\right|$
(MMD)
$D_{KL}(P;Q) = \int_{-\infty}^{\infty} p(x)\log\left(\frac{p(x)}{q(x)}\right)dx$
KL divergence
$\frac{P}{Q}$
$\widehat{D}_{CS}(P;Q) = -2\log(\cos(\mu_P,\mu_Q))$
$\widehat{MMD}^2(P;Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}}^2$
$\left(\int q^{\lambda-1}p\right)^{\frac{1}{1-\lambda}} \left(\int q^{\lambda}\right)^{\frac{1}{\lambda}} \geq \left(\int p^{\lambda}\right)^{\frac{1}{\lambda(1-\lambda)}}$
$\lambda = 2$: $D_{CS}(P;Q)$
$\lambda = 1$: $D_{KL}(P;Q)$
Hölder's inequality ($\lambda \geq 1$)
$D_{CS}(P;Q) = \log\left(\frac{\int p^2(x)dx \int q^2(x)dx}{\left(\int p(x)q(x)dx\right)^2}\right)$
$\int PQ$

Part 2: CS-IB



$$D_{KL}(p;q) \to \infty$$
$$D_{KL}(q;p) \to \infty$$
$$D_{CS}(p;q) \to \infty$$

$$D_{KL}(p;q) \to \infty$$
$$D_{KL}(q;p) \to \infty$$
$$D_{CS}(p;q) \text{ finite}$$

$$D_{KL}(p;q) \to \infty$$
$$D_{KL}(q;p) \text{ finite}$$
$$D_{CS}(p;q) \text{ finite}$$

How to carry out nonlinear Information Bottleneck ***without*** variational approximation or parametric distributional assumption?

original IB Lagrangian

$$\max_{p(t|x)} I(T;Y) - \beta I(T;X)$$

deep IB objective

$$\min_{p(t|x)} D_{\mathrm{KL}}(p(y|\boldsymbol{x}); q_\theta(\hat{y}|\boldsymbol{x})) + \beta I(T;X)$$

Cauchy-Schwarz divergence IB

$$\min_{p(t|x)} \boxed{D_{\mathrm{CS}}(p(y|\boldsymbol{x}); q_\theta(\hat{y}|\boldsymbol{x}))} + \beta \boxed{I_{\mathrm{CS}}(T;X)}$$

Conditional CS Divergence between $p(y|\boldsymbol{x})$ and $q_\theta(\hat{y}|\boldsymbol{x})$

$I_{\mathrm{CS}}(T;X) = D_{\mathrm{CS}}\big(p(\boldsymbol{x},\boldsymbol{t}); p(\boldsymbol{x})p(\boldsymbol{t})\big)$
CS Quadratic Mutual Information

Yu, Shujian, et al., "Cauchy-Schwarz Divergence Information Bottleneck for Regression", ICLR, 2024.

- Estimate $D_{\mathrm{CS}}(p(y|\boldsymbol{x}); q_\theta(\hat{y}|\boldsymbol{x}))$

Proposition 1. Given $\{\boldsymbol{x}_i, y_i, \hat{y}_i\}_{i=1}^N$, let K, L^1 and L^2 denote respectively the kernel matrices for variables $\boldsymbol{x}$, y and $\hat{y}$. Further, let L^{21} denote the cross-kernel matrix for y and $\hat{y}$ (i.e., $L_{ij}^{21} = \kappa(\hat{y}_i, y_j)$). $D_{\mathrm{CS}}(p(y|\boldsymbol{x}); q_\theta(\hat{y}|\boldsymbol{x}))$ is estimated by:

$$D_{\mathrm{CS}}(p(y|\boldsymbol{x}); q_\theta(\hat{y}|\boldsymbol{x})) =$$
$$\log\left(\sum_{j=1}^N\left(\frac{\sum_{i=1}^N K_{ji}L_{ji}^1}{\left(\sum_{i=1}^N K_{ji}\right)^2}\right)\right) + \log\left(\sum_{j=1}^N\left(\frac{\sum_{i=1}^N K_{ji}L_{ji}^2}{\left(\sum_{i=1}^N K_{ji}\right)^2}\right)\right) - 2\log\left(\sum_{j=1}^N\left(\frac{\sum_{i=1}^N K_{ji}L_{ji}^{21}}{\left(\sum_{i=1}^N K_{ji}\right)^2}\right)\right)$$

- Estimate $I_{\text{CS}}(T;X)$

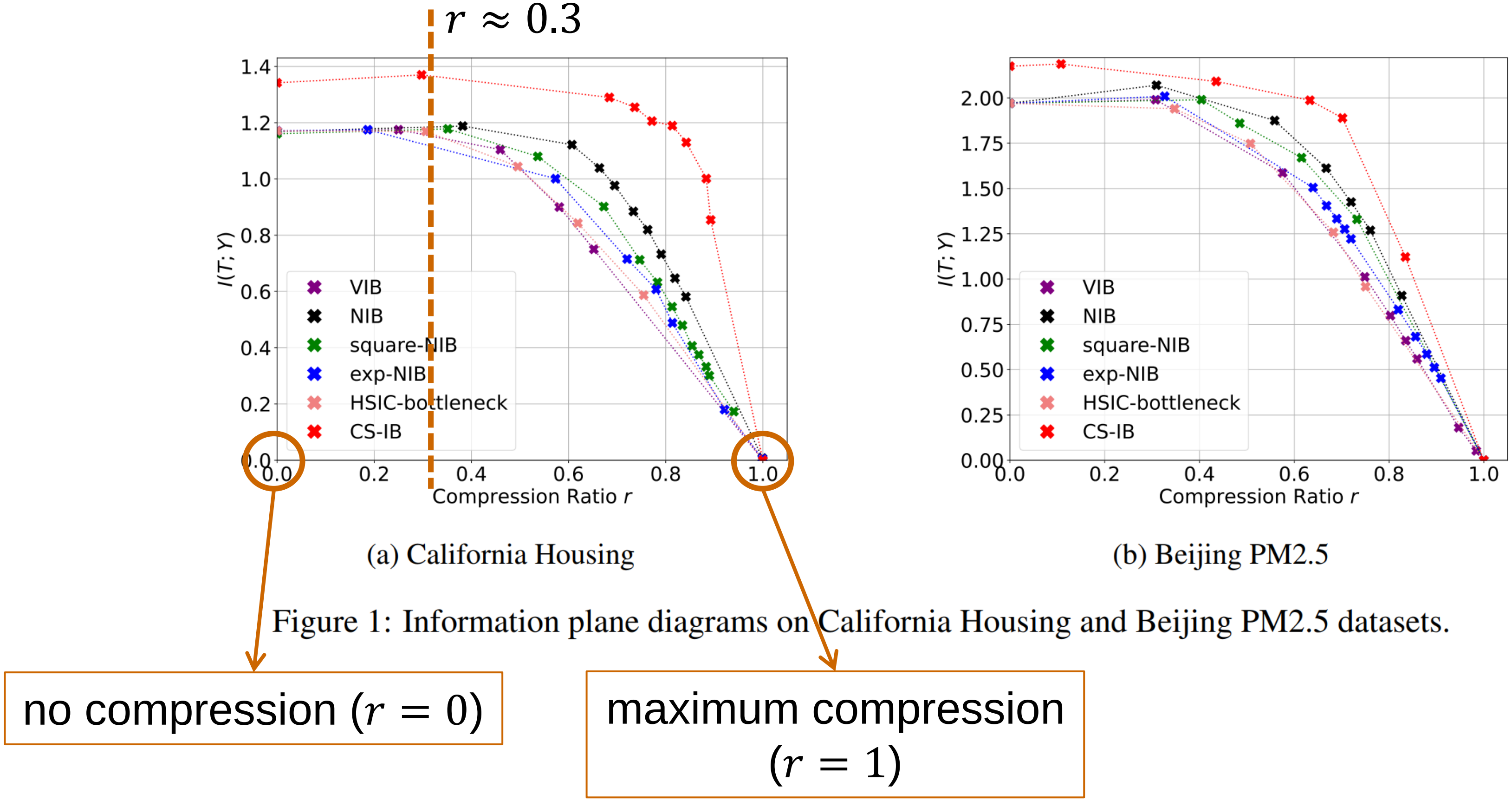
$$I_{\text{CS}}(T;X) = D_{\text{CS}}\big(p(\boldsymbol{x},\boldsymbol{t});p(\boldsymbol{x})p(\boldsymbol{t})\big)$$
$$= \log\left(\frac{\int p^2(x,t)dxdt \int p^2(x)p^2(t)dxdt}{\left(\int p(x,t)p(x)p(t)dxdt\right)^2}\right)$$

Proposition 2. Given $\{\boldsymbol{x}_i,\boldsymbol{t}_i\}_{i=1}^N$, let K and Q denote respectively the kernel matrices for variables $\boldsymbol{x}$ and $\boldsymbol{t}$, and $\mathbf{1}$ denote a $N\times 1$ vector of ones. The empirical estimator of $I_{\text{CS}}(T;X)$ is given by:

$$I_{\text{CS}}(T;X) = \log\left(\frac{1}{N^2}\operatorname{tr}(KQ)\right) + \log\left(\frac{1}{N^4}\mathbf{1}^T K\mathbf{1}\mathbf{1}^T Q\mathbf{1}\right) - 2\log\left(\frac{1}{N^3}\mathbf{1}^T KQ\mathbf{1}\right)$$

Information Bottleneck in DNNs

- Prediction – Compression Tradeoff



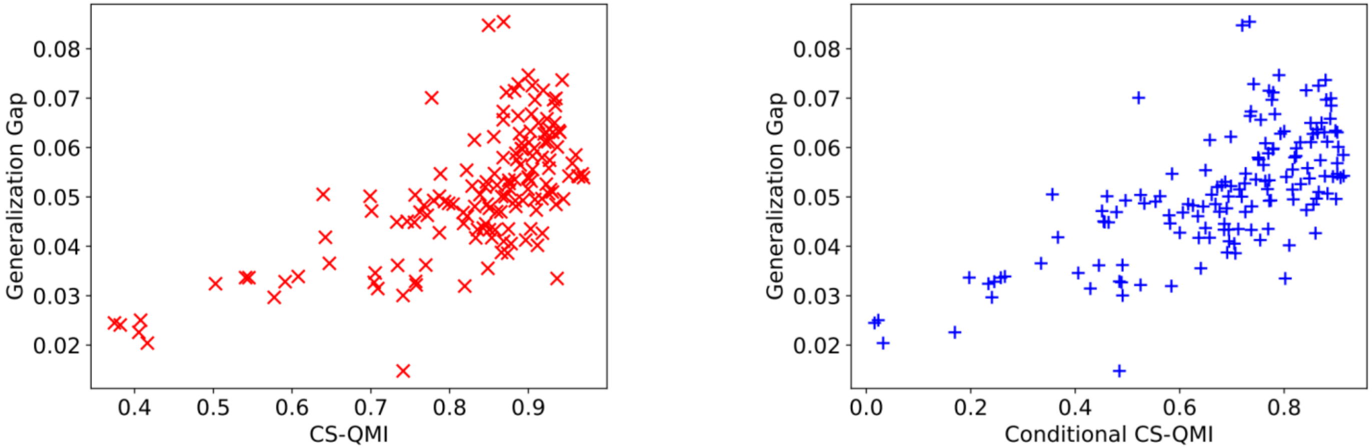
(a) California Housing

(b) Beijing PM2.5

Figure 1: Information plane diagrams on California Housing and Beijing PM2.5 datasets.

Information Bottleneck in DNNs

- Generalization



$I_{\mathrm{CS}}(\boldsymbol{x};\boldsymbol{t})$ (left) and $I_{\mathrm{CS}}(\boldsymbol{x};\boldsymbol{t}|y)$ (right) with respect to the generalization gap in California housing w.r.t. 100 individually trained NNs.

Both $I(X;T|Y)$ and $I(X;T)$ correlate well with empirical generalization error gap.

Does Compression imply generalization?