ROBUST MACHINE LEARNING WITH MATRIX-BASED RÉNYI'S α-ORDER MUTUAL INFORMATION

Shujian Yu[*] NEC Labs Europe

Xi Yu University of Florida

Francesco Alesiani NEC Labs Europe

Jose C. Principe University of Florida

ABSTRACT

We introduce a recently proposed matrix-based Rényi's α-order mutual information (denoted I_α) to measure the dependence (or independence) between two random variables (or vectors). We demonstrate that I_α is automatically differentiable and more statistically powerful than prevalent dependence (or independence) measures in identifying independence and discovering complex dependence patterns. We show the impacts of I_α for the robustness of deep neural networks, and put forward two proposals with both theoretical and empirical justifications.

1 INTRODUCTION

Measuring the amount of dependence (or independence) between random variables (or vectors) is a fundamental problem in statistics. For the general case where the two variables share a non-linear relationship, one of the most well-known dependence measures is the mutual information (MI). MI has been extensively investigated as a key concept in various machine learning problem Principe (2010); Belghazi et al. (2018), including the robustness of deep neural networks (DNNs). For example, using the Determinant based Mutual Information (DMI) Xu et al. (2019) as a loss function, a trained DNN has demonstrated superior robustness to label noise, regardless of noise pattern.

However, MI estimation is notoriously hard, especially in high-dimensional space. Although the nearest neighbor-based (e.g., Kraskov et al. (2004)) or graph-based (e.g., Hero et al. (2002)) MI estimators provide accurate estimates of MI with faster convergence rate Noshad et al. (2019), an unfortunate fact, shared by most estimators based on neighborhoods or graphs, is that they are not differentiable and thus cannot be used for gradient based optimization that is so prevalent in DNNs. In an effort to devise estimators that scale to present-day's machine learning problems, most recent work on estimating MI has focused on variational lower bounds that can be parameterized, for instance the mutual information neural estimator (MINE) Belghazi et al. (2018). These approaches do solve the differentiability issue; however, theoretical results have shown that such high confidence estimators based on the lower bound on MI require a sample size that is exponential in the MI of the data, making reliable estimation impractical in high entropy, high MI scenarios McAllester & Stratos (2020). Moreover, it is quite difficult to assess the tightness of the bound.

This work introduces a recently proposed matrix-based Rényi's α-order mutual information Sanchez Giraldo et al. (2014) and puts forward two proposals for the robustness of DNNs: 1) a new loss function that is robust to covariate shift and non-Gaussian noises, by encouraging the distribution of prediction residual is statistically independent of the distribution of input instances; 2) a new learning objective that generalizes well to test data and is more robust to adversarial attack (than common regularizers like variational information bottleneck (VIB) Alemi et al. (2017)), by parameterizing information bottleneck (IB) principle Tishby et al. (1999) with a deterministic DNN.

2 THE MATRIX-BASED RÉNYI'S α-ORDER MUTUAL INFORMATION I_α

We introduce the recently proposed matrix-based Rényi's α-entropy functional Sanchez Giraldo et al. (2014); Yu et al. (2019) to measure MI between two random variables. The new measure quantifies MI in terms of the normalized eigenspectrum of the Hermitian matrix of the projected data in the reproducing kernel Hilbert space (RKHS).

[*] Email: yusj9011@gmail.com. This work partially summarizes Yu et al. (2021a) and Yu et al. (2021b).

Definition 2.1. *Let $\kappa : \chi \times \chi \mapsto \mathbb{R}$ be a real valued positive definite kernel that is also infinitely divisible Bhatia (2006). Given $\{\mathbf{x}_i\}_{i=1}^{n} \in \chi$, each $\mathbf{x}_i$ can be a real-valued scalar or vector, and the Gram matrix K obtained from evaluating a positive definite kernel κ on all pairs of exemplars, that is $K = \kappa(\mathbf{x}_i, \mathbf{x}_j)$, a matrix-based analogue to Rényi's α-entropy for a normalized positive definite matrix A of size $n \times n$, such that $\mathrm{tr}(A) = 1$, can be given by the following functional:*

$$H_\alpha(A) = \frac{1}{1-\alpha}\log_2\left(\mathrm{tr}(A^\alpha)\right) = \frac{1}{1-\alpha}\log_2\left(\sum_{i=1}^{n}\lambda_i(A)^\alpha\right), \tag{1}$$

where A is the normalized version of K, i.e., $A = K/\mathrm{tr}(K)$, and $\lambda_i(A)$ denotes the i-th eigenvalue of A.

Definition 2.2. *Given n pairs of samples $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^{n}$, each sample contains two different types of measurements $\mathbf{x} \in \chi$ and $\mathbf{y} \in \gamma$ obtained from the same realization, and the positive definite kernels $\kappa_1 : \chi \times \chi \mapsto \mathbb{R}$ and $\kappa_2 : \gamma \times \gamma \mapsto \mathbb{R}$, a matrix-based analogue to Rényi's α-order joint-entropy can be defined as:*

$$H_\alpha(A,B) = H_\alpha\left(\frac{A \circ B}{\mathrm{tr}(A \circ B)}\right), \tag{2}$$

where $A_{ij} = \kappa_1(\mathbf{x}_i, \mathbf{x}_j)$, $B_{ij} = \kappa_2(\mathbf{y}_i, \mathbf{y}_j)$ and $A \circ B$ denotes the Hadamard product between the matrices A and B.

Given Eqs. (1) and (2), the matrix-based Rényi's α-order MI $I_\alpha(A; B)$ in analogy of Shannon's MI is given by Sanchez Giraldo et al. (2014):

$$I_\alpha(A;B) = H_\alpha(A) + H_\alpha(B) - H_\alpha(A,B). \tag{3}$$

Throughout this work, we use the radial basis function (RBF) kernel $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2})$ to obtain the Gram matrices.

I_α demonstrates several appealing properties and empirical observations. We just list two important ones as follows. A detailed explanation and proof is in Appendix A.

Property 1. *I_α has analytical gradients and is automatically differentiable.*

Observation 1. *I_α is more statistically powerful than prevalent dependence (or independence) measures, like Hilbert Schmidt Independence Criterion (HSIC) Gretton et al. (2005), dCov Székely et al. (2007), kernel canonical correlation analysis (KCCA) Akaho (2001) and Cauchy-Schwarz quadratic mutual information (QMI_CS) Principe et al. (2000); Principe (2010), in identifying independence and discovering complex patterns between two random variables (or vectors) $\mathbf{x}$ and $\mathbf{y}$.*

Remark. *I_α measures MI directly from data without PDF estimation. It also avoids additional model training like MINE. The differentiable property of I_α makes it suitable to be used as loss functions to train DNNs. Additionally, the improved statistical power makes I_α a reliable substitute to measures like HSIC in modern deep learning applications, to further improve their performances.*

3 Proposal I: Matrix-based Independence Criterion $\min I_\alpha(\mathbf{x}; e)$[1]

Robust machine learning under domain shift Quionero-Candela et al. (2009) has attracted increasing attentions in recent years. This is just because the training and testing data in reality are collected from different but related domains Wilson & Cook (2020). Let $(\mathbf{x}, y)$ be a pair of random variables with $\mathbf{x} \in \mathbb{R}^p$ and $y \in \mathbb{R}$ (in regression) or $\mathbf{y} \in \mathbb{R}^q$ (in classification), such that $\mathbf{x}$ denotes input instance and y denotes desired signal. We assume $\mathbf{x}$ and y follow a joint distribution $P_{\text{source}}(\mathbf{x}, y)$. Our goal is, given training samples drawn from $P_{\text{source}}(\mathbf{x}, y)$, to learn a model f predicting y from $\mathbf{x}$ that works well on a different, a-priori unknown target distribution $P_{\text{target}}(\mathbf{x}, y)$. We consider here only the covariate shift, in which the assumption is that the conditional label distribution is invariant (i.e., $P_{\text{target}}(y|\mathbf{x}) = P_{\text{source}}(y|\mathbf{x})$) but the marginal distributions of input $P(\mathbf{x})$ are different between source and target domains (i.e., $P_{\text{target}}(\mathbf{x}) \neq P_{\text{source}}(\mathbf{x})$). On the other hand, we also assume that y (in the source domain) may be contaminated with non-Gaussian noises (i.e., $\widetilde{y} = y + \epsilon$). We focus on a fully unsupervised environment, in which we have no access to samples from the target domain.

[1] This section partially summarizes Yu et al. (2021a). Code is available at `https://github.com/SJYuCNEL/Matrix-based-Dependence`.

Our work in this section is directly motivated by Greenfeld & Shalit (2020), which introduces the criterion of minimizing the dependence between the distribution of input $\mathbf{x}$ and that of the prediction residual $e = y - f(\mathbf{x})$ to circumvent the covariate shift, and uses HSIC as the measure to quantify the independence. We provide two contributions over Greenfeld & Shalit (2020). In terms of methodology, we show that by replacing HSIC with I_α, we improve the prediction accuracy in the target domain. Theoretically, we show that this new loss, namely $\min I_\alpha(\mathbf{x}; e)$ is not only robust against covariate shift but also against non-Gaussian noises on y based on Theorem 3.1.

Theorem 3.1. *Minimizing $I(\mathbf{x}; e)$ is equivalent to minimizing error entropy $H(e)$.*

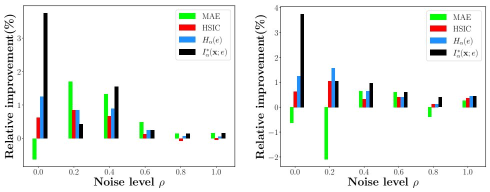
Remark. *The minimum error entropy (MEE) criterion Erdogmus & Principe (2002) has been extensively studied in signal processing and process control to address non-Gaussian noises with both theoretical guarantee and empirical evidence Chen et al. (2009; 2016). We show proof in Appendix B.1 and summarize in Appendix B.2 two insights to further clarify its advantage.*

3.1 Learning under Covariate Shift

We first compare the performances of cross entropy (CE) loss, HSIC loss Goldfeld & Polyanskiy (2020) with our MEE loss $H_\alpha(e)$ and $I_\alpha(\mathbf{x}; e)$ loss under covariate shift. Following Greenfeld & Shalit (2020), the source data is the Fashion-MNIST dataset Xiao et al. (2017), and images which are rotated by an angle θ sampled from a uniform distribution over $[-20^\circ, 20^\circ]$ constitute the target data. The network architecture is specified in Appendix B.3. For $I_\alpha(\mathbf{x}; e)$ and $H_\alpha(e)$, we set $\alpha = 2$. For the HSIC loss, we take the same hyper-parameters as in Greenfeld & Shalit (2020). The results are summarized in Table 1. Our $H_\alpha(e)$ performs comparably to HSIC, but our $I_\alpha(\mathbf{x}; e)$ improves performances in both source and target domains.

Method	Fashion MNIST	
	Source	Target
CE	90.90 ± 0.002	73.73 ± 0.086
HSIC	91.03 ± 0.003	76.56 ± 0.034
$H_\alpha(e)$	91.10 ± 0.013	75.48 ± 0.069
$I_\alpha(\mathbf{x}; e)$	$\mathbf{91.17 \pm 0.040}$	$\mathbf{76.79 \pm 0.040}$

Table 1: Test accuracy (%) on Fashion-MNIST



(a) Laplacian (b) Shifted exponential

Figure 1: Comparisons of models trained with MSE, MAE, HSIC loss, $I_\alpha(\mathbf{x}; e)$ and $H_\alpha(e)$. Each bar denotes the relative performance gain (or loss) over MSE.

3.2 Learning in Noisy Environment

We select the widely used bike sharing data set Fanaee-T & Gama (2014) in UCI repository, in which the task is to predict the number of hourly bike rentals based on features like wind speed and humidity. Consisting of $17,379$ samples, the data was collected over two years, and can be partitioned by year and season. Early studies suggest that this data set contains covariate shift due to the change of time Subbaswamy et al. (2019).

We use the first three seasons samples as source data and the forth season samples as target data. The model of choice is a multi-layered perceptron (MLP) with three hidden layer of size 100, 100 and 10 respectively. We compare our $I_\alpha(\mathbf{x}; e)$ and $H_\alpha(e)$ with mean square error (MSE), mean absolutely error (MAE) and HSIC loss, assuming y is contaminated with additive noise as $\widetilde{y} = y + \epsilon$. We consider two common non-Gaussian noises with the noise level controlled by parameter ρ: the Laplace noise $\epsilon \sim \text{Laplace}(0, \rho)$; and the shifted exponential noise $\epsilon = \rho(1 - \eta)$ with $\eta \sim \exp(1)$. We use batch-size of 32 and the Adam optimizer.

We compared our $I_\alpha(\mathbf{x}; e)$ and $H_\alpha(e)$ against MSE loss, MAE loss and HSIC loss. Fig. 1 demonstrates the averaged performance gain (or loss) of different loss functions over MSE loss in 10 independent runs. In most of cases, $I_\alpha(\mathbf{x}; e)$ improves the most. HSIC is not robust to Laplacian noise, whereas MAE performs poorly under shifted exponential noise. On the other hand, $H_\alpha(e)$ also obtained a consistent performance gain over MSE, which further corroborates our theoretical arguments. More experimental comparisons are shown in Appendix B.4.

4 Proposal II: Deep Deterministic Information Bottleneck[2]

The IB principle considers extracting information about a target signal Y through a correlated observable X. The extracted information is quantified by a variable T, which is (a possibly randomized) function of X, thus forming the Markov chain $Y \leftrightarrow X \leftrightarrow T$. Suppose we know the joint distribution $p(X, Y)$, the objective is to learn a representation T that maximizes its predictive power to Y subject to some constraints on the amount of information that it carries about X:

$$\mathcal{L}_{IB} = I(Y;T) - \beta I(X;T), \tag{4}$$

where β is a Lagrange multiplier that controls the trade-off between the **sufficiency** (the performance on the task, as quantified by $I(Y;T)$) and the **minimality** (the complexity of the representation, as measured by $I(X;T)$).

4.1 Parameterizing IB objective with I_α

The IB objective contains two MI terms: $I(X;T)$ and $I(Y;T)$. When parameterizing IB objective with a DNN, T refers to the latent representation of one hidden layer. The gap between the IB principle and its practical deep learning applications is mainly the result of the challenge in computing mutual information Goldfeld & Polyanskiy (2020); Zaidi & Estella-Aguerri (2020); Alemi et al. (2017). Variational inference offers a natural solution to bridge the gap, as it constructs a lower bound on the IB objective which is tractable with the reparameterization trick Kingma & Welling (2014). Notable examples in this direction include the deep VIB Alemi et al. (2017) and the β-variational autoencoder (β-VAE) Higgins et al. (2017).

In this work, we simply estimate $I(X;T)$ (in a mini-batch) with I_α (i.e., Eq. (3)). The estimation of $I(Y;T)$ is different here. Usually, the maximization of $I(Y;T)$ amounts to the minimization of cross entropy loss CE$(Y, \hat{Y})$ Achille & Soatto (2018); Amjad & Geiger (2019). In this sense, the IB objective can be achieved as a CE loss regularized by a weighted differentiable MI term $I(X;T)$.

4.2 Robustness of DIB to Adversarial Attack

We term our methodology Deep Deterministic Information Bottleneck (DIB) and evaluate the adversarial robustness of model trained with our DIB objective. There are multiple definitions of adversarial robustness in the literature. The most basic one, which we shall use, is accuracy on adversarially perturbed versions of the test set, also called the adversarial examples. One of the most popular attack methods is the Fast Gradient Sign Attack (FGSM) Goodfellow et al. (2015), which uses the gradient of the objective function with respect to the input image to generate an adversarial image maximizing the loss. The FGSM can be summarized by Eq. (5):

$$\hat{x} = x + \epsilon \cdot sign(\nabla_x J(\theta, \mathbf{x}, y)), \tag{5}$$

where x denotes the original clean image, ϵ is the pixel-wise perturbation amount, $\nabla_x J(\theta, \mathbf{x}, \mathbf{y})$ is gradient of the loss with respect to the input image x, and $\hat{x}$ represents the perturbed image.

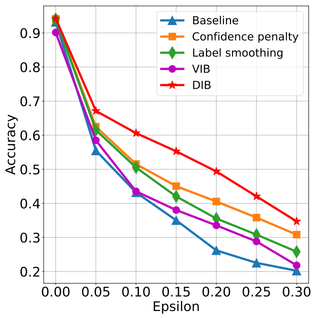


Figure 2: Test accuracy with different ϵ and different methods on CIFAR-10.

We compare behaviors of model trained with different forms of regularizations on CIFAR-10 under FGSM. We elaborate experimental set up in Appendix C, and only add adversarial attack on the test set. The results are shown in Fig. 2. As we can see, our DIB performs much better than VIB and regularizations like Label Smoothing Pereyra et al. (2017) and Confidence Penalty Pereyra et al. (2017) in CIFAR-10.

5 Conclusions and Future Work

We introduced matrix-based Rényi's α-order mutual information I_α and explained its advantage over graph-based or nearest neighbor-based MI estimators and MINE. We put forward two robust learning proposals based on I_α and demonstrated their effectiveness theoretically and empirically.

[2] This section summarizes Yu et al. (2021b). Code is at `https://github.com/yuxi120407/DIB`.

REFERENCES

Martín Abadi et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pp. 265–283, 2016.

Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40 (12):2897–2905, 2018.

S Akaho. A kernel method for canonical correlation analysis. In *Proceedings of the International Meeting of the Psychometric Society (IMPS2001)*. Springer-Verlag, 2001.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

Rana Ali Amjad and Bernhard Claus Geiger. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

Francis R Bach and Michael I Jordan. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48, 2002.

Mohamed Ishmael Belghazi et al. Mutual information neural estimation. In *International Conference on Machine Learning*, pp. 531–540, 2018.

Rajendra Bhatia. Infinitely divisible matrices. *The American Mathematical Monthly*, 113(3):221–235, 2006.

Ba-Dong Chen, Jin-Chun Hu, Yu Zhu, and Zeng-Qi Sun. Information theoretic interpretation of error criteria. *Acta Automatica Sinica*, 35(10):1302–1309, 2009.

Badong Chen, Yu Zhu, Jinchun Hu, and Ming Zhang. A new interpretation on the mmse as a robust mee criterion. *Signal processing*, 90(12):3313–3316, 2010.

Badong Chen, Lei Xing, Bin Xu, Haiquan Zhao, and Jose C Principe. Insights into the robustness of minimum error entropy estimation. *IEEE transactions on neural networks and learning systems*, 29(3):731–737, 2016.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Deniz Erdogmus and Jose C Principe. An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems. *IEEE Transactions on Signal Processing*, 50(7):1780–1786, 2002.

Hadi Fanaee-T and Joao Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2):113–127, 2014.

Xiangbo Feng, Kenneth A Loparo, and Yuguang Fang. Optimal state estimation for stochastic systems: An information theoretic approach. *IEEE Transactions on Automatic Control*, 42(6): 771–785, 1997.

Ziv Goldfeld and Yury Polyanskiy. The information bottleneck problem and its applications in machine learning. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Daniel Greenfeld and Uri Shalit. Robust learning with the hilbert-schmidt independence criterion. In *International Conference on Machine Learning*, pp. 3759–3768. PMLR, 2020.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005.

Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, Alexander J Smola, et al. A kernel statistical test of independence. In *NeurIPS*, volume 20, pp. 585–592, 2007.

Alfred O Hero, Bing Ma, Olivier JJ Michel, and John Gorman. Applications of entropic spanning graphs. *IEEE signal processing magazine*, 19(5):85–95, 2002.

Irina Higgins et al. β-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Ting Hu, Jun Fan, Qiang Wu, and Ding-Xuan Zhou. Learning theory approach to minimum error entropy criterion. *Journal of Machine Learning Research*, 14(2), 2013.

Paul Kalata and Roland Priemer. Linear prediction, filtering, and smoothing: An information-theoretic approach. *Information Sciences*, 17(1):1–14, 1979.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Artemy Kolchinsky, Brendan D. Tracey, and Steven Van Kuyk. Caveats for information bottleneck in deterministic scenarios. In *International Conference on Learning Representations*, 2019.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Jan R Magnus. On differentiating eigenvalues and eigenvectors. *Econometric Theory*, pp. 179–191, 1985.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pp. 875–884. PMLR, 2020.

Morteza Noshad, Yu Zeng, and Alfred O Hero. Scalable mutual information estimation using dependence graphs. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2962–2966. IEEE, 2019.

Adam Paszke et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pp. 8026–8037, 2019.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. In *International Conference on Learning Representations (workshop)*, 2017.

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

Jose C Principe, Dongxin Xu, Qun Zhao, and John W Fisher. Learning from examples with information theoretic criteria. *Journal of VLSI signal processing systems for signal, image and video technology*, 26(1):61–77, 2000.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

Luis Gonzalo Sanchez Giraldo, Murali Rao, and Jose C Principe. Measures of entropy from data using infinitely divisible kernels. *IEEE Transactions on Information Theory*, 61(1):535–548, 2014.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3118–3127. PMLR, 2019.

Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377, 1999.

Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. In *NeurIPS*, pp. 6222–6233, 2019.

Shujian Yu, Luis Gonzalo Sanchez Giraldo, Robert Jenssen, and Jose C Principe. Multivariate extension of matrix-based renyi's α-order entropy functional. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

Shujian Yu, Francesco Alesiani, Xi Yu, Robert Jenssen, and Jose C Principe. Measuring dependence with matrix-based entropy functional. In *AAAI*, 2021a.

Xi Yu, Shujian Yu, and Jose C Principe. Deep deterministic information bottleneck with matrix-based entropy functional. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021b.

Abdellatif Zaidi and Iñaki Estella-Aguerri. On the information bottleneck problems: Models, connections, applications and information theoretic views. *Entropy*, 22(2):151, 2020.

A Elaboration on Properties of I_α

A.1 Analytical Gradient of I_α

It is not hard to verify that $I_\alpha(A;B) = H_\alpha(A) + H_\alpha(B) - H_\alpha(A,B)$ has analytical gradient. In fact, we have:

$$\frac{\partial S_\alpha(A)}{\partial A} = \frac{\alpha}{(1-\alpha)} \frac{A^{\alpha-1}}{\operatorname{tr}(A^\alpha)}, \tag{6}$$

,

$$\frac{\partial S_\alpha(A,B)}{\partial A} = \frac{\alpha}{(1-\alpha)} \left[\frac{(A \circ B)^{\alpha-1} \circ B}{\operatorname{tr}(A \circ B)^\alpha} - \frac{I \circ B}{\operatorname{tr}(A \circ B)} \right] \tag{7}$$

and

$$\frac{\partial I_\alpha(A;B)}{\partial A} = \frac{\partial S_\alpha(A)}{\partial A} + \frac{\partial S_\alpha(A,B)}{\partial A} \tag{8}$$

Since $I_\alpha(A;B)$ is symmetric, the same applies for $\frac{\partial I_\alpha(A;B)}{\partial B}$ with exchanged roles between A and B.

In practice, taking the gradient of the I_α is simple with any automatic differentiation software, like PyTorch Paszke et al. (2019) or Tensorflow Abadi et al. (2016). We recommend PyTorch in this work, because the obtained gradient is consistent with the analytical one. For example, by the Theorem 1 in Magnus (1985), the analytical gradient of the i-th eigenvalue with respect to a real symmetric matrix X is the outer product of the i-th eigenvector (v_i), i.e.,:

$$\frac{\partial \lambda_i}{\partial X} = v_i v_i^T. \tag{9}$$

We noticed that, with Tensorflow, the diagonal entries are the same to their corresponding analytical values, but the off-diagonal entries are just half.

A.2 Experimental Setup on the Statistical Power of I_α

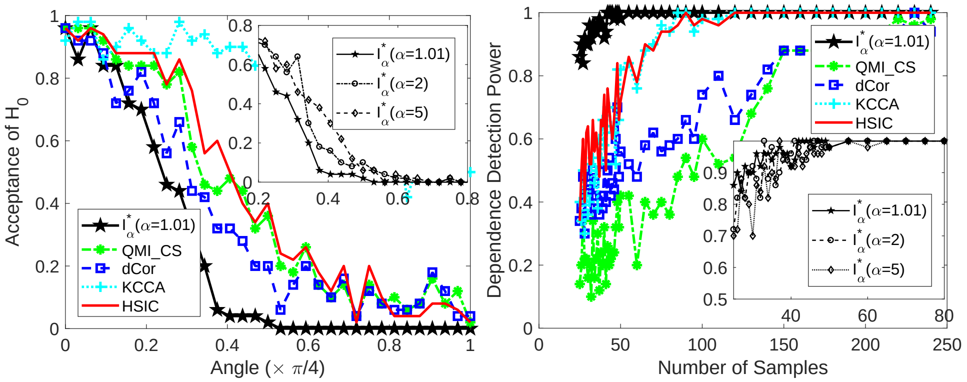
The first test data is generated as follows Gretton et al. (2007). First, we generate N $i.i.d.$ samples from two randomly picked densities in the ICA benchmark densities Bach & Jordan (2002). Second, we mixed these random variables using a rotation matrix parameterized by an angle θ, varying from 0 to $\pi/4$. Third, we added $d-1$ extra dimensional Gaussian noise of zero mean and unit standard deviation to each of the mixtures. Finally, we multiplied each resulting vector by an independent random d-dimensional orthogonal matrix. The resulting random vectors are dependent across all observed dimensions.

The second test data is generated as follows Székely et al. (2007). A matrix $X \in \mathbb{R}^{N\times 5}$ is generated from a multivariate Gaussian distribution with an identity covariance matrix. Then, another matrix $Y \in \mathbb{R}^{N\times 5}$ is generated as $Y_{ml} = X_{ml}\epsilon_{ml}$, $m = 1, 2, \cdots, N$, $l = 1, 2, \cdots, 5$, where ϵ_{ml} are standard normal variables and independent of X.

In each test data, we compare all measures with a threshold computed by sampling a surrogate of the null hypothesis H_0 based on shuffling samples in $\mathbf{y}$ with 100 times. That is, the correspondences between $\mathbf{x}$ and $\mathbf{y}$ are broken by the random permutations. The threshold is the estimated quantile $1-\tau$ where τ is the significance level of the test (Type I error). If the estimated measure is larger than the computed threshold, we reject the null hypothesis and argue the existence of an association between $\mathbf{x}$ and $\mathbf{y}$, and vice versa.

We repeated the above procedure 500 independent trials. Fig. 3 demonstrated the averaged acceptance rate of the null hypothesis H_0 (in test data I with respect to different rotation angle θ) and the averaged detection rate of the alternative hypothesis H_1 (in test data II with respect to different number of samples).

Intuitively, in the first test data, a zero angle means the data are independent, while dependence becomes easier to detect as the angle increases to $\pi/4$. Therefore, a desirable measure is expected to have acceptance rate of H_0 nearly to 1 at $\theta = 0$. But the rate is expected to rapidly decaying as θ increases. In the second test data, a desirable measure is expected to always have a large detection rate of H_1 regardless of the number of samples.



(a) decaying dependence

(b) non-monotonic dependence

Figure 3: Power test against prevalent random vector association measures. Our measure is the most powerful one in a large range of α.

B Supplementary Material to Proposal I

B.1 Proof of Theorem 3.1

Here we prove the equivalence of $\min I(\mathbf{x}; e)$ and $\min H(e)$, where the latter is the well-known minimum error entropy (MEE) criterion Erdogmus & Principe (2002) that has been extensively investigated in signal processing and process control.

Theorem B.1. *Minimizing $I(\mathbf{x}; e)$ is equivalent to minimizing error entropy $H(e)$.*

Proof. We have:

$$\begin{aligned} I(\mathbf{x}; e) &= H(e) - H(e|\mathbf{x}) \\ &= H(e) - H(e + f_\theta(\mathbf{x})|\mathbf{x}) \\ &= H(e) - H(y|\mathbf{x}), \end{aligned} \tag{10}$$

in which the second line is by the property that given two random variables ξ and η, then for any measurable function h, we have $H(\xi|\eta) = H(\xi + h(\eta)|\eta)$. Interested readers can refer to Cover (1999); MacKay (2003) for a detailed account of this property and its interpretation.

Therefore, $\min I(\mathbf{x}; e)$ is equivalent to $\min H(e)$. This is simply by the fact that the conditional entropy of y given $\mathbf{x}$, i.e., $H(y|\mathbf{x})$, is a constant that is purely determined by the training data (regardless of training algorithms). Note that, similar argument has also been claimed in stochastic process control Feng et al. (1997). □

B.2 Deeper Insights into MEE against non-Gaussian Noises

The principle on the reason why encouraging the distribution of prediction residual e is statistically independent of that of input $\mathbf{x}$ is robust against covariate shift has been discussed thoroughly in Greenfeld & Shalit (2020). As a complement, we just present here two insights on the robustness of $\min H(e)$ over the mean square error (MSE) criterion $\min E(e^2)$ against non-Gaussian noises, in which E denotes the expectation. Interested readers can refer to Chen et al. (2009; 2010; 2016); Hu et al. (2013) for more thorough analysis on the advantage of $\min H(e)$.

First, (Chen et al., 2009, Theorem 3) suggests that $\min E(e^2)$ is equivalent to minimizing the error entropy plus a KullbackCLeibler (KL) divergence. Specifically, we have:

$$\min E(e^2) \Leftrightarrow \min H(e) + D_{\mathrm{KL}}(p(e)\|\varphi(e)), \tag{11}$$

in which $p(e)$ is the probability of error, $\varphi(e)$ denotes a zero-mean Gaussian distribution. As the KL-divergence is always nonnegative, minimizing the MSE is equivalent to minimizing an upper bound of the error entropy. Eq. (11) also explains (partially) why in linear and Gaussian cases, the MSE and MEE are equivalent Kalata & Priemer (1979). Nevertheless, in case the error or noise follows a highly non-Gaussian distribution (especially when the signal-to-noise (SNR) value decreases), the MSE solution deviates from the MEE result, but the latter takes full advantage of high-order information of the error Chen et al. (2016).

On the other hand, given the mean-square error $E(e^2)$, the error entropy satisfies Cover (1999):

$$H(e) \leq \max_{E(\zeta^2)=E(e^2)} H(\zeta) = \frac{1}{2} + \frac{1}{2}\log 2\pi + \frac{1}{2}\log\left(E(e^2)\right), \tag{12}$$

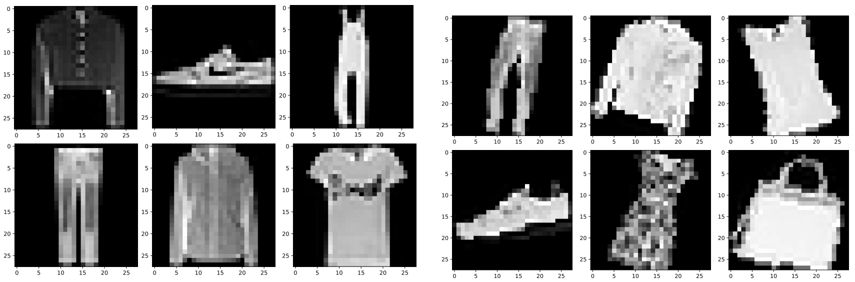
where ζ denotes a random variable whose second moment equals to $E(e^2)$. This implies that the MSE criterion can be recognized as a robust MEE criterion in the minimax sense, because:

$$\begin{aligned} f^*_{\text{MSE}} &= \arg\min_{f\in F} E(e^2) \\ &= \arg\min_{f\in F} \frac{1}{2} + \frac{1}{2}\log 2\pi + \log\left(E(e^2)\right) \\ &= \arg\min_{f\in F} \max_{E(\zeta^2)=E(e^2)} H(\zeta), \end{aligned} \tag{13}$$

where f^*_{MSE} denotes the solution with MSE criterion, $\mathcal{F}$ stands for the collection of all Borel measurable functions. Eq. (13) suggests that minimizing the MSE is equivalent to minimizing an upper bound of the error entropy.

B.3 Network Architecture on Fashion-MNIST

The source and target images for Fashion MNIST is shown in Fig. 4.



(a)

(b)

Figure 4: (a) Fashion MNIST source images; (b) Fashion MNIST target images

The neural network architecture is set as: there are 2 convolutional layers (with, respectively, 16 and 32 filters of size 5×5) and 1 fully connected layers. We add batch normalization and max-pooling layer after each convolutional layer. We choose ReLU activation, batch size 128 and the Adam optimizer Kingma & Ba (2014).

B.4 Additional Results on Synthetic Data

We additionally demonstrate experimental results on synthetic data to demonstrate the robustness of I_α against covariate shift and non-Gaussian noise.

We test on synthetic data with a linear model. The underlying model being considered here is $y = \beta^T\mathbf{x} + \epsilon$ where β is a 100 dimension weight vector, which is drawn from a zero mean Gaussian distribution with standard deviation $\sigma = 0.1$. In the training phase, $\mathbf{x}$ is drawn from a uniform distribution over $[-1, 1]$. To introduce covariate shift in the target test set, we change the marginal distribution of input samples from a uniform distribution to a standard Gaussian distribution. We test with ϵ drawn from two non-Gaussian noises: Laplacian noise and shifted exponential with the form $\epsilon = 1 - e$ where e is drawn from an exponential distribution $exp(1)$. In any case, ϵ is drawn independently from $\mathbf{x}$. The number of training samples is $1,024$ and the test set size is 128.

Table 2: Mean square error for Synthetic data.

Method	Laplace noise		Shifted-exponential noise	
	Source	Target	Source	Target
MSE	1.440 ± 0.013	2.225 ± 0.007	0.813 ± 0.004	1.659 ± 0.006
HSIC	1.463 ± 0.005	2.223 ± 0.021	$\mathbf{0.779 \pm 0.005}$	1.639 ± 0.002
$H_\alpha(e)$	1.450 ± 0.008	$\mathbf{2.193 \pm 0.006}$	0.790 ± 0.009	$\mathbf{1.621 \pm 0.007}$
$I_\alpha(\mathbf{x}; e)$	$\mathbf{1.420 \pm 0.030}$	2.217 ± 0.013	0.781 ± 0.004	1.637 ± 0.009

We use the simple linear neural network with one layer, the number of input units is 100, the output unit is 1. We compare the performance of four loss functions: MSE, HSIC, MIC (i.e., $I_\alpha(\mathbf{x}; e)$) and MEE (i.e., $H_\alpha(e)$). For all of them, we choose Adam optimizer Kingma & Ba (2014) with mini-batch size 32. For MIC, we set α to 2 and $\sigma_\mathbf{x} = \sigma_e = 1$. We repeat each experiment with 10 independent runs. Table 2 summarizes the mean square error for different loss functions. In both cases, $H_\alpha(e)$ enjoys the least prediction error in target data set, but $I_\alpha(\mathbf{x}; e)$ always performs better than HSIC and MSE.

C Supplementary Material to Proposal II

C.1 Experimental Setup in Adversarial Attack

In our experiment, we use VGG16 Simonyan & Zisserman (2015) as the baseline network and compare the performance of VGG16 trained by DIB objective and other regularizations. Again, we view the last fully connected layer before the softmax layer as the bottleneck layer. All models are trained with 400 epochs, a batch-size of 100, and an initial learning rate 0.1. The learning rate was reduced by a factor of 10 for every 100 epochs. We use SGD optimizer with weight decay $5e$-4. We explored β ranging from $1e$-4 to 1, and found that 0.01 works the best.

C.2 Robustness on Image Classification

CIFAR-10 is an image classification dataset consisting of $32 \times 32 \times 3$ RGB images of 10 classes. As a common practice, We use $10k$ images in the training set for hyper-parameter tuning. Test error rates with different methods are shown in Table 3. As can be seen, VGG16 trained with our DIB outperforms other regularizations and also the baseline ResNet50. We also observed, surprisingly, that VIB does not provide performance gain in this example, even though we use the authors' recommended value of β (0.01).

Table 3: Test error (%) on CIFAR-10

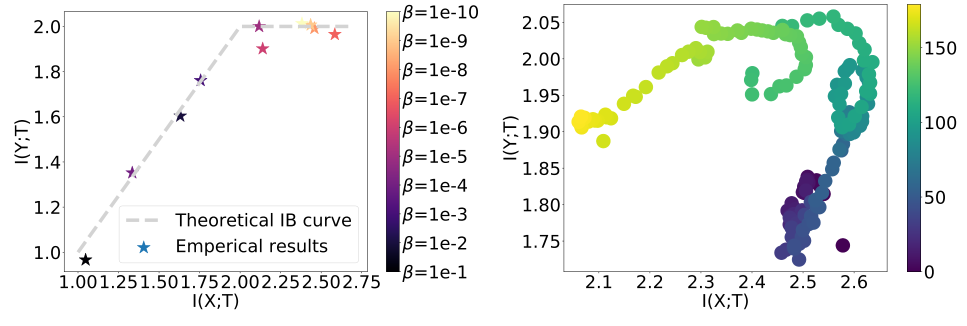
Model	**Test(%)**
VGG16	7.36
ResNet18	6.98
ResNet50	6.36
VGG16+Confidence Penalty	5.75
VGG16+Label smoothing	5.78
VGG16+VIB	9.31
VGG16+DIB (β=1e-2)	**5.66**

C.3 Information Bottleneck (IB) Curve

Given two random variables X and Y, and a bottleneck variable T. IB obeys the Markov condition that $I(X;T) \geq I(Y;T)$ based on the data processing inequality (DPI) Cover (1999), meaning that the bottleneck variable cannot contain more information about Y than it does about X.

According to Kolchinsky et al. (2019), the IB curve in classification scenario is piecewise linear and becomes a flat line at $I(Y;T) = H(Y)$ for $I(X;T) \geq H(Y)$. We obtain both theoretical and empirical IB curve by training a three layer MLP with 256 units in the bottleneck layer on MNIST dataset, as shown in Fig. 5(a). As we can see, when β is approaching to 0, we place no constraint on the **minimality**, a representation learned in this case is sufficient for desired tasks but contains too much redundancy and nuisance factors. However, if β is too large, we are at the risk of sacrificing the performance or representation **sufficiency**. Note that the region below the curve is feasible: for suboptimal mapping $p(t|x)$, solutions will lie in this region. No solution will lie above the curve.

We also plot a representative information plane Shwartz-Ziv & Tishby (2017) (i.e., the values of $I(X;T)$ with respect to $I(Y;T)$ across the whole training epochs) with β=1e-6 in Fig. 5(b). It is very easy to observe the mutual information increase ($a.k.a.$, fitting) phase, followed by the mutual information decrease ($a.k.a.$, compression) phase. This result supports the IB hypothesis in DNNs.



(a)

(b)

Figure 5: (a) Theoretical (the dashed lightgrey line) and empirical IB curve found by maximizing the IB Lagrangian with different values of β; (b) a representative information plane for β=1e-6, different colors denote different training epochs.