Mutual Information Estimation via f-Divergence and Data Derangements

Nunzio A. Letizia, Nicola Novello and Andrea M. Tonello

Mutual information (MI)

- The MI is used in information theory, statistics, representation learning and biology. It measures the amount of information obtained about X from the observation of Y

$$I(X;Y) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{XY}(\mathbf{x},\mathbf{y})}\left[\log \frac{p_{XY}(\mathbf{x},\mathbf{y})}{p_X(\mathbf{x})p_Y(\mathbf{y})}\right]$$

joint

marginals

- Estimating the MI is challenging as we typically do not have access to $p_{XY}(\boldsymbol{x},\boldsymbol{y})$, $p_X(\boldsymbol{x})$ and $p_Y(\boldsymbol{y})$
- We can use a discriminative approach to compute only the density-ratio

$$R(\mathbf{x},\mathbf{y}) = \frac{p_{XY}(\mathbf{x},\mathbf{y})}{p_X(\mathbf{x})p_Y(\mathbf{y})}$$

Related work

- Discriminative approaches maximize a variational lower bound (VLB) of the MI:
  - MINE uses the Donsker-Varadhan dual representation of the KL divergence

$$I(X;Y) \geq I_{MINE}(X;Y) = \sup_{\theta\in\Theta}\left\{\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{XY}(\mathbf{x},\mathbf{y})}[T_\theta(\mathbf{x},\mathbf{y})] - \log(\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_X(\mathbf{x})p_Y(\mathbf{y})}[e^{T_\theta(\mathbf{x},\mathbf{y})}])\right\}$$

  - NWJ exploits a different VLB of the KL (based on f-divergence)

$$I(X;Y) \geq I_{NWJ}(X;Y) = \sup_{\theta\in\Theta}\left\{\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{XY}(\mathbf{x},\mathbf{y})}[T_\theta(\mathbf{x},\mathbf{y})] - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_X(\mathbf{x})p_Y(\mathbf{y})}[e^{T_\theta(\mathbf{x},\mathbf{y})-1}]\right\}$$

  - SMILE clips the partition term to reduce the high-variance estimate in MINE

$$I(X;Y) \geq I_{SMILE}(X;Y) = \sup_{\theta\in\Theta}\left\{\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{XY}(\mathbf{x},\mathbf{y})}[T_\theta(\mathbf{x},\mathbf{y})] - \log(\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_X(\mathbf{x})p_Y(\mathbf{y})}[\mathrm{clip}(e^{T_\theta(\mathbf{x},\mathbf{y})}, e^{-\tau}, e^{\tau})])\right\}$$

Open problems

1. Current neural MI estimators are limited to the VLB of the KL divergence
2. They tend to produce high-variance estimate due to the presence of the partition term

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{XY}(\mathbf{x},\mathbf{y})}[f(\mathbf{x},\mathbf{y})] + \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_X(\mathbf{x})p_Y(\mathbf{y})}[g(\mathbf{x},\mathbf{y})]$$

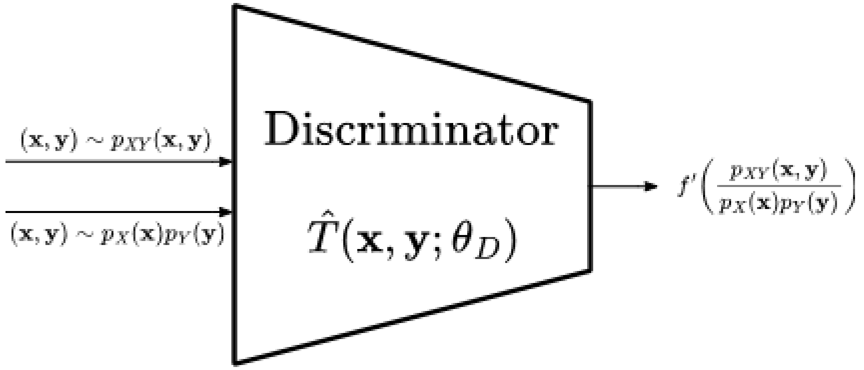
3. It is difficult to correctly estimate large MI values
4. Discriminative approaches require samples from the product of marginals

$$(\mathbf{x},\mathbf{y}) \sim p_{XY}(\mathbf{x},\mathbf{y}) \Rightarrow ?\ (\mathbf{x},\mathbf{y}) \sim p_X(\mathbf{x})p_Y(\mathbf{y})$$

Contributions

1. We use the VLB of the f-divergence to derive an objective function whose maximization leads to a new class of discriminative MI estimators **(f-DIME)**
2. The new MI estimators do not need the evaluation of the partition term, exhibiting an excellent bias-variance trade-off
3. We devise three instantiations which show great performance even for large MI values
4. We prove that permutations lead to upper-bounded estimators and propose a new training sampling strategy based on data derangements

f-DIME



- The objective function reads as

$$\mathcal{J}_f(T) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{XY}(\mathbf{x},\mathbf{y})}\Big[T\big(\mathbf{x},\mathbf{y}\big) - f^*\Big(T\big(\mathbf{x},\sigma(\mathbf{y})\big)\Big)\Big]$$

and

$$\hat{T}(\mathbf{x},\mathbf{y}) = \arg\max_T \mathcal{J}_f(T) = f'\left(\frac{p_{XY}(\mathbf{x},\mathbf{y})}{p_X(\mathbf{x})p_Y(\mathbf{y})}\right)$$

- Extracting the density-ratio, we obtain a new class of MI estimators

$$I_{fDIME}(X;Y) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{XY}(\mathbf{x},\mathbf{y})}\Big[\log\Big(\big(f^*\big)'\big(\hat{T}(\mathbf{x},\mathbf{y})\big)\Big)\Big]$$

where f^* is the Fenchel conjugate of a convex function f and $\sigma(\cdot)$ is a function such that

$$p_{\sigma(Y)}(\sigma(\mathbf{y})|\mathbf{x}) = p_Y(\mathbf{y})$$

f-DIME instantiations

- KL-DIME:

$f(u) = u\log(u)$

$$I_{KL-DIME}(X;Y) := \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{XY}(\mathbf{x},\mathbf{y})}\left[\log\left(\hat{D}(\mathbf{x},\mathbf{y})\right)\right]$$

$$\mathcal{J}_{KL}(D) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{XY}(\mathbf{x},\mathbf{y})}\Big[\log\big(D(\mathbf{x},\mathbf{y})\big)\Big] - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_X(\mathbf{x})p_Y(\mathbf{y})}\Big[D(\mathbf{x},\mathbf{y})\Big] + 1$$

$T(\mathbf{x}) = \log(D(\mathbf{x}))$

- GAN-DIME:

$$f(u) = \log 4 + u\log u - (u+1)\log(u+1)$$

$$I_{GAN-DIME}(X;Y) := \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{XY}(\mathbf{x},\mathbf{y})}\left[\log\left(\frac{1-\hat{D}(\mathbf{x},\mathbf{y})}{\hat{D}(\mathbf{x},\mathbf{y})}\right)\right]$$

$$\mathcal{J}_{GAN}(D) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{XY}(\mathbf{x},\mathbf{y})}\Big[\log\big(1-D(\mathbf{x},\mathbf{y})\big)\Big] + \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_X(\mathbf{x})p_Y(\mathbf{y})}\Big[\log\big(D(\mathbf{x},\mathbf{y})\big)\Big]$$

$T(\mathbf{x}) = \log(1 - D(\mathbf{x}))$

- HD-DIME:

$f(u) = (\sqrt{u}-1)^2$

$$I_{HD-DIME}(X;Y) := \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{XY}(\mathbf{x},\mathbf{y})}\left[\log\left(\frac{1}{\hat{D}^2(\mathbf{x},\mathbf{y})}\right)\right]$$

$$\mathcal{J}_{HD}(D) = 2 - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{XY}(\mathbf{x},\mathbf{y})}\Big[D(\mathbf{x},\mathbf{y})\Big] - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_X(\mathbf{x})p_Y(\mathbf{y})}\left[\frac{1}{D(\mathbf{x},\mathbf{y})}\right]$$

$T(\mathbf{x}) = 1 - D(\mathbf{x})$

Derangements

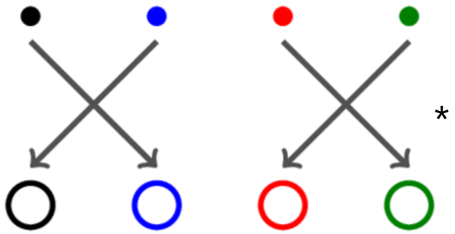
- Practically, $\sigma(\cdot)$ is implemented via data derangements to avoid upper bounded MI estimates

$$(\mathbf{x},\mathbf{y}) \sim p_{XY}(\mathbf{x},\mathbf{y})$$

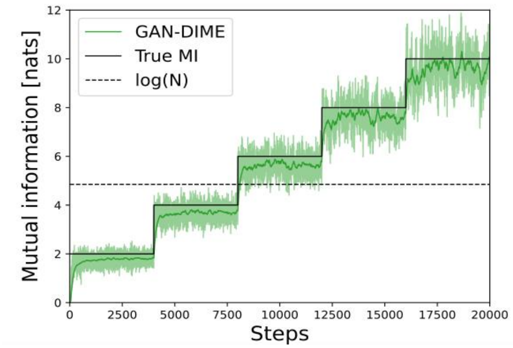
$$\Downarrow$$

$$(\mathbf{x},\mathbf{y}) \sim p_X(\mathbf{x})p_Y(\mathbf{y})$$

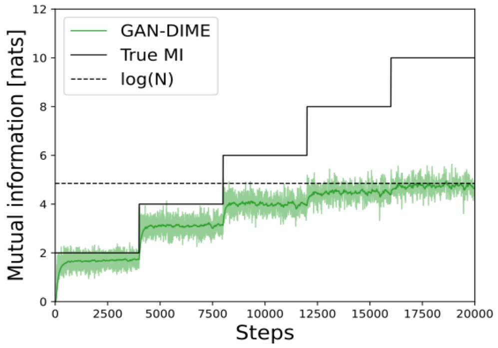
$$\mathbf{y} \sim p_{Y|X}(\mathbf{y}|\mathbf{x})$$



$$\mathbf{y} \sim p_Y(\mathbf{y})$$



a) Derangement strategy

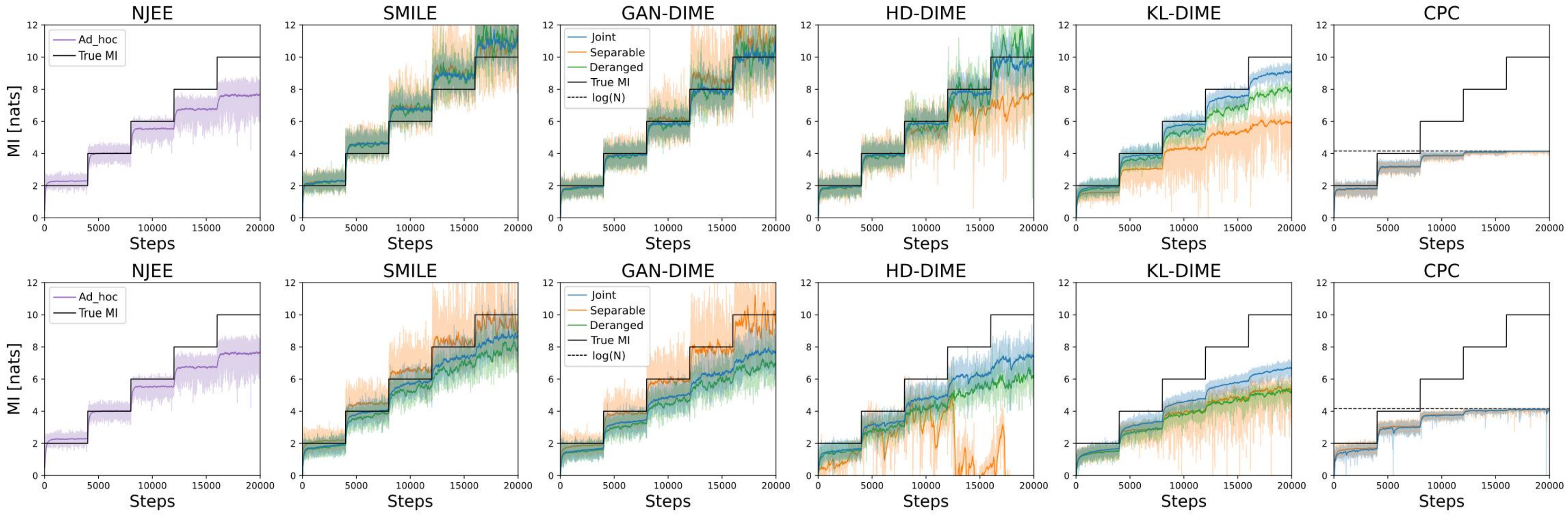


b) Permutation strategy

*https://apieceofthepi.substack.com/p/derangements-and-the-number-e

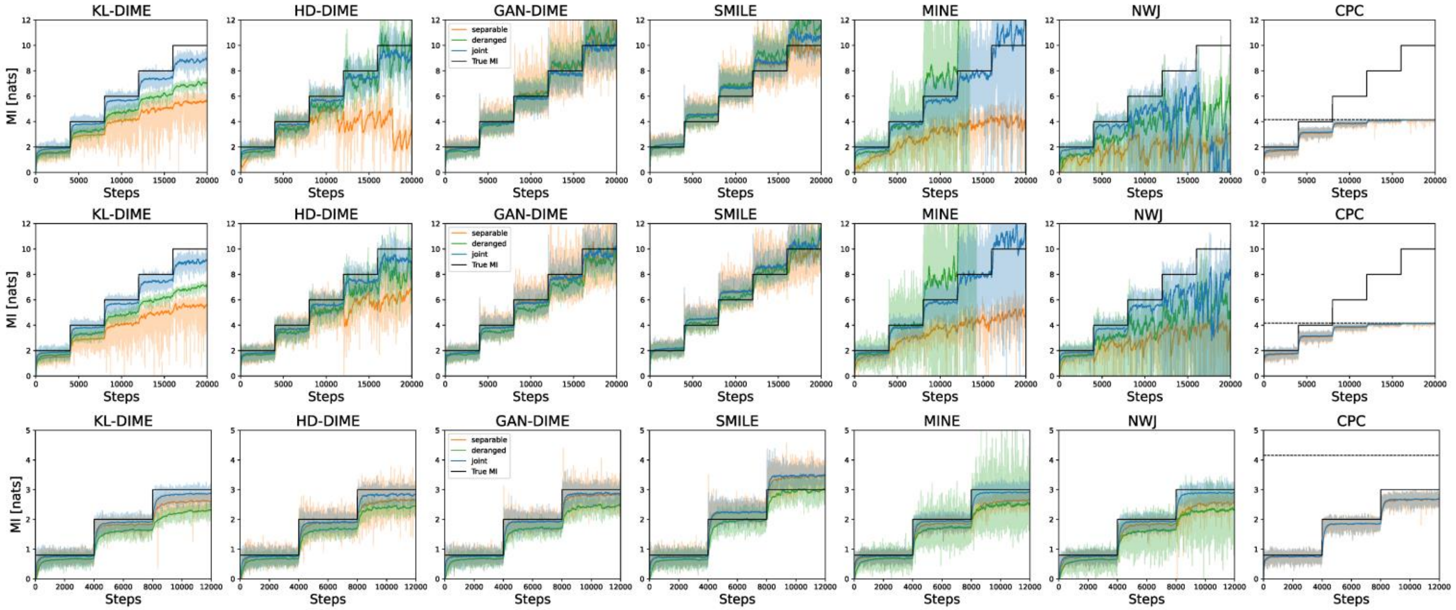
Results – Gaussian scenario

- Staircase comparison for $d = 5$ and batch size $N = 64$
  - Top: Gaussian; bottom: cubic



Results – Non-Gaussian scenario

- Staircase comparison for $d = 5$ and batch size $N = 64$
  - Top: Half-cube scenario; middle: asinh scenario; bottom: Swiss roll scenario as suggested in[1]



[1] Paweł Czy˙z, Frederic Grabowski, Julia E Vogt, Niko Beerenwinkel, and Alexander Marx. Beyond normal: On the evaluation of mutual information estimators. In Thirty-seventh Conference on Neural Information Processing Systems, 2023

Thanks for your interest