MUTUAL INFORMATION CONTINUITY-CONSTRAINED ESTIMATOR

Anonymous authors
Paper under double-blind review

ABSTRACT

The estimation of mutual information (MI) is vital to a variety of applications in machine learning. Recent developments in neural approaches have shown encouraging potential in estimating the MI between high-dimensional variables based on their latent representations. However, these estimators are prone to high variances owing to the inevitable outlier events. Recent approaches mitigate the outlier issue by smoothing the partition function using clipping or averaging strategies; however, these estimators either break the lower bound condition or sacrifice the level of accuracy. Accordingly, we propose Mutual Information Continuity-constrained Estimator (MICE). MICE alternatively smooths the partition function by constraining the Lipschitz constant of the log-density ratio estimator, thus alleviating the induced variances without clipping or averaging. Our proposed estimator outperforms most of the existing estimators in terms of bias and variance in the standard benchmark. In addition, we propose an experiment extension based on the standard benchmark, where variables are drawn from a multivariate normal distribution with correlations between each sample in a batch. The experimental results imply that when the i.i.d. assumption is unfulfilled, our proposed estimator can be more accurate than the existing approaches in which the MI tends to be underestimated. Finally, we demonstrate that MICE mitigates mode collapse in the kernel density estimation task.

1 INTRODUCTION

Mutual information (MI) estimation is essential in various machine learning applications, including learning representations (Oord et al., 2018; Chen et al., 2016; Bachman et al., 2019; Hjelm et al., 2018; Sordoni et al., 2021), feature selection (Battiti, 1994; Estévez et al., 2009), feature disentanglement (Higgins et al., 2018; Esmaeili et al., 2019; Colombo et al., 2021), and reinforcement learning (Oord et al., 2018; Bachman et al., 2019; Li et al., 2016). Some conventional non-parametric approaches have been proposed to estimate MI (Estévez et al., 2009; Fraser & Swinney, 1986; Moon et al., 1995; Kwak & Choi, 2002). Despite promising results, (Belghazi et al., 2018; Poole et al., 2019) indicated that these estimators have limited capability to scale up well with the sample size or dimension (Gao et al., 2015) therefore hard to be utilized in general purpose applications.

Recent studies focus on scalable MI estimation through variational bounds maximization (Oord et al., 2018; Belghazi et al., 2018; Poole et al., 2019) or minimization (Cheng et al., 2020) using neural networks or convex maximum-entropy method(Samo, 2021). These neural estimators have been adopted in some remarkable self-supervised applications, such as computer vision (Chen et al., 2020; He et al., 2020; Grill et al., 2020; Chen & He, 2020; Chen et al., 2021) and speech recognition (Schneider et al., 2019; Baevski et al., 2019), with the aim of maximizing the shared information between different views with respect to space or time. In MI estimation, the neural networks (also known as the *critics*) has been used to approximate the log-density ratio. These MI estimators generally characterize the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951) using a dual representation and subsequently formulate MI lower bounds.

Although multiple applications have attained promising results, two significant issues have not been fully addressed. As the first issue, the existing MI estimators can be debilitated by significant bias and variance owing to inevitable outlier events. It was pointed out by (Poole et al., 2019; Song & Ermon, 2020) that the exponential partition function causes a high-variance issue. It implies

that estimators leveraging f-divergence representations could suffer from the high-variance issue. Numerous studies have been conducted to address this problem. Previous approaches such as Mutual Information Neural Estimation (MINE) (Belghazi et al., 2018) and Contrastive Predictive Coding (CPC) (Oord et al., 2018) reduce the variances by adopting different types of averaging. Based on MINE, the Smoothed Mutual Information Lower-bound Estimator (SMILE) (Song & Ermon, 2020) limits the range of the critic with a hyper-parameter, enabling estimates with low bias and variance. For the second issue, as summarized in (Oord et al., 2018; Belghazi et al., 2018; Poole et al., 2019; Nguyen et al., 2010), most of the existing MI estimators are tested on a standard benchmark where random variables are drawn independently. However, the benchmark is insufficient for an analysis of videos or audio signals in which data frames could be correlated.

In this paper, we address the high variance issue by a novel Mutual Information Continuity-constrained Estimator (MICE) that constrains the Lipschitz constant of the critic by its spectral norm (Miyato et al., 2018), and we block the unstable gradients generated from the partition function. MICE is less underestimated in the extended benchmark because the partition function is smoothed by the scale of the spectral norm instead of hard clipping, which could overly restrict the range of the density ratio. The experimental results show that MICE has a competitive bias-variance trade-off compared to SMILE in the standard benchmark, without selecting a clipping threshold. Based on the standard benchmark, we propose an extension in which random variables are correlated within a batch. Our proposed method is robust when samples are not independent compared to existing variational estimators that underestimate MI drastically when slight correlations are involved. Finally, in the kernel density estimation (KDE) experiment, we demonstrate that using MICE as MI regularization alleviates mode collapse (Che et al., 2016; Dumoulin et al., 2016; Srivastava et al., 2017) in the training of generative adversarial networks (GANs) (Goodfellow et al., 2014). Our contributions are as follows:

- We address the high-variance issue of an existing unbiased estimator by constraining the Lipschitz constant of log-density ratio estimator and gradient stabilization.
- We prove that MICE is a strongly consistent estimator of MI.
- In the proposed experiment extension, the results show that MICE outperforms existing estimators under the condition in which the i.i.d. assumption is not fulfilled.
- A GAN regularized by MICE can capture more modes in the KDE experiment and ease the mode collapse problem.

2 Related Work

For a pair of random variables (X, Y) over the probability space $\mathcal{X} \times \mathcal{Y}$, the mutual information $I(X;Y)$ between X and Y can be defined as the KL divergence of the joint distribution $P_{(X,Y)}$ and the product of the marginals P_X and P_Y:

$$I(X;Y) = D_{\mathrm{KL}}(P_{(X,Y)} \| P_X \otimes P_Y) \tag{1}$$

where D_{KL} is the KL divergence. Next, we start with a common characterization of KL divergence, the Donsker–Varadhan (DV) representation (Donsker & Varadhan, 1983), which is adopted by MINE (Belghazi et al., 2018) and SMILE (Song & Ermon, 2020).

Lemma 1 (Donsker–Varadhan (DV)) *Given two probability distributions P and Q over $\mathcal{X}$:*

$$D_{\mathrm{KL}}(P\|Q) = \sup_{T:\mathcal{X}\to\mathbb{R}} \{\mathbb{E}_P[T] - \log \mathbb{E}_Q[e^T] \triangleq I_{\mathrm{DV}}\} \tag{2}$$

for some bounded function $T : \mathcal{X} \to \mathbb{R}$ such that the expectations are finite.

In particular, if P and Q are specified as $P_{(X,Y)}$ and $P_X \otimes P_Y$, MI can be estimated by maximizing the DV representation. It should be noted that the equation holds when $T = \log dP/dQ + C$ for some constant $C \in \mathbb{R}$.

In (Broniatowski & Keziou, 2009; Nowozin et al., 2016), a general variational estimation of f-divergences is introduced. For any convex, lower-semicontinuous function f, there exists a convex conjugate f^* such that $f(u) = \sup_{t\in\mathrm{dom}(f^*)}\{tu - f^*(t)\}$, where u belongs to the domain of f.

Therefore, f-divergences can be estimated by taking supremum over an arbitrary class of functions $T : \mathcal{X} \to \mathbb{R}$:

$$D_f(P\|Q) = \int_{\mathcal{X}} q(x) \sup_{t \in \mathrm{dom}(f^*)} \left\{ t\frac{p(x)}{q(x)} - f^*(t) \right\} dx \tag{3}$$

$$\geq \sup_{T:\mathcal{X}\to\mathbb{R}} \{\mathbb{E}_P[T] - \mathbb{E}_Q[f^*(T)]\} \tag{4}$$

The derivation form Equation 3 to Equation 4 is based on Jensen's inequality because the supremum is swapped out of the integration. Here, the KL divergence can be obtained by specifying $f(u) = u \log u$, thus $f^*(T) = e^{T-1}$, yielding the Nguyen-Wainright-Jordan (NWJ) lower bound (Nguyen et al., 2010). Similarly, MI can be estimated by setting $P = P_{(X,Y)}$ and $Q = P_X \otimes P_Y$.

Lemma 2 (Nguyen, Wainright, and Jordan (NWJ) (Nguyen et al., 2010)) *Given two probability distributions P and Q over $\mathcal{X}$,*

$$D_{\mathrm{KL}}(P\|Q) \geq \sup_{T_\theta:\mathcal{X}\to\mathbb{R}} \left\{ \mathbb{E}_P[T_\theta] - \mathbb{E}_Q[e^{T_\theta - 1}] \triangleq I_{\mathrm{NWJ}} \right\} \tag{5}$$

where the equation holds when $T_\theta = 1 + \log \frac{dP}{dQ}$.

Note that I_{NWJ} is unbiased since no nonlinear function is taken on the right-hand side out of the expectation. Although I_{DV} and I_{NWJ} are tight with a sufficient large hypothesis set of T_θ, the partition function induces large variances. The following approaches aim to solve the high-variance issue by averaging and clipping on the partition function. For instance, MINE (Belghazi et al., 2018) proposed a neural information measure based on taking supremum of I_{DV} over a neural network $T_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ parameterized by θ.

Lemma 3 (Mutual Information Neural Estimation (MINE) (Belghazi et al., 2018)) *Let P and Q be two probability distributions over $\mathcal{X}$*

$$I(X;Y) \geq \sup_{T_\theta:\mathcal{X}\to\mathbb{R}} \left\{ \mathbb{E}_{P_{(X,Y)}}[T_\theta] - \log \mathrm{EMA}\left(\mathbb{E}_{P_X \otimes P_Y}[e^{T_\theta}]\right) \triangleq I_{\mathrm{MINE}} \right\} \tag{6}$$

In this manner, MINE collects cross-batch statistics to evaluate bias-corrected estimate, reducing the bias and variance simultaneously. In contrast to MINE, which uses the exponential moving average (EMA) to reduce variances induced from the partition function, (Song & Ermon, 2020) proposed to reduce variances by putting limits on the range of the log-density ratio.

Lemma 4 (Smoothed Mutual Information Lower-bound Estimator (SMILE) (Song & Ermon, 2020)) *Let P and Q be two probability distributions over $\mathcal{X}$*

$$I(X;Y) \geq \sup_{T_\theta:\mathcal{X}\to\mathbb{R}} \left\{ \mathbb{E}_{P_{(X,Y)}}[T_\theta] - \log \mathbb{E}_{P_X \otimes P_Y}[e^{\max(\min(T_\theta, \tau), -\tau)}] \triangleq I_{\mathrm{SMILE}} \right\} \tag{7}$$

Another multi-sample estimator, Contrastive Predictive Coding (CPC) (Oord et al., 2018), uses the cross-entropy between the positive and negative samples as an objective

$$\mathbb{E}_{\Pi_j p(x_j, y_j)} \left[\frac{1}{n} \sum_{i=1}^{n} \log \frac{f(x_i, y_i)}{\frac{1}{n}\sum_{j=1}^{n} f(x_i, y_j)} \right] \triangleq I_{\mathrm{CPC}} \tag{8}$$

where $f(x, y) = e^{x^\top W y}$ is a log-bilinear function with a trainable parameter W, and the expectation is taken over the distribution with density $\Pi_j p(x_j, y_j)$. Noted that I_{CPC} is tight when $f(x, y) = \log p(y|x) + c(y)$, where $c(y)$ is an arbitrary function that depends on y. However, (Oord et al., 2018) indicated that this bound is loose when $I(X;Y) > \log n$, requiring an exponentially large batch size to achieve accurate estimates with high confidence (Song & Ermon, 2020).

3 Limitations on DV Representation

3.1 Maximum of Log-Density Ratio Estimate Dominating the Partition Function

According to (Poole et al., 2019; Song & Ermon, 2020), the partition function $\mathbb{E}_Q\left[e^{T_\theta(x,y)}\right]$ is the rationale behind high variances and biases. This expression is highly dependent on the maximum

of the log-density ratio in a batch. We demonstrate this by showing the relationship of LogSumExp (LSE, also known as a smooth approximation to the maximum function) operation and the maximum function as follows

$$\begin{aligned} \mathrm{LSE}(T_\theta(x_1, y_1), \ldots, T_\theta(x_n, y_{n-1})) &> \max\{T_\theta(x_1, y_1), \ldots, T_\theta(x_n, y_{n-1})\} \\ \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n-1} e^{T_\theta(x_i, y_j)} &> \frac{1}{n(n-1)} e^{\max\{T_\theta(x_1, y_1), \ldots, T_\theta(x_n, y_{n-1})\}} \end{aligned} \tag{9}$$

where $T_\theta(x_i, y_j)$ is the estimated log-density ratio $\log dP/dQ$ where x and y are drawn from Q. Note that because Q is the product of marginals, the total number of T_θ sampled from Q is $n(n-1)$. (McAllester & Stratos, 2020) indicated that the partition function is dominated by extremely rare events which are never observed through the sampling from $P_X \otimes P_Y$. They quantified the probability of outlier events using the outlier risk lemma.

Lemma 5 (Outlier risk lemma (McAllester & Stratos, 2020)) *Given n samples ($n \geq 2$) that follow the distribution P_X and a property $\Phi[x]$ such that $P_X(\Phi[x]) \leq 1/n$, the probability that no sample x satisfies $\Phi[x]$is at least 1/4.*

Here, $P_X(\Phi[x])$ is the probability of drawing x from P_X such that statement $\Phi[x]$ holds. Lemma 5 can be easily proved based on the probability of sampling with replacement.

Letting $P = P_{(X,Y)}$ and $Q = P_X \otimes P_Y$, for DV representation, the best estimate of MI is established when

$$\mathbb{E}_P[T_\theta(x, y)] = I(X; Y) \tag{10}$$

$$\mathbb{E}_Q[e^{T_\theta(x,y)}] = 1 \tag{11}$$

The outlier risk lemma indicates that there is at least a probability of $1/4$ that one can draw an unseen variable such that $\mathbb{E}_Q[e^{T_\theta(x,y)}] > 1$. By observing Equation 9, if a pair of unseen variables (x', y') were sampled, the partition function will be larger than $e^{T_\theta(x',y')}/(n(n-1))$; therefore, the estimates of DV representation are of high bias and variance. Similarly, the best estimate of I_{NWJ} is established with the same Equation 10, but Equation 11 should be modified as $\mathbb{E}_Q[e^{T_\theta(x,y)-1}] = 1$.

3.2 Neither Upper Bound nor Lower Bound Estimators

Based on the aforementioned limitations of the DV representation, the I_{MINE} and I_{SMILE} focus on controlling the variance of the partition function. I_{MINE} reduces the variance by applying EMA to the partition function over all previous samples. According to (McAllester & Stratos, 2020), the worst case of the DV representation can be bounded under $\log n$. Because I_{MINE} implicitly enlarges the batch size with the scale of iteration (i.e, the number of covered samples at the i^{th} iteration is $i \times n$, where n is the batch size), it can leverage the linearly increasing batch size to reduce the bias issue. Another method adopted by I_{SMILE} is controlling the range of the partition function by clipping the log-density ratio with a threshold τ in Equation 7.

In (Song & Ermon, 2020), the clipped density ratio $r_\tau = \max(\min(e^{T_\theta(x,y)}, e^\tau), e^{-\tau})$ is estimated by n random variables over the distribution $Q = P_X \otimes P_Y$. The variance of the bounded partition function $\mathbb{E}_Q[r_\tau]$ satisfies $\mathrm{Var}[\mathbb{E}_Q[r_\tau]] \leq (e^\tau - e^{-\tau})^2/4n$. According to (Song & Ermon, 2020), a trade-off of the bias and variance can be determined by a threshold τ. Decreasing τ reduces the variance, but increases bias with such choice.

Although these estimators mitigate the high-variance issue and attain more accurate estimates, they are no longer upper or lower bounds on MI. This is because the modified partition function is no longer a normalizing term. As MINE applies EMA to $\mathbb{E}_Q\left[e^{T_\theta(x,y)}\right]$ across batches, and there is at least $1/4$ chance that the outlier event occurs, the partition function eventually saturates at $e^{T_{\max}}/(4N^2 - N))$, where $T_{\max}$ is the maximum among all T_θ, and N is the amount of training data. As the range of the partition function of I_{SMILE} is limited within $[e^{-\tau}, e^\tau]$, the MI would be overestimated when the log-density ratio is larger than τ and would not be underestimated only if $\tau \to 0$ because of Equation 11.

In a nutshell, although these neither upper bound nor lower bound estimators reached more accurate MI estimates than I_{DV}, these estimators could overestimate MI to some unknown extent as they are

not guaranteed to be bounded below the MI. Moreover, I_{MINE} requires a large batch size to avoid from yielding large errors; in addition, the development of a criterion of selecting a proper threshold for I_{SMILE} is also challenging.

4 METHODOLOGY

4.1 MUTUAL INFORMATION CONTINUITY-CONSTRAINED ESTIMATOR

To alleviate the issue of outlier events dominating the partition function, we adopt two strategies, which are limiting the Lipschitz constant of the log-density ratio estimator and gradient stabilization. The core idea of reducing variances is to smooth the critic. For instance, I_{MINE} and I_{CPC} adopt averaging in different manners on the partition function to achieve a trade-off between the bias and variance, and I_{SMILE} directly truncates the value of density ratio using a hyper-parameter. Clearly, these approaches have certain flaws in that averaging leads to high bias, and it requires prior knowledge to choose the proper thresholds for clipping. To avert these issues, we utilize the spectral normalization that constrains the spectral norm of the parameters in the last layer, and consequently smooth the partition function. In (Miyato et al., 2018), the spectral norm of a weight matrix W is defined as

$$\sigma(W) \coloneqq \max_{h:h\neq 0} \frac{\|Wh\|_2}{\|h\|_2} = \max_{\|h\|_2\leq 1} \|Wh\|_2 \tag{12}$$

where h denotes any non-zero vector. The spectral norm $\sigma(W)$ is equivalent to the largest singular value of W. Therefore, $\sigma(W)$ is independent from h, so the preconceptions regarding the data is no longer required. For the weight matrix W^l in the l^{th} layer of T^l, spectral normalization normalizes W_l with its spectral norm

$$W^l_{\mathrm{SN}} \coloneqq \frac{W^l}{\sigma(W^l)} \tag{13}$$

where W^l_{SN} is the normalized weight matrix such that $\|T^l\|_{\mathrm{Lip}} \leq 1$. Therefore, although we cannot avoid sampling unseen variables, we can still constrain the maximum value of the partition function by limiting the smoothness of the critic.

By leveraging the spectral normalization, we propose the Mutual Information Continuity-constrained Estimator that smooths the critic

$$I(X;Y) = \sup_{T^{\mathrm{SN}}_\theta:\mathcal{X}\to\mathbb{R}} \left\{ \mathbb{E}_{P_{(X,Y)}}\left[T^{\mathrm{SN}}_\theta(x,y)\right] - \mathbb{E}_{P_X\otimes P_Y}\left[e^{T^{\mathrm{SN}}_\theta(x,y)-1}\right] \triangleq I_{\mathrm{MICE}} \right\} \tag{14}$$

where T^{SN}_θ is a critic normalized by the spectral norm of the last layer. In contrast to previous approaches that focus on reducing the variances of the partition function, the proposed I_{MICE} shares the same parameters in both sides of Equation 14, and therefore it is guaranteed to not exceed the MI.

To quantify the maximal variance of the log-density ratio, we assume that $T^{\mathrm{SN}}_\theta : \mathbb{R}^d \to \mathbb{R}$ is a multi-layer perceptron (MLP) with Lipschitz continuous activation functions.

Lemma 6 *Let X be a random variable, and $g(X) : \mathbb{R}^d \to \mathbb{R}$ is an MLP with any Lipschitz continuous activation function. Let L_i be the Lipschitz constant of the i^{th} layer, then*

$$\mathrm{Var}[g(X)] \leq \mathbb{E}\left[\|X - \mathbb{E}(X)\|^2\right] \prod_{i=1}^{I} L_i^2 \tag{15}$$

Here, we defer the proof in Section A.1. Lemma 6 shows that the variance of the critic is bounded above by the product of the square of its Lipschitz constants in each layer. An inequality resembles to Equation 9 that upper bounds the partition function is shown below

$$\begin{aligned} \mathrm{LSE}(T_\theta(x_1,y_1),\ldots,T_\theta(x_n,y_{n-1})) &\leq \max\{T_\theta(x_1,y_1),\ldots,T_\theta(x_n,y_{n-1})\} + \log n(n-1) \\ \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n-1} e^{T_\theta(x_i,y_j)} &\leq \frac{1}{n(n-1)}\left(e^{\max\{T_\theta(x_1,y_1),\ldots,T_\theta(x_n,y_{n-1})\}} + 1\right) \end{aligned} \tag{16}$$

Therefore, by Equation 15 and Equation 16, the variance of the partition function is reduced by limiting the Lipschitz constant L of the critic and controlling the variance of X, and the estimate is

of lower variance with smaller L determined by the network during the optimization. Investigating Equation 14, because the partition function is exponential, its gradient with respect to T_θ^{SN} is still an exponential function, which causes the training to become unstable. Therefore, to further mitigate the high variance issue and stabilize the gradients, we avoid gradients generated by the partition function from back-propagating and consequently stabilize the gradients. The training procedure using gradient stabilization is presented in Algorithm 1.

Algorithm 1: Mutual Information Continuity-constrained Estimator (MICE)

$\theta \leftarrow$ initialize network parameters from uniform distribution $\mathcal{U}\left(-\sqrt{\frac{1}{d}}, \sqrt{\frac{1}{d}}\right)$;

while *not converge* **do**

Draw n pair of samples $(x_1, y_1), \ldots, (x_n, y_n)$ from the joint distribution $P_{(X,Y)}$

Forward pass of MICE:

$T_\theta^{\mathrm{SN}}(x, y) \leftarrow MLP_\theta(x, y)$

$I_{\mathrm{MICE}}(\theta) \leftarrow \frac{1}{n}\sum_{i=1} T_\theta^{\mathrm{SN}}(x_i, y_i) - \log \frac{1}{n(n-1)} \sum_{i \neq j} e^{T_\theta^{\mathrm{SN}}(x_i, y_j)}$

Compute the gradients on the left-hand side of I_{MICE} with respect to θ:

$\mathcal{G}(\theta) \leftarrow \nabla_\theta I_{\mathrm{MICE}}^{left}(\theta)$

Update the network parameters:

$\theta \leftarrow \theta + \mathcal{G}(\theta)$

end

4.2 Consistency

According to (Belghazi et al., 2018), an estimator $I_n(X;Y)$ constructed using a statistics network over n samples is strongly consistent if for all $\epsilon > 0$, and there exists a positive integer N such that

$$\forall n \geq N, \ \ |I(X;Y) - I_n(X;Y)| \leq \epsilon, a.e. \tag{17}$$

Then, the authors separate the consistency question into approximation and estimation problems. In summary, to prove that MICE is strongly consistent, we first prove that there exists a neural network T_θ parameterized by θ in some compact domain $\Theta \in \mathbb{R}$, such that for all $\epsilon > 0$, $|I(X;Y) - I_\Theta(X;Y)| \leq \epsilon, a.e.$ This ensures the existence of neural networks that can approximate the MI with arbitrary accuracy. Second, we prove that given a family of neural networks T_θ in some bounded domain, for all $\epsilon > 0$, there exists an $N \in \mathbb{N}$ such that for all $n \geq N$, $|I_n(X;Y) - I_\Theta(X;Y)| \leq \epsilon, a.e.$, ensuring that given sufficient number of samples, one can estimate the MI with some statistics networks over samples. Combining the above two results with triangular inequality, we conclude that MICE is strongly consistent. We provide the details of the proofs in Section A.2.

5 Experiments

5.1 Standard Benchmark

Dataset. The standard benchmark (Belghazi et al., 2018; Poole et al., 2019; Song & Ermon, 2020) contains two tasks, the *Gaussian* task and the *Cubic* task. For both tasks, we sample n random variables $X, Y \in \mathbb{R}^d$ for a batch from a standard multivariate normal distribution with correlation ρ between X and Y. For the *Cubic* task, to examine how much the MI estimators degrade when a nonlinear transformation involved, we estimate $I(X; Y^3) = I(X;Y)$, which does not change the MI.

Critics. Following previous studies (Belghazi et al., 2018; Poole et al., 2019; Song & Ermon, 2020), we consider two types of critics: the *joint* critic (Belghazi et al., 2018) and the *separable* critic (Oord et al., 2018). The joint critic first lists all combinations of all random variables in a batch and computes the log-density ratio with an MLP $\mathbb{R}^{2d} \to \mathbb{R}$. The separable critic applies nonlinear mapping to the inputs with two MLPs, $f, g : \mathbb{R}^d \to \mathbb{R}^{d'}$, and subsequently estimates log-density ratio by $\langle f, g \rangle$. The joint critic compares all combinations, having the computational complexity of $O(n^2)$, and since the computation of f and g can be paralleled, thus having a complexity of $O(n)$.

In Figure 1, we show the performance of each estimator under different MI. The top row shows the *Gaussian* task, and the bottom row shows the *Cubic* task. As described in Section 2, I_{CPC} is highly

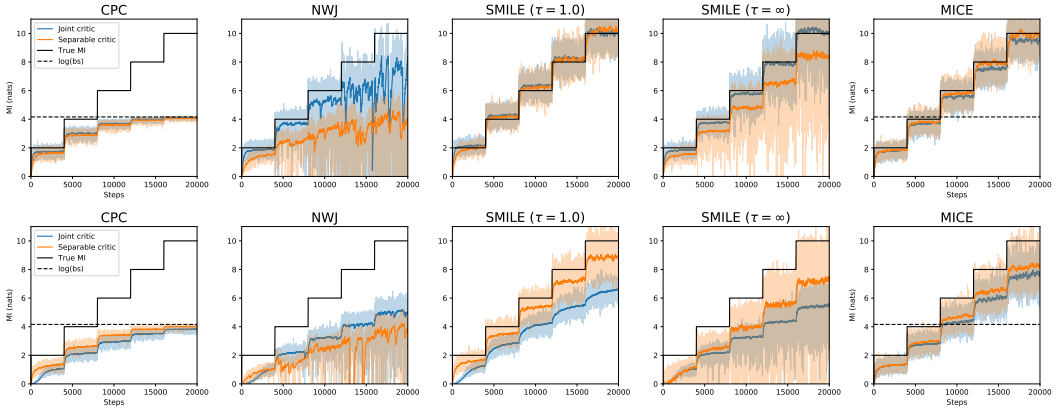

Figure 1: Performance of estimators in the standard benchmark. The top row shows the performance of each estimator in the standard benchmark, and the bottom row shows the performance of MI estimators with nonlinear transformation, namely $Y \mapsto Y^3$. For the two experiments, we increased the MI by 2 every 4000 iterations, and 20000 iterations in total. The ground truth of MI is marked black. The light/dark color lines are the real estimates and their smoothed values.

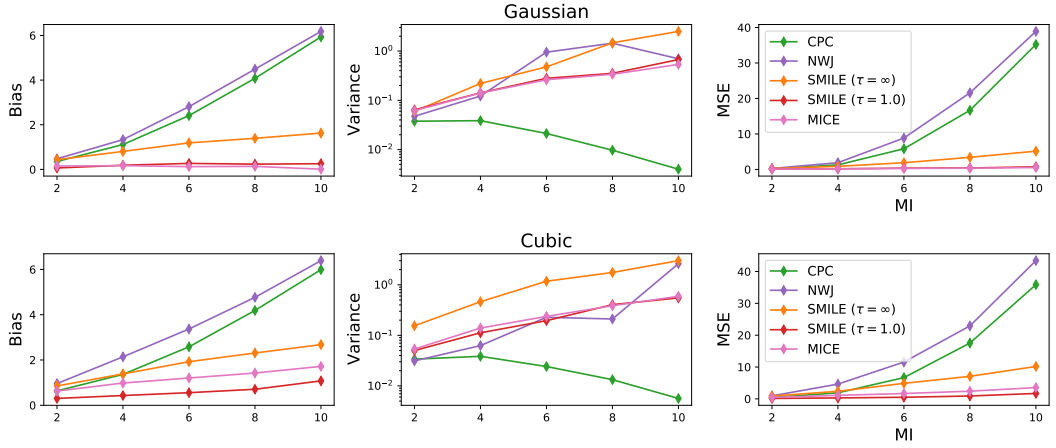

Figure 2: Bias-variance trade-offs of estimators in the standard benchmark. The top row shows the performance of each estimator in the *Gaussian* task, and the bottom row show the performance in the *Cubic* task.

biased and bounded above by $\log n$, and the variance of I_{NWJ} increases along with the ground truth MI. Here, I_{SMILE} ($\tau = 1.0$) and I_{MICE} have overall lower biases and variances, as compared to I_{CPC} and I_{NWJ} using both critics. Because I_{SMILE} is neither an upper bound nor lower bound on MI, MI estimates in the *Gaussian* task are sometimes slightly overestimated, but the moving mean of I_{MICE} is almost not exceeding the ground truth of MI. In the *Cubic* task, the joint critic degrades more severely than the separable critic for most of estimators, except I_{NWJ}.

We show the bias-variance trade-offs of estimators using the separable critic in Figure 2, where the top row illustrates the results of the *Gaussian* task, and the results of the *Cubic* task are shown at the bottom row. It is observed that I_{CPC} is severely biased, but the variance is much lower than all the other approaches. Although I_{NWJ} is theoretically unbiased, it has large bias owing to the inevitable outliers, and the variance grows up exponentially with MI as (Song & Ermon, 2020) pointed out. I_{MICE} leverages the unbiasedness of I_{NWJ} and further reduces the variance by constraining the Lipschitz constant of the critic and gradient stabilization. Comparing result of I_{SMILE} and I_{MICE} using the joint critic, I_{MICE} converges faster than I_{SMILE}. It is possibly benefited from the stabilized

gradients. However, because we limit the Lipschitz constants in some layers of the critic, this could lead to lower flexibility, and thus I_{MICE} is slightly more biased than I_{SMILE} in the *Cubic* task. In brief, I_{MICE} simultaneously guarantees not to exceed MI and remarkably relaxes the high-variance issue of I_{NWJ}.

5.2 Extension of Standard Benchmark

Sampling scheme. Next, we evaluate the MI estimators using an extension experiment based on the standard benchmark. As described in Section 5.1, random variables are sampled independently; that is, no correlations between samples is considered. However, we believe that, for practical scenarios, it is extremely difficult for one to create a batch in which all samples are independent. Therefore, based on the standard benchmark, we established an extension experiment in which random variables are sampled using the scheme below:

$$x_i = \hat{\rho} x_{i-1} + \sqrt{1-\hat{\rho}^2}\epsilon,\ \forall i = 2, \ldots, n \tag{18}$$

$$y_i = \rho x_i + \sqrt{1-\rho^2}\epsilon,\ \forall i = 1, \ldots, n \tag{19}$$

where x_1 and ϵ are d-dimensional random variables following a standard normal distribution $\mathcal{N}(0, I_d)$. Sampling variables using Equation 18 and Equation 19 is equivalent to sample $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_n\}$ from a multivariate normal distribution

$$X, Y \sim \mathcal{N}\left(0, \begin{bmatrix} \Sigma_x & \rho\Sigma_x \\ \rho\Sigma_x & \Sigma_x \end{bmatrix}\right),\ \Sigma_x = \begin{bmatrix} I_d & \hat{\rho}I_d & \hat{\rho}^2 I_d & \ldots & \hat{\rho}^{n-1} I_d \\ \hat{\rho}I_d & I_d & \hat{\rho}I_d & \ldots & \hat{\rho}^{n-2} I_d \\ \hat{\rho}^2 I_d & \hat{\rho}I_d & I_d & \ldots & \hat{\rho}^{n-3} I_d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}^{n-1} I_d & \hat{\rho}^{n-2} I_d & \hat{\rho}^{n-3} I_d & \ldots & I_d \end{bmatrix}$$

where $\hat{\rho}$ is the correlation between each pair of two consecutive samples, i.e., x_i, x_{i+1} and y_i, y_{i+1}. In the extension benchmark, we follow the setting in Section 5.1 with an additional setting $\hat{\rho} = 0.1$, and the ground truth MI is increased by 2 after 4000 iterations during training. In general, the correlation coefficient less than 0.3 is considered to be weak. As shown in Figure 3, data with correlation of 0.1, which is even weaker than 0.3, degenerates other estimators, whereas I_{MICE} still has relatively accurate estimates. To further explore this effect, additional experiments using different settings of $\hat{\rho}$ are presented in Section A.3.

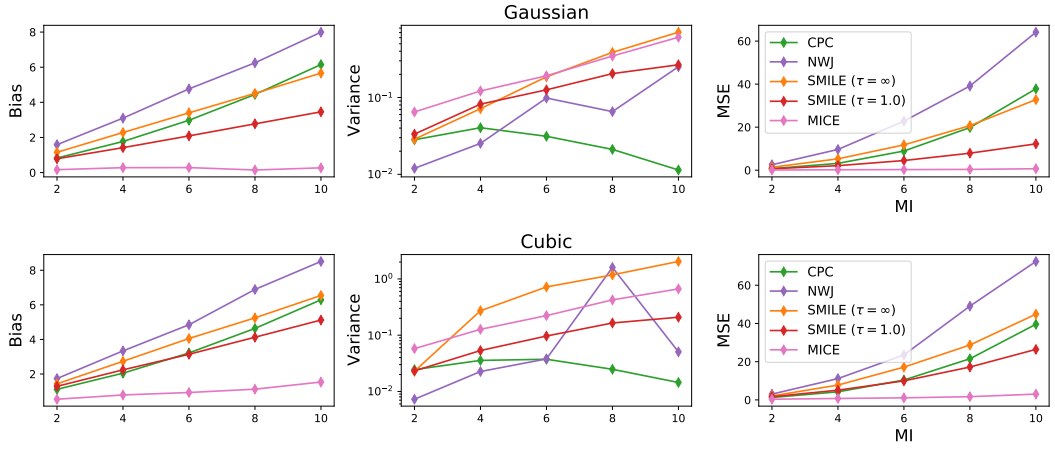

Figure 3: Bias-variance trade-offs of estimators in the extension benchmark. The top row shows the performance of each estimator in the *Gaussian* task, and the bottom row show the performance in the *Cubic* task. I_{MICE} is less biased and more accurate than the other estimators.

In Figure 3, we demonstrate that the bias and MSE of the estimate of I_{MICE} are much lower than those of the other estimators using the separable critic for both the *Gaussian* task and the *Cubic* task.

There are two possible reasons that I_{MICE} outperforms the other approaches. First, as we stated in Section 3, the partition function could be dominated by the nonzero log-density ratios when the correlations between samples are involved. The other reason is that the gradients are stabilized by applying spectral normalization to the critic and blocking the gradients generated by the partition function.

5.3 Regularizing GAN with MICE

GANs (Goodfellow et al., 2014) have recently shown powerful capabilities in real-world data generation. However, the well-known mode collapse agonizes GANs with the consequence of limited diversity. This is because the discriminator does not require the generator to capture all modes to decrease the loss function. (Belghazi et al., 2018) proposed to alleviate mode collapse by involving code variables C, and jointly maximize the MI between the generated data and C. Formally, a GAN regularized by MICE alternately optimizes the following two objectives:

$$\mathcal{L}_D \coloneqq \mathbb{E}_{P_X}[\log D(X)] + \mathbb{E}_{P_Z}[\log(1 - D(G(Z))] \tag{20}$$

$$\mathcal{L}_G \coloneqq \mathbb{E}_{P_Z}[\log(1 - D(G(z)))] - \beta I_{\mathrm{MICE}}(G(Z, C); C) \tag{21}$$

where D, G are the discriminator and the generator, and Z follows a standard uniform distribution. Comparing the results of vanilla GAN and GAN + MICE in Figure 4, a vanilla GAN fails to model the structure, whereas GAN + MICE captures all 25 modes, showing the efficacy of mode collapse mitigation.

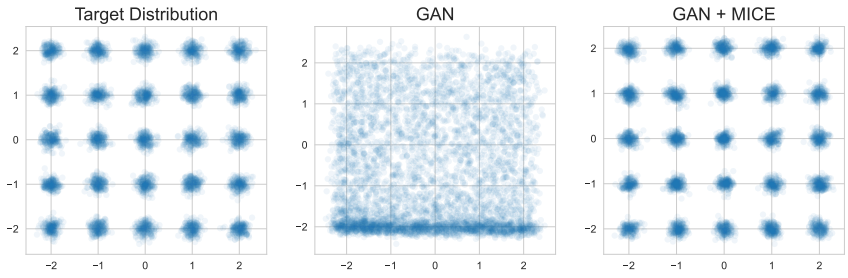

Figure 4: Results of GAN and MICE regluarization on 25 Gaussians dataset. Illustration on the left is the target samples. The middle and the right plots are generated by a vanilla GAN and generated by GAN regularized by MI estimated by MICE.

6 Conclusion

In this study, we comprehensively discuss the attributes and the limitations of existing approaches to variational MI estimation. We show that energy-based estimators such as I_{NWJ}, and I_{DV} are of high variances because they are susceptible to the outlier events. Although neither upper bound nor lower bound estimators achieve much more accurate approximations to MI in the standard benchmark, they are under the risk of overestimating the MI. To address the above mentioned issues, we propose a unbiased and consistent estimator of MI, I_{MICE}, which has been proven free from overestimation of the MI. We also argue that the standard benchmark is insufficient for evaluation since samples can hardly be entirely uncorrelated in general cases. Therefore, we employ an additional benchmark to evaluate the performance of the estimators in which the samples are correlated. In the standard benchmark, the proposed I_{MICE} has a slightly better performance than I_{SMILE} without prior knowledge for selecting clipping threshold. We empirically show that I_{MICE} is more accurate than other estimators in the proposed additional benchmark. Finally, we show that regularizing GANs with MICE improves the ability of the GAN to capture multiple modes and consequently mitigate mode collapse.

REFERENCES

Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*, 2019.

Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*, 2019.

Roberto Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on neural networks*, 5(4):537–550, 1994.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

Michel Broniatowski and Amor Keziou. Parametric estimation and tests through divergences and the duality technique. *Journal of Multivariate Analysis*, 100(1):16–36, 2009.

Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li. Mode regularized generative adversarial networks. *arXiv preprint arXiv:1612.02136*, 2016.

Junya Chen, Zhe Gan, Xuan Li, Qing Guo, Liqun Chen, Shuyang Gao, Tagyoung Chung, Yi Xu, Belinda Zeng, Wenlian Lu, et al. Simpler, faster, stronger: Breaking the log-k curse on contrastive learners with flatnce. *arXiv preprint arXiv:2107.01152*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv preprint arXiv:1606.03657*, 2016.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International Conference on Machine Learning*, pp. 1779–1788. PMLR, 2020.

Pierre Colombo, Chloe Clavel, and Pablo Piantanida. A novel estimator of mutual information for learning to disentangle textual representations. *arXiv preprint arXiv:2105.02685*, 2021.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36(2):183–212, 1983.

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, Narayanaswamy Siddharth, Brooks Paige, Dana H Brooks, Jennifer Dy, and Jan-Willem Meent. Structured disentangled representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2525–2534. PMLR, 2019.

Pablo A Estévez, Michel Tesmer, Claudio A Perez, and Jacek M Zurada. Normalized mutual information feature selection. *IEEE Transactions on neural networks*, 20(2):189–201, 2009.

Andrew M Fraser and Harry L Swinney. Independent coordinates for strange attractors from mutual information. *Physical review A*, 33(2):1134, 1986.

Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Efficient estimation of mutual information for strongly dependent variables. In *Artificial intelligence and statistics*, pp. 277–286. PMLR, 2015.

Sara A Geer and Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Nojun Kwak and Chong-Ho Choi. Input feature selection by mutual information based on parzen window. *IEEE transactions on pattern analysis and machine intelligence*, 24(12):1667–1671, 2002.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *AISTATS*, 2020.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *ICLR*, 2018.

Young-Il Moon, Balaji Rajagopalan, and Upmanu Lall. Estimation of mutual information using kernel density estimators. *Physical Review E*, 52(3):2318, 1995.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *arXiv preprint arXiv:1606.00709*, 2016.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pp. 5171–5180. PMLR, 2019.

Yves-Laurent Kom Samo. Inductive mutual information estimation: A convex maximum-entropy copula approach. In *International Conference on Artificial Intelligence and Statistics*, pp. 2242–2250. PMLR, 2021.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In *ICLR*, 2020.

Alessandro Sordoni, Nouha Dziri, Hannes Schulz, Geoff Gordon, Philip Bachman, and Remi Tachet Des Combes. Decomposed mutual information estimation for contrastive representation learning. volume 139, pp. 9859–9869. PMLR, 2021.

Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 3310–3320, 2017.

A Appendix

A.1 Proof of Lemma 6

Lemma 6 *Let X be a random variable, and $g(X) : \mathbb{R}^d \to \mathbb{R}$ is an MLP with any Lipschitz continuous activation function. Let L_i be the Lipschitz constant of the i^{th} layer, then*

$$\mathrm{Var}[g(X)] \leq \mathbb{E}\left[\|X - \mathbb{E}(X)\|^2\right] \prod_{i=1}^{I} L_i^2 \tag{22}$$

Proof. First, we consider the i-th layer f_i with a Lipschitz continuous activation function, and f_i has Lipschitz constant L_i, then

$$\begin{aligned}
\mathrm{Var}[f_i(X)] &\coloneqq \mathbb{E}\left[(f_i(X) - \mathbb{E}\left[f_i(X)\right])^2\right] && (23)\\
&\leq \mathbb{E}\left[(f_i(X) - f_i(\mathbb{E}\left[X\right]))^2\right] && (24)\\
&\leq L_i^2 \mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] && (25)
\end{aligned}$$

The first inequality stems from the fact that the mean of a random variable is the constant with the smallest MSE. By the definition of Lipschitz continuity, the second inequality holds because L_i is the Lipschitz constant of f_i. Second, let g be the composite function of $f_1, f_2, \ldots, f_I$ that $g = f_1 \circ f_2 \circ \cdots \circ f_I$, where I is the number of layers in g, then

$$\mathrm{Var}[g(X)] \leq \mathbb{E}\left[\|X - \mathbb{E}(X)\|^2\right] \prod_{i=1}^{I} L_i^2 \tag{26}$$

which completes the proof.

A.2 Proof of Consistency

The proof of consistency generally follows the proofs in (Belghazi et al., 2018) yet with some modifications to fit MICE. To prove that MICE is strongly consistent, we first prove that for all $\epsilon > 0$, there exists a class of neural networks T_θ parameterized by θ in some compact domain Θ such that

$$|I(X;Y) - I_\Theta(X;Y)| \leq \epsilon \tag{27}$$

Next, we prove that given $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that

$$|I_n(X;Y) - I_\Theta(X;Y)| \leq \epsilon \tag{28}$$

As consequence, combining the above results with triangular inequality, we have $\forall n \geq N$, $|I(X;Y) - I_n(X;Y)| \leq \epsilon$, which proves the consistency of MICE.

Proof. Let the optimal critic $T^* = \log \frac{dP}{dQ}$, where P and Q denote the joint distribution $P_{(X,Y)}$ and the product of marginals $P_X P_Y$ of the continuous random variables X and Y, respectively. By the definition of I_{NWJ}, we have

$$I(X;Y) - I_\Theta(X;Y) = \mathbb{E}_P[T^* - T] + \mathbb{E}_Q[e^{T^*-1} - e^{T_\theta - 1}] \tag{29}$$

Next, according to the universal approximation theorem (Hornik et al., 1989), one can choose a T_θ such that

$$\mathbb{E}_P|T^* - T_\theta| \leq \frac{\epsilon}{2} \tag{30}$$

$$\mathbb{E}_Q|T^* - T_\theta| \leq \frac{\epsilon}{2} e^{-T_{max}+1} \tag{31}$$

where T^* is upper bounded above by T_{max}. Because $\exp(\cdot)$ is Lipschitz continuous with constant $e^{T_{max}-1}$ on $(-\infty, e^{T_{max}-1}]$, $\mathbb{E}_Q|e^{T^*-1} - e^{T_\theta-1}| \leq e^{T_{max}-1}\mathbb{E}_Q|T^* - T_\theta|$, and consequently we have

$$\mathbb{E}_Q|e^{T^*-1} - e^{T_\theta-1}| \leq \frac{\epsilon}{2} \tag{32}$$

Combine Equation 29, Equation 30, and Equation 32 with triangular inequality, we have

$$|I(X;Y) - I_\Theta(X;Y)| \leq \mathbb{E}_P|T^* - T_\theta| + \mathbb{E}_P|e^{T^*-1} - e^{T_\theta-1}| \leq \epsilon \tag{33}$$

So far we have proved that for $T^* \leq T_{max}$, Equation 27 holds. Next, we consider a subset that $\{T^* > T_{max}\}$ for a suitably chosen large value of T_{max}. Here, let A be the subset belongs to the input domain, we use the indicator function $\mathbf{1_A}$ to partition the input domain. By the Lebesgue dominated convergence theorem, since that T^* and e^{T*} are integrable w.r.t. P and Q, we could choose T_{max} so that

$$\mathbb{E}_P[1_{T^*>T_{max}}(T^*)] \leq \frac{\epsilon}{4} \tag{34}$$

$$\mathbb{E}_Q[1_{T^*>T_{max}}(e^{T^*-1})] \leq \frac{\epsilon}{4} \tag{35}$$

Again, we can choose a function $T_\theta \leq T_{max}$ such that

$$\mathbb{E}_P|T^* - T_\theta| \leq \frac{\epsilon}{2} \tag{36}$$

$$\mathbb{E}_Q 1_{T^*\leq T_{max}}(|T^* - T_\theta|) \leq \frac{\epsilon}{2} e^{-T_{max}+1} \tag{37}$$

Combining Equation 35 and Equation 37 together

$$\begin{aligned}
\mathbb{E}_Q[e^{T^*-1} - e^{T_\theta-1}] &= \mathbb{E}_Q[1_{T^*\leq T_{max}}(e^{T^*-1} - e^{T_\theta-1})] + \mathbb{E}_Q[1_{T^*>T_{max}}(e^{T^*-1} - e^{T_\theta-1})] \\
&\leq e^{T_{max}-1}\mathbb{E}_Q[1_{T^*\leq T_{max}}(T^* - T_\theta)] + \mathbb{E}_Q[1_{T^*>T_{max}}(e^{T^*-1})] \\
&\leq \frac{\epsilon}{2}
\end{aligned} \tag{38}$$

Similar to the derivation of Equation 33, put Equation 36 and Equation 38 together we obtain

$$\forall \epsilon > 0, \;\; |I(X;Y) - I_\Theta(X;Y)| \leq \epsilon \tag{39}$$

For the estimation problem, let $\epsilon > 0$ and given T_θ in some compact domain $\Theta \subset \mathbb{R}^d$, there exists a positive integer N such that

$$\forall n \geq N, \;\; |I_n(X;Y) - I_\Theta(X;Y)| \leq \epsilon \tag{40}$$

Here, we denote P_n and Q_n as the empirical version of P and Q respectively, and I_n is the MI estimation with n samples. By triangular inequality we have

$$|I_n(X;Y) - I_\Theta(X;Y)| \leq \sup_{\theta\in\Theta} \left\{|\mathbb{E}_{P_n}[T_\theta] - \mathbb{E}_P[T_\theta]| + |\mathbb{E}_{Q_n}[e^{T_\theta-1}] - \mathbb{E}_Q[e^{T_\theta-1}]|\right\} \tag{41}$$

Since Θ is compact (therefore bounded) and neural networks are continuous, T_θ and e^{T_θ} satisfiy the uniform law of large numbers (Geer & van de Geer, 2000). Therefore, given $\epsilon > 0$ we can choose a positive integer N such that $\forall n \geq N$ and with probability one, then

$$\sup_{\theta\in\Theta} \{|\mathbb{E}_{P_n}[T_\theta] - \mathbb{E}_P[T_\theta]|\} \leq \frac{\epsilon}{2} \tag{42}$$

$$\sup_{\theta\in\Theta} \left\{|\mathbb{E}_{Q_n}[e^{T_\theta-1}] - \mathbb{E}_Q[e^{T_\theta-1}]|\right\} \leq \frac{\epsilon}{2} \tag{43}$$

According to the three inequalities above we derive Equation 40.

Finally, combining Equation 39 and Equation 40 with triangular inequality, let $\epsilon > 0$ and $\delta = 2\epsilon$, and there exists a positive integer N such that

$$\forall n \geq N, \ \ |I(X;Y) - I_n(X;Y)| \leq |I(X;Y) - I_\Theta(X;Y)| + |I_n(X;Y) - I_\Theta(X;Y)| \leq \delta \quad (44)$$

which completes the proof.

A.3 Additional Experiments

Experimental Settings. The experiments in Section 5.1 and Section 5.2 are established using a GTX 1080 Ti GPU with 11 GB VRAM. For the MLPs utilized in the joint/separable critic have the input dimension of 20, two hidden layers of 256 hidden dimension, and the output dimension is 32. In addition, ReLU Agarap (2018) is used as the activation function for both critics.

Performance of MI Estimators under Specific Correlations. We compare the performance of the MI estimators under specific $\hat{\rho}$ settings ($\hat{\rho}$=0.1, 0.2, 0.3, 0.4, and 0.5) using the separable critic. As shown in Figure 5, the MI estimators are more biased with larger correlation between samples. Among the MI estimators, I_{NWJ} is the most biased since neither the partition function nor the critic is constrained, so the outliers lead to large variances and biases, and this is the same reason that causes $I_{\mathrm{SMILE}(\tau=\infty)}$ to be inaccurate. As mentioned in Section 2, the estimates of $I_{\mathrm{SMILE}(\tau=\infty)}$ are more accurate than that of I_{NWJ} because $I_{\mathrm{SMILE}(\tau=\infty)}$ is equivalent to I_{DV} which is sharper than I_{NWJ}. Despite that I_{CPC} is bounded above by $\log n$, it is consistent under different settings of correlation. $I_{\mathrm{SMILE}(\tau=1.0)}$ is of low variance and bias comparing to itself when $\tau = \infty$, but the improvement is mainly on reducing the variance. Comparing to the other MI estimators, our proposed I_{MICE} is the least biased, and is robust when correlations between samples involved.

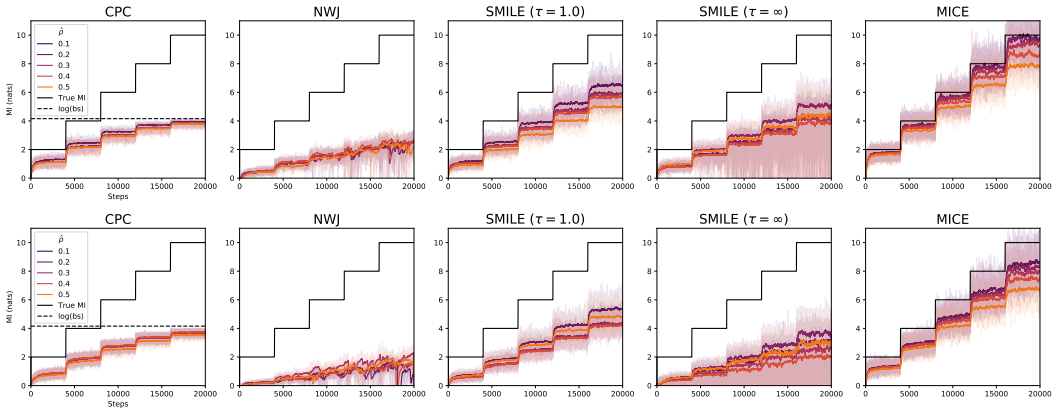

Figure 5: Performance of estimators in the extension benchmark of different $\hat{\rho}$. The top row shows the performance of each estimator in the *Gaussian task*, and the bottom row show the performance in the *Cubic* task. The correlation $\hat{\rho}$ between samples ranges from 0.1 to 0.5.

Randomly Selected $\hat{\rho}$. We provide an experiment that correlations between samples are randomly initialized, which is a more complicated configuration than the extension benchmark in Section 5.2. Here, $\hat{\rho}$ are randomly initialized from a uniform distribution that ranges from 0.0 to 0.5. In Figure 6, each estimator using the separable critic has an average performance in Figure 5. The proposed I_{MICE} benefits the separable critic that it is robust to random correlations. In Figure 7, we also observed that the variance of I_{NWJ} is very sensitive to the data because the right-hand side of Eqn. Equation 5 is an exponential function without logarithm in I_{DV}, and consequently yields high MSE.

By constraining the continuity and gradient stabilization, I_{MICE} is robust when correlation between samples involved as compared with the other estimators, especially for the separable critic. This could benefit large scale training that requires a light weight model structure for the critic.

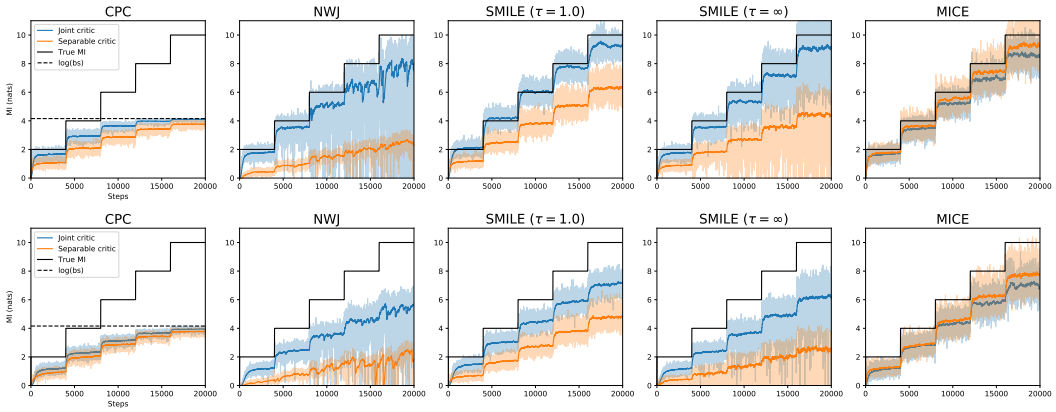

Figure 6: Performance of estimators in the extension benchmark of a random selection of $\hat{\rho}$. The top row shows the performance of each estimator in the *Gaussian task*, and the bottom row show the performance in the *Cubic* task. The correlation $\hat{\rho}$ between samples uniformly ranges from 0.0 to 0.5.

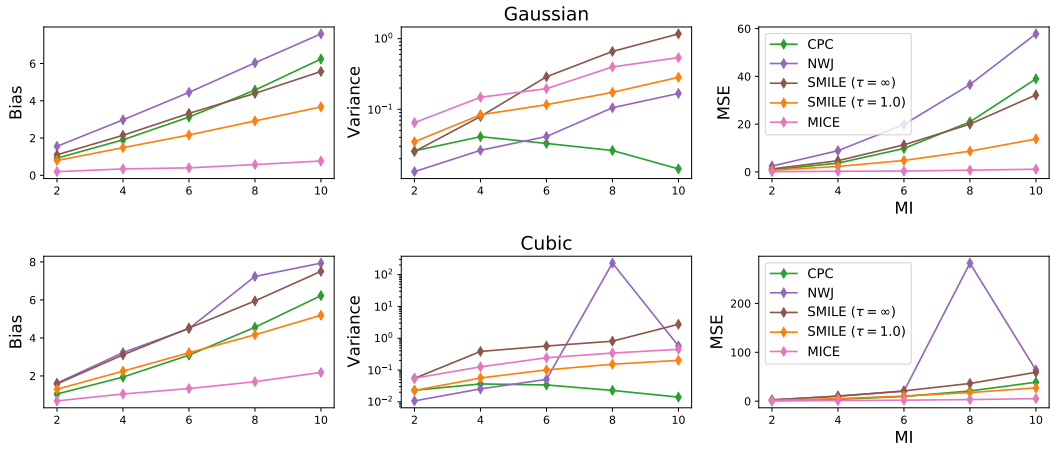

Figure 7: Bias-variance trade-offs of estimators in the extension benchmark with random $\hat{\rho}$. The top row shows the performance of each estimator in the *Gaussian* task, and the bottom row show the performance in the *Cubic* task. I_{MICE} is less biased and more accurate than the other estimators.

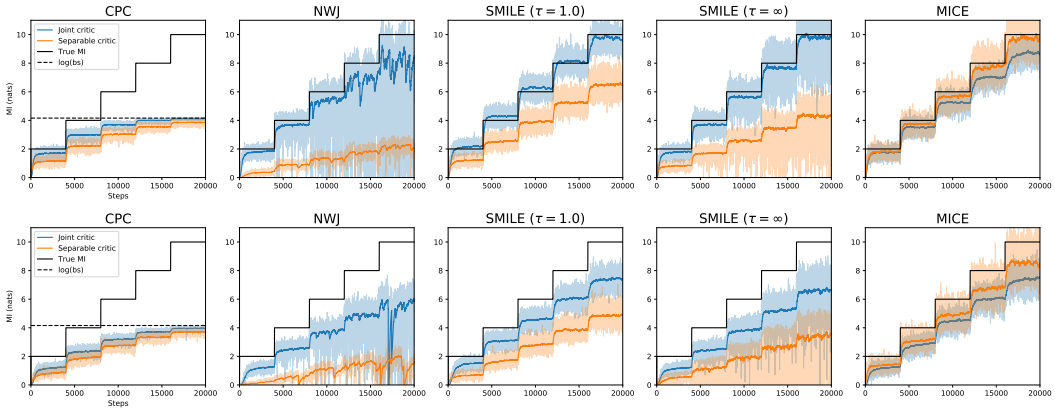

Figure 8: Performance of estimators in the extension benchmark (Section 5.2 related). The top row shows the performance of each estimator in the *Gaussian task*, and the bottom row show the performance in the *Cubic* task.