

Reducing the Variance of Variational Estimates of Mutual Information by Limiting the Critic's Hypothesis Space to RKHS

International Conference on Pattern Recognition 2020

P Aditya Sreekar, Ujjwal Tiwari and Anoop Namboodiri

Center for Visual Information Technology
International Institute of Information Technology, Hyderabad

January 15, 2021

Mutual Information

Definition

Mutual information (MI) between two random variables X and Y , denoted by $I(X; Y)$ is defined as:

$$I(X; Y) = \int_{\mathcal{X} \times \mathcal{Y}} \log \frac{d\mathbb{P}_{XY}}{d\mathbb{P}_X \otimes \mathbb{P}_Y} d\mathbb{P}_{X,Y}$$

Where, \mathbb{P}_{XY} is the joint probability distribution and, \mathbb{P}_X and \mathbb{P}_Y are the corresponding marginal distributions.

- Mutual information is a fundamental information theoretic measure that quantifies the dependency between two random variables
- Mutual information, $I(X; Y)$ between any two RVs ranges from 0 to $+\infty$. $I(X; Y)$ is high when X and Y share considerable information or in other words have a high degree of dependency and vice-versa. It is equal to zero iff X and Y are mutually

Applications of Mutual Information

Mutual information maximization is used in:

- Calculating channel capacity in information theory
- Independent Component Analysis
- Representation learning
- Generative modeling
- Information bottleneck

Mutual Information Estimation

- When the distribution of both the RVs are available, MI can be directly computed using integration.
- This is not possible in real world scenarios where we have access only to samples from the distributions.
- Classical non-parametric MI estimators that used methods like binning, kernel density estimation and K-Nearest Neighbour based entropy estimation are computationally expensive, produce unreliable estimates, and do not conform to mini-batch based optimisation strategies.

Mutual Information Estimation

- Recent estimation methods couple neural networks with variational lower bounds of MI (Nguyen, Wainwright, and Jordan 2010; Donsker and Varadhan 1983) for differential and tractable estimation of MI.
- A critic parameterized as a neural network is trained to approximate unknown density ratios. The approximated density ratios are used to estimate different variational lower bounds of MI.
- These methods consider universal approximation property of the critic neural network to estimate tighter variational lower bounds of MI.

Mutual Information Estimation

- However, universal approximation ability of neural networks comes at the cost of neglecting the effect of critic's unbounded complexity on variational estimation of mutual information, which leads to unstable and highly fluctuating estimates.

Our Contributions

- We argue that these variational lower bound estimators exhibit high sensitivity to the complexity of critic's (Neural Network) hypothesis space when optimised using mini-batch stochastic gradient strategy.
- We use a data-driven measure of hypothesis space complexity called *Rademacher complexity* to bound the generalization error for variational lower bounds of MI. Using these bounds, it is shown that higher complexity of critic's hypothesis space leads to higher generalization error and hence high variance estimates.
- We construct critic's hypothesis space in a smooth family of functions, the *Reproducing Kernel Hilbert Space* (RKHS). This corresponds to learning a kernel using *Automated Spectral Kernel Learning* (ASKL) (Li, Liu, and Wang 2019)

Variational Estimates of Mutual Information

- Parametric probability distribution or critic f_θ with trainable parameters θ is optimised to approximate the likelihood density ratio between the joint and product of marginal distributions ($d\mathbb{P}_{XY}/d\mathbb{P}_X \otimes \mathbb{P}_Y$). The approximated density ratio is used for sample based estimation of MI.

Variational Estimates of Mutual Information

I_{MINE} and I_{NWJ} lower bounds can be derived from Tractable Unnormalized Barber and Argakov (TUBA) lower bound, I_{TUBA} , considering only constant positive baseline in (Poole et al. 2018), that is $a > 0$ in the I_{TUBA} formulation defined as:

$$I(X; Y) \geq I_{TUBA}(f_{\theta}) = \mathbb{E}_{\mathbb{P}_{XY}} [f_{\theta}(x, y)] - \frac{\mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Y} [e^{f_{\theta}(x, y)}]}{a} - \log(a) + 1 \quad (1)$$

I_{MINE} is formulated from I_{TUBA} by fixing the parameter a in the above equation as exponential moving average of $e^{f_{\theta}(x, y)}$ across mini-batches. Similarly, I_{NWJ} is formulated from I_{TUBA} by substituting the parameter $a = e$.

Variational Estimates of Mutual Information

For I_{JS} (Poole et al. 2018) and I_{SMILE} (Song and Ermon 2019) estimates, density ratio is estimated by maximizing GAN discriminator objective defined as:

$$\max_{\theta} \mathbb{E}_{\mathbb{P}_{XY}} [\log (\sigma(f_{\theta}(x, y)))] + \mathbb{E}_{\mathbb{P}_X \times \mathbb{P}_Y} [\log (1 - \sigma(f_{\theta}(x, y)))] \quad (2)$$

I_{JS} is obtained by plugging in $f_{GAN}^* + 1$ into I_{NWJ} lower bound, where f_{GAN}^* is the optimal critic from GAN optimization. I_{SMILE} is obtained by plugging in f_{GAN}^* in Donsker-Vardhan variational lower bound (Donsker and Varadhan 1983), I_{DV} , given by:

$$I(X; Y) \geq I_{DV}(f_{\theta}) = \mathbb{E}_{\mathbb{P}_{XY}} [f_{\theta}(x, y)] - \log \left(\mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Y} \left[e^{f_{\theta}(x, y)} \right] \right) \quad (3)$$

Automated Spectral Kernel Learning

Feature mapping, $\phi(x)$, of a reproducing kernel Hilbert space (RKHS) corresponding to a non-stationary kernel can be represented as following (Li, Liu, and Wang 2019):

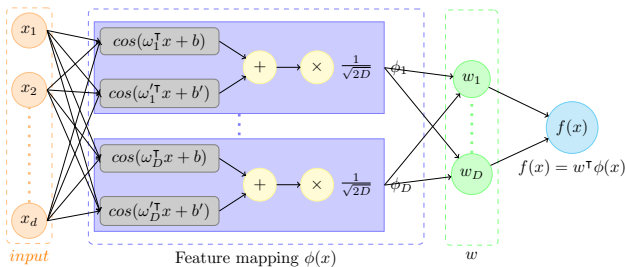
$$\phi(x) = \frac{1}{\sqrt{2D}} [\cos(\Omega^T x + b) + \cos(\Omega'^T x + b')] \quad (4)$$

b and b' are vectors of D uniform samples

$\{b_i\}_{i=1}^D, \{b'_i\}_{i=1}^D \stackrel{iid}{\sim} \mathcal{U}[0, 2\pi]$, and $\Omega = [\omega_1, \dots, \omega_D]$ and

$\Omega' = [\omega'_1, \dots, \omega'_D]$ are optimized during the training procedure. Any function f in the RKHS corresponding to $\phi(x)$ can be represented by $f(x) = w\phi(x)$, where w is D dimensional vector that is learned during training. Critics in this work are restricted to RKHS by using the above form.

Automated Spectral Kernel Learning



Theoretical Guarantees

Given n i.i.d samples, $\{x_i, y_i\}_{i=0}^n$ from joint distribution \mathbb{P}_{XY} and m i.i.d samples, $\{x'_i, y'_i\}_{i=0}^m$ from the product of marginal distributions $\mathbb{P}_X \otimes \mathbb{P}_Y$ empirical approximations of variational lower bounds of MI are defined as:

$$\hat{I}_{TUBA}^{n,m}(f_\theta, S) = \mathbb{E}_{\mathbb{P}_{XY}^n}[f_\theta(x, y)] - \frac{\mathbb{E}_{\mathbb{P}_X^m \otimes \mathbb{P}_Y^m}[e^{f_\theta(x, y)}]}{a} - \log(a) + 1 \quad (5)$$

$$\hat{I}_{DV}^{n,m}(f_\theta, S) = \mathbb{E}_{\mathbb{P}_{XY}^n}[f_\theta(x, y)] - \log\left(\mathbb{E}_{\mathbb{P}_X^m \otimes \mathbb{P}_Y^m}[e^{f_\theta(x, y)}]\right) \quad (6)$$

Theoretical Guarantees

Theorem (Generalization Error Bounds)

Assume, that the hypothesis space \mathcal{F} of the critic is uniformly bounded by M , that is

$|f(x, y)| \leq M \forall f \in \mathcal{F} \ \& \ \forall (x, y) \in \mathcal{X} \times \mathcal{Y}, M < \infty$. For a fixed $\delta > 0$ generalization errors of $\hat{I}_{TUBA}^{n,m}$ and $\hat{I}_{DV}^{n,m}$ can be bounded with probability of at least $1 - \delta$, given by

$$\sup_{f \in \mathcal{F}} \left(I_{TUBA}(f) - \hat{I}_{TUBA}^{n,m}(f) \right) \leq 4\hat{\mathcal{R}}_n(\mathcal{F}) + \frac{8}{a} e^M \hat{\mathcal{R}}_m(\mathcal{F}) + \frac{4M}{n} \log \left(\frac{4}{\delta} \right) + \frac{8Me^M}{am} \log \left(\frac{4}{\delta} \right) + \sqrt{\frac{\left(\frac{4M^2}{n} + \frac{(e^M - e^{-M})^2}{a^2 m} \right) \log \left(\frac{2}{\delta} \right)}{2}} \quad (7)$$

$$\sup_{f \in \mathcal{F}} \left(I_{DV}(f) - \hat{I}_{DV}^{n,m}(f) \right) \leq 4\hat{\mathcal{R}}_n(\mathcal{F}) + 8e^{2M} \hat{\mathcal{R}}_m(\mathcal{F}) + \frac{4M}{n} \log \left(\frac{4}{\delta} \right) + \frac{8Me^{2M}}{m} \log \left(\frac{4}{\delta} \right) + \sqrt{\frac{\left(\frac{4M^2}{n} + \frac{(e^{2M} - 1)^2}{m} \right) \log \left(\frac{2}{\delta} \right)}{2}} \quad (8)$$

Where, sample set S for $\hat{I}_{TUBA}^{n,m}$ and $\hat{I}_{DV}^{n,m}$ is assumed to be known, and $\hat{\mathcal{R}}_n(\mathcal{F})$ and $\hat{\mathcal{R}}_m(\mathcal{F})$ are empirical Rademacher averages of the hypothesis space \mathcal{F} for different sample sizes.

Theoretical Guarantees

Theorem (Rademacher Complexity)

The empirical Rademacher average of the RKHS \mathcal{F} to which ASKL critic belongs can be bounded as following

$$\hat{\mathcal{R}}_n(\mathcal{F}) \leq \frac{B}{n} \sqrt{\sum_{i=1}^n \|\phi(x_i)\|_2^2} \leq \frac{B}{\sqrt{n}} \quad (9)$$

Where $B = \sup_{f \in \mathcal{F}} \|w\|_2$.

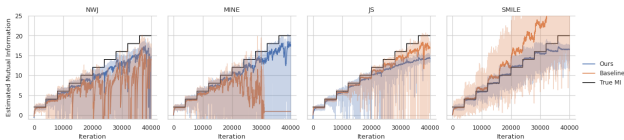
Training Methodology

The critic, f_θ , is optimized to simultaneously maximize empirical estimate of MI and minimize regularization terms defined below. The overall training objective is:

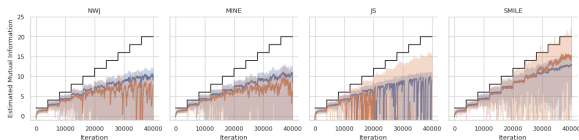
$$\operatorname{argmin}_{\theta} -\hat{I}(f_\theta, S) + \lambda_1 \|w\|_2 + \lambda_2 \|\phi(S; \theta)\|_F \quad (10)$$

Where, \hat{I} can be an empirical estimate of any variational lower bound of MI, $\hat{I}_{NWJ}^{n,m}$, $\hat{I}_{MINE}^{n,m}$, $\hat{I}_{JS}^{n,m}$ or $\hat{I}_{SMILE}^{n,m}$. And θ is the set of trainable parameters w , Ω , and Ω' . GAN discriminator objective is maximized in cases where \hat{I} is $\hat{I}_{JS}^{n,m}$ or $\hat{I}_{SMILE}^{n,m}$. Bias-variance tradeoff is controlled by tuning hyperparameters, λ_1 and λ_2 . We use mini-batch stochastic gradient decent to train the estimator.

Results



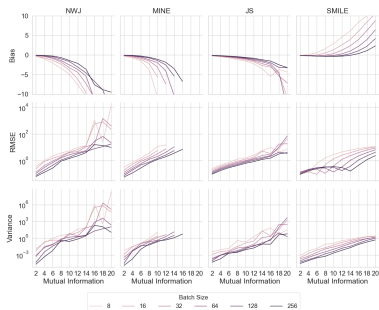
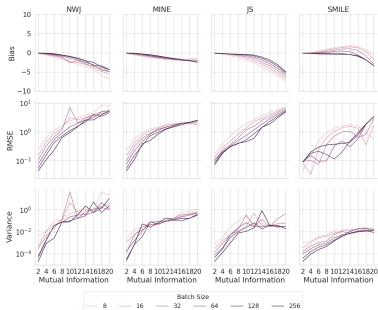
(a) Comparison on 20 dimensional correlated Gaussian dataset



(b) Comparison on cubed 20 dimensional correlated Gaussian dataset

Qualitative comparison between ASKL and baseline critic on four different variational lower bounds of MI, I_{NWJ} , I_{MINE} , I_{JS} , and I_{SMILE} . MI estimates on Gaussian correlated and cubed Gaussian correlated datasets are plotted in (a) and (b), respectively. MI estimate by the proposed ASKL critic are in blue and the estimates of baseline critic are depicted in orange.

Results



(a) Bias, variance and RMSE of ASKL (b) Bias, variance and RMSE of baseline critic estimates for different batch line critic estimates for different batch sizes.





Bias, variance, and RMSE values of ASKL critic and baseline critic estimates averaged over 50 experimental trials are shown in figures (a) and (b), respectively.

Code

- Please checkout our code at
https://github.com/blackPython/mi_estimator

Thank You

Literature I

-  Donsker, Monroe D and SR Srinivasa Varadhan (1983). “Asymptotic evaluation of certain Markov process expectations for large time. IV”. In: *Communications on Pure and Applied Mathematics* 36.2, pp. 183–212.
-  Li, Jian, Yong Liu, and Weiping Wang (2019). “Automated Spectral Kernel Learning”. In: *arXiv preprint arXiv:1909.04894*.
-  Nguyen, XuanLong, Martin J Wainwright, and Michael I Jordan (2010). “Estimating divergence functionals and the likelihood ratio by convex risk minimization”. In: *IEEE Transactions on Information Theory* 56.11, pp. 5847–5861.
-  Poole, Ben et al. (2018). “On variational lower bounds of mutual information”. In: *NeurIPS Workshop on Bayesian Deep Learning*.

Literature II



Song, Jiaming and Stefano Ermon (2019). “Understanding the Limitations of Variational Mutual Information Estimators”. In: *arXiv preprint arXiv:1910.06222*.