7 Supplementary

7.1 Hyper-parameters

We provide the experimental settings and hyper-parameters for ease of reproducibility of the results.

Table 2: Classifier : Hyper-parameters

Hyper-parameter	**Value**
Hidden Units	64
# Hidden Layers	2 (Inp-64-64-Out)
Activation	ReLU
Batch-Size	64
Learning Rate	0.001
Optimizer	Adam $(\beta_1 = 0.90, \beta_2 = 0.999)$
# Epoch	20
Regularizer	L2 (0.001)

Table 3: CGAN : Hyper-parameters

Hyper-parameter	**Value**
Hidden Units	256
# Hidden Layers	2 (Inp-256-256-Out)
Activation	Leaky ReLU(0.2)
Batch-Size	128
Learning Rate	$1e-4$
Optimizer	Adam $(\beta_1 = 0.5, \beta_2 = 0.9)$
# Epoch	100
Noise dimension	20
Noise distribution	$\mathcal{U}(-1.0, 1.0)^{d_s}$

Table 4: CVAE : Hyper-parameters

Hyper-parameter	**Value**
Hidden Units	256
# Hidden Layers	2 (Inp-256-256-Out)
Activation	Leaky ReLU(0.2)
Batch-Size	128
Learning Rate	$1e-4$
Optimizer	Adam $(\beta_1 = 0.5, \beta_2 = 0.9)$
# Epoch	20
Dropout	0.9
Latent dimension	20

Table 5: f-MINE : Hyper-parameters

Hyper-parameter	**Value**
Hidden Units	64
# Hidden Layers	1 (Inp-64-Out)
Activation	ReLU
Batch-Size	128 (512 for DV-MINE)
Learning Rate	$1e-4$
Optimizer	Adam $(\beta_1 = 0.5, \beta_2 = 0.999)$
# Epoch	200

Table 6: AuROC : Flow-Cytometry Data (Mean ± Std. of 5 runs.

Tester	**AuROC**
CCIT	0.6665 ± 0.006
CCMI	**0.7569** ± 0.047

7.2 Calibration Curve

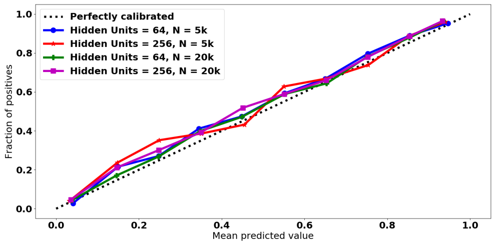
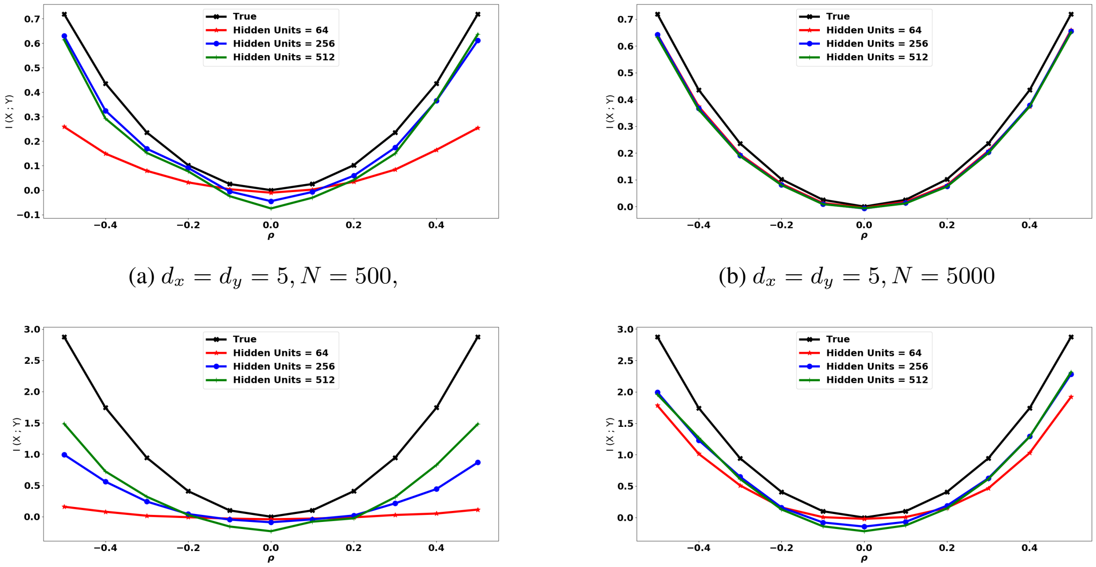


Figure 6: Calibrated Classifiers : We find that our classifiers trained with $L2$-regularization and two hidden layers are well-calibrated. The calibration is obtained for MI Estimation of Correlated Gaussians with $d_x = 10, \rho = 0.5$

While Niculescu-Mizil and Caruana (2005) showed that neural networks for binary classification produce well-calibrated outputs. the authors in Guo et al. (2017) found miscalibration in deep networks with batch-normalization and no L2 regularization. In our experiments, the classifier is shallow, consisting of only 2 layers with relatively small number of hidden units. There is no batch-normalization or dropout used. Instead, we use $L2$-regularization which was shown in Guo et al. (2017) to be favorable for calibration. Figure 6 shows that our classifiers are well-calibrated.



(a) $d_x = d_y = 5, N = 500,$

(b) $d_x = d_y = 5, N = 5000$

(c) $d_x = d_y = 20, N = 500$

(d) $d_x = d_y = 20, N = 5000$

Figure 7: The Donsker-Varadhan Representation provides a lower bound of the true MI. For each hyper-paramter choice, the estimates lie below $I^*(X;Y)$. An optimal estimator would return the maximum estimate from multiple hyper-parameter choices for a given data-set. Estimates are plotted for Correlation Gaussians introduced in Figure 1.

7.3 Choosing Optimal Hyper-parameter

The Donsker-Varadhan representation 1 is a lower bound on the true MI estimate (which is the supremum over all functions). So, for any classifier parameter, the plug-in estimate value computed on the test samples will be less than or equal to the true value $I(X;Y)$ with high probability (Theorem 2). We illustrate this using estimation of MI for Correlated Gaussians in Figure 7. The estimated value lies below the true values of MI. Thus, the optimal hyper-parameter is the one that returns the maximum value of MI estimate on the test set.

Once we have this block that returns the maximum MI estimate after searching over hyper-parameters, CMI estimate in CCMI is the difference of two MI estimates, calling this block twice.

We also plot the AuROC curves for the two choices of number of hidden units in flow-Cytometry data (Figure 9b) and post Non-linear noise synthetic data (Figure 9a). When the number of samples is high, the estimates are pretty robust to hyper-parameter choice (Figure 7 (b), 9a). But in sparse sample regime, proper choice of hyper-parameter can improve performance (Figure 9b).

7.4 Additional Figures and Tables

- For Flow-Cytometry data-set, we used number of hidden units = 64 for Classifier and trained for 10 epochs. Table 6 shows the mean AuROC values for two CIT testers.

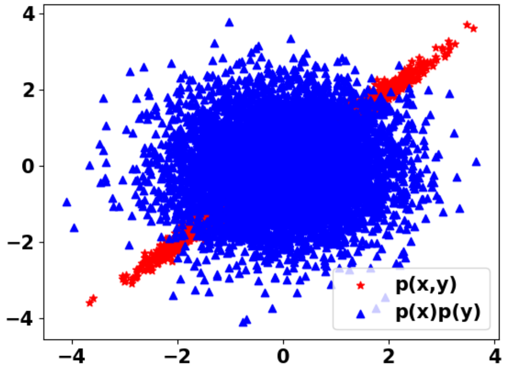
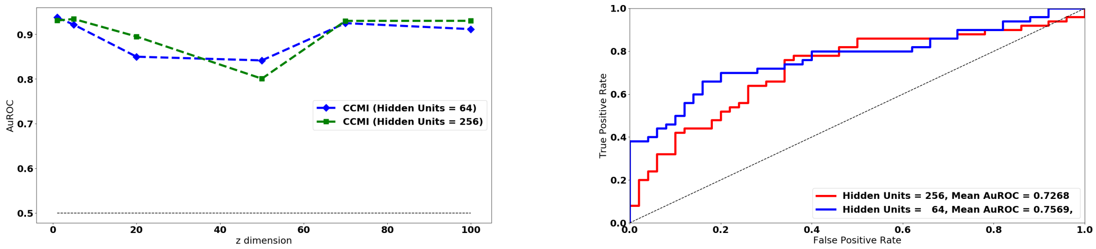


Figure 8: Logistic Regression Fails to Classify points from $p(x_1, x_2)$ (colored red) and those from $p(x_1)p(x_2)$ (colored blue).

- Figure 8 shows the distribution of points from $p(x_1, x_2)$ and $p(x_1)p(x_2)$. Here the classifier would return 0.5 as prediction for either class (leading to $\hat{D}_{KL} = 0$), even though X_1 and X_2 are highly correlated ($\rho = 0.99$) and the mutual information is high.



(a) Post Non-linear noise data-sets

(b) Flow-Cytometry data-sets

Figure 9: Hyper-parameter Sensitivity : We observe the performance in conditional independence testing with number of hidden units as 64 Vs 256, keeping all other hyper-parameters the same in respective cases.

8 Theoretical Properties of CCMI

In this Section, we explore some of the theoretical properties of CCMI. Let the samples $x_i \sim p(x)$ be labeled as $l = 1$ and $x_j \sim q(x)$ be labeled as $l = 0$. Let $Pr(l = 1) = Pr(l = 0) = 0.5$. The positive label probability for a given point x is denoted as $\gamma(x) = Pr(l = 1|x)$. When the prediction is from a classifier with parameter θ, then it is denoted as $\gamma_\theta(x)$. The argument x of γ is dropped when it is understood from the context.

The following assumptions are used throughout this Section.

- Assumption (A1) : The underlying data distributions $p(\cdot)$ and $q(\cdot)$ admit densities in a compact subset $\mathcal{X} \subset \mathbb{R}^{d_x}$.
- Assumption (A2) : $\exists\, \alpha, \beta > 0$, such that $\alpha \leq p(x), q(x) \leq \beta\; \forall x$.
- Assumption (A3) : We clip predictions in algorithm such that $\gamma(x) \in [\tau, 1-\tau]\, \forall x$, with $0 < \tau \leq \alpha/(\alpha+\beta)$.
- Assumption (A4) : The classifier class $\mathcal{C}_\theta$ is parameterized by θ in some compact domain $\Theta \subset \mathbb{R}^h$. $\exists$ constant K, such that $\|\theta\| \leq K$ and the output of the classifier is L-Lipschitz with respect to parameters θ.

Notation and Computation Procedure

- In the case of mutual information estimation $I(U;V)$, $x \in \mathbb{R}^{d_u+d_v}$ represents the concatenated data point (u, v). To be precise, $p(x) = p(u, v)$ and $q(x) = p(u)p(v)$.
- In the proofs below, we need to compute the Lipschitz constant for various functions. The general procedure for those computations are as follows.

  $$|\phi(x) - \phi(y)| \leq L_\phi |x - y|$$

  We compute L_ϕ using $\sup_z |\phi'(z)|, z \in \mathrm{domain}(\phi)$. The functions encountered in the proofs are continuous, differentiable and have bounded domains.
- The binary-cross entropy loss estimated from n samples is

  $$\mathrm{BCE}_n(\gamma) = -\left(\frac{1}{n}\sum_i l_i \log \gamma(x_i) + (1 - l_i)\log(1 - \gamma(x_i))\right) \quad (4)$$

  When computed on the train samples (resp. test samples), it is denoted as $\mathrm{BCE}_n^{\mathrm{ERM}}(\gamma)$ (resp. $\mathrm{BCE}_n(\gamma)$). The population mean over the joint distribution of data and labels is

  $$\mathrm{BCE}(\gamma) = -\left(\mathbb{E}_{XL} L \log \gamma(X) + (1 - L)\log(1 - \gamma(X))\right) \quad (5)$$
- The estimate of MI from n test samples for classifier parameter $\hat{\theta}$ is given by

  $$I_n^{\gamma_{\hat{\theta}}} = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\gamma_{\hat{\theta}(x_i)}}{1 - \gamma_{\hat{\theta}(x_i)}} - \log\left(\frac{1}{n}\sum_{j=1}^{n} \frac{\gamma_{\hat{\theta}}(x_j)}{1 - \gamma_{\hat{\theta}}(x_j)}\right)$$

  The population estimate for classifier parameter $\hat{\theta}$ is given by

  $$I^{\gamma_{\hat{\theta}}} = \mathop{\mathbb{E}}_{x\sim p} \log \frac{\gamma_{\hat{\theta}(x)}}{1 - \gamma_{\hat{\theta}(x)}} - \log\left(\mathop{\mathbb{E}}_{x\sim q} \frac{\gamma_{\hat{\theta}}(x)}{1 - \gamma_{\hat{\theta}}(x)}\right)$$

Theorem 3 (Theorem 1 restated)**.** *Classifier-MI is consistent, i.e., given $\epsilon, \delta > 0, \exists\, n \in \mathbb{N}$, such that with probability at least $1 - \delta$, we have*

$$|I_n^{\gamma_{\hat{\theta}}}(U;V) - I(U;V)| \leq \epsilon$$

Intuition of Proof

The classifier is trained to minimize the empirical risk on the train set and obtains the minimizer as $\hat{\theta}$. From generalization bound of classifier, this loss value ($\mathrm{BCE}(\gamma_{\hat{\theta}})$) on the test set is close to the loss obtained by the best optimizer in the classifier family ($\mathrm{BCE}(\gamma_{\tilde{\theta}})$), which itself is close to the loss from global optimizer γ^* (viz. $\mathrm{BCE}(\gamma^*)$) by Universal Function Approximation Theorem of neural-networks.

The BCE loss is strongly convex in γ. γ links BCE to $I(\cdot\,;\cdot)$, i.e., $|\mathrm{BCE}_n(\gamma_{\hat{\theta}}) - \mathrm{BCE}(\gamma^*)| \leq \epsilon' \implies \|\gamma_{\hat{\theta}} - \gamma^*\|_1 \leq \eta \implies |\hat{I}_n(U;V) - I(U;V)| \leq \epsilon$.

Lemma 3 (Likelihood-Ratio from Cross-Entropy Loss)**.** *The point-wise minimizer of binary cross-entropy loss $\gamma^*(x)$ is related to the likelihood ratio as $\frac{\gamma^*(x)}{1-\gamma^*(x)} = \frac{p(x)}{q(x)}$, where $\gamma^*(x) = Pr(l = 1|x)$ and l is the label of point x.*

Proof. The binary cross entropy loss as a function of gamma is defined in (5). Now,

$$\begin{aligned}\mathbb{E}_{XL} L \log \gamma(X) &= \sum_{x,l} p(x,l) l \log \gamma(x) \\ &= \sum_{x,l=1} p(x|l=1)p(l=1) \log \gamma(x) + 0 \\ &= \frac{1}{2} \sum_{x} p(x) \log \gamma(x)\end{aligned}$$

Similarly,

$$\mathbb{E}_{XL}(1-L)\log(1-\gamma(X)) = \frac{1}{2}\sum_x q(x)\log(1-\gamma(x))$$

Using these in the expression for $\mathrm{BCE}(\gamma)$, we obtain

$$\mathrm{BCE}(\gamma) = -\frac{1}{2}\left(\sum_{x\in\mathcal{X}} p(x)\log\gamma(x) + q(x)\log(1-\gamma(x))\right)$$

The point-wise minimizer γ^* of $\mathrm{BCE}(\gamma)$ gives $\frac{\gamma^*(x)}{1-\gamma^*(x)} = \frac{p(x)}{q(x)}$. □

Lemma 4 (Function Approximation)**.** *Given $\epsilon' > 0$, $\exists\, \tilde{\theta} \in \Theta$ such that*

$$\mathrm{BCE}(\gamma_{\tilde{\theta}}) \leq \mathrm{BCE}(\gamma^*) + \frac{\epsilon'}{2}$$

Proof. The last layer of the neural network being sigmoid (followed by clipping to $[\tau, 1-\tau]$) ensures that the outputs are bounded. So by the Universal Function Approximation Theorem for multi-layer feed-forward neural networks (Hornik et al. 1989), $\exists$ parameter $\tilde{\theta}$ such that $|\gamma^* - \gamma_{\tilde{\theta}}| \leq \epsilon''\,\forall\, x$, where $\gamma_{\tilde{\theta}}$ is the estimated classifier prediction function with parameter $\tilde{\theta}$. So,

$$|\mathrm{BCE}(\gamma_{\tilde{\theta}}) - \mathrm{BCE}(\gamma^*)| \leq \frac{1}{\tau}\epsilon''$$

since $\log$ is Lipshitz continuous with constant $\frac{1}{\tau}$. Choose $\epsilon'' = \frac{\epsilon'\tau}{2}$ to complete the proof. □

Lemma 5 (Generalization)**.** *Given $\epsilon', \delta > 0$, $\forall\, n \geq \frac{18M^2}{\epsilon'^2}(h\log(96KL\sqrt{d}/\epsilon') + \log(2/\delta))$, such that with probability at least $1-\delta$, we have*

$$\mathrm{BCE}_n(\gamma_{\hat{\theta}}) \leq \mathrm{BCE}(\gamma_{\tilde{\theta}}) + \frac{\epsilon'}{2}$$

Proof. Let $\hat{\theta} \leftarrow \arg\min_{\theta} \mathrm{BCE}_n^{\mathrm{ERM}}(\gamma_\theta)$.

From Hoeffding's inequality,

$$Pr\left(|\mathrm{BCE}_n^{\mathrm{ERM}}(\gamma_\theta) - \mathrm{BCE}(\gamma_\theta)| \geq \mu\right) \leq 2\exp\left(\frac{-2n\mu^2}{M^2}\right)$$

where $M = \log\left(\frac{1-\tau}{\tau}\right)$.

Similarly, for the test samples,

$$Pr\left(|\mathrm{BCE}_n(\gamma_\theta) - \mathrm{BCE}(\gamma_\theta)| \geq \mu\right) \leq 2\exp\left(\frac{-2n\mu^2}{M^2}\right) \tag{6}$$

We want this to hold for all parameters $\theta \in \Theta$. This is obtained using the covering number of the compact domain $\Theta \subset \mathbb{R}^h$. We use small balls $B_r(\theta_j)$ of radius r centered at θ_j so that $\Theta \subset \cup_j B_r(\theta_j)$ The covering number $\kappa(\Theta, r)$ is finite as Θ is compact and is bounded as

$$\kappa(\Theta, r) \leq \left(\frac{2K\sqrt{h}}{r}\right)^h$$

Using the union bound on these finite hypotheses,

$$\begin{aligned}&Pr\left(\max_\theta |\mathrm{BCE}_n^{\mathrm{ERM}}(\gamma_\theta) - \mathrm{BCE}(\gamma_\theta)| \geq \mu\right) \\ &\qquad \leq 2\kappa(\Theta, r)\exp\left(\frac{-2n\mu^2}{M^2}\right)\end{aligned} \tag{7}$$

Choose $r = \frac{\mu}{8L}$ (Mohri et al. 2018). Solving for number of samples n with $2\kappa(\Theta, r)\exp\left(\frac{-2n\mu^2}{M^2}\right) \leq \delta$, we obtain $n \geq \frac{M^2}{2\mu^2}(h\log(16KL\sqrt{d}/\mu) + \log(2/\delta))$.

So for $n \geq \frac{M^2}{2\mu^2}(h\log(16KL\sqrt{d}/\mu) + \log(2/\delta))$, with probability at least $1-\delta$,

$$\begin{aligned}\mathrm{BCE}_n(\gamma_{\hat\theta}) &\overset{(a)}{\leq} \mathrm{BCE}(\gamma_{\hat\theta}) + \mu \overset{(b)}{\leq} \mathrm{BCE}_n^{\mathrm{ERM}}(\gamma_{\hat\theta}) + 2\mu \\ &\overset{(c)}{\leq} \mathrm{BCE}_n^{\mathrm{ERM}}(\gamma_{\tilde\theta}) + 2\mu \overset{(d)}{\leq} \mathrm{BCE}(\gamma_{\tilde\theta}) + 3\mu\end{aligned}$$

(a) follows from (6). (b) and (d) follow from (7). (c) is due to the fact that $\hat\theta$ is the minimizer of train loss. Choosing $\mu = \epsilon'/6$ completes the proof. □

Lemma 6 (Convergence to minimizer)**.** *Given* $\epsilon' > 0$, $\exists\, \eta \left(= (1-\tau)\sqrt{\frac{2\lambda(\mathcal{X})\epsilon'}{\alpha}}\right) > 0$ *such that whenever* $\mathrm{BCE}(\gamma_\theta) - \mathrm{BCE}(\gamma^*) \leq \epsilon'$, *we have*

$$\|\vec{\gamma_\theta} - \vec{\gamma^*}\|_1 \leq \eta$$

where $\vec\gamma = [\gamma(x)]_{x\in\mathcal{X}}$ *and* $\lambda(\mathcal{X})$ *is the Lebesgue measure of compact set* $\mathcal{X} \subset \mathbb{R}^{d_x}$.

Proof.

$$\mathrm{BCE}(\gamma) = -\frac{1}{2}\left(\sum_{x\in\mathcal{X}} p(x)\log\gamma(x) + q(x)\log(1-\gamma(x))\right)$$

is α'-strongly convex as a function of $\vec\gamma$ under Assumption (A2), where $\alpha' = \frac{\alpha}{(1-\tau)^2}$. So $\forall\gamma$, $\frac{\partial^2\mathrm{BCE}}{\partial\gamma(x_k)\partial\gamma(x_l)} \geq \alpha'$ for $k = l$ and 0 otherwise. Using the Taylor expansion for strongly convex functions, we have

$$\begin{aligned}\mathrm{BCE}(\vec{\gamma_\theta}) \geq \mathrm{BCE}(\vec{\gamma^*}) &+ \langle\nabla\mathrm{BCE}(\vec{\gamma^*}), \vec{\gamma_\theta} - \vec{\gamma^*}\rangle \\ &+ \frac{\alpha'}{2}\|\vec{\gamma_\theta} - \vec{\gamma^*}\|_2^2\end{aligned}$$

Since $\vec{\gamma^*}$ is the minimizer, $\nabla\mathrm{BCE}(\vec{\gamma^*}) = 0$. So,

$$\begin{aligned}&\|\vec{\gamma^*} - \vec{\gamma_\theta}\|_2 \\ &\leq (1-\tau)\sqrt{\frac{2}{\alpha}\left(\mathrm{BCE}(\vec{\gamma_\theta}) - \mathrm{BCE}(\vec{\gamma^*})\right)} \\ &\implies \|\vec{\gamma^*} - \vec{\gamma_\theta}\|_2 \leq (1-\tau)\sqrt{\frac{2}{\alpha}\epsilon'}\end{aligned}$$

From Holder's inequality in finite measure space,

$$\begin{aligned}\|\vec{\gamma^*} - \vec{\gamma_\theta}\|_1 &\leq \sqrt{\lambda(\mathcal{X})}\|\vec{\gamma^*} - \vec{\gamma_\theta}\|_2 \\ &\leq (1-\tau)\sqrt{\frac{2}{\alpha}\lambda(\mathcal{X})\epsilon'} = \eta\end{aligned}$$

□

Lemma 7 (Estimation from Samples)**.** *Given* $\epsilon > 0$, *for any classifier with parameter* $\theta \in \Theta$, $\exists\, n \in \mathbb{N}$ *such that with probability* 1,

$$|I_n^{\gamma_\theta}(U;V) - I^{\gamma_\theta}(U;V)| \leq \frac{\epsilon}{2}$$

Proof. We denote the empirical estimates as $\underset{x\sim p_n}{\mathbb{E}}(\cdot)$ and $\underset{x\sim q_n}{\mathbb{E}}(\cdot)$ respectively. The proof essentially relies on the empirical mean of functions of independent random variables converging to the true mean. More specifically, we consider the functions $f^\theta(x) = \log\frac{\gamma^\theta(x)}{1-\gamma^\theta(x)}$ and $g^\theta(x) = \frac{\gamma^\theta(x)}{1-\gamma^\theta(x)}$. Since $\gamma(x) \in [\tau, 1-\tau]$, both $f(x)$ and $g(x)$ are bounded. ($f \in [\log\frac{\tau}{1-\tau}, \log\frac{1-\tau}{\tau}]$ and $g \in [\frac{\tau}{1-\tau}, \frac{1-\tau}{\tau}]$). Functions of independent random variables are independent. Also, since the functions are bounded, they have finite mean and variance. Invoking the law of large numbers, $\exists\, n \geq n_1'(\epsilon)$ such that with probability 1

$$|\underset{x\sim p_n}{\mathbb{E}} f^\theta - \underset{x\sim p}{\mathbb{E}} f^\theta| \leq \frac{\epsilon}{4} \tag{8}$$

and $\exists\, n \geq n_2'(\epsilon)$ such that with probability 1

$$|\underset{x\sim q_n}{\mathbb{E}} g^\theta - \underset{x\sim q}{\mathbb{E}} g^\theta| \leq \frac{\epsilon\tau}{4(1-\tau)} \tag{9}$$

Then, for $n \geq \max(n_1'(\epsilon), n_2'(\epsilon))$, we have with probability 1

$$\begin{aligned}&|I_n^{\gamma_\theta}(U;V) - I^{\gamma_\theta}(U;V)| \\ &\leq |\underset{x\sim p_n}{\mathbb{E}} f^\theta - \underset{x\sim p}{\mathbb{E}} f^\theta| + |\log\underset{x\sim q_n}{\mathbb{E}} g^\theta - \log\underset{x\sim q}{\mathbb{E}} g^\theta| \\ &\leq |\underset{x\sim p_n}{\mathbb{E}} f^\theta - \underset{x\sim p}{\mathbb{E}} f^\theta| + \frac{1-\tau}{\tau}|\underset{x\sim q_n}{\mathbb{E}} g^\theta - \underset{x\sim q}{\mathbb{E}} g^\theta| \\ &= \frac{\epsilon}{4} + \frac{\epsilon}{4} = \frac{\epsilon}{2}\end{aligned}$$

where in the last inequality, we use the Lipschitz constant for $\log$ with the bounded function g as argument. □

Proof of Theorem 3

Using Proposition 1, $I^{\gamma^*}(U;V) = I(U;V)$, where γ^* is the unique global minimizer of $\mathrm{BCE}(\gamma)$.

The empirical risk minimizer of BCE loss is $\hat\theta$. For a rich enough class Θ and large enough samples n, Lemma 5 and Lemma 4 combine to give $\mathrm{BCE}_n(\gamma_{\hat\theta}) - \mathrm{BCE}(\gamma^*) \leq \epsilon'$. Applying Lemma 6 with $\epsilon' = \frac{\alpha}{8\lambda(\mathcal{X})}\left(\frac{\eta}{\beta(1-\tau}\right)^2$, we have $\|\vec{\gamma^*} - \vec{\gamma_{\hat\theta}}\|_1 \leq \frac{\eta}{2\beta}$. This further implies that

$$\underset{x\sim p}{\mathbb{E}}|\gamma^* - \hat\gamma_{\hat\theta}| \leq \frac{\eta}{2} \tag{10}$$

and

$$\underset{x\sim q}{\mathbb{E}}|\gamma^* - \hat\gamma_{\hat\theta}| \leq \frac{\eta}{2} \tag{11}$$

We now compute the Lipschitz constant for $f = \log\frac{\gamma}{1-\gamma}$ as a function of γ, which links the classifier predictions to Donsker-Varadhan representation.

$$|f^* - \hat{f}^{\hat{\theta}}| = |\log \frac{\gamma^*}{1-\gamma^*} - \log \frac{\hat{\gamma}_{\hat{\theta}}}{1-\hat{\gamma}_{\hat{\theta}}}| \leq \frac{1}{\tau^2}|\gamma^* - \hat{\gamma}_{\hat{\theta}}|$$

and

$$|e^{f^*} - e^{\hat{f}^{\hat{\theta}}}| = |\frac{\gamma^*}{1-\gamma^*} - \frac{\hat{\gamma}_{\hat{\theta}}}{1-\hat{\gamma}_{\hat{\theta}}}| \leq \frac{1}{\tau^2}|\gamma^* - \hat{\gamma}_{\hat{\theta}}|$$

For $\gamma \in [\tau, 1-\tau]$, the function $f \in [\log \frac{\tau}{1-\tau}, \log \frac{1-\tau}{\tau}]$ is continuous and bounded with Lipschitz constant $\frac{1}{\tau^2}$. So, using (10) and (11),

$$\underset{x\sim p}{\mathbb{E}} |f^* - \hat{f}^{\hat{\theta}}| \leq \frac{1}{\tau^2}\frac{\eta}{2} \text{ and } \underset{x\sim q}{\mathbb{E}} |e^{f^*} - e^{\hat{f}^{\hat{\theta}}}| \leq \frac{1}{\tau^2}\frac{\eta}{2}$$

Finally, from the Donsker-Varadhan representation 1,

$$\begin{aligned} &|I(U;V) - I^{\gamma_{\hat{\theta}}}(U;V)| \leq |\underset{x\sim p}{\mathbb{E}} f^* - \underset{x\sim p}{\mathbb{E}} \hat{f}^{\hat{\theta}}| + \\ &|\log \underset{x\sim q}{\mathbb{E}} e^{f^*} - \log \underset{x\sim q}{\mathbb{E}} e^{\hat{f}^{\hat{\theta}}}| \\ &\leq \underset{x\sim p}{\mathbb{E}} |f^* - \hat{f}^{\hat{\theta}}| + \underset{x\sim q}{\mathbb{E}} |e^{f^*} - e^{\hat{f}^{\hat{\theta}}}| \\ &= \frac{\eta}{2\tau^2} + \frac{\eta}{2\tau^2} = \frac{\eta}{\tau^2} \end{aligned} \tag{12}$$

where we use the inequality $\log(t) \leq t - 1$ coupled with the fact that $\underset{x\sim q}{\mathbb{E}} e^{f^*} = 1$. Given $\epsilon > 0$, we choose $\eta = \tau^2 \frac{\epsilon}{2}$.

To complete the proof, we combine the above result (12) with Lemma 7 using Triangle Inequality,

$$\begin{aligned} &|I_n^{\gamma_{\hat{\theta}}}(U;V) - I(U;V)| \\ \leq &|I_n^{\gamma_{\hat{\theta}}}(U;V) - I^{\gamma_{\hat{\theta}}}(U;V)| + |I^{\gamma_{\hat{\theta}}}(U;V) - I(U;V)| \\ &\frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \end{aligned}$$

Corollary 1. *CCMI is consistent.*

Proof. For each individual MI estimation, we can obtain the classifier parameter θ_1(resp. θ_2) $\in \Theta$ such that Theorem 1 holds with approximation accuracy $\epsilon/2$. So, $\exists n \geq n_1(\epsilon/2)$ such that with probability at least $1-\delta$

$$|\hat{I}_n^{\gamma_{\theta_1}}(X;YZ) - I(X;YZ)| \leq \frac{\epsilon}{2}$$

and $n \geq n_2(\epsilon/2)$ such that with probability at least $1-\delta$

$$|\hat{I}_n^{\gamma_{\theta_2}}(X;Z) - I(X;Z)| \leq \frac{\epsilon}{2}$$

Using Triangle inequality, for $n \geq \max(n_1, n_2)$, with probability at least $1-\delta$, we have

$$\begin{aligned} &|\hat{I}_n(X;Y|Z) - I(X;Y|Z)| \\ &= |\hat{I}_n^{\gamma_{\theta_1}}(X;Y,Z) - \hat{I}_n^{\gamma_{\theta_2}}(X;Z) - I(X;Y,Z) + I(X;Z)| \\ &\leq |\hat{I}_n^{\gamma_{\theta_1}}(X;Y,Z) - I(X;Y,Z)| + |\hat{I}_n^{\gamma_{\theta_2}}(X;Z) - I(X;Z)| \\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \end{aligned}$$

□

Theorem 4 (Theorem 2 restated)**.** *The finite sample estimate from Classifier-MI is a lower bound on the true MI value with high probability, i.e., given n test samples and the trained classifier parameter $\hat{\theta}$, we have for $\epsilon > 0$*

$$Pr(I(U;V) + \epsilon \geq I_n^{\gamma_{\hat{\theta}}}(U;V)) \geq 1 - 2\exp(-Cn)$$

where C is some constant independent of n and the dimension of the data.

Proof.

$$I(U;V) = \max_{\gamma} I^{\gamma}(U;V)) \geq \max_{\theta} I^{\gamma_\theta}(U;V)) \geq I^{\gamma_{\hat{\theta}}}(U;V))$$

We apply one-sided Hoeffding's inequality to (8) and (9) with given $\epsilon > 0$,

$$\begin{aligned} &Pr(\underset{x\sim p_n}{\mathbb{E}} f^{\hat{\theta}} - \underset{x\sim p}{\mathbb{E}} f^{\hat{\theta}} \leq \frac{\epsilon}{2}) \\ &\geq 1 - \exp\left(-\frac{n\epsilon^2}{8(\log((1-\tau)/\tau))^2}\right) \\ &= 1 - \exp(-C_1 n\epsilon^2) \end{aligned}$$

$$\begin{aligned} &Pr\left(\underset{x\sim p}{\mathbb{E}} g^{\hat{\theta}} - \underset{x\sim p_n}{\mathbb{E}} g^{\hat{\theta}} \leq \frac{\epsilon\tau}{2(1-\tau)}\right) \\ &\geq 1 - \exp\left(-\frac{n\epsilon^2}{2}\left(\frac{\tau}{1-\tau}\right)^4\right) = 1 - \exp(-C_2 n\epsilon^2) \end{aligned}$$

$$Pr\left(I_n^{\gamma_{\hat{\theta}}}(U;V)) \leq I^{\gamma_{\hat{\theta}}}(U;V)) + \epsilon\right) \geq 1 - 2\exp(-Cn)$$

where $C = \epsilon^2 \min(C_1, C_2)$.

□