CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information

Pengyu Cheng [1] **Weituo Hao** [1] **Shuyang Dai** [1] **Jiachang Liu** [1] **Zhe Gan** [2] **Lawrence Carin** [1]

Abstract

There has been considerable recent interest in mutual information (MI) minimization for various machine learning tasks. However, estimating and minimizing MI in high-dimensional spaces remains a challenging problem, especially when only samples are accessible, rather than the underlying distribution forms. Previous works mainly focus on MI lower bound approximation, which is not applicable to MI minimization problems. In this paper, we propose a novel Contrastive Log-ratio Upper Bound (CLUB) of mutual information. We provide a theoretical analysis of the properties of CLUB and its variational approximation. Based on this upper bound, we introduce an accelerated MI minimization training scheme, that bridges MI minimization with *negative sampling*. Simulation studies on Gaussian distributions show that CLUB provides reliable estimates. Real-world MI minimization experiments, including domain adaptation and the information bottleneck, further demonstrate the effectiveness of the proposed method.

1. Introduction

Mutual information (MI) is a fundamental measure of the dependence between two random variables. Mathematically, the definition of MI between variables $\boldsymbol{x}$ and $\boldsymbol{y}$ is

$$\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}\right]. \tag{1}$$

This important tool has been applied in a wide range of scientific fields, including statistics (Granger & Lin, 1994; Jiang et al., 2015), bioinformatics (Lachmann et al., 2016; Zea et al., 2016), robotics (Julian et al., 2014; Charrow et al., 2015), and machine learning (Chen et al., 2016; Alemi et al., 2016; Hjelm et al., 2018; Cheng et al., 2020b).

In machine learning, especially in deep learning frameworks, MI is typically utilized as a criterion or a regularizer in loss functions, to encourage or limit the dependence between variables. MI maximization has been studied extensively in various tasks, *e.g.*, representation learning (Hjelm et al., 2018; Hu et al., 2017), generative models (Chen et al., 2016), information distillation (Ahn et al., 2019), and reinforcement learning (Florensa et al., 2017). Recently, MI minimization has received increased attention for its applications in disentangled representation learning (Chen et al., 2018), style transfer (Kazemi et al., 2018), domain adaptation (Gholami et al., 2018), fairness (Kamishima et al., 2011), and the information bottleneck (Alemi et al., 2016).

However, only in a few special cases can one calculate the exact value of mutual information, since the calculation requires closed forms of density functions and a tractable log-density ratio between the joint and marginal distributions. In most machine learning tasks, only samples from the joint distribution are accessible. Therefore, sample-based MI estimation methods have been proposed. To approximate MI, most previous works focused on lower-bound estimation (Chen et al., 2016; Belghazi et al., 2018; Oord et al., 2018), which is inconsistent with MI minimization tasks. In contrast, MI upper bound estimation lacks extensive exploration in the literature. Among the existing MI upper bounds, Alemi et al. (2016) fixes one of the marginal distribution to a standard Gaussian, and obtains a variational upper bound in closed form. However, the Gaussian marginal distribution assumption is unduly strong, which makes the upper bound fail to estimate MI with low bias. Poole et al. (2019) develops a leave-one-out upper bound, that provides tighter MI estimation when the sample size is large. However, it suffers from high numerical instability in practice when applied to MI minimization models.

To overcome the defects of previous MI estimators, we introduce a **C**ontrastive **L**og-ratio **U**pper **B**ound (CLUB). Specifically, CLUB bridges mutual information estimation with contrastive learning (Oord et al., 2018), where MI is estimated by the difference of conditional probabilities between positive and negative sample pairs. Further, we develop a variational form of CLUB (vCLUB) into scenar-

[1]Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina, USA [2]Microsoft, Redmond, Washington, USA. Correspondence to: Pengyu Cheng <pengyu.cheng@duke.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

ios where the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ is unknown, by approximating $p(\boldsymbol{y}|\boldsymbol{x})$ with a neural network. We prove that, with good variational approximation, vCLUB can either provide reliable MI estimation or remain a valid MI upper bound. Based on this new bound, we propose an MI minimization algorithm, and further accelerate it via a negative sampling strategy. The main contributions of this paper are summarized as follows.

- We introduce a Contrastive Log-ratio Upper Bound (CLUB) of mutual information, which is not only reliable as a mutual information estimator, but also trainable in gradient-descent frameworks.
- We extend CLUB with a variational network approximation, and provide theoretical analysis to the good properties of this variational bound.
- We develop a CLUB-based MI minimization algorithm, and accelerate it with a negative sampling strategy.
- We compare CLUB with previous MI estimators on both simulation studies and real-world applications, demonstrating that CLUB is not only better in the bias-variance estimation trade-off, but also more effective when applied to MI minimization.

2. Background

Although it has widespread use in numerous applications, mutual information (MI) remains challenging to estimate accurately, especially when the closed forms of distributions are unknown or intractable. Earlier MI estimation approaches include non-parametric binning (Darbellay & Vajda, 1999), kernel density estimation (Härdle et al., 2004), likelihood-ratio estimation (Suzuki et al., 2008), and K-nearest neighbor entropy estimation (Kraskov et al., 2004). These methods fail to provide reliable approximations when the data dimension increases (Belghazi et al., 2018). Also, the gradient of these estimators is difficult to calculate, which makes them inapplicable to back-propagation frameworks for MI optimization tasks.

To obtain differentiable and scalable MI estimation, recent approaches utilize deep neural networks to construct variational MI estimators. Most of these estimators focus on problems involving MI maximization, and provide MI lower bounds. Specifically, Barber & Agakov (2003) replaces the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ with an auxiliary distribution $q(\boldsymbol{y}|\boldsymbol{x})$, and obtains the Barber-Agakov (BA) bound:

$$\mathrm{I}_{\text{BA}} := \mathrm{H}(\boldsymbol{x}) + \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log q(\boldsymbol{x}|\boldsymbol{y})] \leq \mathrm{I}(\boldsymbol{x};\boldsymbol{y}), \quad (2)$$

where $\mathrm{H}(\boldsymbol{x})$ is the entropy of variable $\boldsymbol{x}$. Belghazi et al. (2018) introduces a Mutual Information Neural Estimator (MINE), that treats MI as the Kullback-Leibler (KL) divergence (Kullback, 1997) between the joint and marginal distributions, and converts it into the dual representation:

$$\mathrm{I}_{\text{MINE}} := \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[f(\boldsymbol{x},\boldsymbol{y})] - \log(\mathbb{E}_{p(\boldsymbol{x})p(\boldsymbol{y})}[e^{f(\boldsymbol{x},\boldsymbol{y})}]), \quad (3)$$

where $f(\cdot,\cdot)$ is a score function (or, a critic) approximated by a neural network. Nguyen, Wainwright, and Jordan (NWJ) (Nguyen et al., 2010) derives another lower bound based on the MI f-divergence representation:

$$\mathrm{I}_{\text{NWJ}} := \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[f(\boldsymbol{x},\boldsymbol{y})] - \mathbb{E}_{p(\boldsymbol{x})p(\boldsymbol{y})}[e^{f(\boldsymbol{x},\boldsymbol{y})-1}]. \quad (4)$$

More recently, based on Noise Contrastive Estimation (NCE) (Gutmann & Hyvärinen, 2010), an MI lower bound, called InfoNCE, was introduced in Oord et al. (2018):

$$\mathrm{I}_{\text{NCE}} := \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{f(\boldsymbol{x}_i,\boldsymbol{y}_i)}}{\frac{1}{N}\sum_{j=1}^{N}e^{f(\boldsymbol{x}_i,\boldsymbol{y}_j)}}\right], \quad (5)$$

where the expectation is over N samples $\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^{N}$ drawn from the joint distribution $p(\boldsymbol{x},\boldsymbol{y})$.

Unlike the above MI lower bounds that have been studied extensively, MI upper bounds are still lacking extensive published exploration. Most existing MI upper bounds require the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ to be known. For example, Alemi et al. (2016) introduces a variational marginal approximation $r(\boldsymbol{y})$ to build a variational upper bound (VUB):

$$\begin{aligned}\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) &= \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log\frac{p(\boldsymbol{y}|\boldsymbol{x})}{p(\boldsymbol{y})}] \\ &= \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log\frac{p(\boldsymbol{y}|\boldsymbol{x})}{r(\boldsymbol{y})}] - \mathrm{KL}(p(\boldsymbol{y})\|r(\boldsymbol{y})) \\ &\leq \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log\frac{p(\boldsymbol{y}|\boldsymbol{x})}{r(\boldsymbol{y})}] = \mathrm{KL}(p(\boldsymbol{y}|\boldsymbol{x})\|r(\boldsymbol{y})). \quad (6)\end{aligned}$$

The inequality is based on the fact that the KL-divergence is always non-negative. To be a good MI estimation, this upper bound requires a well-learned density approximation $r(\boldsymbol{y})$ to $p(\boldsymbol{y})$, so that the difference $\mathrm{KL}(p(\boldsymbol{y})\|r(\boldsymbol{y}))$ is small. However, learning a good marginal approximation $r(\boldsymbol{y})$ without any additional information, recognized as the distribution density estimation problem (Magdon-Ismail & Atiya, 1999), is challenging, especially when variable $\boldsymbol{y}$ is in a high-dimensional space. In practice, Alemi et al. (2016) fixes $r(\boldsymbol{y})$ as a standard normal distribution, $r(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}|\mathbf{0},\mathbf{I})$, which results in a high-bias MI estimation. With N sample pairs $\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^{N}$, Poole et al. (2019) replaces $r(\boldsymbol{y})$ with a Monte Carlo approximation $r_i(\boldsymbol{y}) = \frac{1}{N-1}\sum_{j\neq i}p(\boldsymbol{y}|\boldsymbol{x}_j) \approx p(\boldsymbol{y})$ and derives a leave-one-out upper bound (L1Out):

$$\mathrm{I}_{\text{L1Out}} := \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{p(\boldsymbol{y}_i|\boldsymbol{x}_i)}{\frac{1}{N-1}\sum_{j\neq i}p(\boldsymbol{y}_i|\boldsymbol{x}_j)}\right]\right]. \quad (7)$$

This bound does not require any additional parameters, but depends highly on a sufficient sample size to achieve a satisfying Monte Carlo approximation. In practice, L1Out suffers from numerical instability when applied to real-world MI minimization problems.

To compare our method with the aforementioned MI upper bounds in more general scenarios (*i.e.*, $p(\boldsymbol{y}|\boldsymbol{x})$ is unknown), we use a neural network $q_\theta(\boldsymbol{y}|\boldsymbol{x})$ to approximate $p(\boldsymbol{y}|\boldsymbol{x})$, and develop variational versions of VUB and L1Out as:

$$\mathrm{I}_{\mathrm{vVUB}} = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log \frac{q_\theta(\boldsymbol{y}|\boldsymbol{x})}{r(\boldsymbol{y})}\right], \tag{8}$$

$$\mathrm{I}_{\mathrm{vL1Out}} = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left[\log \frac{q_\theta(\boldsymbol{y}_i|\boldsymbol{x}_i)}{\frac{1}{N-1}\sum_{j\neq i} q_\theta(\boldsymbol{y}_i|\boldsymbol{x}_j)}\right]\right]. \tag{9}$$

We discuss theoretical properties of these two variational bounds in the Supplementary Material. In a simulation study (Section 4.1), we find that variational L1Out reaches better performance than previous lower bounds for MI estimation. However, the problem of numerical instability still remains for variational L1Out in real-world applications (as shown in Section 4.4). To the best of our knowledge, we provide the first variational version of VUB and L1Out upper bounds, and study their properties in both a theoretical analysis and wrt empirical performance.

3. Proposed Method

Suppose we have sample pairs $\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^N$ drawn from an unknown or intractable distribution $p(\boldsymbol{x},\boldsymbol{y})$. We aim to derive an upper bound estimator of the mutual information $\mathrm{I}(\boldsymbol{x};\boldsymbol{y})$ based on the given samples. In a range of machine learning tasks (*e.g.*, information bottleneck), one of the conditional distributions between variables $\boldsymbol{x}$ and $\boldsymbol{y}$ (as $p(\boldsymbol{x}|\boldsymbol{y})$ or $p(\boldsymbol{y}|\boldsymbol{x})$) can be known. To efficiently utilize this additional information, we first derive a mutual information (MI) upper bound with the assumption that one of the conditional distributions is provided (we suppose $p(\boldsymbol{y}|\boldsymbol{x})$ is provided, without loss of generality). Then, we extend the bound into more general cases where no conditional distribution is known. Finally, we develop a MI minimization algorithm based on the derived bound.

3.1. CLUB with $p(\boldsymbol{y}|\boldsymbol{x})$ Known

With the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$, our MI Contrastive Log-ratio Upper Bound (CLUB) is defined as:

$$\begin{aligned}\mathrm{I}_{\mathrm{CLUB}}(\boldsymbol{x};\boldsymbol{y}) :=& \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{x})] \\ &- \mathbb{E}_{p(\boldsymbol{x})}\mathbb{E}_{p(\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{x})].\end{aligned} \tag{10}$$

To show that $\mathrm{I}_{\mathrm{CLUB}}(\boldsymbol{x};\boldsymbol{y})$ is an upper bound of $\mathrm{I}(\boldsymbol{x};\boldsymbol{y})$, we calculate the gap Δ between them:

$$\begin{aligned}\Delta :=& \mathrm{I}_{\mathrm{CLUB}}(\boldsymbol{x};\boldsymbol{y}) - \mathrm{I}(\boldsymbol{x};\boldsymbol{y}) \\ =& \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{x})] - \mathbb{E}_{p(\boldsymbol{x})}\mathbb{E}_{p(\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{x})] \\ &- \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}\left[\log p(\boldsymbol{y}|\boldsymbol{x}) - \log p(\boldsymbol{y})\right] \\ =& \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log p(\boldsymbol{y})] - \mathbb{E}_{p(\boldsymbol{x})}\mathbb{E}_{p(\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{x})] \\ =& \mathbb{E}_{p(\boldsymbol{y})}\left[\log p(\boldsymbol{y}) - \mathbb{E}_{p(\boldsymbol{x})}\left[\log p(\boldsymbol{y}|\boldsymbol{x})\right]\right].\end{aligned} \tag{11}$$

By the definition of the marginal distribution, we have $p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \mathbb{E}_{p(\boldsymbol{x})}[p(\boldsymbol{y}|\boldsymbol{x})]$. Note that $\log(\cdot)$ is a concave function, and by Jensen's Inequality, we have $\log p(\boldsymbol{y}) = \log\left(\mathbb{E}_{p(\boldsymbol{x})}[p(\boldsymbol{y}|\boldsymbol{x})]\right) \geq \mathbb{E}_{p(\boldsymbol{x})}[\log p(\boldsymbol{y}|\boldsymbol{x})]$. Applying this inequality to (11), we conclude that the gap Δ is always non-negative. Therefore, $\mathrm{I}_{\mathrm{CLUB}}(\boldsymbol{x};\boldsymbol{y})$ is an upper bound of $\mathrm{I}(\boldsymbol{x};\boldsymbol{y})$. The bound is tight when $p(\boldsymbol{y}|\boldsymbol{x})$ has the same value for any $\boldsymbol{x}$, which means variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent. Consequently, we summarize the above discussion into the following Theorem 3.1.

Theorem 3.1. *For two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$,*

$$\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) \leq \mathrm{I}_{CLUB}(\boldsymbol{x};\boldsymbol{y}). \tag{12}$$

Equality is achieved if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent.

With sample pairs $\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^N$, $\mathrm{I}_{\mathrm{CLUB}}(\boldsymbol{x};\boldsymbol{y})$ has an unbiased estimate as:

$$\begin{aligned}\hat{\mathrm{I}}_{\mathrm{CLUB}} =& \frac{1}{N}\sum_{i=1}^{N}\log p(\boldsymbol{y}_i|\boldsymbol{x}_i) - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\log p(\boldsymbol{y}_j|\boldsymbol{x}_i) \\ =& \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\log p(\boldsymbol{y}_i|\boldsymbol{x}_i) - \log p(\boldsymbol{y}_j|\boldsymbol{x}_i)\right].\end{aligned} \tag{13}$$

In the estimator $\hat{\mathrm{I}}_{\mathrm{CLUB}}$, $\log p(\boldsymbol{y}_i|\boldsymbol{x}_i)$ provides the conditional log-likelihood of positive sample pair $(\boldsymbol{x}_i,\boldsymbol{y}_i)$; $\{\log p(\boldsymbol{y}_j|\boldsymbol{x}_i)\}_{i\neq j}$ provide the conditional log-likelihood of negative sample pair $(\boldsymbol{x}_i,\boldsymbol{y}_j)$. The difference between $\log p(\boldsymbol{y}_i|\boldsymbol{x}_i)$ and $\log p(\boldsymbol{y}_j|\boldsymbol{x}_i)$ is the contrastive probability log-ratio between two conditional distributions. Therefore, we name this novel MI upper bound estimator **C**ontrastive **L**og-ratio **U**pper **B**ound (CLUB). Compared with previous MI neural estimators, CLUB has a simpler form, as a linear combination of log-ratios between positive and negative sample pairs. The linear form of log-ratios improves the numerical stability for calculation of CLUB and its gradient, which we discuss in detail in Section 3.3.

3.2. CLUB with Conditional Distributions Unknown

When the conditional distributions $p(\boldsymbol{y}|\boldsymbol{x})$ or $p(\boldsymbol{x}|\boldsymbol{y})$ is provided, the MI can be directly upper-bounded by (13) with samples $\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^N$. Unfortunately, in a large number of machine learning tasks, the conditional relation between variables is unavailable.

To further extend the CLUB estimator into more general scenarios, we use a variational distribution $q_\theta(\boldsymbol{y}|\boldsymbol{x})$ with parameter θ to approximate $p(\boldsymbol{y}|\boldsymbol{x})$. Consequently, a variational CLUB term (vCLUB) is defined by:

$$\begin{aligned}\mathrm{I}_{\mathrm{vCLUB}}(\boldsymbol{x};\boldsymbol{y}) :=& \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log q_\theta(\boldsymbol{y}|\boldsymbol{x})] \\ &- \mathbb{E}_{p(\boldsymbol{x})}\mathbb{E}_{p(\boldsymbol{y})}[\log q_\theta(\boldsymbol{y}|\boldsymbol{x})].\end{aligned} \tag{14}$$

Similar to the MI upper bound estimator $\hat{\mathrm{I}}_{\mathrm{CLUB}}$ in (13), the unbiased estimator for vCLUB with samples $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$ is:

$$\hat{\mathrm{I}}_{\text{vCLUB}} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}[\log q_\theta(\boldsymbol{y}_i|\boldsymbol{x}_i) - \log q_\theta(\boldsymbol{y}_j|\boldsymbol{x}_i)]$$
$$= \frac{1}{N}\sum_{i=1}^{N}\Big[\log q_\theta(\boldsymbol{y}_i|\boldsymbol{x}_i) - \frac{1}{N}\sum_{j=1}^{N}\log q_\theta(\boldsymbol{y}_j|\boldsymbol{x}_i)\Big]. \quad (15)$$

Using the variational approximation $q_\theta(\boldsymbol{y}|\boldsymbol{x})$, vCLUB no longer guarantees an upper bound of $\mathrm{I}(\boldsymbol{x};\boldsymbol{y})$. However, the vCLUB shares good properties with CLUB. We claim that with good variational approximation $q_\theta(\boldsymbol{y}|\boldsymbol{x})$, vCLUB can still hold a MI upper bound or become a reliable MI estimator. The following analyses support this claim.

Let $q_\theta(\boldsymbol{x},\boldsymbol{y}) = q_\theta(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})$ be the variational joint distribution induced by $q_\theta(\boldsymbol{y}|\boldsymbol{x})$. Generally, we have the following Theorem 3.2. Note that when $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent, $\mathrm{I}_{\text{vCLUB}}$ has exactly the same value as $\mathrm{I}(\boldsymbol{x};\boldsymbol{y})$, without requiring any additional assumption on $q_\theta(\boldsymbol{y}|\boldsymbol{x})$. However, unlike in Theorem 3.1 as a sufficient and necessary condition, independence between $\boldsymbol{x}$ and $\boldsymbol{y}$ becomes sufficient but not necessary to conclude $\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) = \mathrm{I}_{\text{vCLUB}}(\boldsymbol{x};\boldsymbol{y})$, due to the variation approximation $q_\theta(\boldsymbol{y}|\boldsymbol{x})$.

Theorem 3.2. *Denote $q_\theta(\boldsymbol{x},\boldsymbol{y}) = q_\theta(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})$. If*

$$\mathrm{KL}\left(p(\boldsymbol{x},\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y})\right) \le \mathrm{KL}\left(p(\boldsymbol{x})p(\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y})\right),$$

then $\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) \le \mathrm{I}_{vCLUB}(\boldsymbol{x};\boldsymbol{y})$. *The equality holds when $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent.*

Theorem 3.2 provides insight that vCLUB remains a MI upper bound if the variational joint distribution $q_\theta(\boldsymbol{x},\boldsymbol{y})$ is "closer" to $p(\boldsymbol{x},\boldsymbol{y})$ than to $p(\boldsymbol{x})p(\boldsymbol{y})$. Therefore, minimizing $\mathrm{KL}(p(\boldsymbol{x},\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y}))$ will facilitate the condition in Theorem 3.2 to be achieved. We show that $\mathrm{KL}(p(\boldsymbol{x},\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y}))$ can be minimized by maximizing the log-likelihood of $q_\theta(\boldsymbol{y}|\boldsymbol{x})$, because of the following equation:

$$\min_\theta \mathrm{KL}(p(\boldsymbol{x},\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y}))$$
$$= \min_\theta \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log(p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})) - \log(q_\theta(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}))]$$
$$= \min_\theta \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log p(\boldsymbol{y}|\boldsymbol{x})] - \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log q_\theta(\boldsymbol{y}|\boldsymbol{x})]. \quad (16)$$

Equation (16) equals $\min_\theta \mathrm{KL}(p(\boldsymbol{y}|\boldsymbol{x})\|q_\theta(\boldsymbol{y}|\boldsymbol{x}))$, in which the first term has no relation with parameter θ. Therefore, $\min_\theta \mathrm{KL}(p(\boldsymbol{x},\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y}))$ is equivalent to the maximization of the second term, $\max_\theta \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log q_\theta(\boldsymbol{y}|\boldsymbol{x})]$. With samples $\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^N$, we can maximize the log-likelihood function $\mathcal{L}(\theta) := \frac{1}{N}\sum_{i=1}^N \log q_\theta(\boldsymbol{y}_i|\boldsymbol{x}_i)$, which is the unbiased estimate of $\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log q_\theta(\boldsymbol{y}|\boldsymbol{x})]$.

In practice, the variational distribution $q_\theta(\boldsymbol{y}|\boldsymbol{x})$ is usually implemented with neural networks. By enlarging the network capacity (*i.e.*, adding layers and neurons) and applying gradient-ascent to the log-likelihood $\mathcal{L}(\theta)$, we can obtain far more accurate approximation $q_\theta(\boldsymbol{y}|\boldsymbol{x})$ to $p(\boldsymbol{y}|\boldsymbol{x})$, thanks to the high expressiveness of neural networks (Hu et al., 2019; Oymak & Soltanolkotabi, 2019). Therefore, to further discuss the properties of vCLUB, we assume the neural network approximation q_θ achieves $\mathrm{KL}(p(\boldsymbol{y}|\boldsymbol{x})\|q_\theta(\boldsymbol{y}|\boldsymbol{x})) \le \varepsilon$ with a small number $\varepsilon > 0$. In the Supplementary Material, we quantitatively discuss the reasonableness of this assumption. Consider the KL-divergence between $p(\boldsymbol{x})p(\boldsymbol{y})$ and $q_\theta(\boldsymbol{x},\boldsymbol{y})$. If $\mathrm{KL}(p(\boldsymbol{x})p(\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y})) \ge \mathrm{KL}(p(\boldsymbol{x},\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y}))$, by Theorem 3.2, vCLUB is already a MI upper bound. Otherwise, if $\mathrm{KL}(p(\boldsymbol{x})p(\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y})) < \mathrm{KL}(p(\boldsymbol{x},\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y}))$, we have the following corollary:

Corollary 3.3. *Given $\mathrm{KL}(p(\boldsymbol{y}|\boldsymbol{x})\|q_\theta(\boldsymbol{y}|\boldsymbol{x})) \le \varepsilon$, if*

$$\mathrm{KL}(p(\boldsymbol{x},\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y})) > \mathrm{KL}(p(\boldsymbol{x})p(\boldsymbol{y})\|q_\theta(\boldsymbol{x},\boldsymbol{y})),$$

then $|\mathrm{I}(\boldsymbol{x};\boldsymbol{y}) - \mathrm{I}_{vCLUB}(\boldsymbol{x};\boldsymbol{y})| < \varepsilon$.

Combining Corollary 3.3 and Theorem 3.2, we conclude that with a good variational approximation $q_\theta(\boldsymbol{y}|\boldsymbol{x})$, vCLUB can either remain a MI upper bound, or become a MI estimator whose absolute error is bounded by the approximation performance $\mathrm{KL}(p(\boldsymbol{y}|\boldsymbol{x})\|q_\theta(\boldsymbol{y}|\boldsymbol{x}))$.

3.3. CLUB in MI Minimization

One of the major applications of MI upper bounds is for mutual information minimization. In general, MI minimization aims to reduce the correlation between two variables $\boldsymbol{x}$ and $\boldsymbol{y}$ by selecting an optimal parameter σ of the joint variational distribution $p_\sigma(\boldsymbol{x},\boldsymbol{y})$. Under some application scenarios, additional conditional information between $\boldsymbol{x}$ and $\boldsymbol{y}$ is known. For example, in the information bottleneck task, the joint distribution between input $\boldsymbol{x}$ and bottleneck representation $\boldsymbol{y}$ is $p_\sigma(\boldsymbol{x},\boldsymbol{y}) = p_\sigma(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})$. Then the MI upper bound I_{CLUB} can be calculated directly based on (13).

Algorithm 1 MI Minimization with vCLUB

```
for each training iteration do
  Sample {(x_i, y_i)}_{i=1}^N from p_σ(x, y)
  Log-likelihood L(θ) = 1/N Σ_{i=1}^N log q_θ(y_i|x_i)
  Update q_θ(y|x) by maximizing L(θ)
  for i = 1 to N do
    if use sampling then
      Sample k'_i uniformly from {1, 2, ..., N}
      U_i = log q_θ(y_i|x_i) − log q_θ(y_{k'_i}|x_i)
    else
      U_i = log q_θ(y_i|x_i) − 1/N Σ_{j=1}^N log q_θ(y_j|x_i)
    end if
  end for
  Update p_σ(x, y) by minimize Î_vCLUB = 1/N Σ_{i=1}^N U_i
end for
```

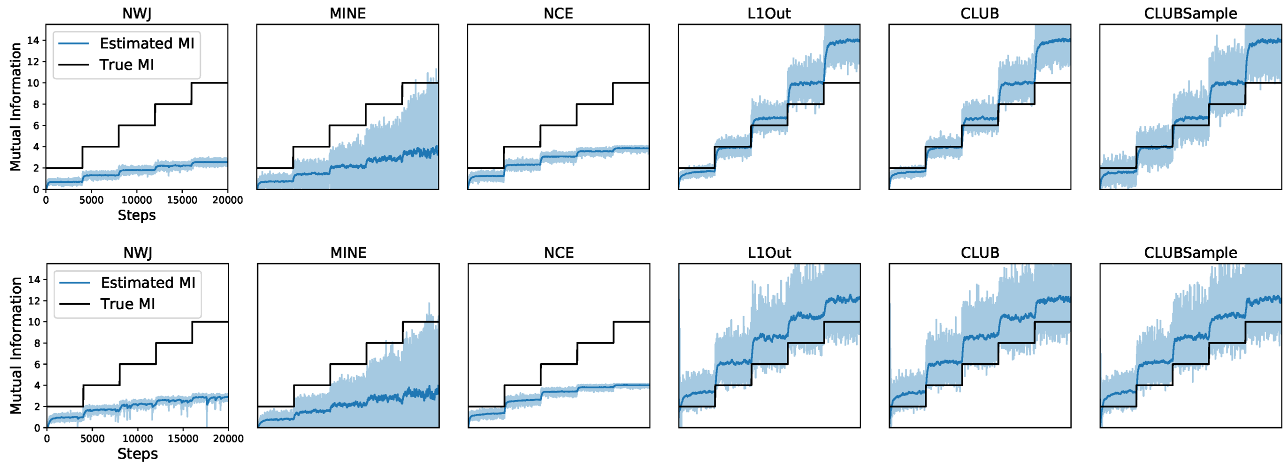

Figure 1. Simulation performance of MI estimators. In the **top** row, data are from joint Gaussian distributions with the MI true value stepping over time. In the **bottom** row, a cubic transformation is further applied to the Gaussian samples as $\boldsymbol{y}$. In each figure, the true MI values is a step function shown as the black line. The estimated values are displayed as shadow blue curves. The dark blue curves shows the local averages of estimated MI, with a bandwidth equal to 200.

For cases in which the conditional information between $\boldsymbol{x}$ and $\boldsymbol{y}$ remains unclear, we propose an MI minimization algorithm using the vCLUB estimator. At each training iteration, we first obtain a batch of samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$ from $p_\sigma(\boldsymbol{x}, \boldsymbol{y})$. We then update the variational approximation $q_\theta(\boldsymbol{y}|\boldsymbol{x})$ by maximizing the log-likelihood $\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^N \log q_\theta(\boldsymbol{y}_i|\boldsymbol{x}_i)$. After $q_\theta(\boldsymbol{y}|\boldsymbol{x})$ is updated, we calculate the vCLUB estimator as described in (15). Finally, the gradient of $\hat{\mathrm{I}}_{\text{vCLUB}}$ is calculated and back-propagated to parameters of $p_\sigma(\boldsymbol{x}, \boldsymbol{y})$. The reparameterization trick (Kingma & Welling, 2013) ensures the gradient back-propagates through the sampled embeddings $(\boldsymbol{x}_i, \boldsymbol{y}_i)$. Updating joint distribution $p_\sigma(\boldsymbol{x}, \boldsymbol{y})$ will lead to the change of conditional distribution $p_\sigma(\boldsymbol{y}|\boldsymbol{x})$. Therefore, we need to update the approximation network $q_\theta(\boldsymbol{y}|\boldsymbol{x})$ again. Consequently, $q_\theta(\boldsymbol{y}|\boldsymbol{x})$ and $p_\sigma(\boldsymbol{x}, \boldsymbol{y})$ are updated alternately during the training (as shown in Algorithm 1 without *sampling*).

In each training iteration, the vCLUB estimator requires calculation of all conditional distributions $\{p_\sigma(\boldsymbol{y}_j|\boldsymbol{x}_i)\}_{i,j=1}^N$, which leads to $\mathcal{O}(N^2)$ computational complexity. To accelerate the training, we use stochastic sampling to approximate the mean of conditional probabilities in $\hat{\mathrm{I}}_{\text{vCLUB}}$ (Eqn. (15)), and obtain a sampled vCLUB estimator:

$$\begin{aligned} &\log q_\theta(\boldsymbol{y}_i|\boldsymbol{x}_i) - \tfrac{1}{N}\textstyle\sum_{j=1}^N \log q_\theta(\boldsymbol{y}_j|\boldsymbol{x}_i) \\ \approx &\log q_\theta(\boldsymbol{y}_i|\boldsymbol{x}_i) - \log q_\theta(\boldsymbol{y}_{k_i'}|\boldsymbol{x}_i), \end{aligned} \tag{17}$$

with k_i' uniformly selected from indices $\{1, 2, \dots, N\}$. With this sampling strategy, the computational complexity in each iteration can be reduced to $\mathcal{O}(N)$ (as shown in Algorithm 1 using *sampling*). A similar sampling strategy can also be applied to CLUB when $p(\boldsymbol{y}|\boldsymbol{x})$ is known. This stochastic sampling estimator not only provides an unbiased estimation to $\hat{\mathrm{I}}_{\text{vCLUB}}$, but bridges the MI minimization with *negative sampling*, a commonly used training strategy (Grover & Leskovec, 2016; Chen et al., 2019; Cheng et al., 2020a), in which for each positive data pair $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, a negative pair $(\boldsymbol{x}_i, \boldsymbol{y}_{k_i'})$ is sampled. The mutual information is minimized by reducing the positive conditional probability, while enlarging the negative conditional probability. Although previous MI upper bounds also utilize the negative data pairs (*e.g* L1Out in (7)), they do not yield an unbiased estimate when accelerated with negative sampling, because of the non-linear log function applied after the linear probability summation. The unbiasedness of our sampled CLUB is manifested thanks to the form of linear log-ratio summation. In the experiments, we find the sampled vCLUB not only provides comparable MI estimation performance, but also improves the model generalization abilities.

4. Experiments

We first show the performance of CLUB as a MI estimator on tractable toy (simulated) cases. Then we evaluate the minimization ability of CLUB on two real-world applications: Information Bottleneck (IB) and Unsupervised Domain Adaptation (UDA). In the information bottleneck, the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ is known, so we compare both CLUB and variational CLUB (vCLUB) estimators. In other experiments for which $p(\boldsymbol{y}|\boldsymbol{x})$ is unknown, all the tested upper bounds require variational approximation. Without ambiguity, we abbreviate all variational upper bounds (*e.g.*, vCLUB) with their original names (*e.g.*, CLUB) for simplicity.

4.1. MI Estimation Quality

Following the setup from Poole et al. (2019), we apply CLUB as an MI estimator in two toy tasks: (i) estimating MI

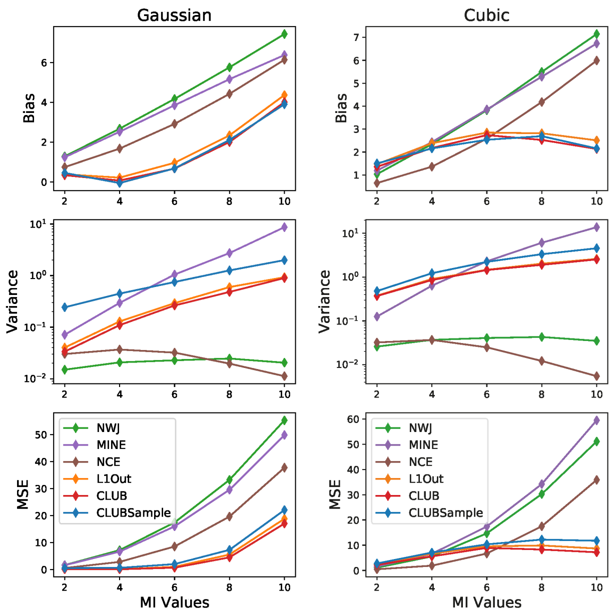

Figure 2. Estimation quality comparison of MI estimators. The **left** column shows the results of estimations under a Gaussian distribution, while the **right** column is under Cubic setup. In each column, estimation metrics are reported as bias, variance, and mean-square-error (MSE). In each plot, the evaluation metric is reported with different true MI values varying from 2 to 10.

with samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$ drawn jointly from a multivariate Gaussian distribution with correlation ρ; (ii) estimating MI with samples $\{(\boldsymbol{x}_i, (\boldsymbol{W}\boldsymbol{y}_i)^3)\}$, where $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ still comes from a Gaussian with correlation ρ, and $\boldsymbol{W}$ is a full-rank matrix. Since the transformation $\boldsymbol{y} \to (\boldsymbol{W}\boldsymbol{y})^3$ is smooth and bijective, the mutual information is invariant (Kraskov et al., 2004), and $\mathrm{I}(\boldsymbol{x}; \boldsymbol{y}) = \mathrm{I}(\boldsymbol{x}; (\boldsymbol{W}\boldsymbol{y})^3)$. For both of the tasks, the dimension of samples $\boldsymbol{x}$ and $\boldsymbol{y}$ is set to $d = 20$. Under Gaussian distributions, the MI true value can be calculated as $\mathrm{I}(\boldsymbol{x}, \boldsymbol{y}) = -\frac{d}{2}\log(1-\rho^2)$, and therefore we set the MI true value in the range $\{2.0, 4.0, 6.0, 8.0, 10.0\}$ by varying the value of ρ. At each MI true value, we sample data batches 4000 times, with batch size equal to 64, for the training of variational MI estimators.

We compare our method with baselines including MINE (Belghazi et al., 2018), NWJ (Nguyen et al., 2010), InfoNCE (Oord et al., 2018), VUB (Alemi et al., 2016) and L1Out (Poole et al., 2019). Since the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ is unknown in this simulation setup, all upper bounds (VUB, L1Out, CLUB) are calculated with an auxiliary approximation network $q_\theta(\boldsymbol{y}|\boldsymbol{x})$. The approximation network has the same structure for all upper bounds, parameterized in a Gaussian family, $q_\theta(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{\sigma}^2(\boldsymbol{x}) \cdot \mathbf{I})$ with mean $\boldsymbol{\mu}(\boldsymbol{x})$ and variance $\boldsymbol{\sigma}^2(\boldsymbol{x})$ inferred by neural networks. On the other hand, all the MI lower bounds (MINE, NWJ, InfoNCE) require learning of a value function $f(\boldsymbol{x}, \boldsymbol{y})$. To make a fair comparison, we

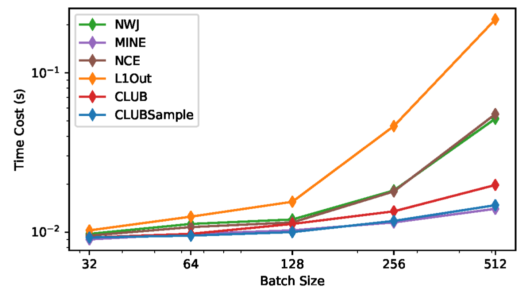

Figure 3. Estimator speed comparison with different batch size. Both the axes have a logarithm scale.

set the value function and the neural approximation to have one hidden layer and the same hidden units. For Gaussian setup, the number of hidden units is 20; for Cubic setup, the number of hidden units is 40. On the top of hidden layer outputs, we add the ReLU activation function. The learning rate for all estimators is set to 1×10^{-4}.

We report in Figure 1 the estimated MI values in each training step. The estimation of VUB has incomparably large bias, so we provide its results in the Supplementary Material. Lower bound estimators, such as NWJ, MINE, and InfoNCE, provide estimated values mainly under the true MI values step function, while L1Out, CLUB and Sampled CLUB (CLUBSample) estimate values above the step function, which supports our theoretical analysis about CLUB with variational approximation. The numerical results of bias and variance in the estimation are reported in Figure 2. Among these methods, CLUB and CLUBSample have the lowest bias. The bias difference between CLUB and CLUBSample is insignificant, supporting our claim in Section 3.3 that CLUBSample is an unbiased stochastic approximation of CLUB. L1Out also provides small bias estimation which is slightly worse than CLUB. NWJ and InfoNCE have the lowest variance under both setups. CLUBSample has larger variance than CLUB and L1Out due to the use of the sampling strategy. When considering the bias-variance trade-off as the mean square estimation error (MSE, equals bias^2+variance), CLUB outperforms other estimators, while L1Out and CLUBSample also provide competitive performance.

Although the L1Out estimator reaches similar estimation performance as our CLUB on toy examples, we find L1Out fails to effectively reduce the MI when applied as a critic in real-world MI minimization tasks. The numerical results in Sections 4.3 and 4.4 support this claim.

4.2. Time Efficiency of MI Estimators

Besides the estimation quality comparison, we further study the time efficiency of different MI estimators. We conduct the comparison under the same experimental setup as the

Gaussian case in Section 4.1. Each MI estimator is tested with a different batch size, from 32 to 512. We count the total time cost of the whole estimation process and average it into each estimation step. In Figure 3, we report the average estimation time costs of different MI estimators. MINE and CLUBSample have the best computational efficiency; both have $\mathcal{O}(N)$ computational complexity with respect to the sample size N, because of the negative sampling strategy. Among other computational $\mathcal{O}(N^2)$ methods, CLUB has the highest estimation speed, thanks to its simple form as mean of log-ratios, which can be easily accelerated by matrix multiplication. Leave-one-out (L1out) has the highest time cost, because it requires "leaving out" the positive sample pair each time in the denominator of equation (7).

4.3. MI Minimization in Information Bottleneck

The Information Bottleneck (Tishby et al., 2000) (IB) is an information-theoretical method for latent representation learning. Given an input source $\boldsymbol{x} \in \mathcal{X}$ and a corresponding output target $\boldsymbol{y} \in \mathcal{Y}$, the information bottleneck aims to learn an encoder $p_\sigma(\boldsymbol{z}|\boldsymbol{x})$, such that the compressed latent code $\boldsymbol{z}$ is highly relevant to the target $\boldsymbol{y}$, with irrelevant source information from $\boldsymbol{x}$ being filtered. In other words, IB seeks to find the sufficient statistics of $\boldsymbol{x}$ with respect to $\boldsymbol{y}$ (Alemi et al., 2016), with minimum information used from $\boldsymbol{x}$. To address this task, an objective is introduced as

$$\min_{p_\sigma(\boldsymbol{z}|\boldsymbol{x})} -\mathrm{I}(\boldsymbol{y};\boldsymbol{z}) + \beta \mathrm{I}(\boldsymbol{x};\boldsymbol{z}) \tag{18}$$

where hyper-parameter $\beta > 0$. Following the setup from Alemi et al. (2016), we apply the IB technique in the permutation-invariant MNIST classification. The input $\boldsymbol{x}$ is a vector converted from a 28×28 image of a hand-written number, and the output $\boldsymbol{y}$ is the class label of this number. The stochastic encoder $p_\sigma(\boldsymbol{z}|\boldsymbol{x})$ is implemented in a Gaussian variational family, $p_\sigma(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}|\mu_\sigma(\boldsymbol{x}), \Sigma_\sigma(\boldsymbol{x}))$, where μ_σ and Σ_σ are two fully-connected neural networks.

For the first part of the IB objective (18), the MI between target $\boldsymbol{y}$ and latent code $\boldsymbol{z}$ is maximized. We use the same strategy as in the deep variational information bottleneck (DVB) (Alemi et al., 2016), where a variational classifier $q_\phi(\boldsymbol{y}|\boldsymbol{z})$ is introduced to implement a Barber-Agakov MI lower bound (Eqn. (2)) of $\mathrm{I}(\boldsymbol{y};\boldsymbol{z})$. The second term in the IB objective requires the MI minimization between input $\boldsymbol{x}$ and the latent representation $\boldsymbol{z}$. DVB (Alemi et al., 2016) utilizes the MI variation upper bound (VUB) (Eqn. (6)) for the minimization of $\mathrm{I}(\boldsymbol{x};\boldsymbol{z})$. Since the closed form of $p_\sigma(\boldsymbol{z}|\boldsymbol{x})$ is already known as a Gaussian distribution parameterized by neural networks, we can directly apply our CLUB estimator for minimizing $\mathrm{I}(\boldsymbol{x};\boldsymbol{z})$. Alternatively, the variational CLUB can be also applied under this scenario. Besides CLUB and vCLUB, we compare previous methods such as MINE, NWJ, InfoNCE, and L1Out. The misclassification rates for different MI estimators are reported in Table 1.

Method	**Misclass. rate(%)**
NWJ (Nguyen et al., 2010)	1.29
MINE (Belghazi et al., 2018)	1.17
InfoNCE (Oord et al., 2018)	1.24
DVB (VUB) (Alemi et al., 2016)	1.13
L1Out (Poole et al., 2019)	-
CLUB	1.12
CLUB (Sample)	1.10
vCLUB	1.10
vCLUB (Sample)	**1.06**

Table 1. Performance on the Permutation invariant MNIST classification. Different MI estimators are applied for the minimization of $\mathrm{I}(\boldsymbol{x};\boldsymbol{z})$ in the Information Bottleneck. Misclassification rates of learned latent representation $\boldsymbol{z}$ are reported. The top three methods are MI lower bounds, while the rest are MI upper bounds.

MINE achieves the lowest misclassification error among lower bound estimators. Although providing good MI estimation in the Gaussian simulation study, L1Out suffers from numerical instability in MI optimization and fails during training. Both CLUB and vCLUB estimators outperform previous methods in bottleneck representation learning, with lower misclassification rates. Note that sampled versions of CLUB and vCLUB improve the accuracy compared with the original CLUB and vCLUB, respectively, which verify the claim that a negative sampling strategy improves model robustness. Besides, using variational approximation $q_\theta(\boldsymbol{y}|\boldsymbol{x})$ even attains higher accuracy than using ground truth $p_\sigma(\boldsymbol{y}|\boldsymbol{x})$ for CLUB. Although $p_\sigma(\boldsymbol{y}|\boldsymbol{x})$ provides more accurate MI estimation, the variational approximation $p_\sigma(\boldsymbol{y}|\boldsymbol{x})$ can add noise into the gradient of CLUB. Both the sampling and the variational approximation increase the randomness in the model, which helps to increase the model generalization ability (Hinton et al., 2012; Belghazi et al., 2018).

4.4. MI Minimization in Domain Adaptation

Another important application of MI minimization is disentangled representation learning (DRL) (Kim & Mnih, 2018; Chen et al., 2018; Locatello et al., 2019). Specifically, we aim to encode the data into several separate embedding parts, each with different semantic meaning. The semantically disentangled representations help improve the performance of deep learning models, especially in the fields of conditional generation (Ma et al., 2018), style transfer (John et al., 2019), and domain adaptation (Gholami et al., 2018). To learn (ideally) independent disentangled representations, one effective solution is to minimize the mutual information among different latent embedding parts.

We compare performance of MI estimators for learning disentangled representations in unsupervised domain adap-

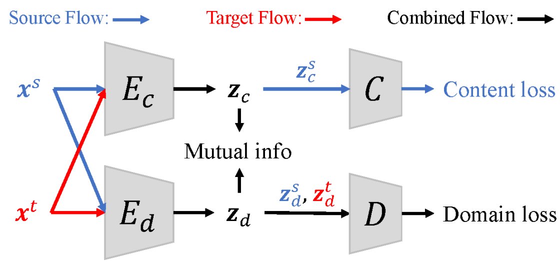

Figure 4. The information-theoretical framework for unsupervised domain adaptation. The input data $\boldsymbol{x}$ (including $\boldsymbol{x}^s$ and $\boldsymbol{x}^t$) are passed to a content encoder E_c and a domain encoder E_d, with output feature $\boldsymbol{z}_c$ and $\boldsymbol{z}_d$, respectively. C is the content classifier, and D is the domain discriminator. The mutual information between $\boldsymbol{z}_c$ and $\boldsymbol{z}_d$ is minimized.

tation (UDA) tasks. In UDA, we have images $\boldsymbol{x}^s \in \mathcal{X}^s$ from the source domain $\mathcal{X}^s$ and $\boldsymbol{x}^t \in \mathcal{X}^t$ from the target domain $\mathcal{X}^t$. While each source image $\boldsymbol{x}^s$ has a corresponding label y^s, no label information is available for observations in the target domain. The objective is to learn a model based on data $\{\boldsymbol{x}^s, y^s\}$ and $\{\boldsymbol{x}^t\}$, which not only performs well in source domain classification, but also provides satisfying predictions in the target domain.

To solve this problem, we use the information-theoretical framework inspired from Gholami et al. (2018). Specifically, two feature extractors are introduced: the domain encoder E_d and the content encoder E_c. The former encodes the domain information from an observation $\boldsymbol{x}$ into a domain embedding $\boldsymbol{z}_d = E_d(\boldsymbol{x})$; the latter outputs a content embedding $\boldsymbol{z}_c = E_c(\boldsymbol{x})$ based on an input data point $\boldsymbol{x}$. As shown in Figure 4, the content embedding $\boldsymbol{z}_c^s$ from the source domain is further used as an input to a content classifier $C(\cdot)$ to predict the corresponding class label, with a content loss defined as $\mathcal{L}_c = \mathbb{E}[-y^s \log C(\boldsymbol{z}_c^s)]$. The domain embedding $\boldsymbol{z}_d$ (including $\boldsymbol{z}_d^s$ and $\boldsymbol{z}_d^t$) is input to a domain discriminator $D(\cdot)$ to predict whether the observation comes from the source domain or target domain, with a domain loss defined as $\mathcal{L}_d = \mathbb{E}_{\boldsymbol{x}\in\mathcal{X}^s}[\log D(\boldsymbol{z}_d)] + \mathbb{E}_{\boldsymbol{x}\in\mathcal{X}^t}[\log(1-D(\boldsymbol{z}_d))]$. Since the content information and the domain information should be independent, we minimize the mutual information $\mathrm{I}(\boldsymbol{z}_c, \boldsymbol{z}_d)$ between the content embedding $\boldsymbol{z}_c$ and domain embedding $\boldsymbol{z}_d$. The final objective is (shown in Figure 4):

$$\min_{E_c, E_d, C, D} \mathrm{I}(\boldsymbol{z}_c, \boldsymbol{z}_d) + \lambda_c \mathcal{L}_c + \lambda_d \mathcal{L}_d, \quad (19)$$

where $\lambda_c, \lambda_d > 0$ are hyper-parameters.

We apply different MI estimators to the framework (19), and evaluate the performance on several DA benchmark datasets, including MNIST, MNIST-M, USPS, SVHN, CIFAR-10, and STL. A detailed description of the datasets and model setups are provided in the Supplementary Material. Besides the proposed information-theoretical UDA model, we also compare the performance with other UDA frameworks: DANN (Ganin et al., 2016), DSN (Bousmalis et al., 2016), and MCD (Saito et al., 2018). The numerical results are shown in Table 2. From the results, we find our MI-based disentangling shows competitive results with previous UDA methods. Among different MI estimators, the Sampled CLUB uniformly outperforms other competitive methods on four DA tasks. The stochastic sampling in CLUBSample improves the model generalization ability and helps the model avoid overfitting. The other two MI upper bounds, VUB and L1Out, fail to train a satisfying UDA model, whose results are worse than the MI lower bound estimators. With L1Out, the training loss cannot even decrease on the most challenging SVHN→MNIST task, due to the numerical instability.

Method	M→MM	M→U	U→M	SV→M	C→S	S→C
Source-Only	59.9	76.7	63.4	67.1	-	-
MI-based Disentangling Framework						
NWJ	83.3	98.3	91.1	86.5	78.2	71.0
MINE	88.4	98.1	94.8	83.4	77.9	70.5
InfoNCE	85.5	98.3	92.7	84.1	77.4	69.4
VUB	76.4	97.1	96.3	81.5	-	-
L1Out	76.2	96.3	93.9	-	77.8	69.2
CLUB	93.7	**98.9**	97.7	89.7	78.7	71.8
CLUB-S	**94.6**	**98.9**	**98.1**	**90.6**	**79.1**	**72.3**
Other Frameworks						
DANN	81.5	77.1	73.0	71.1	-	-
DSN	83.2	91.3	-	76.0	-	-
MCD	93.5	94.2	94.1	**92.6**	78.1	69.2

Table 2. Performance comparison on UDA. Datasets are MNIST (M), MNIST-M (MM), USPS (U), SVHN (SV), CIFAR-10 (C), and STL (S). Classification accuracy on target domain is reported. Among results in MI-based disentangling framework, the top three are MI lower bounds, while the rest are MI upper bounds. CLUB-S refers to Sampled CLUB.

5. Conclusions

We have introduced a novel mutual information upper bound called Contrastive Log-ratio Upper Bound (CLUB). This novel MI estimator can be extended to a variational version for general scenarios when only samples of the joint distribution are available. Based on the variational CLUB, we have proposed a new MI minimization algorithm, and further accelerated it with a negative sampling strategy. We have studied the good properties of CLUB both theoretically and empirically. Experimental results on simulation studies and real-world applications show the attractive performance of CLUB on both MI estimation and MI minimization tasks. This work provides insight into the connection between mutual information and widespread machine learning training strategies, including contrastive learning and negative sampling. We believe the proposed CLUB estimator will have significant applications for reducing the correlation of different model parts, especially in the domains of interpretable machine learning, controllable generation, and fairness.

Acknowledgements

The portion of this work performed at Duke University was supported in part by DARPA, DOE, NIH, NSF and ONR.

References

Ahn, S., Hu, S. X., Damianou, A., Lawrence, N. D., and Dai, Z. Variational information distillation for knowledge transfer. In *CVPR*, 2019.

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Barber, D. and Agakov, F. V. The im algorithm: a variational approach to information maximization. In *NeurIPS*, 2003.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Hjelm, D., and Courville, A. Mutual information neural estimation. In *ICML*, 2018.

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., and Erhan, D. Domain separation networks. In *NeurIPS*, 2016.

Charrow, B., Liu, S., Kumar, V., and Michael, N. Information-theoretic mapping using cauchy-schwarz quadratic mutual information. In *ICRA*, 2015.

Chen, L., Wang, G., Tao, C., Shen, D., Cheng, P., Zhang, X., Wang, W., Zhang, Y., and Carin, L. Improving textual network embedding with global attention via optimal transport. In *ACL*, 2019.

Chen, T. Q., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. In *NeurIPS*, 2018.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NeurIPS*, 2016.

Cheng, P., Li, Y., Zhang, X., Cheng, L., Carlson, D., and Carin, L. Dynamic embedding on textual networks via a gaussian process. In *AAAI*, 2020a.

Cheng, P., Min, M. R., Shen, D., Malon, C., Zhang, Y., Li, Y., and Carin, L. Improving disentangled text representation learning with information-theoretic guidance. *arXiv preprint arXiv:2006.00693*, 2020b.

Dai, S., Cheng, Y., Zhang, Y., Gan, Z., Liu, J., and Carin, L. Contrastively smoothed class alignment for unsupervised domain adaptation. *arXiv preprint arXiv:1909.05288*, 2019.

Darbellay, G. A. and Vajda, I. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 1999.

Florensa, C., Duan, Y., and Abbeel, P. Stochastic neural networks for hierarchical reinforcement learning. *arXiv preprint arXiv:1704.03012*, 2017.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *JMLR*, 2016.

Gholami, B., Sahu, P., Rudovic, O., Bousmalis, K., and Pavlovic, V. Unsupervised multi-target domain adaptation: An information theoretic approach. *arXiv preprint arXiv:1810.11547*, 2018.

Granger, C. and Lin, J.-L. Using the mutual information coefficient to identify lags in nonlinear models. *Journal of time series analysis*, 1994.

Grover, A. and Leskovec, J. node2vec: Scalable feature learning for networks. In *KDD*, 2016.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010.

Härdle, W. K., Müller, M., Sperlich, S., and Werwatz, A. *Nonparametric and Semiparametric Models*. Springer Science & Business Media, 2004.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Hu, W., Miyato, T., Tokui, S., Matsumoto, E., and Sugiyama, M. Learning discrete representations via information maximizing self-augmented training. In *ICML*, 2017.

Hu, W., Li, Z., and Yu, D. Understanding generalization of deep neural networks trained with noisy labels. *arXiv preprint arXiv:1905.11368*, 2019.

Jiang, B., Ye, C., and Liu, J. S. Nonparametric k-sample tests via dynamic slicing. *Journal of the American Statistical Association*, 2015.

John, V., Mou, L., Bahuleyan, H., and Vechtomova, O. Disentangled representation learning for non-parallel text style transfer. In *ACL*, 2019.

Julian, B. J., Karaman, S., and Rus, D. On mutual information-based control of range sensing robots for mapping applications. *The International Journal of Robotics Research*, 2014.

Kamishima, T., Akaho, S., and Sakuma, J. Fairness-aware learning through regularization approach. In *IEEE 11th International Conference on Data Mining Workshops*, 2011.

Kazemi, H., Soleymani, S., Taherkhani, F., Iranmanesh, S., and Nasrabadi, N. Unsupervised image-to-image translation using domain-specific variational information bound. In *NeurIPS*, 2018.

Kim, H. and Mnih, A. Disentangling by factorising. In *ICML*, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kraskov, A., Stögbauer, H., and Grassberger, P. Estimating mutual information. *Physical review E*, 2004.

Kullback, S. *Information theory and statistics*. Courier Corporation, 1997.

Lachmann, A., Giorgi, F. M., Lopez, G., and Califano, A. Aracne-ap: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics*, 2016.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, 2019.

Ma, L., Sun, Q., Georgoulis, S., Van Gool, L., Schiele, B., and Fritz, M. Disentangled person image generation. In *CVPR*, 2018.

Magdon-Ismail, M. and Atiya, A. F. Neural networks for density estimation. In *NeurIPS*, 1999.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 2010.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Oymak, S. and Soltanolkotabi, M. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *ICML*, 2019.

Poole, B., Ozair, S., Van Den Oord, A., Alemi, A., and Tucker, G. On variational bounds of mutual information. In *ICML*, 2019.

Saito, K., Watanabe, K., Ushiku, Y., and Harada, T. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018.

Suzuki, T., Sugiyama, M., Sese, J., and Kanamori, T. Approximating mutual information by maximum likelihood density ratio estimation. In *New challenges for feature selection in data mining and knowledge discovery*, 2008.

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Zea, D. J., Anfossi, D., Nielsen, M., and Marino-Buslje, C. Mitos. jl: mutual information tools for protein sequence analysis in the julia language. *Bioinformatics*, 2016.