

A. Proofs of Theorems

Proof of Theorem 3.2. We calculate the gap between $I_{\text{vCLUB}}(\mathbf{x}; \mathbf{y})$ and $I(\mathbf{x}; \mathbf{y})$:

$$\begin{aligned}
 \tilde{\Delta} &:= I_{\text{vCLUB}}(\mathbf{x}; \mathbf{y}) - I(\mathbf{x}; \mathbf{y}) \\
 &= \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} [\log q_\theta(\mathbf{y}|\mathbf{x})] - \mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{p(\mathbf{y})} [\log q_\theta(\mathbf{y}|\mathbf{x})] - \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} [\log p(\mathbf{y}|\mathbf{x}) - \log p(\mathbf{y})] \\
 &= [\mathbb{E}_{p(\mathbf{y})} [\log p(\mathbf{y})] - \mathbb{E}_{p(\mathbf{x})p(\mathbf{y})} [\log q_\theta(\mathbf{y}|\mathbf{x})]] - [\mathbb{E}_{p(\mathbf{x}, \mathbf{y})} [\log p(\mathbf{y}|\mathbf{x})] - \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} [\log q_\theta(\mathbf{y}|\mathbf{x})]] \\
 &= \mathbb{E}_{p(\mathbf{x})p(\mathbf{y})} [\log \frac{p(\mathbf{y})}{q_\theta(\mathbf{y}|\mathbf{x})}] - \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} [\log \frac{p(\mathbf{y}|\mathbf{x})}{q_\theta(\mathbf{y}|\mathbf{x})}] \\
 &= \mathbb{E}_{p(\mathbf{x})p(\mathbf{y})} [\log \frac{p(\mathbf{x})p(\mathbf{y})}{q_\theta(\mathbf{y}|\mathbf{x})p(\mathbf{x})}] - \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} [\log \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{q_\theta(\mathbf{y}|\mathbf{x})p(\mathbf{x})}] \\
 &= \text{KL}(p(\mathbf{x})p(\mathbf{y})\|q_\theta(\mathbf{x}, \mathbf{y})) - \text{KL}(p(\mathbf{x}, \mathbf{y})\|q_\theta(\mathbf{x}, \mathbf{y})).
 \end{aligned}$$

Therefore, $I_{\text{vCLUB}}(\mathbf{x}; \mathbf{y})$ is an upper bound of $I(\mathbf{x}; \mathbf{y})$ if and only if $\text{KL}(p(\mathbf{x})p(\mathbf{y})\|q_\theta(\mathbf{x}, \mathbf{y})) \geq \text{KL}(p(\mathbf{x}, \mathbf{y})\|q_\theta(\mathbf{x}, \mathbf{y}))$.

If \mathbf{x} and \mathbf{y} are independent, $p(\mathbf{x})p(\mathbf{y}) = p(\mathbf{x}, \mathbf{y})$. Then, $\text{KL}(p(\mathbf{x})p(\mathbf{y})\|q_\theta(\mathbf{x}, \mathbf{y})) = \text{KL}(p(\mathbf{x}, \mathbf{y})\|q_\theta(\mathbf{x}, \mathbf{y}))$ and $\tilde{\Delta} = 0$. Therefore, $I_{\text{vCLUB}}(\mathbf{x}; \mathbf{y}) = I(\mathbf{x}; \mathbf{y})$, the equality holds. \square

Proof of Corollary 3.3. If $\text{KL}(p(\mathbf{y}|\mathbf{x})\|q_\theta(\mathbf{y}|\mathbf{x})) \leq \epsilon$, then

$$\text{KL}(p(\mathbf{x}, \mathbf{y})\|q_\theta(\mathbf{x}, \mathbf{y})) = \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} [\log \frac{p(\mathbf{x}, \mathbf{y})}{q_\theta(\mathbf{x}, \mathbf{y})}] = \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} [\log \frac{p(\mathbf{y}|\mathbf{x})}{q_\theta(\mathbf{y}|\mathbf{x})}] = \text{KL}(p(\mathbf{y}|\mathbf{x})\|q_\theta(\mathbf{y}|\mathbf{x})) \leq \epsilon.$$

By the condition $\text{KL}(p(\mathbf{x}, \mathbf{y})\|q_\theta(\mathbf{x}, \mathbf{y})) > \text{KL}(p(\mathbf{x})p(\mathbf{y})\|q_\theta(\mathbf{x}, \mathbf{y}))$, we have $\text{KL}(p(\mathbf{x})p(\mathbf{y})\|q_\theta(\mathbf{x}, \mathbf{y})) < \epsilon$.

Note that the KL-divergence is always non-negative. From the proof of Theorem 3.2,

$$\begin{aligned}
 |I_{\text{vCLUB}}(\mathbf{x}; \mathbf{y}) - I(\mathbf{x}; \mathbf{y})| &= |\text{KL}(p(\mathbf{x})p(\mathbf{y})\|q_\theta(\mathbf{x}, \mathbf{y})) - \text{KL}(p(\mathbf{x}, \mathbf{y})\|q_\theta(\mathbf{x}, \mathbf{y}))| \\
 &< \max \{ \text{KL}(p(\mathbf{x})p(\mathbf{y})\|q_\theta(\mathbf{x}, \mathbf{y})), \text{KL}(p(\mathbf{x}, \mathbf{y})\|q_\theta(\mathbf{x}, \mathbf{y})) \} \leq \epsilon,
 \end{aligned}$$

which supports the claim. \square

B. Network Expressiveness in Variational Inference

In Section 3.2, when analyze the properties of the vCLUB estimator, we claim a reasonable assumption that with high expressiveness of the neural network $q_\theta(\mathbf{y}|\mathbf{x})$, we can achieve $\text{KL}(p(\mathbf{y}|\mathbf{x})\|q_\theta(\mathbf{y}|\mathbf{x})) < \epsilon$. Here we provide a analysis under the scenario that the conditional distribution is a Gaussian distribution, $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}^*(\mathbf{x}), \mathbf{I})$. The variational approximation $q_\theta(\mathbf{y}|\mathbf{x})$ is parameterized by $q_\theta(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\theta(\mathbf{x}), \mathbf{I})$.

Then training samples pair $(\mathbf{x}_i, \mathbf{y}_i)$ can be treated as $(\mathbf{x}_i, \boldsymbol{\mu}^*(\mathbf{x}_i) + \boldsymbol{\xi}_i)$, where $\boldsymbol{\xi}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Then

$$\begin{aligned}
 \log p(\mathbf{y}|\mathbf{x}) &= \log \prod_{d=1}^D [\frac{1}{\sqrt{2\pi}} e^{(y^{(d)} - \mu^{*(d)}(\mathbf{x}))^2/2}] = -\frac{D}{2} \log(2\pi) - \frac{1}{2} \|\mathbf{y} - \boldsymbol{\mu}^*(\mathbf{x})\|^2, \\
 \log q_\theta(\mathbf{y}|\mathbf{x}) &= \log \prod_{d=1}^D [\frac{1}{\sqrt{2\pi}} e^{(y^{(d)} - \mu_\theta^{(d)}(\mathbf{x}))^2/2}] = -\frac{D}{2} \log(2\pi) - \frac{1}{2} \|\mathbf{y} - \boldsymbol{\mu}_\theta(\mathbf{x})\|^2.
 \end{aligned}$$

The log-ratio between $p(\mathbf{y}_i|\mathbf{x}_i)$ and $q_\theta(\mathbf{y}_i|\mathbf{x}_i)$ is

$$\log \frac{p(\mathbf{y}_i|\mathbf{x}_i)}{q_\theta(\mathbf{y}_i|\mathbf{x}_i)} = \log p(\mathbf{y}_i|\mathbf{x}_i) - \log q_\theta(\mathbf{y}_i|\mathbf{x}_i) = [\boldsymbol{\mu}^*(\mathbf{x}_i) - \boldsymbol{\mu}_\theta(\mathbf{x}_i)]^T [\mathbf{y}_i - \boldsymbol{\mu}_\theta(\mathbf{x}_i) + \boldsymbol{\xi}_i].$$

We further assume $\|\boldsymbol{\mu}^*(\mathbf{x}) - \boldsymbol{\mu}_\theta(\mathbf{x})\| < A$ is bounded. Then $|\log p(\mathbf{y}_i|\mathbf{x}_i) - \log q_\theta(\mathbf{y}_i|\mathbf{x}_i)| < A \|\mathbf{y}_i - \boldsymbol{\mu}_\theta(\mathbf{x}_i) + \boldsymbol{\xi}_i\|$.

Denote a loss function $l(\boldsymbol{\mu}_\theta(\mathbf{x}_i), \mathbf{y}_i) = \|\mathbf{y}_i - \boldsymbol{\mu}_\theta(\mathbf{x}_i) + \boldsymbol{\xi}_i\|$. With all reasonable assumptions in Hu et al. (2019), and applying the Theorem 5.1 in Hu et al. (2019), we know that when the number of samples $n \rightarrow \infty$, the expected error $\mathbb{E}_{p(\mathbf{x}, \mathbf{y})} [l(\boldsymbol{\mu}_\theta(\mathbf{x}), \mathbf{y})] \rightarrow \infty$ with probability $1 - \delta$.

$$\text{KL}(p(\mathbf{y}|\mathbf{x})\|q_\theta(\mathbf{y}|\mathbf{x})) = \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} [\log p(\mathbf{y}|\mathbf{x}) - \log q_\theta(\mathbf{y}|\mathbf{x})] < A \cdot \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} [l(\boldsymbol{\mu}_\theta(\mathbf{x}), \mathbf{y})].$$

Therefore, when given a small number $\varepsilon > 0$, having the sample size n large enough, we can guarantee that $\text{KL}(p(\mathbf{y}|\mathbf{x})\|q_\theta(\mathbf{y}|\mathbf{x}))$ is smaller than ε .

C. Properties of Variational Upper Bounds

In the Section 2, we introduce two variational MI upper bounds with neural network approximation $q_\theta(\mathbf{y}|\mathbf{x})$ to $p(\mathbf{y}|\mathbf{x})$:

$$\begin{aligned} I_{\text{vVUB}}(\mathbf{x}; \mathbf{y}) &= \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} \left[\log \frac{q_\theta(\mathbf{y}|\mathbf{x})}{r(\mathbf{y})} \right], \\ I_{\text{vL1Out}}(\mathbf{x}; \mathbf{y}) &= \mathbb{E} \left[\frac{1}{N} \sum_{i=1}^N \left[\log \frac{q_\theta(\mathbf{y}_i|\mathbf{x}_i)}{\frac{1}{N-1} \sum_{j \neq i} q_\theta(\mathbf{y}_i|\mathbf{x}_j)} \right] \right]. \end{aligned}$$

With the neural approximation $q_\theta(\mathbf{y}|\mathbf{x})$, I_{vVUB} and I_{vL1Out} no longer guarantee to be the MI upper bounds. However, both of the two estimators have good properties with a good approximation $q_\theta(\mathbf{y}|\mathbf{x})$.

Theorem C.1. *If $q_\theta(\mathbf{y}|\mathbf{x})$ satisfies $\text{KL}(p(\mathbf{y}|\mathbf{x})\|q_\theta(\mathbf{y}|\mathbf{x})) \leq \text{KL}(p(\mathbf{y})\|r(\mathbf{y}))$, then $I(\mathbf{x}; \mathbf{y}) \leq I_{\text{vVUB}}(\mathbf{x}; \mathbf{y})$.*

Proof of Theorem C.1. With the conditional $\text{KL}(p(\mathbf{y}|\mathbf{x})\|q_\theta(\mathbf{y}|\mathbf{x})) \leq \text{KL}(p(\mathbf{y})\|r(\mathbf{y}))$,

$$\begin{aligned} I(\mathbf{x}; \mathbf{y}) &= \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} \left[\log \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})} \right] = \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} \left[\log \left(\frac{p(\mathbf{y}|\mathbf{x})}{q_\theta(\mathbf{y}|\mathbf{x})} \cdot \frac{q_\theta(\mathbf{y}|\mathbf{x})}{r(\mathbf{y})} \cdot \frac{r(\mathbf{y})}{p(\mathbf{y})} \right) \right] \\ &= \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} \left[\log \frac{q_\theta(\mathbf{y}|\mathbf{x})}{r(\mathbf{y})} \right] + \text{KL}(p(\mathbf{y}|\mathbf{x})\|q_\theta(\mathbf{y}|\mathbf{x})) - \text{KL}(p(\mathbf{y})\|r(\mathbf{y})) \leq \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} \left[\log \frac{q_\theta(\mathbf{y}|\mathbf{x})}{r(\mathbf{y})} \right]. \end{aligned}$$

□

Theorem C.2. *Given $N - 1$ samples $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{N-1}$ from the marginal $p(\mathbf{x})$, If*

$$\text{KL}(p(\mathbf{y}|\mathbf{x})\|q_\theta(\mathbf{y}|\mathbf{x})) \leq \mathbb{E}_{\mathbf{x}_i \sim p(\mathbf{x})} \left[\text{KL} \left(p(\mathbf{y}) \left\| \frac{1}{N-1} \sum_{i=1}^{N-1} q_\theta(\mathbf{y}|\mathbf{x}_i) \right. \right) \right],$$

then $I(\mathbf{x}; \mathbf{y}) \leq I_{\text{vL1Out}}(\mathbf{x}; \mathbf{y})$.

Proof. Assume we have N sample pairs $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ drawn from $p(\mathbf{x}, \mathbf{y})$, then

$$\begin{aligned} I(\mathbf{x}; \mathbf{y}) &= \mathbb{E}_{(\mathbf{x}_i, \mathbf{y}_i) \sim p(\mathbf{x}, \mathbf{y})} \left[\frac{1}{N} \sum_{i=1}^N \left[\log \frac{p(\mathbf{y}_i|\mathbf{x}_i)}{p(\mathbf{y}_i)} \right] \right] \\ &= \mathbb{E}_{(\mathbf{x}_i, \mathbf{y}_i) \sim p(\mathbf{x}, \mathbf{y})} \left[\frac{1}{N} \sum_{i=1}^N \left[\log \left(\frac{p(\mathbf{y}_i|\mathbf{x}_i)}{q_\theta(\mathbf{y}_i|\mathbf{x}_i)} \cdot \frac{q_\theta(\mathbf{y}_i|\mathbf{x}_i)}{\frac{1}{N-1} \sum_{j \neq i} q_\theta(\mathbf{y}_i|\mathbf{x}_j)} \cdot \frac{\frac{1}{N-1} \sum_{j \neq i} q_\theta(\mathbf{y}_i|\mathbf{x}_j)}{p(\mathbf{y}_i)} \right) \right] \right] \\ &= \text{KL}(p(\mathbf{y}|\mathbf{x})\|q_\theta(\mathbf{y}|\mathbf{x})) + I_{\text{vVUB}}(\mathbf{x}; \mathbf{y}) - \mathbb{E} \left[\frac{1}{N} \sum_{i=1}^N \text{KL} \left(p(\mathbf{y}) \left\| \frac{1}{N-1} \sum_{j \neq i} q_\theta(\mathbf{y}|\mathbf{x}_j) \right. \right) \right]. \end{aligned}$$

Apply the condition in Theorem C.2 to each $N - 1$ combination of $\{\mathbf{x}_j\}_{j \neq i}$, we conclude $I(\mathbf{x}; \mathbf{y}) \leq I_{\text{vL1Out}}(\mathbf{x}; \mathbf{y})$. □

Theorem C.1 and Theorem C.2 indicate that if the approximation $q_\theta(\mathbf{y}|\mathbf{x})$ is good enough, the estimators I_{vVUB} and I_{vL1Out} can remain as MI upper bounds. Based on the analysis in Section B, when implemented with neural networks, the approximation can be far more accurate to preserve the variational estimators as MI upper bounds.

D. Detailed Experimental Setups

Information Bottleneck: For the experiment on information bottleneck, we follow the setup from Alemi et al. (2016). The parameters $\mu_\sigma(x)$ and $\Sigma_\sigma(x)$ are the output from a MLP with layers $784 \rightarrow 1024 \rightarrow 1024 \rightarrow 2K$, where K is the size of the bottleneck. We set $K = 256$. For the variational classifier to implement the Barber-Agakov MI lower bound, the structure is set to a one-layer MLP. The batch size is 100. We set our learning rate to 10^{-4} , with an exponential decay rate of 0.97 and a decay step of 1200.

Domain Adaptation: The network is constructed as follows. Both feature extractors (*i.e.*, E_c and E_d) are nine-layer convolutional neural network with leaky ReLU non-linearities. The content classifier C and the domain discriminator D are a one-layer and a two-layer MLPs, respectively. Images from each domain are normalized using Gaussian normalization.

Classifier C	Discriminator D	Extractor (both E_c and E_d)
Content feature z_c^s	Domain feature z_d	Input data x
MLP output $C(z_c^s)$ with shape 10	MLP, 64 ReLU MLP output $D(z_d)$ with shape 2	3×3 conv. 64 lReLU, stride 1 3×3 conv. 64 lReLU, stride 1 3×3 conv. 64 lReLU, stride 1 2×2 max pool, stride 2, dropout, $p = 0.5$, Gaussian noise, $\sigma = 1$ 3×3 conv. 64 lReLU, stride 1 3×3 conv. 64 lReLU, stride 1 3×3 conv. 64 lReLU, stride 1 2×2 max pool, stride 2, dropout, $p = 0.5$, Gaussian noise, $\sigma = 1$ 3×3 conv. 64 lReLU, stride 1 3×3 conv. 64 lReLU, stride 1 3×3 conv. 64 lReLU, stride 1 global average pool, output feature with shape 64

Table 3. Model architecture for the unsupervised domain adaptation experiments.

E. Numerical Results of MI Estimation

MI	Gaussian					Cubic				
	2	4	6	8	10	2	4	6	8	10
VUB	3.85	15.33	34.37	61.25	95.70	2.09	10.38	25.56	47.84	77.59
NWJ	1.67	7.20	17.46	33.26	55.34	1.10	5.54	14.68	30.25	51.07
MINE	1.61	6.66	16.01	29.60	49.87	1.53	6.58	17.4	34.20	59.46
NCE	0.59	2.85	8.56	19.66	37.79	0.45	1.89	6.70	17.48	35.86
L1Out	0.13	0.11	0.75	4.65	17.08	2.30	5.58	8.92	8.27	7.19
CLUB	0.15	0.12	0.70	4.53	16.57	2.22	5.89	8.25	8.23	6.93
CLUBSample	0.38	0.44	1.31	5.30	17.63	2.37	5.89	8.07	8.87	7.54

Table 4. MSE of MI estimation