Information Theory beyond Communications: Generalization, Explainability and Robustness

Pablo Piantanida

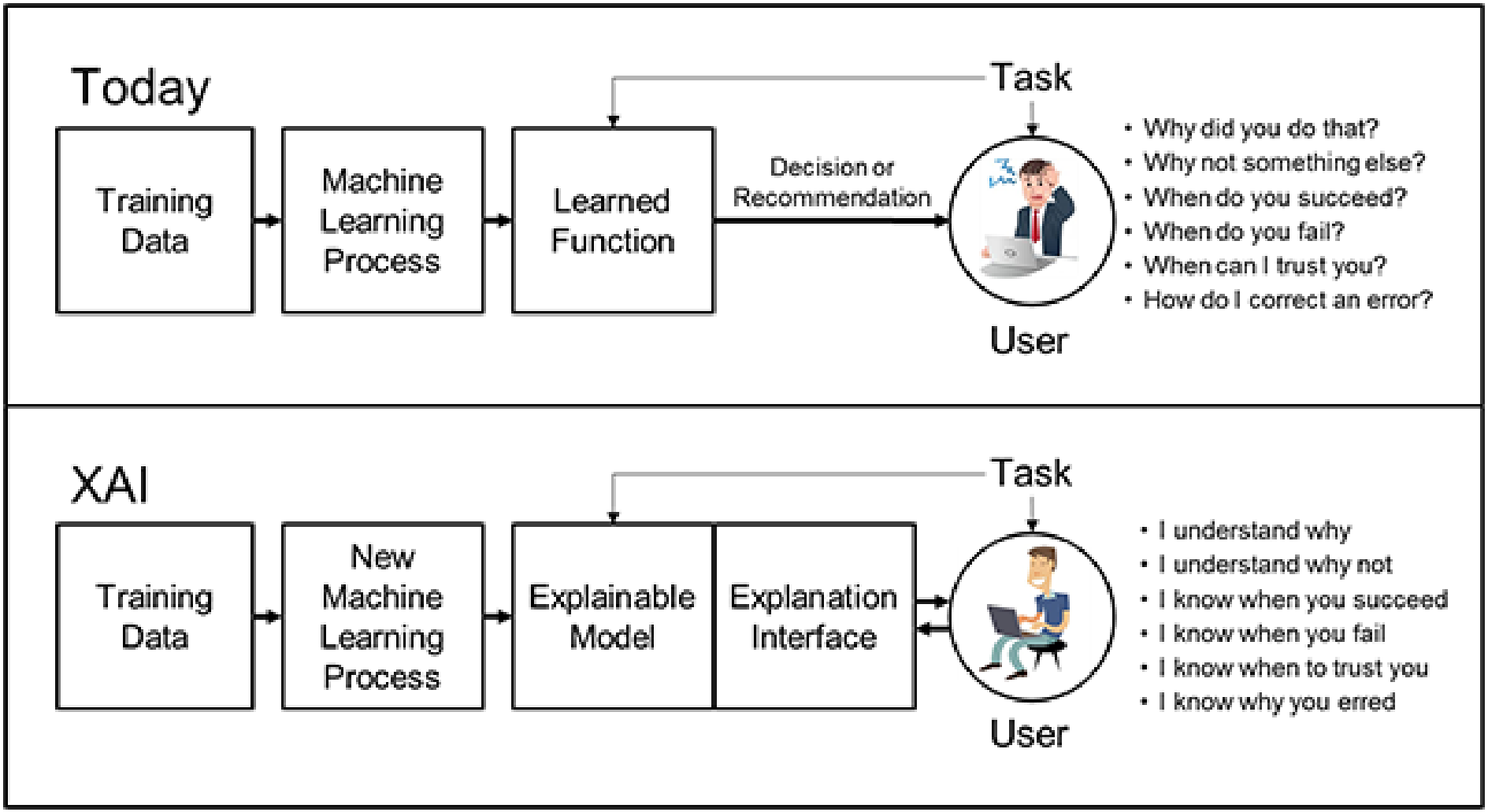
ppiantani@mila.quebec

Laboratoire des Signaux et Systèmes (L2S), CentraleSupélec-CNRS-Université Paris Sud &
Montreal Institute for Learning Algorithms (Mila), Université de Montréal

DATAIA - Workshop on Safety AI, Inria Saclay, Palaiseau, September 11th, 2019

Outline

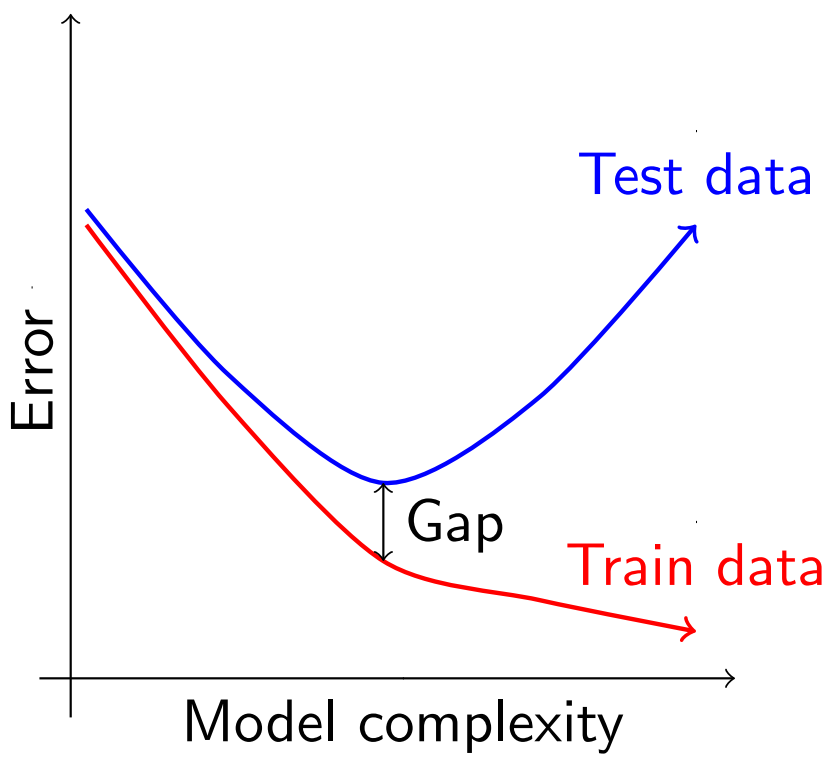
Motivation


Today
Training Data
Machine Learning Process
Learned Function
Decision or Recommendation
Task
User
• Why did you do that?
• Why not something else?
• When do you succeed?
• When do you fail?
• When can I trust you?
• How do I correct an error?
XAI
Training Data
New Machine Learning Process
Explainable Model
Explanation Interface
Task
User
• I understand why
• I understand why not
• I know when you succeed
• I know when you fail
• I know when to trust you
• I know why you erred

Objectives

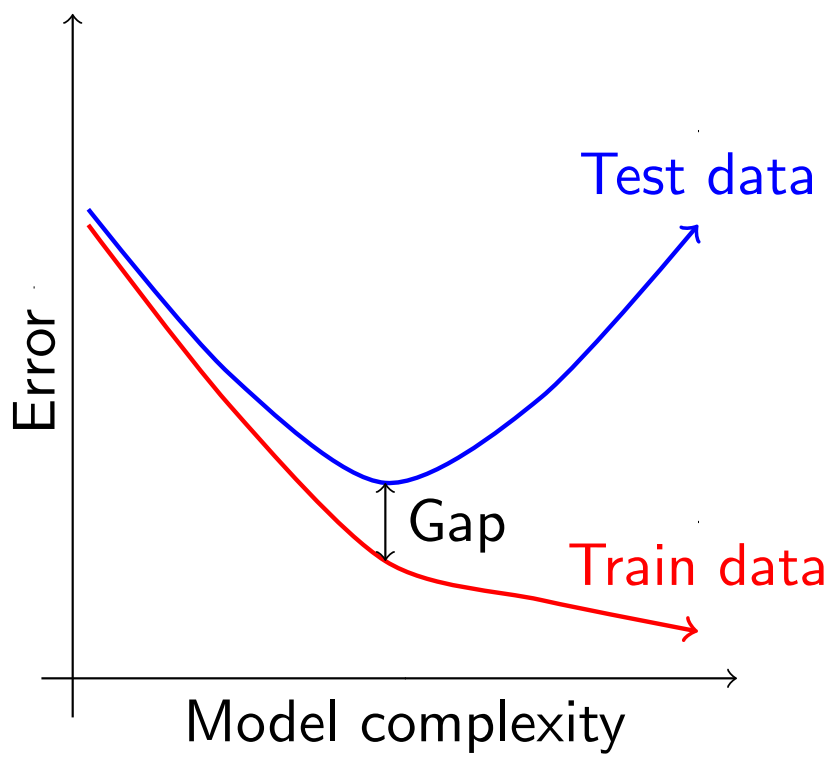
- The overall research goal here is to develop a fundamental theoretical understanding of modern machine-learning architectures through the lens of information concepts and quantities.
- Main goals:
  - **Improvement of existing algorithms and approaches:** we must ensure ML algorithms do not take on and amplify our biases present in the datasets, systems that pursue long-term goals, understanding generalization and beyond the training distribution, non-stationarities,...
  - **Explainability and intrinsic reliability:** knowing how an AI system arrives at an outcome is key to trust, designing predictors that can provide meaningful, calibrated notions of their uncertainty, that can explain their decisions...
  - **Safety and application related reliability:** we need to ensure the security and reliability of AI systems, exposing and fixing their vulnerabilities, identifying new attacks and defense, developing new metric to evaluate robustness...

Generalization beyond the Training Distribution



A better understanding of the information-theoretic principles of generalization beyond the training distribution (i.e., the observed distribution changes) may be fundamentally important to predict the uncertainty of an AI system and to know how it arrives at an outcome

Generalization beyond the Training Distribution



A better understanding of the information-theoretic principles of generalization beyond the training distribution (i.e., the observed distribution changes) may be fundamentally important to predict the uncertainty of an AI system and to know how it arrives at an outcome

How to control generalization?

- Regularization term: A weight penalty is included in the cost function;
- Noise injection: The generalization ability is improved adding noise;
- Implicit regularization: Deep learning algorithms reduce the generalization gap naturally.

Main Definitions

Definition (Classification rule and misclassification probability)

Let $Q_{\hat{Y}|X}$ be a classifier, the misclassification probability is defined by:

$$P_{\mathcal{E}}(Q_{\hat{Y}|X}) = 1 - \sum_{\forall\,(x,y)} Q_{\hat{Y}|X}(y|x) P_{XY}(x,y)$$

Definition (Randomized encoders and decoders)

The classifier can be divided into that of finding an encoder $Q_{U|X} \in \mathcal{F}_E$ (representation model) and a soft decoder $Q_{\hat{Y}|U} \in \mathcal{F}_D$ simultaneously:

$$Q_{\hat{Y}|X}(y|x) = \mathbb{E}_{q_{U|X}}\left[Q_{\hat{Y}|U}(y|U)|X = x\right],$$

Definition (Cross-entropy risk)

$$\mathcal{L}(Q) = \mathbb{E}_{P_{XY}Q_{U|X}}\left[-\log Q_{\hat{Y}|U}(Y|U)\right], \qquad Q \equiv \left(Q_{U|X}, Q_{\hat{Y}|U}\right)$$

Main Definitions

Definition (Classification rule and misclassification probability)

Let $Q_{\hat{Y}|X}$ be a classifier, the misclassification probability is defined by:

$$P_{\mathcal{E}}(Q_{\hat{Y}|X}) = 1 - \sum_{\forall\,(x,y)} Q_{\hat{Y}|X}(y|x) P_{XY}(x,y)$$

Definition (Randomized encoders and decoders)

The classifier can be divided into that of finding an encoder $Q_{U|X} \in \mathcal{F}_E$ (representation model) and a soft decoder $Q_{\hat{Y}|U} \in \mathcal{F}_D$ simultaneously:

$$Q_{\hat{Y}|X}(y|x) = \mathbb{E}_{q_{U|X}}\left[Q_{\hat{Y}|U}(y|U)|X = x\right],$$

Definition (Cross-entropy risk)

Data distribution is unknown

$$\mathcal{L}(Q) = \mathbb{E}_{P_{XY}Q_{U|X}}\left[-\log Q_{\hat{Y}|U}(Y|U)\right], \qquad Q \equiv \left(Q_{U|X}, Q_{\hat{Y}|U}\right)$$

Main Definitions (Cont'd.)

Definition (Empirical risk)

Let $\hat{P}_{XY}$ denote the empirical distribution over the random training set $\mathcal{S}_n = \{(X_1, Y_1) \cdots (X_n, Y_n)\}$. The empirical risk is defined by:

$$\mathcal{L}_{\text{emp}}(Q, \mathcal{S}_n) = \mathbb{E}_{\hat{P}_{XY} Q_{U|X}} \left[-\log Q_{\hat{Y}|U}(Y|U) \right]$$

Main Definitions (Cont'd.)

Definition (Empirical risk)

Let $\hat{P}_{XY}$ denote the empirical distribution over the random training set $\mathcal{S}_n = \{(X_1, Y_1) \cdots (X_n, Y_n)\}$. The empirical risk is defined by:

$$\mathcal{L}_{\text{emp}}(Q, \mathcal{S}_n) = \mathbb{E}_{\hat{P}_{XY} Q_{U|X}} \left[-\log Q_{\hat{Y}|U}(Y|U) \right]$$

Lemma (Optimality of empirical decoders)

$$\mathcal{L}_{\text{emp}}(Q_{U|X}, Q_{\hat{Y}|U}, \mathcal{S}_n) \geq \mathcal{L}_{\text{emp}}(Q_{U|X}, \hat{Q}_{Y|U}, \mathcal{S}_n) \qquad a.e.,$$

where $\hat{Q}_{Y|U}(y|u) = \frac{\sum_{x \in \mathcal{X}} Q_{U|X}(u|x) \hat{P}_{XY}(x,y)}{\sum_{x \in \mathcal{X}} Q_{U|X}(u|x) \hat{P}_X(x)}$

Main Definitions (Cont'd.)

Definition (Empirical risk)

Let $\hat{P}_{XY}$ denote the empirical distribution over the random training set $\mathcal{S}_n = \{(X_1, Y_1) \cdots (X_n, Y_n)\}$. The empirical risk is defined by:

$$\mathcal{L}_{\text{emp}}(Q, \mathcal{S}_n) = \mathbb{E}_{\hat{P}_{XY} Q_{U|X}} \left[-\log Q_{\hat{Y}|U}(Y|U) \right]$$

Lemma (Optimality of empirical decoders)

$$\mathcal{L}_{\text{emp}}(Q_{U|X}, Q_{\hat{Y}|U}, \mathcal{S}_n) \geq \mathcal{L}_{\text{emp}}(Q_{U|X}, \hat{Q}_{Y|U}, \mathcal{S}_n) \qquad a.e.,$$

where $\hat{Q}_{Y|U}(y|u) = \frac{\sum_{x \in \mathcal{X}} Q_{U|X}(u|x) \hat{P}_{XY}(x,y)}{\sum_{x \in \mathcal{X}} Q_{U|X}(u|x) \hat{P}_X(x)}$

Definition (Generalization Gap)

$$\mathcal{E}_{\text{gap}}(Q, \mathcal{S}_n) = |\mathcal{L}(Q) - \mathcal{L}_{\text{emp}}(Q, \mathcal{S}_n)|$$

Information Theoretic Bounds on the Gap

- Gap bounds: The goal is to find the learning rate ϵ_n satisfying

$$\mathbb{P}\left(\mathcal{E}_{\text{gap}}(Q, \mathcal{S}_n) > \epsilon_n(Q, \mathcal{S}_n, \gamma_n)\right) \leq \gamma_n$$

- PAC style bounds: Compute the sample dependent ϵ_n such that

$$\mathbb{P}\left(\mathcal{L}(Q) \leq \mathcal{L}_{\text{emp}}\left(Q, \mathcal{S}_n\right) + \epsilon_n(Q, \mathcal{S}_n, \gamma_n)\right) \geq 1 - \gamma_n$$

- Regularized risk: We study properties and implications for related algorithms that minimize:

$$\mathcal{L}_{\text{emp}}\left(Q, \mathcal{S}_n\right) + \lambda \cdot \epsilon_n(Q, \mathcal{S}_n, \gamma_n), \qquad \lambda \geq 0$$

Outline

Information-Theoretic Bound

Theorem

For each $Q_{U|X} \in \mathcal{F}_E$ and $Q_{\hat{Y}|U} \in \mathcal{F}_D$,

$$\mathcal{E}_{\text{gap}}(Q, \mathcal{S}_n) \leq \inf_{K \in \mathbb{N}} 2\epsilon(K) + A_\delta \sqrt{\mathcal{I}(p_X; q_{U|X})} \cdot \frac{\log(n)}{\sqrt{n}} r(K) + \frac{D_\delta \cdot \mathcal{D}_{\text{HL}}\left(Q^D_{Y|U} \| Q_{\hat{Y}|U} | q^D_U\right) + C_\delta}{\sqrt{n}} + \mathcal{O}\left(\frac{\log(n)}{n}\right),$$

with probability at least $1 - \delta$ over the choice of $\mathcal{S}_n \sim P^n_{XY}$.

Information-Theoretic Bound

Theorem

For each $Q_{U|X} \in \mathcal{F}_E$ and $Q_{\hat{Y}|U} \in \mathcal{F}_D$,

$$\mathcal{E}_{\text{gap}}(Q, \mathcal{S}_n) \leq \inf_{K \in \mathbb{N}} 2\epsilon(K) + A_\delta \sqrt{\mathcal{I}(p_X; q_{U|X})} \cdot \frac{\log(n)}{\sqrt{n}} r(K) + \frac{D_\delta \cdot \mathcal{D}_{\text{HL}}\left(Q^D_{Y|U} \| Q_{\hat{Y}|U} | q^D_U\right) + C_\delta}{\sqrt{n}} + \mathcal{O}\left(\frac{\log(n)}{n}\right),$$

with probability at least $1 - \delta$ over the choice of $\mathcal{S}_n \sim P^n_{XY}$.

$\mathcal{D}_{\text{HL}}$ is the Hellinger distance

$$\mathcal{D}_{\text{HL}}\left(Q^D_{Y|U} \| Q_{\hat{Y}|U} | q^D_U\right) = \sqrt{\frac{1}{2} \cdot \mathbb{E}_{q^D_U}\left[\sum_{y \in \mathcal{Y}} \left(\sqrt{Q_{\hat{Y}|U}(y|U)} - \sqrt{Q^D_{Y|U}(y|U)}\right)^2\right]}$$

Information-Theoretic Bound (Cont'd.)

constants are defined as

- $A_\delta := \sqrt{2}B_\delta$,
- $B_\delta := \left(1 + \sqrt{\log\left(\frac{|\mathcal{Y}|+4}{\delta}\right)}\right)$,
- $C_\delta := 2\mathrm{Vol}\left(\mathcal{U}\right)e^{-1} + B_\delta\sqrt{|\mathcal{Y}|}\log\left(\frac{\mathrm{Vol}(\mathcal{U})}{P_Y(y_{\min})}\right)$,

Information-Theoretic Bound (Cont'd.)

constants are defined as

- $A_\delta := \sqrt{2}B_\delta$,
- $B_\delta := \left(1 + \sqrt{\log\left(\frac{|\mathcal{Y}|+4}{\delta}\right)}\right)$,
- $C_\delta := 2\mathrm{Vol}\left(\mathcal{U}\right)e^{-1} + B_\delta\sqrt{|\mathcal{Y}|}\log\left(\frac{\mathrm{Vol}(\mathcal{U})}{P_Y(y_{\min})}\right)$,
- $D_\delta = Q_{\hat{Y}|U}^{-1/4}(y_{\min}|u_{\min})\sqrt{8\frac{|\mathcal{Y}|+4}{\delta}}$; and

$$\epsilon(K) = \sup_{\substack{k,x,y:\\ 1\le k\le K\\ y\in\mathcal{Y}\\ x\in\mathcal{K}_k^{(y)}}} \left|\ell(x,y) - \ell(x^{(k,y)},y)\right|, \quad r(K) = \frac{1}{\min\limits_{\substack{k,y:\\ 1\le k\le K\\ y\in\mathcal{Y}}} \int_{\mathcal{K}_k^{(y)}} p_X(x)dx}.$$

- $\left(\{\mathcal{K}_k^{(y)}\}_{k=1}^K, \{x^{(k,y)}\}_{k=1}^K\right)_{y\in\mathcal{Y}}$ are $|\mathcal{Y}|$ partitions of $\mathcal{X}$ with resp. centroids, s.t. for each $y \in \mathcal{Y}$: $\bigcup_{k=1}^K \mathcal{K}_k^{(y)} = \mathcal{X}$, $\mathcal{K}_i^{(y)} \cap \mathcal{K}_j^{(y)} = \emptyset$ $\forall 1 \le i < j \le K$; $q_U^D(u)$ and $Q_{Y|U}^D(y|u)$ are distributions functions induced by the quantization of $p_{XY}(x,y)$ by these partitions.

Information-Theoretic Bound (Cont'd.)

- Encoder and decoder are given but the bound can be further averaged w.r.t randomness of the training samples and the algorithm itself.
- $\mathcal{I}(p_X; q_{U|X})$: Mutual information between raw data X and its representation U is a measure of information complexity.
- $\mathcal{D}_{\mathsf{HL}}\left(Q^D_{Y|U} \| Q_{\hat{Y}|U}\right)$: Hellinger distance is a measure of the decoder efficiency w.r.t. the decoder $Q^D_{Y|U}$, induced by $q_{U|X}$ and the quantized testing distribution. $Q_{\hat{Y}|U} = Q^D_{Y|U}$ makes this term zero.

Information-Theoretic Bound (Cont'd.)

- Encoder and decoder are given but the bound can be further averaged w.r.t randomness of the training samples and the algorithm itself.
- $\mathcal{I}(p_X; q_{U|X})$: Mutual information between raw data X and its representation U is a measure of information complexity.
- $\mathcal{D}_{\mathsf{HL}}\left(Q^D_{Y|U} \| Q_{\hat{Y}|U}\right)$: Hellinger distance is a measure of the decoder efficiency w.r.t. the decoder $Q^D_{Y|U}$, induced by $q_{U|X}$ and the quantized testing distribution. $Q_{\hat{Y}|U} = Q^D_{Y|U}$ makes this term zero.
- $\epsilon(K)$ and $r(K)$: These functions define, for $y \in \mathcal{Y}$, an artificial discretization of $\mathcal{X}$ into cells. While $\epsilon(K)$ is associated with the robustness of the loss over the partition element, $r(K)$ is the minimum probability of falling into a cell.
- Tradeoff: $\epsilon(K)$ is a decreasing function (when the number of cells is increased), $r(K)$ is increasing (smaller cells enclose less probability).
- Vol($\mathcal{U}$): If ReLU activations are implemented (the volume is limited for bounded entries), Vol($\mathcal{U}$) will be larger than for the case of sigmoid activations, and the mutual information.

Outline

Experimental Framework

Goal

Our experiments show a relationship between $\mathcal{E}_{\text{gap}}(Q, \mathcal{S}_n)$ and $\sqrt{\mathcal{I}(\hat{P}_X; Q_{U|X})}$, which potentially implies that the gap is characterized from training samples.

Experimental Framework

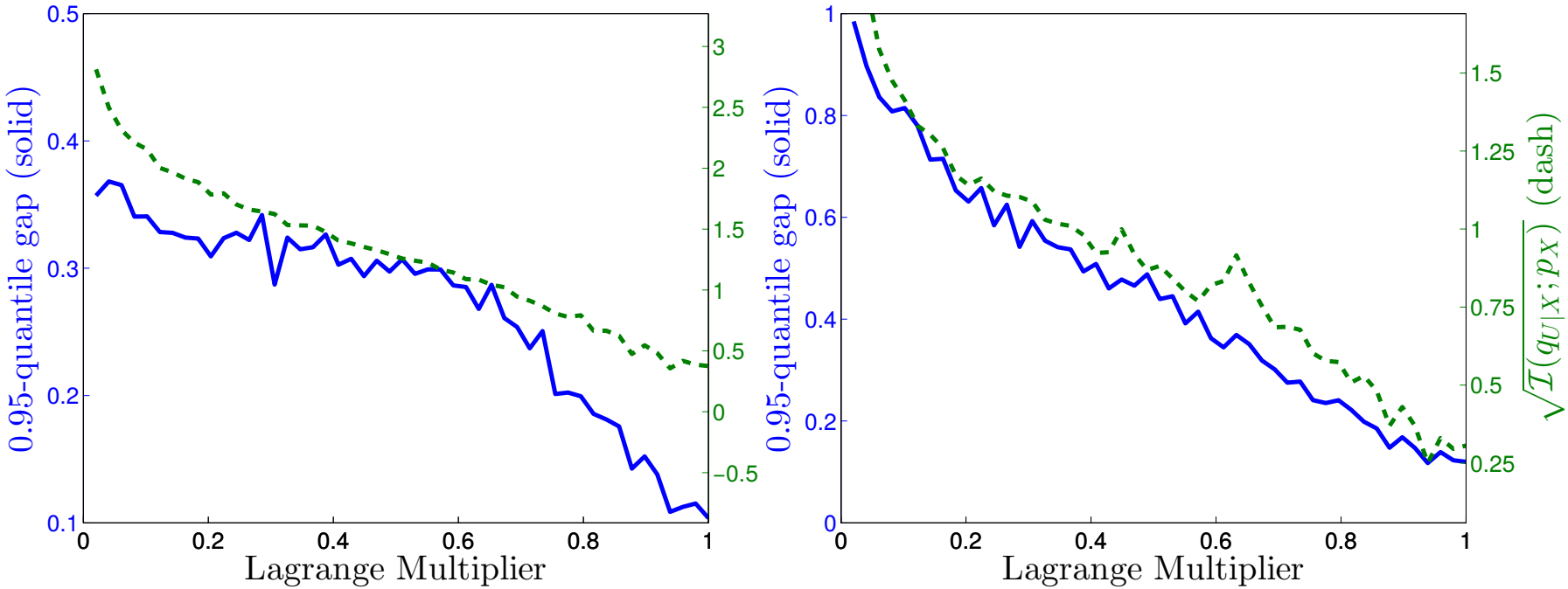
Goal

Our experiments show a relationship between $\mathcal{E}_{\text{gap}}(Q, \mathcal{S}_n)$ and $\sqrt{\mathcal{I}(\hat{P}_X; Q_{U|X})}$, which potentially implies that the gap is characterized from training samples.

Technical details

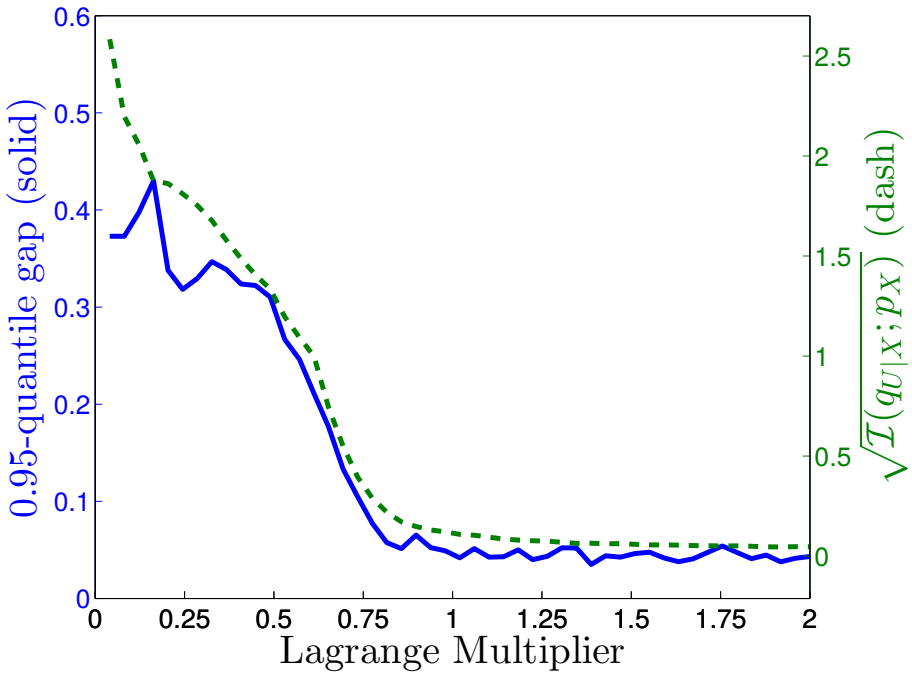
- Training data set: 5K MNIST and 5K CIFAR-10;
- Early stopping in the misclassification error according a random selection of the validation set composed of 500 samples;
- 3-layer feed-forward with based on Gaussian and Log-Normal Variational Auto-Encoders (VAEs);
- Random translations are drawn from an uniform distribution between -5 and 5 (quantized) for each axis and random rotations are drawn from an uniform distribution $(-\frac{\pi}{4}, \frac{\pi}{4})$ for the angle;
- Experiments are repeated several times and then averaged.

Comparison on MNIST: $\mathcal{E}_{\text{gap}}$ vs MI variational bound

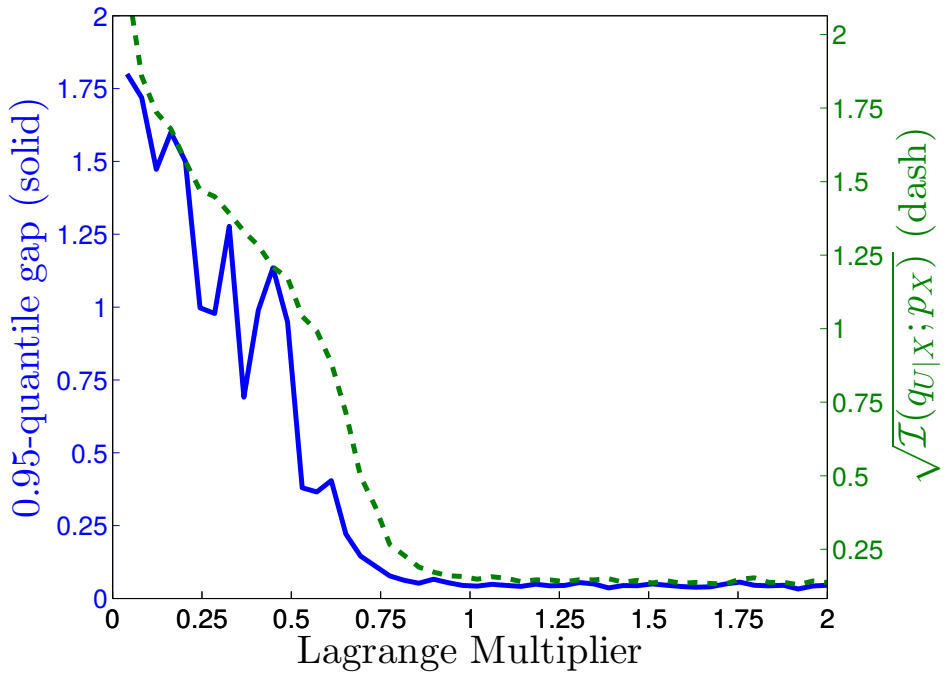


Comparison between 0.95-quantile of $\mathcal{E}_{\text{gap}}$ and the mutual information upper bound for normal encoder and testing with: (a) Images generated from training distribution, (b) Images generated with other distribution.

Comparison on MNIST: $\mathcal{E}_{\text{gap}}$ vs MI variational bound



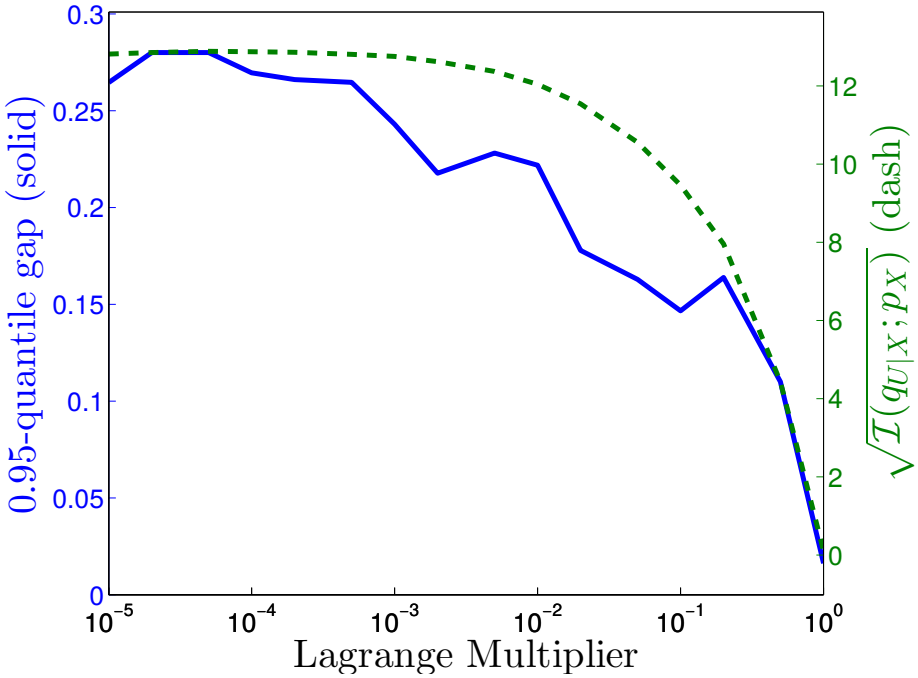
(a)



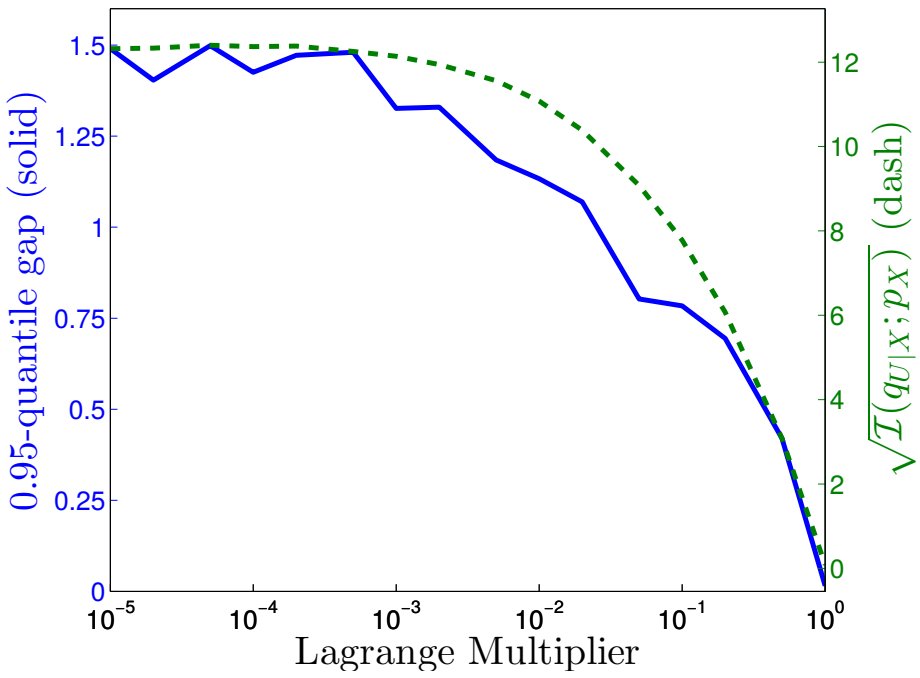
(b)

Comparison between 0.95-quantile of $\mathcal{E}_{\text{gap}}$ and the mutual information variational bound for Log-Normal encoder and testing with: (a) Images generated from training distribution, (b) Images generated with other distribution.

Comparison on MNIST: $\mathcal{E}_{\text{gap}}$ vs MI variational bound



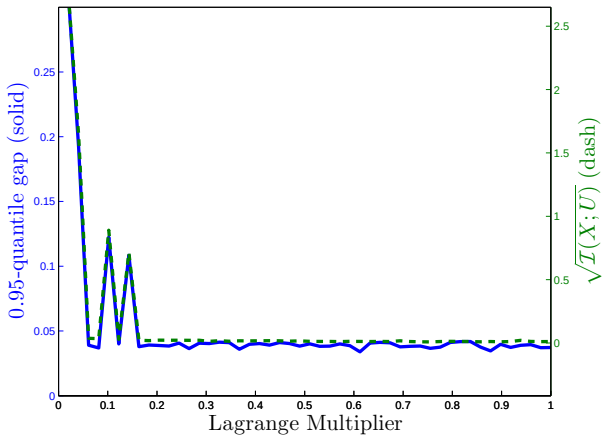
(a)



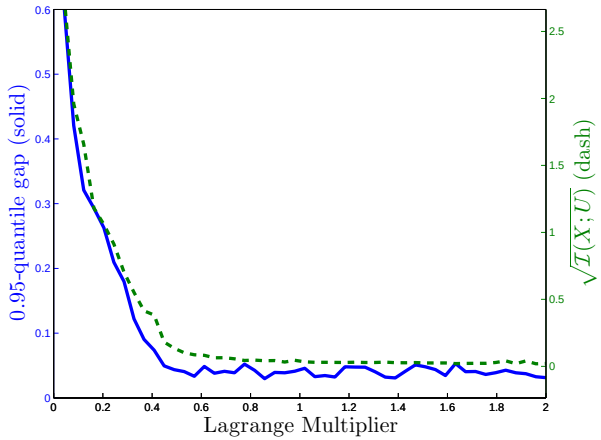
(b)

Comparison between 0.95-quantile of $\mathcal{E}_{\text{gap}}$ and the mutual information variational bound for RBM encoder and testing with: (a) Images generated with the training distribution, (b) Images generated with other distribution.

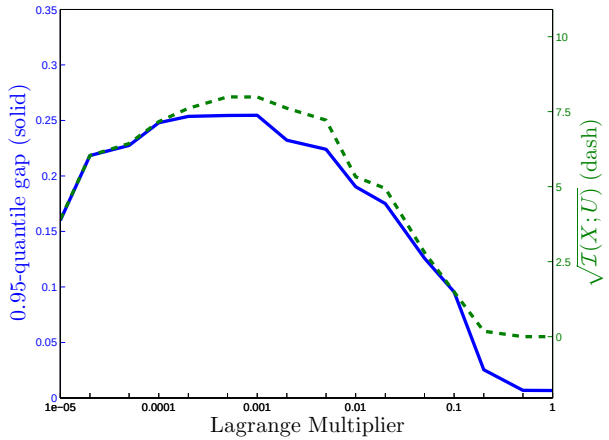
Comparison on CIFAR-10: $\mathcal{E}_{\text{gap}}$ vs MI variational bound



(a)



(b)



(c)

Comparison between 0.95-quantile of $\mathcal{E}_{\text{gap}}$ and the mutual information variational bound for CIFAR-10 dataset and: (a) Normal encoder, (b) LogNormal encoder, (c) RBM encoder.

Outline

Summary and Open Problems

- We presented a theoretical investigation of a typical classification task in which we have training data from a source domain, but we wish the testing gap measured w.r.t. a possible different probability law to be as small as possible.
- Our main result is that this gap can be bounded by the mutual information between the input testing samples and the corresponding representations and the Hellinger distance which measures the decoder efficiency and other less relevant constants.
- Empirical study suggests that the mutual information may be a good measure to capture the dynamic of the gap with respect to important training parameters.
- Further work is needed to provide strong support to these numerical results in presence of other sources of non-stationarities between training and testing datasets.

On-going Work and Perspectives

- Learning the dynamic of information measures on high-dimensional data, i.e., differential entropy, KL divergence and MI, based on variational bounds on information measures and gradient descent over neural networks.
- Statistical confidence bounds for the parametric estimation of those information measures have been recently derived for the case of MI and entropies with at least one discrete variable.
- These tools may have a major impact in various problems:
  - Maximizing entropy of generative models, e.g., increasing diversity of generative models,
  - Simple methods for detecting dataset shift, measuring uncertainty, explaining decisions,
  - Investigating dependences in various problems, e.g., to study causal effects in networks,
  - Learning disentangled representations, e.g., improving unsupervised methods,
  - ...

Outline

Learning the Dynamic of Mutual Information

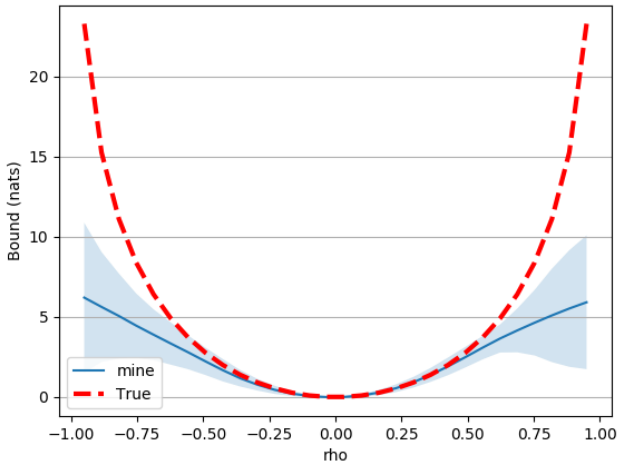
- We study the simple case of two multivariate correlated gaussians random variables $X, Z \in \mathbb{R}^n$, where:

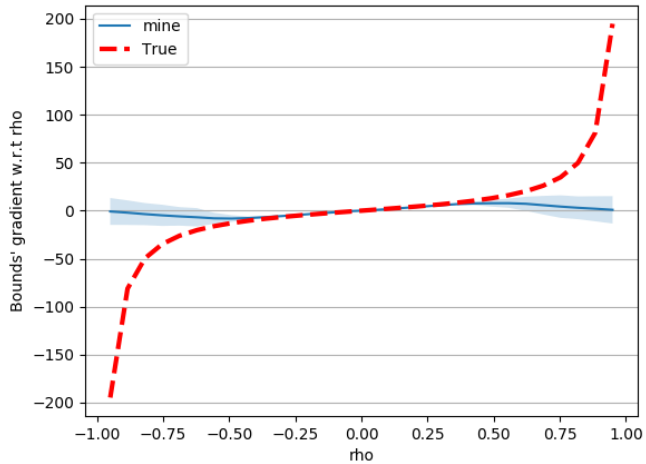
$$\begin{bmatrix} X \\ Z \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} I_n & \rho I_n \\ \rho I_n & I_n \end{bmatrix}\right)$$

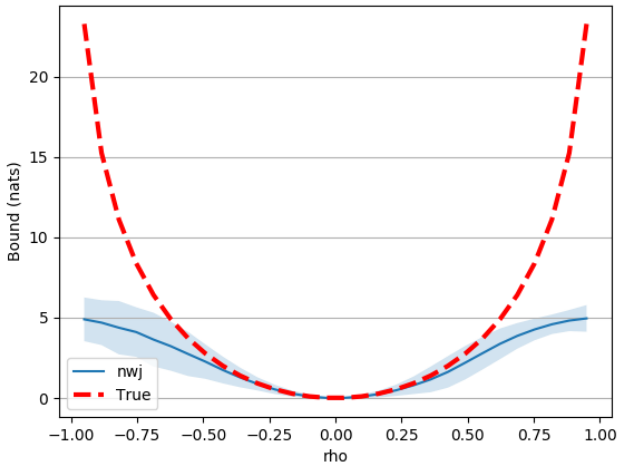
- With ρ controlling the amount of correlation.
- We train the bounds from scratch for 30 values of ρ equally spaced in $[-0.95, 0.95]$ where the differential mutual information is given by:

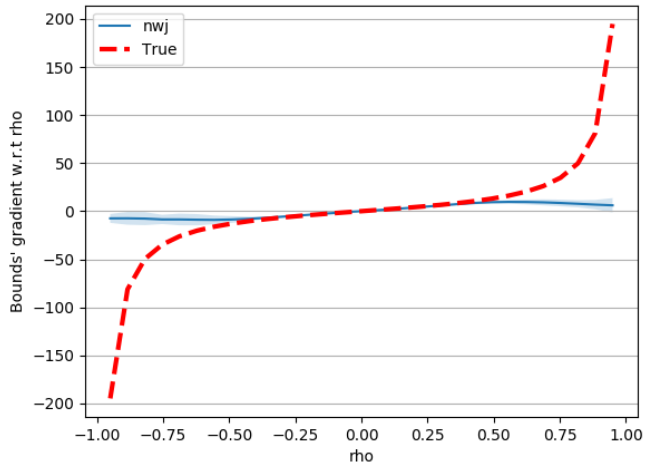
$$\begin{aligned} \mathcal{I}(X;Z) &= \mathcal{H}(X) + \mathcal{H}(Z) - \mathcal{H}(X,Z) \\ &= \frac{1}{2}\log((2\pi e)^n) + \frac{1}{2}\log(2\pi e)^n) - \frac{1}{2}\log((2\pi e)^{2n}|\Sigma|). \end{aligned}$$

Learning the Dynamic of Mutual Information (Cont'd.)



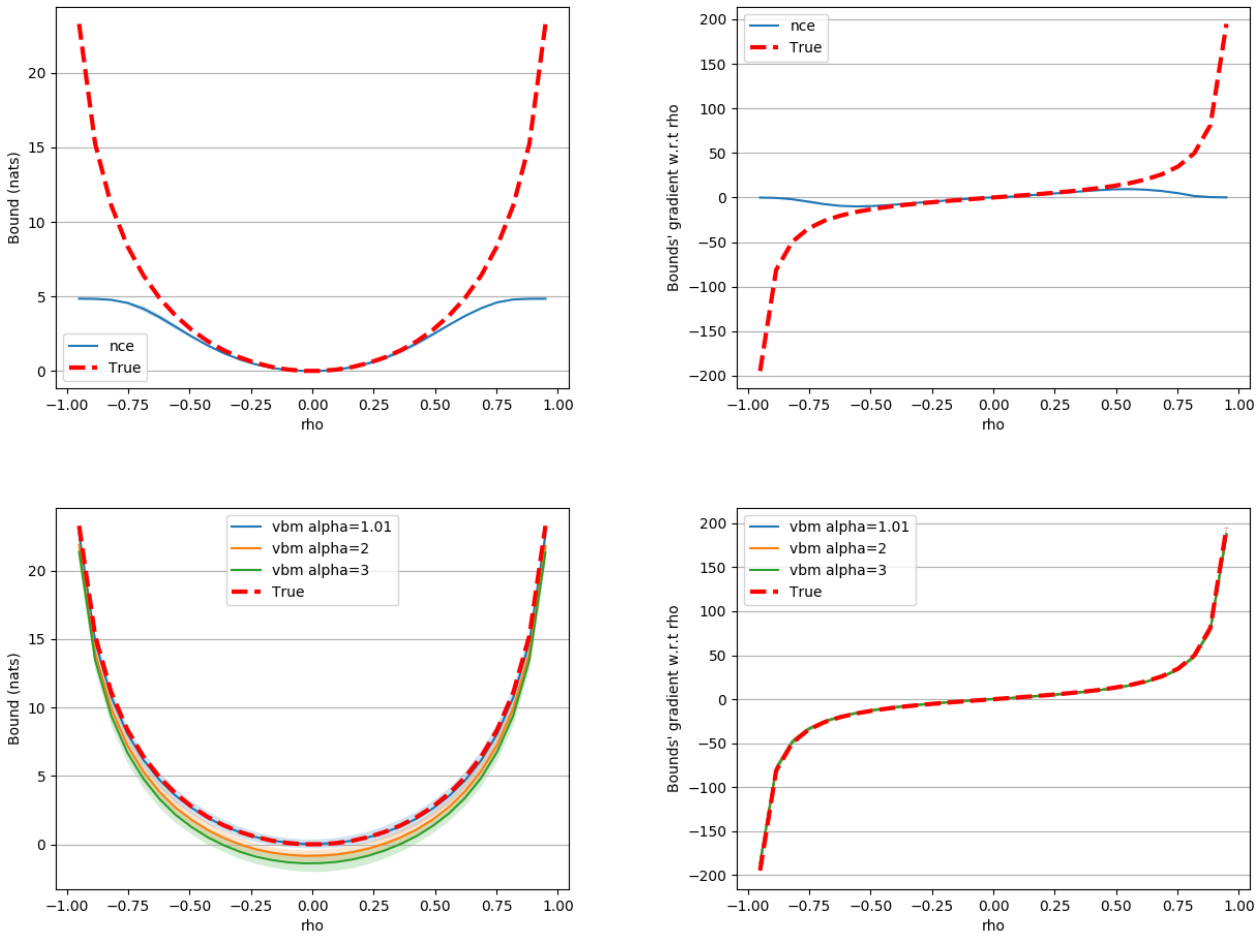






Plots for MI $\mathcal{I}(X;Z)$, for $X, Z \in \mathbb{R}^{20}$. *Left:* Estimated value $\hat{\mathcal{I}}(X;Z)$. *Right:* Estimated gradient $\nabla_\rho \hat{\mathcal{I}}(X;Z)$ *From Top to Bottom:* MINE and NWJ. Results are averaged over 10 seeds, and shadowed areas represent 95% of confidence.

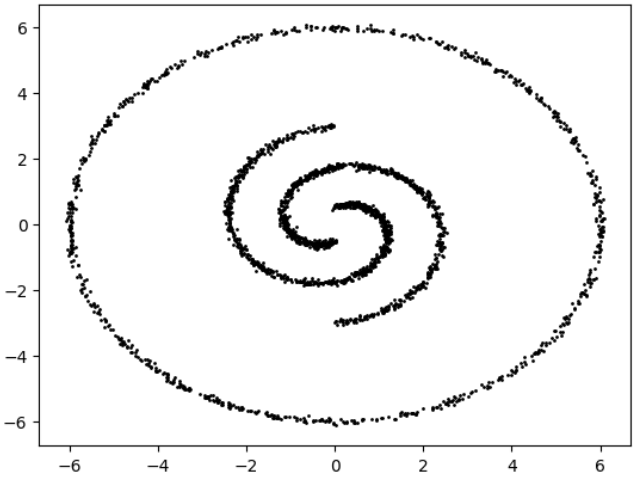
Learning the Dynamic of Mutual Information (Cont'd.)



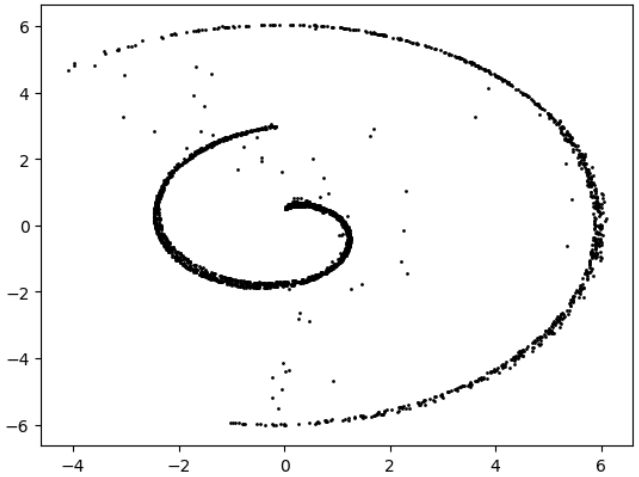
Plots for MI $\mathcal{I}(X;Z)$, for $X, Z \in \mathbb{R}^{20}$. *Left:* Estimated value $\hat{\mathcal{I}}(X;Z)$. *Right:* Estimated gradient $\nabla_\rho \hat{\mathcal{I}}(X;Z)$ *From Top to Bottom:* NCE and VBM (novel). Results are averaged over 10 seeds, and shadowed areas represent 95% of confidence.

GAN experiments

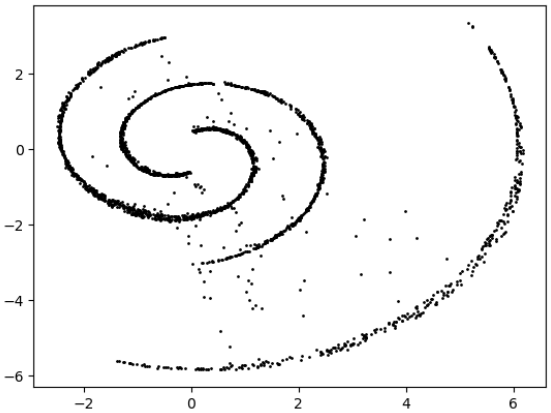
Target



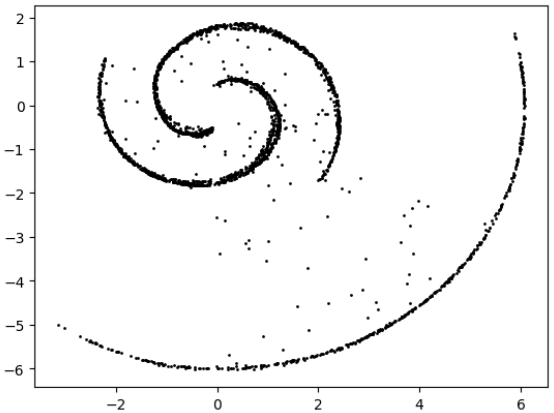
GAN



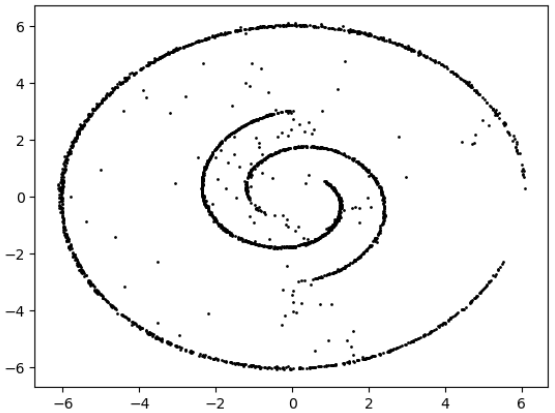
GAN + NWJ



GAN + MINE



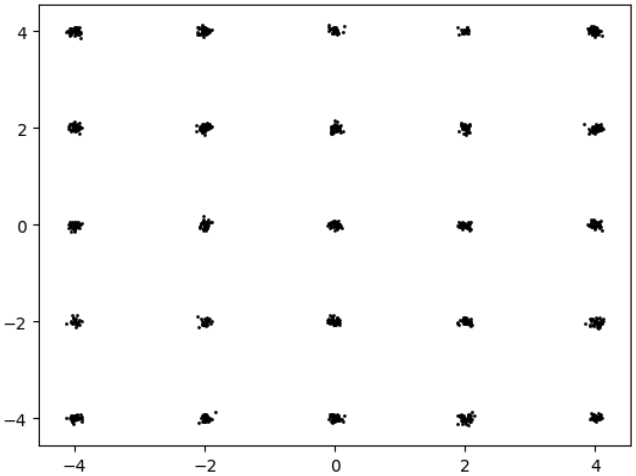
GAN + VBM



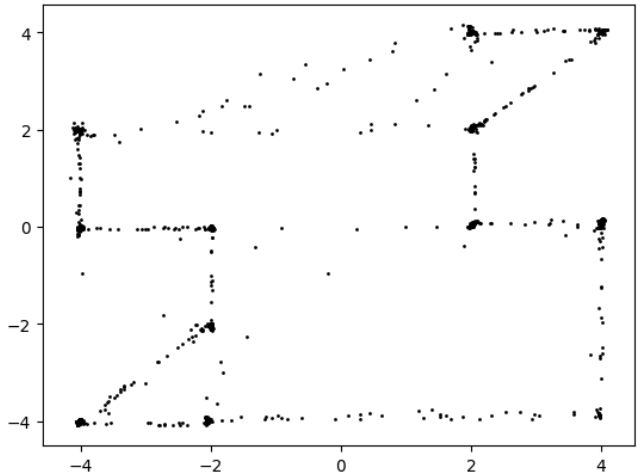
View of 2500 generated samples after 400 epochs of training on the same randomly chosen seed.

GAN experiments (Cont'd.)

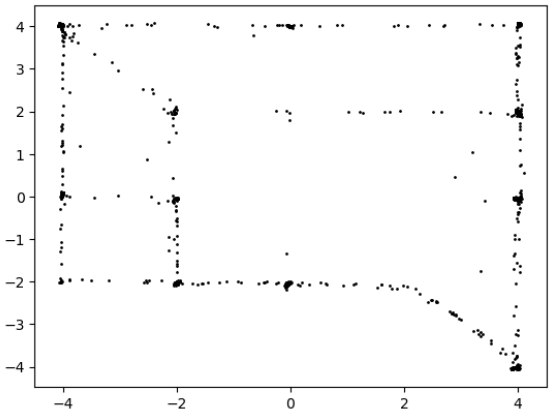
Target distribution



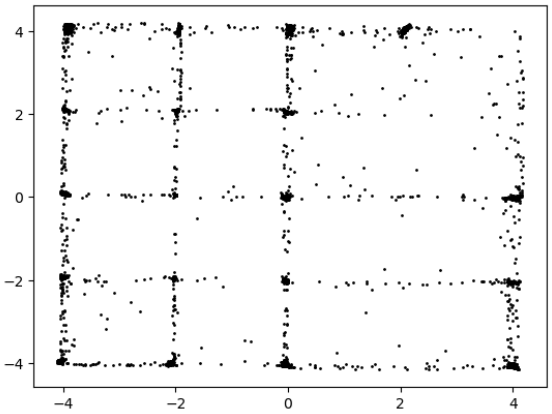
GAN



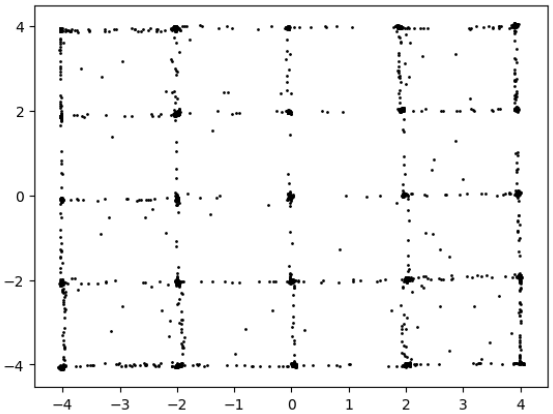
GAN + NWJ



GAN + MINE

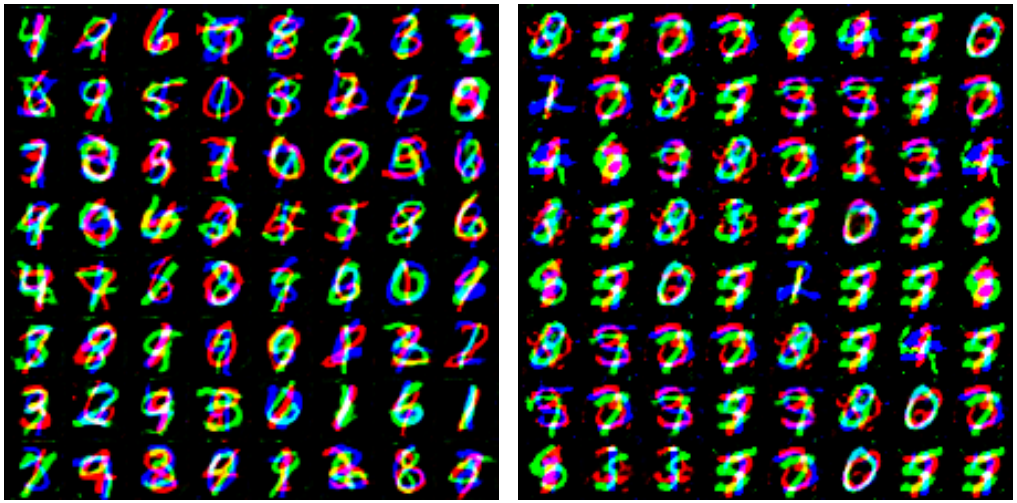


GAN + VBM



View of 2500 generated samples after 400 epochs of training on the same randomly chosen seed.

GAN experiments (Cont'd.)



Stacked4 MNIST: *Left:* VBM. Out of the 4 channels, we only plot 3 channels in oder to keep clear RGB images.*Right:* MINE.

Stacked4 MNIST		
Method	Modes (Max 10000)	KL
DCGAN	222.25 ± 85	$5.49 \pm 2.5e^{-1}$
DCGAN+MINE	5182 ± 4042	2.23 ± 2.33
DCGAN+VBM ($\alpha \approx 1$)	8858 ± 50	$0.29 \pm 1.0e^{-2}$
DCGAN+VBM ($\alpha = 1.8$)	8872 ± 61	$0.29 \pm 1.4e^{-2}$

Results on stacked4 MNIST dataset, averaged over 10 randomly chosen seeds.

Thank you for your attention !

https://arxiv.org/abs/1905.11972

Related Work

- Ishmael Belghazi, Sai Rajeswar, Aristide Baratin, R. Devon Hjelm, and Aaron C. Courville. **"MINE: mutual information neural estimation"**. CoRR, abs/1801.04062, 2018.
- Tu Nguyen, Trung Le, Hung Vu, and Dinh Phung. **"Dual discriminator generative adversarial nets"**. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Process- ing Systems 30, pages 2670â2680. Curran Associates, Inc., 2017.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. **"Representation learning with contrastive predictive coding"**. CoRR, abs/1807.03748, 2018.
- P. Piantanida and L. Rey Vega, **"Information Bottleneck and Representation Learning"**, Cambridge Press, 2019.
- N. Tishby, F. C. Pereira, and W. Bialek, **"The information bottleneck method,"** Allerton Conference on Communication, Control and Computing, 1999.
- O. Shamir, S. Sabato, and N. Tishby, **"Learning and generalization with the information bottleneck,"** Theoretical Computer Science, vol. 411, no. 29-30, pp. 2696-2711, 2010.
- Y. Bengio, A. Courville, and P. Vincent, **"Representation learning: A review and new perspective,"** IEEE Transactions on PAMI, vol. 35, no. 8, pp. 1798-1828, 2013.
- D. Kingma and M. Welling, **"Auto-encoding variational bayes,"** ICML, 2014.
- N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov **"Dropout: A simple way to prevent neural networks form overfitting,"** JMLR, vol. 15, no. 1, pp. 1929-1958, 2014.
- B. Neyshabur, R. Tomioka, and N. Srebro, **"In search of the real inductive bias: On the role of implicit regularization in deep learning,"** ICLR, 2015.
- D. Russo and J. Zou, **"How much does your data exploration overfit? Controlling bias via information usange,"** Arxiv, 2015.
- C. Zhang, S. Bengio, M. Hardt, B. Recht and O. Vinyals, **"Understanding deep learning requires rethinking generalization,"** Arxiv, 2016.
- A. Xu, and M. Raginsky, **"Information-theoretic analysis of generalization capability of learning algorithms,"** Arxiv, 2017.
- P. Bartlett, D. Foster, and M. Telgarsky, **"Spectrally-normalized margin bounds for neural networks,"** NIPS, 2017.
- M. Vera, L. Rey Vega, and P. Piantanida, **"Compression-based regularization with an application to multi-task learning,"** Arxiv, 2017.
- ...