On Study of Mutual Information and Its Estimation Methods

Marshal Arijona S.
Faculty of Computer Science
University of Indonesia
Depok, Indonesia
Email: marshal.arijona01@ui.ac.id

Abstract—**The presence of mutual information in the research of deep learning has grown significantly. It has been proven that mutual information can be a good objective function to build a robust deep learning model. Most of the researches utilize estimation methods to approximate the true mutual information. This technical report delivers an extensive study about definitions as well as properties of mutual information. This article then delivers some reviews and current drawbacks of mutual information estimation methods afterward.**

Keyword – **Mutual Information, KL-Divergence, Entropy, Variational Distribution, Deep Learning**

I. Introduction

Mutual information (MI) is viewed as one of the most fundamental measurements to quantify the dependence of two random variables [1]. Evidently, mutual information has been applied in wide spectrums, including statistics [1]–[3], biostatistics [1], [4], [5], robotics [1], [6], [7], and machine learning [8], [9]. This shows that mutual information can capture the notion of dependence on nature universally.

For machine learning applications (especially deep learning), MI is used as an objective function or a regularizer in loss function [1]. The objective function is either maximizing the MI or minimizing the MI. MI maximization is applied in various tasks, including representation learning [1], [9], [14], generative models [1], [8], and reinforcement learning [1], [15]. Meanwhile, MI minimization has taken parts in disentangled representation learning, style transfer [1], [11], and information bottleneck [1], [12].

Almost all MI maximization or MI minimization do not use the exact MI but rather compute the estimation. This due to the required closed form of the density function and tractable log-density ratio between the joint distribution and the product of marginal distribution [1]. In the real world, it is not always possible to have all access to the required distributions. Commonly, we only have samples from the joint distribution [1]. Therefore, the estimation methods are proposed to solve the problems. Info-GAN for example is using Barber-Agakov lower bound [13] to estimate the mutual information between the latent factor and the generated images [8]. Another example is the contrastive predictive model, which uses noise contrastive estimation to estimate mutual information between the current context and the data at the time steps ahead [16]. Mutual information estimation is currently active research in machine learning and still opens a huge possibility to improve.

This article aims to deliver a theoretical study about mutual information. Especially, the article focus on discussing MI from an information theory perspective. Aside from that, the article also reviews some MI estimation methods. The article is represented as follows. In the beginning, the article discusses the background of this article. The preliminaries section helps the reader to understand the basic concepts of information theory. The MI: definitions and properties section is divided into several subsections. The first subsection talks about the definition of mutual information in general. The rest of the subsections talk about the properties of MI, including the convexity and continuity of MI, the consequences of Jensen inequality for MI, the relations between MI and conditional independence distribution, geometric interpretation of MI, and variational form of MI. The MI: estimation methods section delivers a review of several mutual information methods and their current drawbacks.

II. Preliminaries

Sufficient knowledge about entropy and divergence is needed to have a better understanding of mutual information.

A. Entropy

Entropy can be viewed as a tool to measure the uncertainty of random variable (RV) [17]. Let X be a discrete random variable on space $\mathcal{X}$ with distribution P_X. Also, let $x \in \mathcal{X}$ be an element from space $\mathcal{X}$.The entropy of X can be written as:

$$\begin{aligned} H(X) &= -\mathbb{E}\left[\log P_X(x)\right] \\ &= -\sum_{x \in \mathcal{X}} P_X(x) \log P_X(x) \end{aligned}$$

Note that the equation is also hold for continuous random variable. The logarithm term in the equation uses either base 2 (*bit*) or base e (*nat*) [17]. Furthermore, it is easy to see that $H(X) \geq 0$ is satisfied since $0 \geq P_X(x) \geq 1$.

Entropy can also be used to measure the uncertainty for more than 1 random variable. Let Y be another discrete random variable on space $\mathcal{Y}$ with distribution P_Y. At first, we review joint entropy between random variables X and Y. Joint entropy $H(X, Y)$ with a joint distribution $P_{X,Y}$ is defined by:

$$\begin{aligned} H(X, Y) &= -\mathbb{E}\left[\log P_{X,Y}(x, y)\right] \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{X,Y}(x, y) \log P_{X,Y}(x, y) \end{aligned}$$

Then, we define conditional entropy of X given Y with conditional distribution $P_{X|Y}$ as:

$$\begin{aligned} H(X|Y) = \mathbb{E}_{y\in\mathcal{Y}}\left[H(P_{X|Y=y})\right] &= -\mathbb{E}\left[\log P_{X|Y}(x|y)\right] \\ &= -\sum_{y\in\mathcal{Y}} P_Y(y) \sum_{x\in\mathcal{X}} P_{X|Y}(x|y) \log P_{X|Y}(x|y) \\ &= -\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} P_{X,Y}(x, y) \log P_{X|Y}(x|y) \end{aligned}$$

The conditioning impacts on the reduction on entropy means that $H(X) \geq H(X|Y)$ [17]. We discuss about this inequality in the later section.

Joint entropy $H(X, Y)$ can be derived from marginal entropy $H(X)$ and conditional entropy $H(Y|X)$.

Theorem II.1. *Both $H(X, Y)$ and $H(X|Y)$ derive chain rule property written as:*

$$H(X, Y) = H(X) + H(Y|X) \leq H(X) + H(Y)$$

Note that the inequality holds from the conditioning of $H(Y|X)$. We also can extend the relations for more than two random variables as we call conditional joint entropy. Let us specify another random variable Z on space $\mathcal{Z}$. We can write conditional joint entropy $H(X, Y|Z)$ as:

$$H(X, Y|Z) = H(X|Z) + H(Y|X, Z) \leq H(X) + H(Y)$$

with the inequality $H(Y|X, Z) \leq H(Y)$ holds for the equation.

B. Divergence

Divergence (also known as Kullback-Leibler (KL) divergence or relative entropy) is a measurement of the distance between two distributions over a random variable [18]. We already specified random variable X on space $\mathcal{X}$ and distribution P_X. Then, let Q_X be another distribution function quantifying RV X. KL-Divergence between P_X and Q_X is defined by:

$$\begin{aligned} D_{KL}(P_X||Q_X) &= \mathbb{E}\left[\log \frac{P_X(x)}{Q_X(x)}\right] \\ &= \sum_{x\in\mathcal{X}} P_X(x) \log \frac{P_X(x)}{Q_X(x)} \; (discrete) \\ &= \int P_X(x) \log \frac{P_X(x)}{Q_X(x)} dx \; (continuous) \end{aligned}$$

There are two constraints for the above definitions:

- $0. \log \frac{0}{0} = 0$
- $\exists x \; : \; Q_X(x) = 0$ and $P_X(x) > 0 \implies D_{KL}(P_X||Q_X) = \infty$

Note that KL divergence is not symmetric means $D_{KL}(P_X||Q_X) \neq D_{KL}(Q_X||P_X)$. Furthermore, we can also extend KL-divergence into conditional case where probability function P_X is given. In particular, KL-divergence between $P_{Y|X}$ and $Q_{Y|X}$ (not symmetric) given P_X can be written by:

$$\begin{aligned} & D_{KL}(P_{Y|X}||Q_{Y|X}|P_X) = \mathbb{E}_{P_X}\left[D(P_{Y|X=x}||Q_{Y|X=x})\right] \\ &= \sum_{x\in\mathcal{X}} P_X(x) D(P_{Y|X=x}||Q_{Y|X=x}) \; (disc.) \\ &= \int P_X(x) D(P_{Y|X=x}||Q_{Y|X=x}) dx \; (cont.) \end{aligned}$$

Evidently, KL divergence is a special case of f-divergence [18]. With $P_X \ll Q_X$, f-divergence is defined by:

$$\begin{aligned} D_f(P_X||Q_X) &= \mathbb{E}_{Q_X}\left[f\left(\frac{dP_X}{dQ_X}\right)\right] \\ &= \sum_{x\in\mathcal{X}} Q_X(x) f\left(\frac{P_X(x)}{Q_X(x)}\right) \; (discrete) \\ &= \int Q_X(x) f\left(\frac{P_X(x)}{Q_X(x)}\right) dx \; (continuous) \end{aligned}$$

Using the definition above, we can rewrite $D_{KL}(P_X||Q_X)$ as:

$$D_f(P_X||Q_X) = \mathbb{E}_{P_X}\left[\log \frac{P_X}{Q_X}\right] = \mathbb{E}_{Q_X}\left[\frac{P_X}{Q_X} \log \frac{P_X}{Q_X}\right]$$

with $f\left(P_X/Q_x\right) = P_X/Q_X \log P_X/Q_X$. Another case of f-divergence including Jensen-Shannon divergence ($JS(P_X||Q_X) = D_{KL}(P_X||(P_X + Q_X)/2) + D_{KL}(Q_X||(P_X + Q_X)/2)$), total variation ($T(P_X, Q_X) = 1/2\, \mathbb{E}_{Q_X}\left[|P_X/Q_X - 1|\right]$), etc [18].

III. Mutual Information : Definitions and Properties

A. General Definition of Mutual Information

We have discussed entropy in the previous section. We then define mutual information (MI) which quantifies the amount of information of a particular random variable given another random variable [17], [19]. Given joint probability $P_{X,Y}$ and marginal probability P_X & P_Y, mutual information between random variable X and Y is written by:

$$\begin{aligned} I(X; Y) &= \mathbb{E}_{P_{X,Y}} \log \frac{P_{X,Y}(x, y)}{P_X(x)P_Y(y)} && (1) \\ &= D(P_{Y|X}||P_Y|P_X) && (2) \\ &= D(P_{X|Y}||P_X|P_Y) && (3) \\ &= D_{KL}(P_{X,Y}(x, y)||P_X(x)P_Y(y)) && (4) \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{X,Y}(x, y) \log \frac{P_{X,Y}(x, y)}{P_X(x)P_Y(y)} && (5) \end{aligned}$$

Following the same constraint as entropy, MI can also be applied to a continuous random variable [17]. In contrast to KL-divergence which is not symmetric, MI results in symmetric form means that $I(X; Y) = I(Y; X)$.

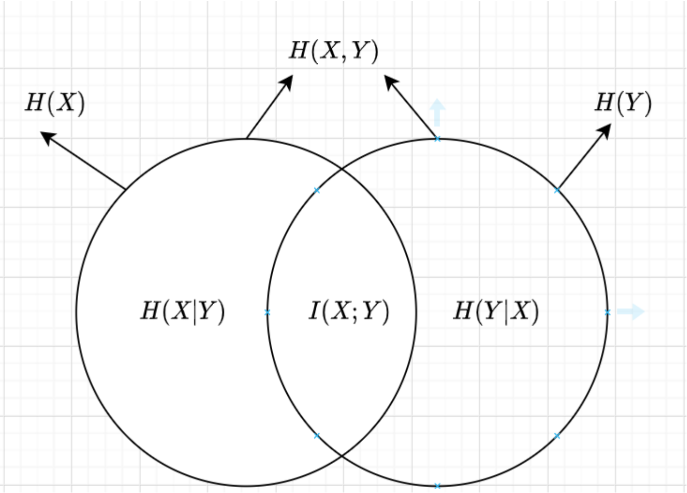
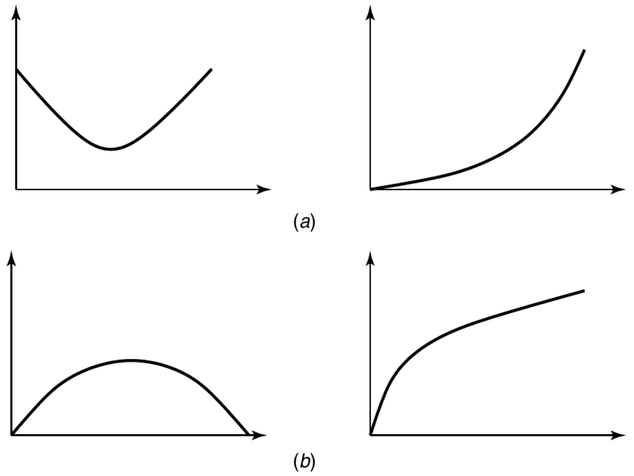


Fig. 1. Venn diagram that shows the relationship between RV X and Y. Observe that MI between X and Y is lied on the intersection between marginal entropy $H(X)$ and $H(Y)$ [17].


(*a*)

(*b*)

Fig. 2. **a.** Convex functions represented by a an upward-opening. **b.** Concave function represented by a downward-opening curve. [17]

In the previous section we already elaborate the entropy of joint distribution and conditional distribution as well. Evidently, those entropies have relationship with mutual information. Figure 1 shows the relationship between two random variables from information theory perspective. From the figure, we can derive the definition of mutual information $I(X;Y)$ in term of $H(X)$, $H(Y)$, $H(X,Y)$, $H(Y|X)$, and $H(X|Y)$.

Theorem III.1.

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) && (6)\\ &= H(Y) - H(Y|X) && (7)\\ &= H(X) + H(Y) - H(X,Y) && (8) \end{aligned}$$

Observe that $I(X;X) = H(X)$ for discrete RV (since $H(X,X) = H(X)$), otherwise it results ∞. We also can use the entropy to define the conditional mutual information. In particular, conditional MI of RV X and Y given Z is defined by:

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) && (9)\\ &= \mathbb{E}_{P_{X,Y,Z}} \log \frac{P_{X,Y|Z}(x,y|z)}{P_{X|Z}(x|z)P_{Y|Z}(y|z)} && (10) \end{aligned}$$

. Mutual information also satisfied a chain rule theorem.

Theorem III.2.

$$I(X_1,...,X_n;Y) = \sum I(X_i;Y|X_{i-1},...,X_1) \qquad (11)$$

We have discuss about the definition of MI in term of entropy and KL-divergence as well. In the next sections, we discuss about some properties of MI.

B. Convexity and Continuity of Mutual Information

We begin this section by defining convex and concave function. A function $f(x)$ is a convex function for interval (u,v) if for every $x_i, x_j \in (u,v)$ and $0 \leq \alpha \leq 1$ holds $f(\alpha x_i + (1-\alpha)x_j) \leq \alpha f(x_i) + (1-\alpha)f(x_j)$ [17]. We then call f as strictly convex if equality is satisfied when $\alpha = 0$ or $\alpha = 1$. Meanwhile, a function f is said to be concave when the negation $-f$ is convex. Figure 2 shows the examples of convex and concave function.

We then have three theorems about the convexity and concavity of KL-divergence, entropy, and mutual information.

Theorem III.3. *$D_{KL}(P_X||Q_X)$ is convex function. In particular given the pair of distribution functions (P_Xi, Q_Xi) and (P_Xj, Q_Xj) then*

$$\begin{aligned} D_{KL}(\alpha P_{Xi} + (1-\alpha)P_{Xj}||\alpha Q_{Xi} + (1-\alpha)Q_{Xj}) \leq \\ \alpha D_{KL}(P_{Xi}||Q_{Xi} + (1-\alpha)D_{KL}(P_{Xj}||Q_{Xj})) \quad (12) \end{aligned}$$

for $0 \leq \alpha \leq 1$ [17]

Theorem III.4. *Given a probability distribution P_X of RV X on space $\mathcal{X}$, entropy $H(P_X)$ is concave [17].*

Theorem III.5. *Let $(X,Y) \sim P_{X,Y}(x,y) = P_X(x)P_{Y|X}(y|x)$. The mutual information $I(X;Y)$ is a concave function of $P_X(x)$ for fixed $P_{Y|X}(y|x)$ and a convex function of $P_{Y|X}(y|x)$ for fixed P_X [17].*

Besides being convex, MI also possesses continuity property. We show this property by first seeing that KL divergence and entropy are continuous. Formally, for a fix distribution Q_X on space $\mathcal{X}$ with $Q(x) > 0 \quad \forall x \in \mathcal{X}$ then $D_{KL}(P_X||Q_X)$ is continuous. In particular, $H(P_X)$ is continuous [18]. We then define MI by $I(X;Y) = H(X) + H(Y) - H(X,Y)$. Since $H(X)$ is continuous, then $I(X;Y)$ is assured to be continuous.

C. Jensen Inequality and The Consequences for Mutual Information

The Jensen inequality requires a function to be convex.

Theorem III.6. *Jensen's inequality: if g is a convex function and X is a random variable then*

$$\mathbb{E}(g(X)) \geq g(\mathbb{E}X) \tag{13}$$

with equality hold when the function is strictly convex

This inequality is used to discover the property of KL-divergence. Note that we have shown that KL-divergence is a convex function (Equation 12).

Theorem III.7. *Divergence inequality: Given distribution function P_X and Q_X over $x \in \mathcal{X}$. Then it applies that*

$$D(P_X||Q_X) \geq 0 \tag{14}$$

with equality hold when $P_X(x) = Q_X(x)$

We then use the theorem above to imply the property of MI. Since $I(X;Y) = D(P_{X,Y}(x,y)||P_X(x)P_Y(y))$ then it implies that $I(X;Y) \geq 0$ with equality hold when $P_{X,Y}(x,y) = P_X(x)P_Y(y)$. Second implication is $I(X;Y|Z) \geq 0$ since we can transform it into the form of D_{KL} as well. The last implication already being stated in preliminary section which is $H(X|Y) \leq H(X)$. Recall that $I(X;Y) = H(X) - H(X|Y)$. Since $I(X;Y) \geq 0$ then $H(X) - H(X|Y) \geq 0$.

D. Relations between Conditional Independence and Mutual Information

In this section, we show that some conditional independent forms of distribution results in inequality of MI. Random variable X, Y, Z are said to be conditional independent if:

$$\begin{aligned} P_{X,Z|Y}(x,z|y) &= \frac{P_{X,Y,Z}(x,y,z)}{P_Y(y)} \\ &= \frac{P_{X,Y}(x,y)P_{Z|Y}(z|y)}{P_Y(y)} \\ &= P_{X|Y}(x|y)P_{Z|Y}(z|y) \end{aligned} \tag{15}$$

From the graphical model perspective, random variable X, Z are conditionally independent given Y if and only if X, Y, Z forms a Markov chain denoted by $X \rightarrow Y \rightarrow Z$ [18]. Under the circumstance, joint probability X, Y, Z is defined by:

$$P_{X,Y,Z} = P_X(x)P_{Y|X}(y|x)P_{Z|Y}(z|y) \tag{16}$$

Furthermore, Markov chain $X \rightarrow Y \rightarrow Z$ also implies $Z \rightarrow Y \rightarrow X$ [18]. Another form of Markov chain that satisfies conditional independence is $X \leftarrow Y \rightarrow Z$ [18] where the joint probability is defined by:

$$P_{X,Y,Z} = P_Y(y)P_{X|Y}(x|y)P_{Z|Y}(z|y) \tag{17}$$

Having the definitions, we derive inequality theorem constrained by the Markov chain form.

Theorem III.8. *if $X \rightarrow Y \rightarrow Z$ then $I(X;Y) \geq I(X;Z)$*

Using the above theorem, we can derive two properties. First, if $Z = g(Y)$ then we have $I(X;Y) \geq I(X;g(Y))$ since $X \rightarrow Y \rightarrow g(Y)$ will follows Markov chain. We also have $(X;Y|Z) \leq I(X;Y)$. This property comes by noticing that $I(X;Y|Z) = 0$ and $I(X;Z) \geq 0$ [18].

E. Geometric Interpretation of Mutual Information

We know elaborate mutual information from the perspective of geometry. First, we examine mutual information as conditional divergence. Recall Equation 2, we write it into discrete form as:

$$\begin{aligned} I(X;Y) &= D_{KL}(P_{Y|X}||P_Y|P_X) \\ &= \sum_x D_{KL}(P_{Y|X=x}||P_Y)P_X(x) \end{aligned}$$

We can see that each outcome x is weighted by probability distribution $P_X(x)$. Hence, we can say that MI is a weighted distance measure between two distributions.

In this section, we specify an auxiliary distribution Q to redefine MI.

Theorem III.9. $\forall Q_Y$ *such that* $D_{KL}(P_Y||Q_Y) < \infty$

$$I(X;Y) = D_{KL}(P_{X|Y}||Q_X|P_Y) - D(P_X||Q_X) \tag{18}$$

If Q_X is optimum such that $Q_X = P_X$ then the second term can be removed, thus $I(X;Y) = \underset{Q_X}{\operatorname{argmin}} D_{KL}(P_{X|Y}||Q_X|P_Y)$ [18]. Intuitively, the auxiliary distribution Q_X will be moving towards the real distribution P_X in some probability measure space during the optimization.

We can scale up the utilization of auxiliary/variational distribution for two RV X, Y. In the theorem below, we specify a new auxiliary distribution Q_Y.

Theorem III.10. *We can see mutual information as a distance to product distribution [18].*

$$I(X;Y) = \underset{Q_X,Q_Y}{\operatorname{argmin}} D_{KL}(P_{X,Y}||Q_XQ_Y) \tag{19}$$

We can generalize the theorem above to conditional mutual information as $I(X;Z|Y) = \underset{Q_{X,Y,Z}:X\rightarrow Y\rightarrow Z}{\operatorname{argmin}} D_{KL}(P_{X,Y,Z}||Q_{X,Y,Z})$ [18].

F. Variational Form of Mutual Information

In the previous section, we have discussed one of the variational form of MI (Equation 18). This section provides another two variational forms of MI. These forms are based on characterizations KL-divergence : Donsker-Varadhan and Gelfand-Yaglom-Perez.

We begin by introducing the Donsker-Varadhan form of KL-divergence.

Theorem III.11. *Donsker-Varadhan: Let P_X, Q_X be a probability measures of RV X on space $\mathcal{X}$ and $\mathcal{C}$ be the set of function $g : \mathcal{X} \rightarrow \mathbb{R}$ such that $\mathbb{E}_{Q_X}[e^{g(X)}] < \infty$. If $D_{KL}(P_X||Q_X) < \infty$ then for all $f \in \mathcal{C}$ expectation $\mathbb{E}_{P_X}[g(X)]$ exists and also [18]:*

$$D(P_X||Q_X) = \sup_{g\in\mathcal{C}} \mathbb{E}_{PX}[g(X)] - \log \mathbb{E}_{Q_X}[e^{g(X)}] \tag{20}$$

We then apply the theorem above to find the Donsker-Varadhan form of MI. Using the Equation 4 and Equation 20 we get:

$$I(X;Y) = \sup_{g} \mathbb{E}[g(X,Y)] - \log \mathbb{E}[e^{g(X,\hat{Y})}] \tag{21}$$

with $\hat{Y}$ is a duplicate of Y which is independent of X and the supremum is over bounded or even bounded by continuous functions g.

The next theorem introducing Gelfand-Yaglom-Perez form of KL-divergence which involves σ-space.

Theorem III.12. *Gelfand-Yaglom-Perez: Let P_X, Q_X be a probability measures on space $\mathcal{X}$ with σ-algebra $\mathcal{F}$. Then:*

$$D(P_X||Q_X) = \sup_{\{E_1,\dots,E_n\}} \sum_{i=1}^{n} P_X[E_i] \log \frac{P_X[E_i]}{Q_X[E_i]} \tag{22}$$

with the supremum is over all finite $\mathcal{F}$-measurable partitions: $\cup_{j=1}^{n} E_j = \mathcal{X}$, $E_j \cap E_i = \emptyset$.

with $0 \log \frac{1}{0} = 0$ and $\log \frac{1}{0} = \infty$ for conventions. We then apply the theorem above to find the Donsker-Varadhan form of MI. Using the Equation 4 and Equation 22 we get:

$$I(X;Y) = \sup_{\{E_i\}\times\{F_j\}} \sum_{i,j} P_{X,Y}\left[E_i \times F_j\right] \log \frac{P_{X,Y}\left[E_i \times F_j\right]}{P_X\left[E_i\right] P_Y\left[F_j\right]} \tag{23}$$

with supremum is over finite partitions space $\mathcal{X}$ and $\mathcal{Y}$.

IV. Mutual Information: Estimation Methods

We already know that mutual information can capture the dependence of random variables. But often times we can not directly use the closed function of mutual information. Recall that in the Equation 1, we need the access to $P_{X,Y}(x,y)$, P_X, and $P_Y(y)$ which are not always guaranteed. The mutual information estimation then come to bound the true MI. The estimation is either upper-bounding or lower-bounding the true MI. The idea of MI estimations come from variational form of MI. In the previous section we already discuss three variational forms of MI. We try to approximate the MI estimation by using an auxilary distribution or a critic function.

In this section, we review several MI estimation methods. The review has been conducted before by Poole et al., (2019). Figure 3 shows the schematic of variational bounds of mutual information proposed by Poole et al., 2019 [20]. In this article, we divide the reviews into three sections: normalized bounds, unnormalized bounds, and improved bounds.

A. Normalized Bounds

In this section, we discuss two versions of normalized bounds, upper bound and lower bound MI estimation. The bounds were firstly introduced by Agakov [13]. Recall the definition of MI in Equation 1. We then rewrite $P_{X,Y}(x,y) = P_{Y|X}(y|x)P_Y(y)$. Subsequently, we apply Theorem III.10 by replacing $P_Y(y)$ with a variational distribution $Q_Y(y)$. Mathematically, we can write:

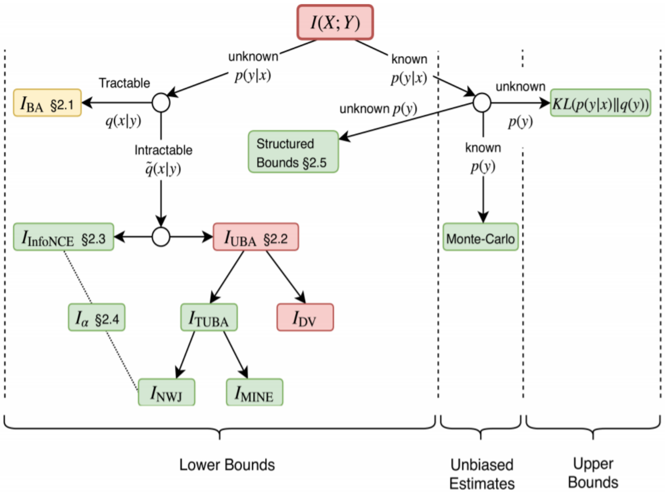


Fig. 3. Schematic of variational bounds of mutual information proposed by Poole et al., 2019. The schematic is based on the presence of the available distributions [20]

$$\begin{aligned}
I(X,Y) &= \mathbb{E}_{P_{X,Y}(x,y)}\left[\log \frac{P_{Y|X}(y|x)}{P_Y(y)}\right] \\
&= \mathbb{E}_{P_{X,Y}(x,y)}\left[\log \frac{P_{Y|X}(y|x)Q_Y(y)}{P_Y(y)Q_Y(y)}\right] \\
&= \mathbb{E}_{P_{X,Y}(x,y)}\left[\log \frac{P_{Y|X}(y|x)}{Q_Y(y)}\right] - KL(P_Y(y)||Q_Y(y)) \\
&\geq \mathbb{E}_{P_{X,Y}(x,y)}\left[\log \frac{P_{Y|X}(y|x)}{Q_Y(y)}\right] \triangleq I_R
\end{aligned} \tag{24}$$

Thus, we upper-bounding the MI. Note that in Theorem III.10, we can assure equality since we assumed that we can find the optimum Q_Y. We also need to constraint $Q_Y(y)$ to be intractable. However, the assumption is not assured in the real world. One of the applications of the bound is for deep information bottle-neck model [12].

In contrast, we derive lower-bound by applying Theorem III.10 into the numerator $P_{X|Y}(x|y)$ [13]. We replace $P_{X|Y}(x|y)$ with $Q_{X|Y}(x|y)$:

$$\begin{aligned}
I(X,Y) &= \mathbb{E}_{P_{X,Y}(x,y)}\left[\log \frac{P_{X|Y}(x|y)}{P_X(x)}\right] \\
&= \mathbb{E}_{P_{X,Y}(x,y)}\left[\log \frac{Q_{X|Y}(x|y)}{P_X(x)}\right] + \\
&\quad \mathbb{E}_{P_Y(y)}\left[KL(P_{X|Y}(x|y)||Q_{X|Y}(x|y))\right] \\
&\geq \mathbb{E}_{P_{X,Y}(x,y)}\left[\log Q_{X|Y}(x|y)\right] + h(X) \triangleq I_{BA}
\end{aligned} \tag{25}$$

with $h(X)$ is the marginal entropy of X. The objective is tractable if $h(X)$ is known. However, $h(X)$ is often to be unknown. This bound has been applied as regularizer of Info-GAN objective function [8].

B. Unnormalized Bounds

We can solve the intractibility problem from the previous section by using the unnormalized form of $Q_{X|Y}(x|y)$. We write the distribution in terms of a critic function $g(x, y)$ and marginal distribution $P_X(x)$:

$$Q_{X|Y}(x|y) = \frac{P_X(x)}{Z(y)} e^{g(x,y)};\ Z(y) = \mathbb{E}_{P_X(x)} \left[e^{g(x,y)} \right] \quad (26)$$

By applying the equation above into Equation 25, we get unnormalized BA estimation (I_{UBA}):

$$\mathbb{E}_{P_{X,Y}(x,y)}[g(x, y)] - \mathbb{E}_{P_Y(y)} [\log Z(y)] \triangleq I_{UBA} \quad (27)$$

Note that in the equation above, the entropy $H(X)$ is no longer involved. However, the term $\log Z(y)$ is still intractable. Since log function is convex, by applying the Jensen inequality we have Donsker-Varadhan lower bound [21]:

$$\mathbb{E}_{P_{X,Y}(x,y)}[g(x, y)] - \log \mathbb{E}_{P_Y(y)} [Z(y)] \triangleq I_{DKV} \quad (28)$$

Note that $I_{BA} \geq I_{DKV}$ (by Jensen inequality). We have seen this form from Theorem III.11, except without confirming the equality. This bound is also still intractable. By upper-bounding the log partition $\log Z(y)$, we can form a tractable bound. We specify an inequality $\log(x) \leq \frac{x}{a} + \log(a) - 1, \forall x, a > 0$. Applying the inequality into the second term of Equation 27 will give $\log(Z(y)) \leq \frac{Z(y)}{a(y)} + \log(a(y)) - 1$. Finally, we can rewrite the bound as:

$$\begin{aligned} &\mathbb{E}_{P_{X,Y}(x,y)}[g(x, y)] - \\ &\mathbb{E}_{P_Y(y)} \left[\frac{\mathbb{E}_{P_X(x)} \left[e^{g(x,y)} \right]}{a(y)} + \log(a(y)) - 1 \right] \triangleq I_{TUBA} \end{aligned} \quad (29)$$

The bound is optimized with respect to $a(y)$ and g. Both are optimized simultaneously. Furthermore, we can simplify Equation 29 by set $a(y) = e$ which leads to Nguyen-WainWright-Jordan estimation [22]:

$$\mathbb{E}_{P_{X,Y}(x,y)}[g(x, y)] - e^{-1} \mathbb{E}_{P_Y(y)} [Z(y)] \triangleq I_{NWJ} \quad (30)$$

Generally, unnormalized bounds suffer from the high variance problem due to the log partition function.

C. Improved Bounds

In this section, we discuss several improvements that have been made to respond the current drawbacks of normalized and unnormalized bound.

Info-NCE extends the NWJ estimations by using Monte Carlo estimation on multiple samples [16]:

$$I(X, Y) \geq \mathbb{E} \left[\frac{1}{k} \sum_{i=1}^{K} \log \frac{e^{f(x_i, y_i)}}{\sum_{j=1}^{K} e^{f(x_j, y_j)}} \right] \triangleq I_{NCE} \quad (31)$$

However, this estimation tends to have a higher bias compared to NWJ estimation.

Barber-Agakov upper bound estimation also have a problem with the variational distribution $Q_Y(y)$. Evidently, learning distribution $Q_y(y)$ without any prior knowledge is extremely difficult especially when RV Y is high dimensional [1], [23]. The distribution $Q_Y(y_i)$ can be replaced with Monte Carlo approximation $Q_y(y) = \frac{1}{K-1} \sum_{j \neq i} P_{Y|X}(y|x_j)$ [20], we derive one left out (L1-out) upper bound estimation:

$$\mathbb{E} \left[\frac{1}{K} \sum_{i=1}^{K} \left[\log \frac{P_{Y|X}(y_i|x_i)}{\frac{1}{K-1} \sum_{j \neq i} P_{Y|X}(y_i|x_j)} \right] \right] \triangleq I_{L1-out} \quad (32)$$

The estimation method is called one left out because we discard one sample on the denumerator inside the sum. The drawback of this method lies to its numerical instability especially when RV Y is high dimensional [1].

Given all existing MI estimations, the current methods still have several drawbacks. MI estimation is currently active research. For example, current research shows that we can estimate MI by using optimal transport concept that is Wasserstein distance [24]. Another research using clipping method to reduce the variance of NWJ estimation [25].

V. Conclusion

The article discussed the definitions of mutual information in the form of KL-divergence and entropy as well. The article then delivered some properties of mutual information including concavity, the continuity, Jensen inequality, conditional independence, and variational form. Later, the article reviewed several mutual information estimation methods. The estimation methods are useful whenever we have an unaccessible probability (commonly marginal distribution). We also mention that the current mutual information estimation methods also have drawbacks.

References

[1] P. Cheng, “CLUB: A contrastive log-ratio upper bound of mutual information, ”, In International Conference on Machine Learning, PMLR, 2020, pp.1779-1788.

[2] D. R. Brillinger, “Second-order moments and mutual information in the analysis of time series, ”, in Recent Advances in Statistical Methods (Ed. Y. P. Chaubey), London: Imperial College Press, 2002, Pp. 64-76.

[3] P. Viola, “Alignment by Maximization of Mutual Information, ”, PhD thesis, Massachusetts Institute of Technology, 1995.

[4] L. Song, and S. Horvath, “Comparison of co-expression measures: mutual information, correlation, and model based indices, ”, in BMC Bioinformatics 13, 328, 2012, https://doi.org/10.1186/1471-2105-13-328.

[5] I. Priness, and O. Maimon, and I. Ben-Gal “Evaluation of gene-expression clustering via mutual information distance measure, ”, in BMC Bioinformatics 8, 111, 2007, https://doi.org/10.1186/1471-2105-8-111.

[6] B. J. Julian, “Mutual information-based gradient-ascent control for distributed robotics, ”, Doctoral Disertation, Massachusetts Institute of Technology, 2013.

[7] B. J. Julian, S. Karaman and D. Rus, “On mutual information-based control of range sensing robots for mapping applications, ”, 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2013, pp. 5156-5163, doi: 10.1109/IROS.2013.6697102.

[8] X. Chen, and Y. Duan, and R. Houthooft, and J. Schulman, and I. Sutskever, and P. Abbeel, “Infogan: Interpretable representation learning by information maximizing generative adversarial nets, ”, arXiv preprint, arXiv:1606.03657, 2016.

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

[10] T. Q. Chen, and X. Li, and R. B. Grosse, and K. D. Duvenaud, “Isolating sources of disentanglement in variational au- toencoders, ”, In NeurIPS, 2018.

[11] H. Kazemi, and S. Soleymani, and F. Taherkhani, and S. Iranmanesh, and N. Nasrabadi, “Unsupervised image-to-image translation using domain-specific variational information bound, ”, In NeurIPS, 2018.
[12] A. Alemi, and I. Fischer, and J. V. Dillon, and K. Murphy, “Deep variational information bottleneck, ”, arXiv preprint arXiv:1612.00410, 2016.
[13] D. B. F. Agakov, “The im algorithm: a variational approach to information maximization, ” Advances in neural information processing systems, 16, 2004.
[14] W. Hu, and T. Miyato, and S. Tokui, and E. Matsumoto, and Sugiyama, “M. Learning discrete representations via information maximizing self-augmented training, ”, In ICML, 2017.
[15] C. Florensa, and Y. Duan, and P. Abbeel, “Stochastic neural networks for hierarchical reinforcement learning, ”. arXiv preprint, arXiv:1704.03012, 2017.
[16] A. V. D. Oord, and Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding, ”, arXiv preprint, arXiv:1807.03748, 2018.
[17] T. M. Cover, and J. A. Thomas, “Elements of Information Theory, 2nd Edition, ” US:Wiley-Interscience, 2006, pp. 13-37.
[18] Y. Polyanskiy, and Y. Wu, “Lecture Notes on Information Theory, ” MIT, 2012. Accessed on: June 2, 2021. [Online]. http://people.lids.mit.edu/yp/homepage/papers.html.
[19] C. M. Bishop, “Pattern Recognition and Machine Learning, ” Berlin, Heidelberg : Springer-Verlag, 2006, pp.55-57.
[20] B. Poole, and S. Ozair, and A. V. D. Oord, and A. Alemi, and G. Tucker, “On variational bounds of mutual information, ” In International Conference on Machine Learning, PMLR, 2019, (pp. 5171-5180).
[21] M. D. Donsker, and S. S. Varadhan, “Asymptotic evaluation of certain markov process expectations for large time, ”, iv. Communications on Pure and Applied Mathematics, 1983, 36 (2):183–212.
[22] X. Nguyen, and M. J. Wainwright, and M. I. Jordan, “Estimating divergence functionals and the likelihood ratio by convex risk minimization, ”, IEEE Transactions on Information Theory, 2010, , 56(11):5847–5861.
[23] M. Magdon-Ismail, and A. F. Atiya, “Neural networks for density estimation, ” In NeurIPS, 1999.
[24] S. Ozair, C. Lynch, and Y. Bengio, and A. V. D. Oord, and S. Levine, and P. Sermanet, “Wasserstein dependency measure for representation learning, ” ICLR, 2019.
[25] J. Song, and S. Ermon, “Understanding the limitations of variational mutual information estimators, ”, arXiv preprint, arXiv:1910.06222, 2019.

Appendix

A. Proof of Theorem II.1

We prove the theorem in the discrete form of random variables [17]

$$\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{X,Y}(x,y)\log P_{X,Y}(x,y) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{X,Y}(x,y)\log P_X(x)P_{Y|X}(y|x) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{X,Y}(x,y)\log P_X(x) \\
&\quad -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{X,Y}(x,y)\log P_{Y|X}(y|x) \\
&= -\sum_{x\in\mathcal{X}} P_X \log P_X(x) \\
&\quad -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{X,Y}(x,y)\log P_{Y|X}(y|x) \\
&= H(X) + H(Y|X)
\end{aligned}$$

B. Proof of Theorem II.1

$$\begin{aligned}
I(X_1,...,X_n;Y) &= H(X_1,...,X_n) - H(X_1,...,X_n|Y) \\
&= \sum_{i=1}^{n} H(X_i|X_{i-1}),...,X_1) \\
&\quad - \sum_{i=1}^{n} H(X_i|X_{i-1},...,X1,Y) \\
&= \sum_{i=1}^{n} I(X_i;Y|X1,...,X_{i-1})
\end{aligned}$$

C. Proof of Theorem III.3

In order to prove the theorem, we apply log sum inequality on the left-hand side [17].

$$\begin{aligned}
&(\alpha P_1(x) + (1-\alpha)P_2(x)\log\frac{\alpha P_1(x) + (1-\alpha)P_2(x)}{\alpha Q_1(x) + (1-\alpha)Q_2(x)}) \\
&\le \alpha P_1(x)\log\frac{\alpha P_1(x)}{\alpha Q_1(x)} + (1-\alpha)P_2(x)\log\frac{(1-\alpha)P_2(x)}{(1-\alpha)Q_2(x)}
\end{aligned}$$

D. Proof of Theorem III.4

The result comes from the fact that $H(P_X) = log|\mathcal{X}| - D_{KL}(P_X||U_X)$ where U_X is an uniform distribution of $x \in \mathcal{X}$. The negative term of KL-divergence of the equation then implies its concavity [17].

E. Proof of Theorem III.5

We recall the definition of MI to prove the theorem:

$$\begin{aligned}
I(X;Y) &= H(Y) - H(Y|X) \\
&= H(Y) - \sum_x P_X(x)H(Y|X=x)
\end{aligned}$$

First, we proof the first argument of the theorem. Given $P_{Y|X}(y|x)$, then $P_Y(y)$ is linear function of $P_X(x)$. Since $H(Y)$ is a convex function of $P_Y(y)$, then we can say that $H(Y)$ is a concave function of $P_X(x)$. We can see the second term of as a function of $P_X(x)$. Thus, the difference is a concave function of $P_X(x)$ [17].

For the second argument, we specify two conditional distributions $P_{1\,Y|X}, P_{2\,Y|X}$. The corresponding joint distributions given the conditional distributions are $P_{1\,X,Y}(x,y) = P_X(x)P_{1\,Y|X}(y|x)$ and $P_{2\,X,Y}(x,y) = P_X(x)P_{2\,Y|X}(y|x)$ with respective marginals $P_X(x)$, $P_{1\,Y}(y)$ and $P_X(x)$, $P_{1\,Y}(y)$ We then specify a conditional distribution which is a mixture of $P_{1\,Y|X}(y|x)$ and $P_{2\,Y|X}(y|x)$:

$$P_{\alpha Y|X}(y|x) = \alpha P_{1\,Y|X}(y|x) + (1-\alpha)P_{2\,Y|X}(y|x)$$

$0 \le \alpha \le 1$. We can easily see that the corresponding joint distribution is also a mixture joint distribution,

$$P_{\alpha X,Y}(x,y) = \alpha P_{1\,X,Y}(x,y) + (1-\alpha)P_{2\,X,Y}(x,y)$$

and the marginal distribution Y is also a mixture,

$$P_{\alpha Y}(y) = \alpha P_{1\,Y}(y) + (1-\alpha)P_{2\,Y}(y)$$

If we let $Q_{\alpha\,X,Y}(x,y) = P_X(x)P_{\alpha\,Y}(y)$ be the product of the marginal distributions, then we have:

$$Q_{\alpha X,Y}(x,y) = \alpha Q_{1\,X,Y}(x,y) + (1-\alpha)Q_{2\,X,Y}(x,y)$$

We already know that MI can be thought as KL-divergence between joint distribution and the product of marginal distributions, hence:

$$I(X;Y) = D_{KL}(P_{\alpha\,X,Y}(x,y)||Q_{\alpha|:X,Y}(x,y))$$

Since KL-divergence is a convex function, thus the MI is convex function of conditional distribution [17].

F. Proof of Theorem III.6

The proof is for discrete distribution by using induction on the number of mass point. At first, we settle the base case which is the inequality of two-mass distribution (x_1 and x_2) [17]. Let w_1 and w_2 be the weights for x_1 and x_2 respectively, the inequality becomes:

$$w_1\,g(x_1) + w_2\,g(x_2) \geq g(w_1\,x_1 + w_2\,x_2)$$

Note that this inequality is similar with the definition of convex function. Suppose that the inequality is true for $k-1$ points. If we write $\hat{w_i} = w_i/(1-w_k)$ then :

$$\begin{aligned}
\sum_{i=1}^{k} w_i\,g(x_i) &= w_k\,g(x_k) + (1-w_k)\sum_{i=1}^{k-1}\hat{w}_i\,g(x_i) \\
&\geq w_k g(x_k) + (1-w_k) g\left(\sum_{i=1}^{k-1}\hat{w}_i x_i\right) \\
&\geq g\left(w_k\,x_k + (1-w_k)\sum_{i=1}^{k-1}\hat{w}_i x_i\right) \\
&= g\left(\sum_{i=1}^{k} w_i\,x_i\right)
\end{aligned}$$

G. Proof of Theorem III.3

Let $\mathcal{A}$ be the support of $P_x(x)$

$$\begin{aligned}
-D_{KL}(P_X||Q_X) &= -\sum_{x\in\mathcal{A}} P_X(x)\log\frac{P_X(x)}{Q_X x} \\
&= \sum_{x\in\mathcal{A}} P_X(x)\log\frac{Q_X(x)}{P_X(x)} \\
&\leq \log\sum_{xin\mathcal{A}} P_X(x)\log\frac{Q_X(x)}{P_X(x)} \\
&= \log\sum_{x\in\mathcal{A}} Q_X(x) \\
&\leq \log\sum_{x\in\mathcal{X}} Q_X(x) \\
&= \log 1 \\
&= 0
\end{aligned}$$

H. Proof of Theorem III.8

$$\begin{aligned}
I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \\
&= I(X;Y) + I(X;Z|Y)
\end{aligned}$$

Since X and Z are conditionally independent given Y, thus $I(X;Z|Y)$. Moreover, $I(X;Y|Z) \geq 0$ implies:

$$I(X;Y) \geq I(X;Z)$$

I. Proof of Theorem III.9

$$\begin{aligned}
I(X;Y) &= \mathbb{E}_{P_{X,Y}}\left[\log\frac{P_{Y|X}(y|x)}{P_Y(y)}\right] \\
&= \mathbb{E}_{P_{X,Y}}\left[\log\frac{P_{Y|X}(y|x)Q_Y(y)}{P_Y(y)Q_Y(y)}\right] \\
&= \mathbb{E}_{P_{X,Y}}\left[\log\frac{P_{Y|X}(y|x)}{Q_Y(y)}\right] + \\
&\quad \mathbb{E}_{P_{X,Y}}\left[\log\frac{Q_Y(y)}{P_Y(y)}\right] \\
&= \mathbb{E}_{P_X}\mathbb{E}_{P_Y}\left[\log\frac{P_{Y|X}(y|x)}{Q_Y(y)}\right] - \\
&\quad \mathbb{E}_{P_{X,Y}}\left[\log\frac{P_Y(y)}{Q_Y(y)}\right] \\
&= \mathbb{E}_{P_X}\left[D_{KL}(P_{Y|X}||Q_Y)\right] - D_{KL}(P_Y||Q_Y) \\
&= D_{KL}(P_{Y|X}||Q_Y|P_X) - D_{KL}(P_Y||Q_Y)
\end{aligned}$$

J. Proof of Theorem III.10

Since Q_X and Q_Y minimum, we have $Q_X = P_X$ and $Q_Y = P_Y$.

$$\begin{aligned}
I(X;Y) &= \mathbb{E}_{P_{X,Y}}\left[\log\frac{P_{X,Y}(x,y)Q_X(x)Q_Y(y)}{P_X(x)P_Y(y)Q_X(x)Q_Y(y)}\right] \\
&= \mathbb{E}_{P_{X,Y}}\left[\log\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}\right] + \mathbb{E}_{P_{X,Y}}\left[\log\frac{Q_X(x)}{P_X(x)}\right] + \\
&\quad \mathbb{E}_{P_{X,Y}}\left[\log\frac{Q_Y(y)}{P_Y(y)}\right] \\
&= D_{KL}(P_{X,Y}||Q_X Q_Y) + 0 + 0 \\
&= D_{KL}(P_{X,Y}||Q_X Q_Y)
\end{aligned}$$