Lower Bounds for Mutual Information in Representation Learning

Sheng Liu

May 2020

1 Introduction

Say you have two oranges, three apples, and a kitten, and you are asked to place these six objects into three groups. You will most likely put the apples, oranges, and the kitten in three different groups based on their color, shape, or smell without knowing the names (i.e. labels) of the objects. This impressive ability of humans to learn different concepts and perform different tasks in the absence of any supervised labels, called *unsupervised learning*, is already known in cognitive science. Infants learn to categorize objects based on surface appearances and functional features [1] without clearly specifying the tasks (*task-agnostic*) and with only minimal external supervision (*unsupervised*). It is generally hypothesized that the brain classifies objects by encoding the sensory input that reduces the amount of information processing [2]. Hence for humans to classify objects, some useful encoding (also called *representation*) for different objects is often explicitly learned. On the other hand, most machine learning algorithms categorize objects by using many object-label pairs (*supervised learning*) rather than by explicitly learning the representative encoding. When the labels are expensive to obtain, these algorithms are highly infeasible. Therefore, it is extremely valuable to propose classification algorithms that can discover representations without access to supervised labels during training. These algorithms are called *unsupervised representation learning*.

Formally, unsupervised representation learning is to learn a function g which maps the input data $x \in \mathbb{R}^d$ (images, audios, texts, etc.) into a lower dimensional space where one can solve some target supervised tasks more efficiently. Unlike supervised learning tasks where the discrepancy between the algorithm's conjectures and the labels is minimized, unsupervised learning has no natural objective to optimize. But recently, the *mutual information* (MI) between the input data x and the corresponding representations $z := g(x)$ are proposed to be the objective by researchers. The assumption is that a good representation should be the one that possesses most information in the input. MI is a fundamental quantity for measuring the amount of information obtained from a random variable X by observing some other random variable Z. It has found applications in a wide range of domains and tasks in data science for biomedical applications [3], information bottle neck [4], feature selection [5], and causality [6].

To be more specific, The MI of two random variables X and Z, with marginal distributions $p(x)$, $p(z)$ and joint density $p(x, z)$ is

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

In contrast to correlation, mutual information captures non-linear statistical dependencies between variables, and thus can act as a measure of true dependence.

Despite being a pivotal quantity across data science, MI is notoriously difficult to compute, particularly in continuous and high dimensional settings. Computing the MI exactly is only tractable for discrete variables (as the sum in the expectation can be computed exactly), or for a limited family of problems where the probability distributions are known. However, the lower bounds of MI are often more accessible. As we aim to maximize MI, maximizing its lower bounds could also be effective when the bounds are tight. In fact, several recent works [7, 8, 9] have demonstrated promising empirical results in unsupervised representation learning by maximizing the lower bounds of MI. In this report, we will explain how to derive the lower bounds of MI and how these lower bounds are related to each other.

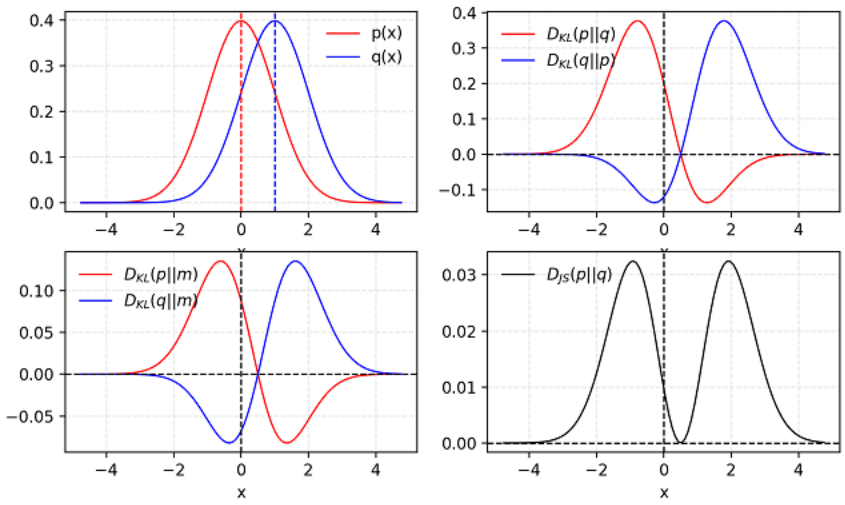

Figure 1: Figure on the top left shows two Gaussian density functions $p(x)$ and $q(x)$ with the same variance but different means. The forward KL divergences $D_{KL}(p||q)$ between $p(x)$ and $q(x)$ are shown as the integral of the red curve in the top right figure while the backward KL divergence $D_{KL}(q||p)$ is the integral of the blue curve in the same figure, KL divergence is not symmetric. Figure on the bottom left shows the KL divergence between the Gaussian distributions and their average $m := \frac{p(x)+q(x)}{2}$, JS divergence is the average of $D_{KL}(p||m)$ and $D_{KL}(q||m)$ and is symmetric. JS divergence will be the integral of the black curve in the bottom right figure.

2 Preliminary Concepts in Information Theory

In this section, we will recall some preliminary concepts which will be used to derive the lower bounds of MI in the later section. The mutual information is equivalent to the *Kullback–Leibler* (KL) divergence between the joint distribution $p(x, z)$ and the product of the marginals distributions $p(x)p(z)$:

$$I(X; Z) = D_{KL}(p(x, z)||p(x)p(z)). \tag{2}$$

Where the KL divergence D_{KL} between any two probability distributions $p(x)$ and $q(x)$ is defined as,

$$D_{KL}(p||q) = \mathbb{E}_{p(x)} \log \frac{p(x)}{q(x)}.$$

Note that KL divergence is non-symmetric, i.e. $D_{KL}(p||q)$ often does not equals to $D_{KL}(q||p)$. KL divergence is not a true distance metric between two distributions as it is non-symmetric. Unlike KL divergence, *Jensen-Shannon* (JS) divergence is defined as

$$D_{JS}(p||q) = \frac{1}{2} D_{KL}(p||\frac{p+q}{2}) + \frac{1}{2} D_{KL}(q||\frac{p+q}{2}), \tag{3}$$

and is symmetric. Figure 1 shows the differences between KL divergence and JS divergence. We can see from the figure that $D_{KL}(p||q)$ and $D_{KL}(q||p)$ are not equal.

There is a larger class of divergences of which the KL divergence, JS divergence are the special cases. This larger class of divergences is defined by some function f and thus is called *f-divergence*. The definition of f-divergence is

$$D_f(p||q) = \mathbb{E}_{q(x)} f\left(\frac{p(x)}{q(x)}\right), \tag{4}$$

where $f : \text{dom} f \subset \mathbb{R}_+ \to \mathbb{R}$ is a convex, lower-semicontinuous function satisfying $f(1) = 0$. Note that f with such properties will have a convex conjugate function also known as the *Fenchel conjugate* [10]. This

conjugate is defined as

$$f^*(t) := \sup_{u \in \mathrm{dom} f} \{ut - f(u)\}. \tag{5}$$

The function f^* is again convex and lower-semicontinuous and satisfying $(f^*)^* = f$. Hence we can represent f in terms of f^* as

$$f(u) = \sup_{t \in \mathrm{dom} f^*} \{tu - f^*(t)\}. \tag{6}$$

Notice that KL divergence is obtained by setting $f(u) = u \log u$, and JS divergence is obtained by setting $f(u) = -(u+1) \log \frac{1+u}{2} + u \log u$.

3 Lower Bounds on Mutual Information

Here, we review existing lower bounds on MI in unsupervised representation learning. One of the most widely used methods of deriving lower bounds of MI comes from the fact that MI can be represented as KL divergence (See Eq. (2)). Some subsequent works then replace the KL divergence by f-divergences we mentioned in the previous section to extend this idea. Technically, the lower bounds obtained from the f-divergence is not the lower bounds for MI defined in Eq. (1) anymore, but these "lower bounds" are still effective in learning representations in practice. So for completeness, we will cover them in the second part of this section as well.

3.1 Lower bounds obtained from KL divergence

First, observe that KL divergence can be represented by its Donsker-Varadhan (DV) dual representation:

Theorem 1 (Donsker-Varadhan representation)**.** *The KL divergence admits the following dual representation:*

$$D_{KL}(p \,||\, q) = \sup_{T: \Omega \to \mathbb{R}} \mathbb{E}_{p(x)}[T] - \log(\mathbb{E}_{q(x)}[e^T]), \tag{7}$$

where the supremum is taken over all functions T such that the two expectations are finite. The compact set $\Omega \subset \mathbb{R}^d$ is the support of density functions $p(x)$ and $q(x)$.

Proof. See the proof in the Appendix.

A straightforward consequence of this dual representation is as follows. Let $\mathcal{F}$ be any class of functions $T : \Omega \to \mathbb{R}$ satisfying the integrability constraints of the theorem. We then have the lower-bound:

$$D_{KL}(p||q) \geq \sup_{T \in \mathcal{F}} \mathbb{E}_{p(x)}[T] - \log(\mathbb{E}_{q(x)}[e^T]).$$

As function T can be any function from an unknown class of functions $\mathcal{F}$, T is not accessible in real life. Therefore, we assume that the algorithm will only search through a specific family of functions parameterized by θ (e.g. linear functions that parametrized by the slope and intercept), and we denote this specific family of functions as T_θ. Then using both Eq. (2) and the lower bound above we get the lower bound of the MI in practice as

$$I(X;Z) := D_{KL}(p(x,z)||p(x)p(z)) \geq \mathbb{E}_{p(x,z)}[T_\theta(x,z)] - \log(\mathbb{E}_{p(x)}\mathbb{E}_{p(z)}[e^{T_\theta(x,z)}]). \tag{8}$$

During learning, computer algorithms take N input data points $x_i,\ i = 1, 2, \ldots, N$, e.g. pictures of multiple objects, and produce representations z for each picture by finding the optimal θ that maximizing the empirical formulation of the lower bound.

3.2 Lower bounds obtained from f-divergence

As we previously mentioned, the MI can only be expressed as the KL divergence and not as any other f-divergences. However, as we are primarily interested in maximizing mutual information and not its precise value, KL divergence are replaced with some other f-divergences in the literature [7]. The relationship between lower bounds obtained from the f-divergence and the MI remains unclear (they are not necessary the lower bounds for MI). But in [8], authors shows that they have an approximately monotonic relationship for some specific distributions. Therefore, we call the f-divergence (not KL divergence) between the joint $p(x, z)$ and the product of marginals $p(x)p(z)$ as "Pseudo Mutual Information" which is similar to Eq. (2). To be specific, the "Pseudo Mutual Information" is defined as:

$$I^f(X; Z) = D_f(p(x, z)|| \, p(x)p(z)). \tag{9}$$

Similar to DV representation of KL divergence in Eq. (7), f-divergence also has a dual representation:

Theorem 2 (The f-divergence representation)**.** *The f-divergence admits the following dual representation:*

$$D_f(p||q) = \sup_{T:\Omega \to dom f^*} \mathbb{E}_{p(x)}[T] - \mathbb{E}_{q(x)}[f^*(T)], \tag{10}$$

where the supremum is taken over all functions T such that the two expectations are finite. The compact set $\Omega \subset \mathbb{R}^d$ is the support of density functions $p(x)$ and $q(x)$. Function f^* is the convex conjugate of f.

Proof. The proof is in the Appendix.

Because KL divergence is also a special case of f-divergence, it would be interesting to derive the lower bound using the f-divergence dual representation and compare it to the previous one derived from the DV representation. That is when $f(u) = u \log u$, we have the corresponding conjugate function $f^*(t) = e^{t-1}$, thus the lower bound obtained using the f-divergence representation for KL divergence is

$$I^{KL}(X; Z) \geq \mathbb{E}_{p(x,z)}[T(x, z)] - \mathbb{E}_{p(x)}\mathbb{E}_{p(z)}[e^{T(x,z)-1}]. \tag{11}$$

Since $x \geq e \log x$, for all $x > 0$, we have $I^{KL} \leq I^{DV}$ for any fixed T, in this sense, the DV bound for MI is tighter, so in practice, people often use the one derived from the DV representation.

Another function $f(u)$ that is often used is $-(u+1) \log \frac{1+u}{2} + u \log u$, and the f-divergence now becomes JS divergence. The conjugate for f in this case is $f^*(t) = -\log(2 - e^t)$. Because the domain of the conjugate is $t < \log(2)$, we design $T_\theta(\cdot)$ so that it is a composition of two functions $g_f(V_\theta(\cdot))$. Following Ref. [11], the function V_θ does not have any range constraint, while $g_f : \mathbb{R} \to \text{dom}_{f^*}$ is an output activation function to transform any outputs from V_θ to the domain of the conjugate f^*. Here, we use the activation function $g_f(v) := \log(2) - \log(1 + e^{-v})$ as suggested in Ref. [11]. We will skip the detailed calculation here, but one can use Theorem 2 and perform calculation and paramatrization similar to the lower bounds' derivations before to get:

$$\begin{aligned} I^{JSD}(X; Z) &\propto \mathbb{E}_{p(x,z)}[-sp(-T(x, z))] - \mathbb{E}_{p(x)}\mathbb{E}_{p(z)}[sp(T(x, z))] \\ &\propto \mathbb{E}_{p(x,z)}[\log \sigma(T(x, z))] + \mathbb{E}_{p(x)}\mathbb{E}_{p(z)}[\log \sigma(-T(x, z))], \end{aligned} \tag{12}$$

where $sp(t) := \log(1 + e^t)$ and $\sigma(t) = \frac{1}{1+e^{-t}}$.

The relationship between the value of the lower bound in Eq. (12) and MI remains unclear for arbitrary distributions, but the authors of [8] shows that the value of Eq. (12) has an approximately monotonic relationship with MI for some specific distributions.

4 Discussion and Future Directions

In the previous sections, we discussed the idea of unsupervised learning by maximizing the MI between the inputs and the representations. Because direct calculation of the MI is difficult, we instead used the lower bounds and introduced how they are derived. We also extended the KL divergence definition of the MI to other f-divergences. Even though the lower bounds derived based on the dual representation of f-divergence

is not necessary lower bounds for the MI anymore, these lower bounds still work effectively in practice. Hence it would be very interesting to understand the relationship between the lower bound of MI and the lower bounds obtained from other f-divergence.

By comparing Eq. (8) and Eq. (12), it is clear that the key difference between the DV, JSD formulations is whether an expectation over the marginals $p(x)$ or $p(z)$ is inside or outside of the logarithm. Part of my research in the future will be on understanding how does this small difference affect the algorithm's learning process and why the lower bounds obtained from f-divergences are also successful in practice.

References

[1] Elizabeth A Ware and Amy E Booth. Form follows function: Learning about function helps children learn about shape. *Cognitive Development*, 25(2):124–137, 2010.

[2] David H Rakison and Yevdokiya Yermolayeva. Infant categorization. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(6):894–905, 2010.

[3] Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging*, 16(2):187–198, 1997.

[4] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

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

[6] Atul J Butte and Isaac S Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Biocomputing 2000*, pages 418–429. World Scientific, 1999.

[7] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning*, pages 531–540, 2018.

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

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

[10] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2012.

[11] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pages 271–279, 2016.

Appendix

Proof of Theorem 1

Proof. A simple proof goes as follows. For a given function T, consider the Gibbs distribution defined by $g(x) = \frac{1}{Z} e^T q(x)$, where $Z = \mathbb{E}_{q(x)}[e^T]$. By construction,

$$\mathbb{E}_{p(x)}[T] - \log Z = \mathbb{E}_{p(x)}\left[\log \frac{g(x)}{q(x)}\right] \tag{13}$$

Let Δ be the gap,

$$\Delta := D_{KL}(p \,||\, q) - \left(\mathbb{E}_{p(x)}[T] - \log(\mathbb{E}_{q(x)}[e^T])\right) \tag{14}$$

Using Eq. (13), we can write Δ as a KL-divergence:

$$\Delta = \mathbb{E}_{p(x)}\left[\log\frac{p(x)}{q(x)} - \log\frac{g(x)}{q(x)}\right] = \mathbb{E}_{p(x)}\log\frac{p(x)}{g(x)} = D_{KL}(p \,||\, g) \tag{15}$$

The positivity of the KL-divergence gives $\Delta \geq 0$. We have thus shown that for any T,

$$D_{KL}(p \,||\, q) \geq \mathbb{E}_{p(x)}[T] - \log(\mathbb{E}_{q(x)}[e^T]) \tag{16}$$

and the inequality is preserved upon taking the supremum over the right-hand side. Finally, the identity (15) also shows that this bound is *tight* whenever $p(x) = q(x)$, $\forall x$, namely for optimal functions T^* taking the form $T^* = \log\frac{p(x)}{q(x)} + C$ for some constant $C \in \mathbb{R}$. □

Proof of Theorem 2

Proof. Since f is convex and lower semi-continuous, Fenchel convex duality [10] guarantees that we can write f in terms of its conjugate dual as $f(u) = \sup_{t\in \mathrm{dom} f^*}\{tu - f^*(t)\}$. Consequently, we have

$$\begin{aligned} D_f(\mathbb{P}||\mathbb{Q}) &= \int_\Omega q(x) \sup_{t\in\mathrm{dom}f^*}\left\{t\frac{p(x)}{q(x)} - f^*(t)\right\}dx && (17)\\ &\geq \sup_{T\in\mathcal{T}}\left(\int_\Omega p(x)T(x)dx - \int_\Omega q(x)f^*(T(x))dx\right) && (18)\\ &= \sup_{T\in\mathcal{T}}\left[\mathbb{E}_{p(x)}T(x) - \mathbb{E}_{q(x)}f^*(T(x))\right], && (19)\end{aligned}$$

where $\mathcal{T}$ is an arbitrary class of functions $T : \Omega \to \mathrm{dom} f^*$. The above derivation yields a lower bound for two reasons: first, because of Jensen's inequality when swapping the integration and supremum operations. Second, the class of functions $\mathcal{T}$ may contain only a subset of all possible functions. □