On Slicing Optimality for Mutual Information

Ammar FAYAD Majd IBRAHIM

NeurIPS 2023

Dependence Measures

- The mutual information between two random variables X and Y:

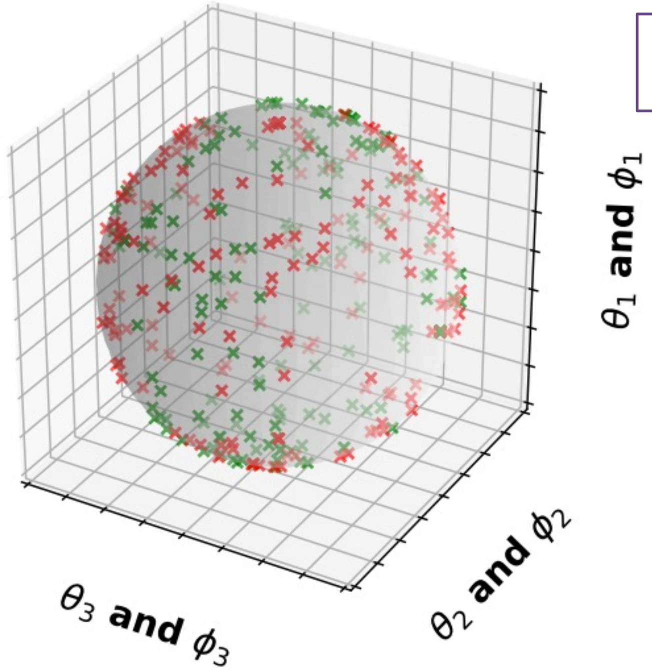
$$I(X;Y) = KL(P_{X,Y}||P_X \otimes P_Y) = \int_{X\times Y} \log\left(\frac{dP_{X,Y}}{dP_X \otimes P_Y}\right) dP_{X,Y}$$

- The sliced mutual information SI [1] is:

$$SI(X;Y) = \oint_{S^{d_x-1}\times S^{d_y-1}} I(\theta^T X; \phi^T Y)\, d\gamma(\theta) \otimes \gamma(\phi)$$

Where γ is the uniform distribution.

[1] Goldfeld, Z. and Greenewald, K. (2021). Sliced mutual information: A scalable measure of statistical dependence. Advances in Neural Information Processing Systems, 34.

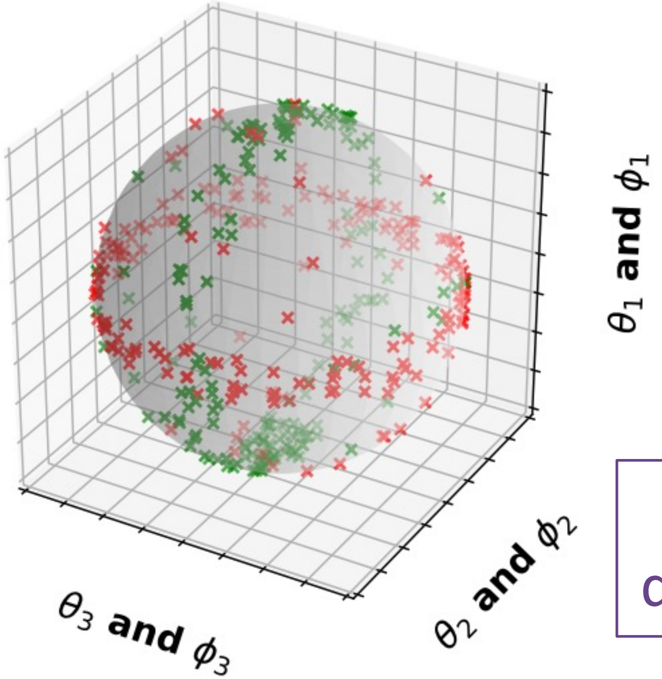


Random slices

- How to reach an optimal slicing distribution?

I. The projection directions are mainly concentrated into areas where the one-dimensional variables contain the maximum mutual information possible.

II. The slicing directions are also diversified over the whole sphere, ensuring that all regions with relevant information are visited.



Optimally distributed slices

Definition of SI^*

The optimal sliced mutual information SI^* between random variables $X \in R^{d_x}$ and $Y \in R^{d_y}$ can be expressed as:

$$SI^*(X;Y) = \sup_{\sigma} \oint_{S^{d_x-1} \times S^{d_y-1}} I(\theta^T X; \varphi^T Y)\, d\sigma(\theta,\varphi) \quad : \sigma_\Theta \in \Sigma_{d_x,\omega_x}, \sigma_\Phi \in \Sigma_{d_y,\omega_y}$$

- Where $\Sigma_{d,\omega} = \{\mu : \mu \in P(S^{d-1}),\ E_{x,y\sim\mu}[\arccos|x^T y|] \geq \omega\}$
- We prove that for any $\omega_X, \omega_Y \in [0, \pi/2]$ there exists an optimal slicing policy σ^* such that the term is maximized.

Properties of SI^*

- $SI^*(X;Y)$ is nonnegative and symmetric.
- $SI^*(X;Y) = 0$ if and only if X and Y are independent.
- If X_n and Y_n are sequences of random variables with joint distribution $P_{X,Y}^{(n)}$ Y that converges pointwise to the joint distribution $P_{X,Y}$ then $\lim_{n\to\infty} SI^*(X_n;Y_n) = SI^*(X;Y)$.
- Similar to MI, SI^* has a relative entropy form, a variational representation, and a discriminator-based form.

Estimation of SI^*

- $\{(X_n, Y_n)\}$ are i.i.d. data points drawn from some P_{XY}.

- $\hat{I}_n$ is a one-dimensional MI estimator over n samples.

- $\{(\theta_m^*, \varphi_m^*)\}$ are i.i.d slicing directions drawn from the optimal policy $\sigma_{\Theta\Phi}^*$.

$$\widehat{SI^*}_{n,m}(X;Y) = \frac{1}{m}\sum_{j=1}^{m}\left[\hat{I}_n\left(\theta_j^{*T}X; \varphi_j^{*T}Y\right)\right]$$

Estimation of SI^*

How to obtain $\sigma^*_{\Theta\Phi}$?

➢Slicing directions can be expressed as $(\theta, \phi) = \big(f_1(\psi, \nu), f_2(\psi, \nu)\big)$ with

$$(\psi, \nu) \sim Uniform\left(S^{d_x-1}\right) \otimes Uniform\left(S^{d_y-1}\right).$$

➢Estimate f_1, f_2 using NNs.

To train the NNs:

- Sample ψ and ν independently and uniformly on spheres S^{d_x-1} and S^{d_y-1}
- Feed the random slices to f_1 and f_2: $\theta = f_1(\psi, \nu), \varphi = f_2(\psi, \nu)$
- Calculate average MI over output slices: $A = \frac{1}{m}\sum_{j=1}^{m} \widehat{I_n}\left(\theta_j^T X; \varphi_j^T Y\right)$
- Calculate $\mathcal{L} = A - \lambda_1\left(\frac{1}{m^2}\sum_{k,j} \arccos\left|{f_1^{(k)}}^T f_1^{(j)}\right| - \omega_X\right) - \lambda_2\left(\frac{1}{m^2}\sum_{k,j} \arccos\left|{f_2^{(k)}}^T f_2^{(j)}\right| - \omega_Y\right)$
- Update f_1 and f_2 in the direction of increasing $\mathcal{L}$.

Convergence Rate

The uniform error bound of $\widehat{SI}^*{}_{n,m}(X;Y)$ is:

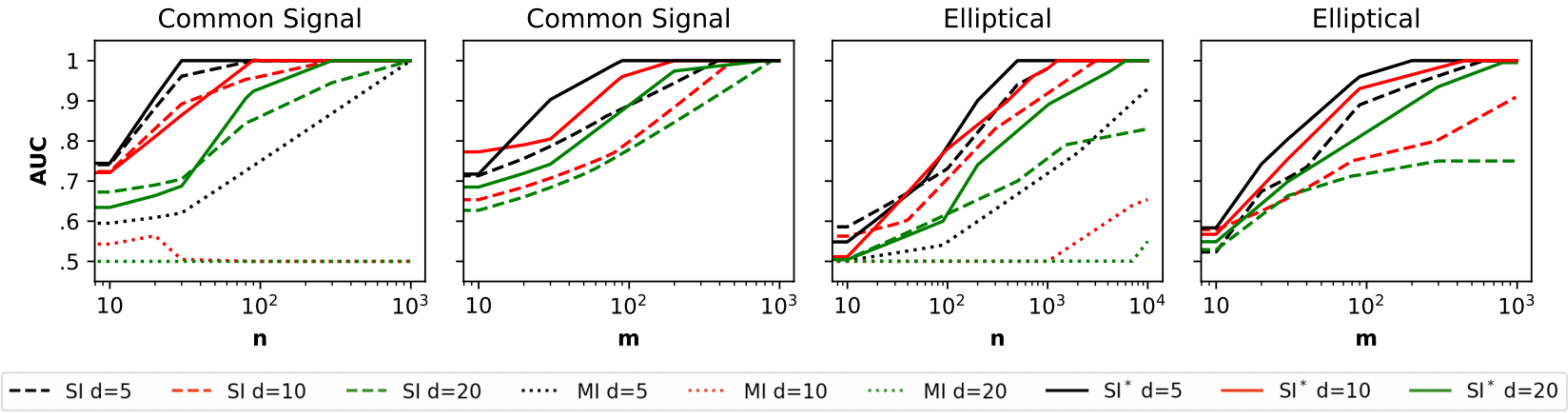
$$\sup_{P_{X,Y}} E\left[\left|SI(X;Y) - \widehat{SI}^*{}_{n,m}(X;Y)\right|\right] \le \delta(n) + \frac{U}{2\sqrt{m}}$$

- Where $\delta(n)$ is the absolute error that uniformly bounds the one-dimensional mutual information estimation, and $U \propto \left(d_x^{-1} + d_y^{-1}\right)^{1/2}$

$\Longrightarrow$ Significantly better than MI

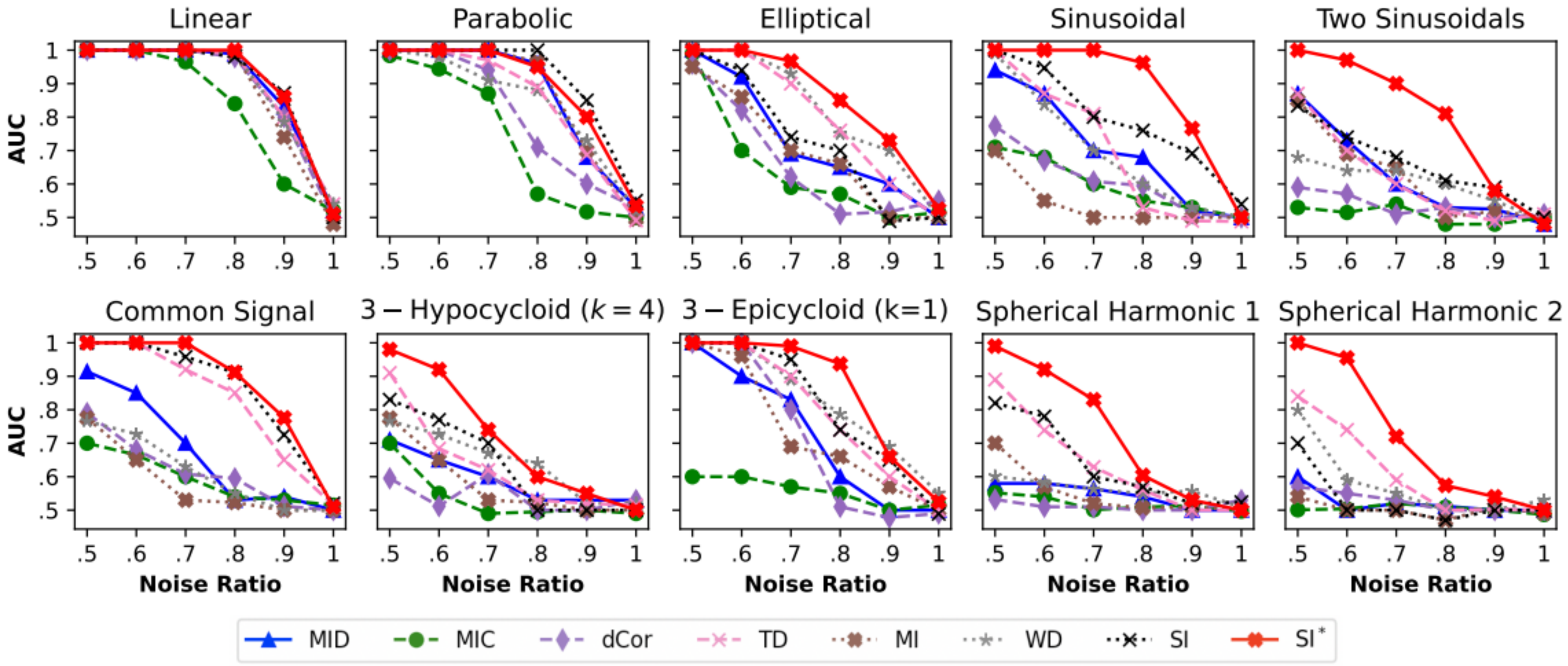
Evaluation of SI^*

Optimal slicing distribution → Slices and samples efficiency



Evaluation of SI^*

- Performance in varied relationships between variables and noise ratios



Evaluation of SI^*

- Performance in very high-dimensional Representation Learning tasks

STL-10 (96×96 images)

	conv	fc	Y
BiGAN	71.53	67.18	58.48
DIM (MI)	69.15	63.81	61.92
DIM (SI)	74.54	71.34	68.90
DIM (SI*)	**76.89**	**71.67**	**70.04**

CIFAR 10

	conv	fc	Y
BiGAN	62.57	62.74	52.54
DIM (MI)	72.66	**70.66**	64.71
DIM (SI)	74.37	**70.23**	65.99
DIM (SI*)	**77.01**	**70.39**	**69.04**

- We compare SI^* against SI and MI using the algorithm Deep InfoMax (DIM) [2] on two baseline datasets, along with the results of BiGAN method [3].

[2] Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. (2019). Learning deep representations by mutual information estimation and maximization. In International Conference on Learning Representations.
[3] Donahue, J., Krähenbühl, P., and Darrell, T. (2016). Adversarial feature learning. arXiv preprint arXiv:1605.09782.

Evaluation of SI^*

- Performance in very high-dimensional Representation Learning tasks

STL-10 (96×96 images)

	conv	fc	Y
BiGAN	71.53	67.18	58.48
DIM (MI)	69.15	63.81	61.92
DIM (SI)	74.54	71.34	68.90
DIM (SI*)	**76.89**	**71.67**	**70.04**

CIFAR 10

	conv	fc	Y
BiGAN	62.57	62.74	52.54
DIM (MI)	72.66	**70.66**	64.71
DIM (SI)	74.37	**70.23**	65.99
DIM (SI*)	**77.01**	**70.39**	**69.04**

- We compare SI^* against SI and MI using the algorithm Deep InfoMax (DIM) [2] on two baseline datasets, along with the results of BiGAN method [3].

Thank you!

Come visit us at the poster!

[2] Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. (2019). Learning deep representations by mutual information estimation and maximization. In International Conference on Learning Representations.
[3] Donahue, J., Krähenbühl, P., and Darrell, T. (2016). Adversarial feature learning. arXiv preprint arXiv:1605.09782.