
Connecting Jensen–Shannon and Kullback-Leibler Divergences: A New Bound for Representation Learning

Reuben Dorent
Inria
reuben.dorent@inria.fr

Polina Golland
MIT
polina@csail.mit.edu

William Wells III
Harvard, MIT
sw@bwh.harvard.edu

Abstract

Mutual Information (MI) is a fundamental measure of statistical dependence widely used in representation learning. While direct optimization of MI via its definition as a Kullback-Leibler divergence (KLD) is often intractable, many recent methods have instead maximized alternative dependence measures, most notably, the Jensen-Shannon divergence (JSD) between joint and product of marginal distributions via discriminative losses. However, the connection between these surrogate objectives and MI remains poorly understood. In this work, we bridge this gap by deriving a new, tight, and tractable lower bound on KLD as a function of JSD in the general case. By specializing this bound to joint and marginal distributions, we demonstrate that maximizing the JSD-based information increases a guaranteed lower bound on mutual information. Furthermore, we revisit the practical implementation of JSD-based objectives and observe that minimizing the cross-entropy loss of a binary classifier trained to distinguish joint from marginal pairs recovers a known variational lower bound on the JSD. Extensive experiments demonstrate that our lower bound is tight when applied to MI estimation. We compared our lower bound to state-of-the-art neural estimators of variational lower bound across a range of established reference scenarios. Our lower bound estimator consistently provides a stable, low-variance estimate of a tight lower bound on MI. We also demonstrate its practical usefulness in the context of the Information Bottleneck framework. Taken together, our results provide new theoretical justifications and strong empirical evidence for using discriminative learning in MI-based representation learning.

1 Introduction

Estimating and optimizing Mutual Information (MI) is central to many representation learning frameworks [4, 18, 8]. The MI between two random variables U and V is defined as the Kullback-Leibler Divergence (KLD) between their joint distribution and the product of their marginals,

$$I[U; V] \doteq D_{\text{KL}}[p_{UV} \parallel p_U \otimes p_V]. \quad (1)$$

Mutual information serves as a natural measure of dependence between U and V . In representation learning, a common goal is to learn compact yet informative representations by maximizing the MI between encoded variables.

A popular class of methods for estimating MI involves learning a discriminator that distinguishes between samples from the joint distribution and those from the product of marginals. However, because the MI objective depends on a discriminator trained through a separate optimization problem, maximizing MI leads to a bilevel optimization problem, which is intractable in practice. To bypass this challenge, tractable Variational Lower Bounds (VLBs) [3, 16, 20, 24] could be used, such as the

Donsker–Varadhan bound [3], to optimize a lower bound of the MI using data-estimated expectations. Indeed, KLD belongs to the family of f -divergences, which have standard VLBs. However, VLBs are often unstable and underperforming in the context of representation learning [8, 18, 13].

In practice, many successful methods optimize alternative f -divergences, such as the Jensen-Shannon divergence (JSD) [8] between the joint p_{UV} and the product of marginals $p_U \otimes p_V$, without knowing about the relationship between KLD and JSD. We refer to this measure as (I_{JS}) ,

$$I_{JS}[U; V] \doteq D_{JS}[p_{UV} \parallel p_U \otimes p_V] . \quad (2)$$

While optimizing JSD has been shown to reach competitive results, the reason why optimizing a JSD-based dependence measure would increase standard MI, defined via KLD, remains unclear.

In this work, we bridge this gap by showing that maximizing JSD is equivalent to maximizing a lower bound on MI. Our main contribution is a new and optimal lower bound on KLD as a function of JSD,

$$\Xi(D_{JS}[p \parallel q]) \leq D_{KL}[p \parallel q] \quad (3)$$

where $\Xi(\cdot)$ is a strictly increasing function, analytically described below, that resembles a Logit function. In particular, this result shows that maximizing I_{JS} also increases a guaranteed lower bound on mutual information.

Furthermore, we revisit the practical optimization of I_{JS} . We show that the f -divergence VLB for JSD is equivalent to optimizing the cross-entropy (CE) loss of a specific discriminator. This discriminator is trained to distinguish between true pairs of data and false pairs generated by scrambling the true pairs, a connection originally noted in the GAN paper [6]:

$$\log 2 - \mathcal{L}_{CE} \leq D_{JS}[p \parallel q] . \quad (4)$$

While this bound is equivalent to the standard f -divergence VLB for JSD, we provide a new information-theoretic derivation that highlights the discriminator perspective. Beyond offering a more intuitive interpretation, this formulation also makes explicit the gap due to suboptimal discriminators.

Together, these two results imply that minimizing the CE loss of a discriminator trained to separate joint from marginal samples leads to an increase in a lower bound on the mutual information:

$$\Xi(\log 2 - \mathcal{L}_{CE}) \leq \Xi(I_{JS}[U; V]) \leq I[U; V] . \quad (5)$$

This work therefore provides a principled theoretical justification for existing representation learning approaches that maximize discriminative JSD-based losses to increase mutual information, and it introduces a theoretically grounded alternative to commonly used variational lower bounds.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 introduces the joint range of f -divergences. Section 4 presents our new lower bound relating JSD and KLD, along with a bound connecting cross-entropy loss and JSD. Section 5 provides empirical evidence of the bound’s tightness and its utility for representation learning tasks. Section 6 concludes this work.

2 Related work

Estimating Mutual Information: MI is a core concept in information theory, but its estimation remains notoriously difficult, particularly in high-dimensional or continuous settings with limited data. Traditional estimators can be broadly categorized as parametric or non-parametric. Parametric methods rely on strong distributional assumptions (e.g., Gaussianity), which often do not hold in practice. Non-parametric approaches, such as histogram binning, k -nearest neighbors [10], and kernel density estimation [15, 11], are assumption-free but suffer from poor scalability with respect to data dimensionality and sample size, limiting their applicability in modern machine learning contexts.

To overcome these limitations, recent work has introduced variational lower bounds (VLBs) on MI that leverage the expressiveness of deep neural networks. A key example is MINE [3], which leverages the Donsker-Varadhan (DV) representation of the KLD to train a neural network to maximize a lower bound on MI. A related estimator, NWJ [16], arises from the dual representation of f -divergences and shares similar objectives. However, theoretical work by McAllester & Stratos [14] shows that such estimators are fundamentally upper-bounded by $\log(n)$, where n is the number of data samples. Moreover, both MINE and NWJ exhibit high variance, in part due to unstable estimation of the

partition function. To tackle this, Poole *et al.* [20] introduce hybrid VLB methods that generalize existing methods to trade off bias and variance. SMILE [24] further improves upon MINE by clipping the density ratio, which helps reducing the estimator variance. It also introduces a consistency check that exposes weaknesses in existing VLB estimators, especially in high MI regimes, and highlights that optimizing VLBs does not necessarily imply an accurate estimation of a lower bound of MI.

Another lower bound approach is InfoNCE, introduced via contrastive predictive coding (CPC) [18]. InfoNCE considers MI estimation as a multi-way classification problem, producing stable and low-variance estimates. However, this estimator is biased and upper-bounded by $\log(b)$, where b is the batch size, which limits its ability to capture high MI values.

Beyond VLB methods, another class of estimators decouples the estimation and optimization stages. NJEE [22] extends theoretical work by McAllester & Stratos [14] and estimates MI through entropy subtraction using neural density models. In contrast, DEMI [13] and PCM [26] reformulate MI estimation as a binary classification task: a model is trained to distinguish paired from unpaired samples, and discriminator outputs are then mapped to MI estimates. Interestingly, SMILE [24] can also be interpreted within this discriminator-based framework. As pointed out in [12], its implementation¹ relies on maximizing an estimator of the JSD [17], with MI subsequently estimated from the outputs of a clipped discriminator. Overall, by disentangling optimization and estimation using a two-step procedure, these methods alleviate some instability issues of VLBs. More recently, Letizia *et al.* [12] extended this setup by incorporating general f -divergences and empirically observed that the JSD (referred as GAN-DIME) provides the best tradeoff between bias and variance, outperforming alternatives based on KLD or Hellinger distance with a relatively low-variance MI estimation. While these two-step estimators are currently among the most accurate MI estimates in practice, they require retraining whenever the underlying joint distribution changes. This limitation is particularly restrictive in representation learning, where the latent variables evolve during optimization. As a result, these methods cannot be reused as plug-in MI estimators during training. In contrast, our bound can be directly optimized via JSD, using a single discriminator, enabling stable, end-to-end training.

Maximizing Mutual Information in representation learning: Several recent methods use MI maximization as a principle for representation learning via Information Bottleneck [25]. VLB-based methods such as MINE and NWJ have also been applied to representation learning, although their practical limitations, such as high variance, make them less attractive compared to contrastive approaches [8, 18]. Alternatively, contrastive predictive coding (CPC) [18] and its InfoNCE objective have shown strong empirical results across domains such as audio, vision, and NLP. However, since the estimation is upper-bounded by $\log(b)$, where b is the batch size, it requires a large batch size, which may be impractical computationally.

Another popular approach is Deep InfoMax from Hjelm *et al.* [8], which shifts the focus from strictly maximizing MI to maximizing other measures of statistical dependence, particularly the JSD-based objective I_{JS} . Unlike KLD-based bounds used in MINE or NWJ, the JSD is symmetric and bounded, often resulting in more stable optimization, and, unlike CPC and MINE, does not require a large batch size. Deep InfoMax leverages a JSD-based MI estimator introduced by Nowozin *et al.* [17]. To support the use of JSD-based optimization objective, Hjelm *et al.* [8] empirically show that for randomly sampled discrete sparse distributions, I_{JS} correlates well with true MI, suggesting it can act as a meaningful proxy. In this work, we go further and provide a theoretical justification: we show that maximizing I_{JS} provably increases a lower bound on the true mutual information.

Jensen-Shannon divergence in GANs: The link between JSD and discriminator training has its roots in Generative Adversarial Networks (GANs), where Goodfellow *et al.* [6] showed that, under an optimal discriminator, the GAN objective minimizes an affine transformation of the JSD between the real and generated data distributions. This was later generalized in the f -GAN framework [17], where GAN training corresponds to a variational lower bound on a general f -divergence, estimated via a learned discriminator. Chen *et al.* [4] proposed InfoGAN, which incorporates MI maximization within the GAN framework using a variational approximation originally introduced by Barber & Agakov [2]. Rather than estimating the MI between joint and marginal distributions directly, InfoGAN targets the MI between a subset of latent variables and the generator’s output, approximated using an auxiliary network and a variational lower bound from Barber & Agakov [2]. While InfoGAN brings information-theoretic principles into adversarial training, it does so in a generative setting with a specific structural assumption on the latent space.

¹<https://github.com/ermongroup/smile-mi-estimator>

3 Joint range of f -divergences

To establish a novel lower bound on mutual information, we build on the concept of the *joint range* of f -divergences, which we summarize here. While this concept has been covered in classical information theory [27, 7], to our knowledge, it has not been utilized in the context of representation learning or mutual information estimation.

Let p and q be two probability distributions defined on the measurable space $(\mathcal{X}, \mathcal{F})$. For any convex function $f : (0, \infty) \rightarrow \mathbb{R}$ such that $f(1) = 0$, the f -divergence between p and q is defined as:

$$D_f(p||q) \doteq \mathbb{E}_q \left[f \left(\frac{dp}{dq} \right) \right]. \quad (6)$$

Common examples include the Kullback-Leibler divergence, total variation distance, and Jensen-Shannon divergence, each corresponding to a different choice of f .

Given two such convex functions f and g , the joint range of the pair (D_f, D_g) is the set of all achievable values:

$$\mathcal{R}_{f,g} = \{(D_f(p||q), D_g(p||q)) : p, q \in \mathcal{P}\}, \quad (7)$$

where \mathcal{P} denotes the set of all probability distributions on some measurable space.

In addition, the joint range over all k -dimensional categorical distributions $[k]$ is defined as:

$$\mathcal{R}_{k;f,g} = \{(D_f(p||q), D_g(p||q)) : p, q \text{ are probability measures over } [k]\}. \quad (8)$$

The joint range $\mathcal{R}_{f,g}$ has several useful properties; it is convex, and the lower envelope of the joint range is the optimal lower bound relating the two divergences.

Remarkably, to characterize the entire joint range $\mathcal{R}_{f,g}$, it is sufficient to restrict to binary distributions, i.e. Bernoulli distributions [7, 27].

Theorem 3.1 (Harremoës-Vajda, 2011 [7]).

$$\mathcal{R}_{f,g} = \text{co}(\mathcal{R}_{2;f,g}) \quad (9)$$

where co denotes the convex hull.

4 Maximizing discriminator cross-entropy increases a lower bound on Mutual Information

In this section, we use f -divergence properties, including the previously introduced joint range, to demonstrate that training a discriminator to differentiate samples from the joint and product of the marginal distributions leads to maximizing a lower bound of the Mutual Information.

4.1 General lower bound on Kullback–Leibler divergence by Jensen–Shannon divergence

We introduce here a new optimal lower bound on the KLD in terms of the JSD, obtained by analyzing the joint range of the two divergences.

The KLD (D_{KL}) and JSD (D_{JS}) between two probability distributions p and q are defined as

$$D_{\text{KL}}[p || q] \doteq \mathbb{E}_p \left[\log \frac{p}{q} \right], \quad D_{\text{JS}}[p || q] \doteq \frac{1}{2} D_{\text{KL}}[p || m] + \frac{1}{2} D_{\text{KL}}[q || m], \quad (10)$$

where $m \doteq \frac{1}{2}(p + q)$ is a mixture distribution of p and q .

To derive the tightest lower bound of KLD in terms of JSD, we analyze the joint range of the two divergences, i.e., the set of all possible pairs $(x, y) \in [0, \log 2) \times \mathbb{R}^+$ such that:

$$x = D_{\text{JS}}[p || q] \quad \text{and} \quad y = D_{\text{KL}}[p || q] \quad (11)$$

for *all* pairs of distributions p, q that are compatible with the divergences.

Following Theorem 3.1 by Harremoës & Vajda [7], this joint range can be fully characterized by restricting p, q to be pairs of Bernoulli distributions compatible with JSD and KLD,

$$x = D_{\text{JS}}[B(\mu) || B(\nu)] \quad \text{and} \quad y = D_{\text{KL}}[B(\mu) || B(\nu)], \quad (12)$$

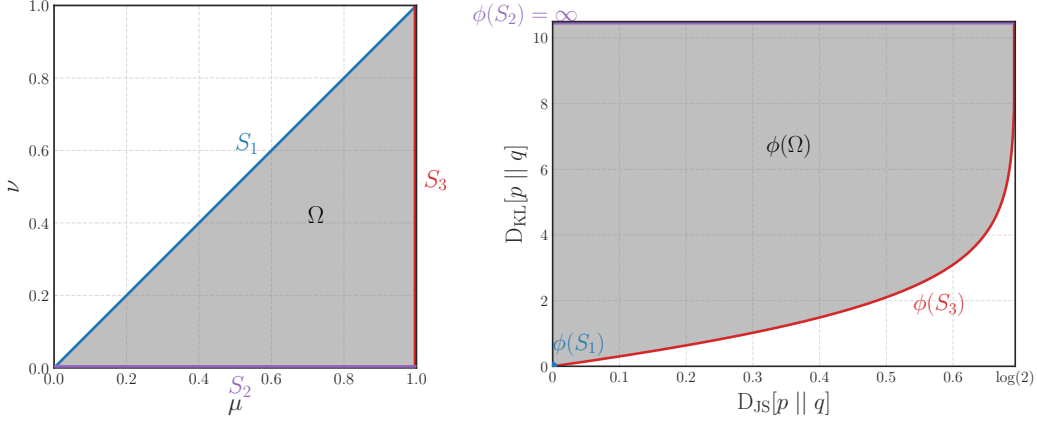


Figure 1: Left: Parameter space of two Bernoulli distributions. Right: Image (in gray) of the mapping ϕ , showing the joint range of the Jensen–Shannon and the Kullback–Leibler divergences. The red curve is our new lower bound.

for $(\mu, \nu) \in [0, 1] \times (0, 1) \cup \{(0, 0), (1, 1)\}$, where $B(\theta)$ is a Bernoulli distribution with parameter θ .

We are therefore interested in the image of the mapping ϕ defined as:

$$\phi : (\mu, \nu) \mapsto (D_{JS}[B(\mu) || B(\nu)], D_{KL}[B(\mu) || B(\nu)]) \quad (13)$$

from the unit square $[0, 1] \times (0, 1) \cup \{(0, 0), (1, 1)\}$ into \mathbb{R}^2 .

Exploiting the symmetry of the divergences:

$$D_{KL}[B(\mu) || B(\nu)] = D_{KL}[B(1 - \mu) || B(1 - \nu)], \quad (14)$$

$$D_{JS}[B(\mu) || B(\nu)] = D_{JS}[B(1 - \mu) || B(1 - \nu)], \quad (15)$$

we restrict our analysis to the image $\phi(\Omega)$ of the lower triangle $\Omega = \{(\mu, \nu) \in [0, 1] \times (0, 1) \mid \nu \leq \mu\}$.

Let us analyze the behavior of ϕ on the three edges of the triangle as shown in Fig. 1.

1. **Diagonal** ($S_1 : \mu = \nu$):

$$D_{KL}[B(\mu) || B(\mu)] = 0, \quad D_{JS}[B(\mu) || B(\mu)] = 0.$$

Thus, all points on the diagonal map to the origin: $\phi(\mu, \mu) = (0, 0)$.

2. **Lower Edge** ($S_2 : \nu \rightarrow 0$):

$$D_{KL}[B(\mu) || B(0)] \rightarrow +\infty \text{ for any } \mu > 0.$$

The JSD remains finite, so points on this edge map to $(0, \log 2) \times \{\infty\}$.

3. **Right edge** ($S_3 : \mu = 1$): For this segment, we can obtain:

$$D_{KL}[B(1) || B(\nu)] = -\log \nu, \quad (16)$$

$$D_{JS}[B(1) || B(\nu)] = \log 2 - \frac{1}{2} \left(\log(1 + \nu) - \nu \log \left(\frac{\nu}{1 + \nu} \right) \right). \quad (17)$$

As $\nu \rightarrow 0$, $D_{KL}[B(1) || B(\nu)] \rightarrow +\infty$, while $D_{JS}[B(1) || B(\nu)]$ increases to its maximum, i.e. $\log 2$.

We now present the main result, which shows that maximizing the JSD increases a lower bound on the KLD and that this lower bound corresponds to the image of S_3 .

Theorem 4.1. *The optimal lower bound (i.e., equality is achieved for a family of distributions) on Kullback-Leibler divergence established by the Jensen-Shannon divergence is:*

$$\Xi(D_{JS}[p || q]) \leq D_{KL}[p || q] \quad (18)$$

where $\Xi : [0, \log 2] \mapsto \mathbb{R}^+$ is a strictly increasing function defined implicitly by its inverse,

$$\Xi^{-1} : y \mapsto D_{JS}[B(1) || B(\exp(-y))], \quad (19)$$

where $B(\exp(-y))$ is a Bernoulli distribution with parameter $\exp(-y) \in [0, 1/2]$. Ξ^{-1} has a closed-form expression related to Eq. 17 and defined in Appendix (41).

In particular, this result holds when p and q are respectively the joint and the product of the marginals, leading to a novel lower bound in information theory between I_{JS} and MI. Theorem 4.1’s proof relies on a conjectured functional property, supported by extensive numerical evidence (see Appendix B.1).

Approximation: Although Ξ does not have a closed-form expression, it can be approximated point-wise using numerical solvers (example in C.2). Alternatively, we introduce a smooth differentiable approximation using the Logit function, $\mathbb{L}(x) \doteq \log \frac{x}{1-x}$ (details in Appendix C.3):

$$\Xi(x) \approx 1.15 * \mathbb{L} \left(.5 \left(\frac{x}{\log 2} + 1.0 \right) \right) . \quad (20)$$

4.2 f -divergence lower bound on JSD from discriminator cross-entropy

Having established that the JS-based information (I_{JS}) provides a lower bound on the MI, we now show that, in turn, the I_{JS} can be lower bounded using the expected cross-entropy (CE) loss of a discriminator trained to distinguish between dependent and independent pairs. Therefore, optimizing a binary cross-entropy classifier effectively increases a lower bound on MI via the inequality chain.

We adopt a standard binary classification setup to estimate the JSD between the joint distribution p_{UV} and the product of its marginals $p_U \otimes p_V$. Consider the mixture model:

$$(U, V) \mid Z = 1 \sim p_{UV}, \quad (U, V) \mid Z = 0 \sim p_U \otimes p_V, \quad \text{with } Z \sim B(1/2). \quad (21)$$

This defines a joint distribution over (u, v, z) , denoted $\tilde{p}(u, v, z)$, with marginal over (u, v) :

$$\tilde{p}(u, v) = m(u, v) \doteq \frac{1}{2}p_{UV}(u, v) + \frac{1}{2}p_U(u)p_V(v). \quad (22)$$

Since the JSD is a f -divergence, it admits a variational representation [17, 27], expressible as:

$$I_{JS}[U; V] = D_{JS}[p_{UV} \parallel p_U \otimes p_V] = \frac{1}{2} \max_t \left[\mathbb{E}_{p_{UV}}[t] - \mathbb{E}_{p_U \otimes p_V}[-\log(2 - \exp(t))] \right] \quad (23)$$

where $t : \mathcal{U} \times \mathcal{V} \rightarrow (-\infty, \log 2)$ is a variational function.

Now, define the following reparameterization,

$$t(u, v) \doteq \log(2q_\theta(z = 1 \mid u, v)), \quad (24)$$

where $q_\theta(z = 1 \mid u, v)$ is meant to approximate $\tilde{p}(z = 1 \mid u, v)$, the optimal classifier.

Substituting Eq. (24) into Eq. (23) (details in B.2), we obtain the following lower bound:

$$I_{JS}[U; V] = \frac{1}{2} \max_\theta \left[\log 4 + \mathbb{E}_{p_{UV}}[\log q_\theta(z = 1 \mid u, v)] + \mathbb{E}_{p_U \otimes p_V}[\log(1 - q_\theta(z = 1 \mid u, v))] \right] \quad (25)$$

Thus, we obtain

$$I_{JS}[U; V] \geq \log 2 - \min_\theta \mathcal{L}_{CE}(\theta), \quad (26)$$

where $\mathcal{L}_{CE}(\theta)$ is the expected binary cross-entropy loss between q_θ and the true posterior.

This bound is tight in the non-parametric limit (infinite data and model capacity). An equivalent result was shown in the context of GAN discriminators [6] and can also be derived from [23] via the reparameterization $d(u, v) \doteq \mathbb{L}(q_\theta(z=1 \mid u, v))$. In Appendix B.3, we present an alternative proof that quantifies the gap between the JSD lower bound and the one obtained from cross-entropy.

Combining (26) and (18), we obtain our new VLB $I_{CE}(\theta)$ on MI as a function of a discriminator CE:

$$I_{CE}(\theta) \doteq \Xi(\log 2 - \mathcal{L}_{CE}(\theta)) \leq \Xi(I_{JS}[U; V]) \leq I[U; V]. \quad (27)$$

4.3 Estimating Mutual Information via discriminators

While we introduced a new VLB on MI (Eq. (27)) that can be estimated from data, we now show that it also enables the construction of an estimator of the true MI, as detailed below. Beginning with the joint model of Eq. (21), we derive $\tilde{p}(z \mid u, v)$ and then $\frac{\tilde{p}(z=1 \mid u, v)}{\tilde{p}(z=0 \mid u, v)}$. From this, it is easy to show that

$$I[U; V] = \mathbb{E}_{p_{UV}}[\mathbb{L}(\tilde{p}(z = 1 \mid u, v))] , \quad (28)$$

where $\mathbb{L}(\cdot)$ is the Logit transform. Details can be found in Appendix B.4.

This result suggests a two-step estimation procedure: 1) train a discriminator q_θ to distinguish between joint and marginal samples; 2) estimate the MI by plugging the approximate q_θ into Eq. (28), and approximate expectations by data averages. This two-step strategy is equivalent to GAN-DIME [12].

5 Experiments

In this section, we evaluate the tightness of our proposed MI lower bound, derived from the JSD, and compare it to widely used variational MI estimators. Our evaluation is three-fold. First, we provide an empirical analysis of the lower bound in the discrete setting, where both MI and JSD can be computed exactly. This allows us to assess the tightness of the bound independently of neural approximation errors. Second, we assess the quality of our bound as a variational objective function in neural estimation tasks, using synthetic benchmarks introduced in prior work [20, 24, 12]. These include Gaussian and non-Gaussian settings designed to challenge both the bias and variance of MI estimators across regimes of increasing complexity and mutual dependence. Finally, we demonstrate the practical usefulness of optimizing JSD as a proxy of MI in the context of the Information Bottleneck (IB) framework.

5.1 Tightness of the JSD lower bound in the discrete setting

We begin by analyzing the proposed MI lower bound in a controlled discrete setup where both the true MI and the JSD between the joint distribution and the product of marginals are tractable. This allows us to assess whether our general optimal lower bound, defined for arbitrary distributions, remains tight when specialized to the MI setting, i.e. restricted to joint and their product of marginals.

We construct a parameterized family of joint distributions $P_{UV}^{(\alpha)} \in \mathbb{R}^{k \times k}$ that interpolate between the independent and perfectly dependent cases. The marginal distributions are uniform categorical $P_U = P_V = \frac{1}{k} \mathbf{1}_k$, where k is the number of categories. The joint $P_{UV}^{(\alpha)}$ distribution is defined as:

$$P_{UV}^{(\alpha)} = (1 - \alpha)P_U \otimes P_V + \alpha \text{diag}(P_U), \quad (29)$$

where $\alpha \in [0, 1]$ controls the strength of dependence. For each value of α , we compute the ground-truth MI $D_{\text{KL}}[P_{UV}^{(\alpha)} \parallel P_U \otimes P_V]$ and the JSD $D_{\text{JS}}[P_{UV}^{(\alpha)} \parallel P_U \otimes P_V]$. We then compare $I[U; V]$ with our bound $\Xi(I_{\text{JS}}[U; V])$, where Ξ is the transformation defined in (19).

Figure 2 shows the resulting trajectories as α varies from 0 (independent variables) to 1 (maximally dependent variables) across set sizes $k \in \{2, 3, \dots, 500\}$. For any given JSD value x , we observe that there exists a discrete distribution with JS-based information $I_{\text{JS}} \approx x$ whose true mutual information nearly coincides with the lower bound $\Xi(x)$. This empirically demonstrates the tightness of our proposed bound. This result is particularly remarkable as it suggests that the optimal lower bound between JSD and KLD introduced in Theorem 4.1, originally derived for arbitrary pairs of distributions, remains tight when restricted to mutual information estimation, i.e., when comparing joint distributions to the product of their marginals. In Appendix D, we further show that there exists an *infinite family* of discrete joint–marginal distribution pairs that lie exactly on our bound, thereby demonstrating that the proposed lower bound remains tight even under this restriction. This implies that our JSD-based lower bound offers a tight proxy when maximizing mutual information.

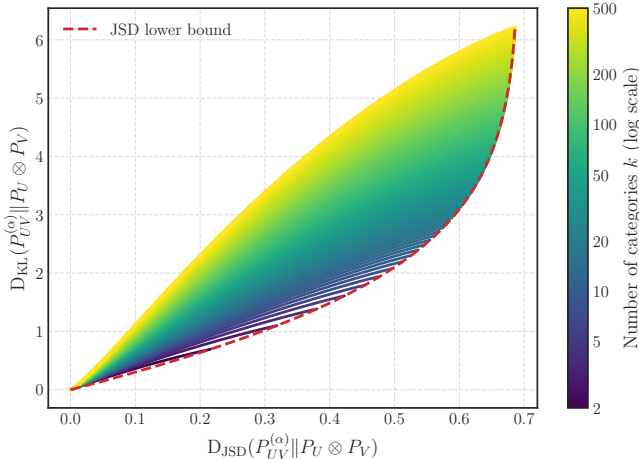


Figure 2: Mutual information and its JSD-based lower bound $\Xi(I_{\text{JS}})$ for a parameterized family of discrete joint distributions with known MI and JSD, varying in dependence strength (α) and number of categories (k). The presence of MIs near the lower bound across settings empirically demonstrates the tightness of our JSD-based estimate.

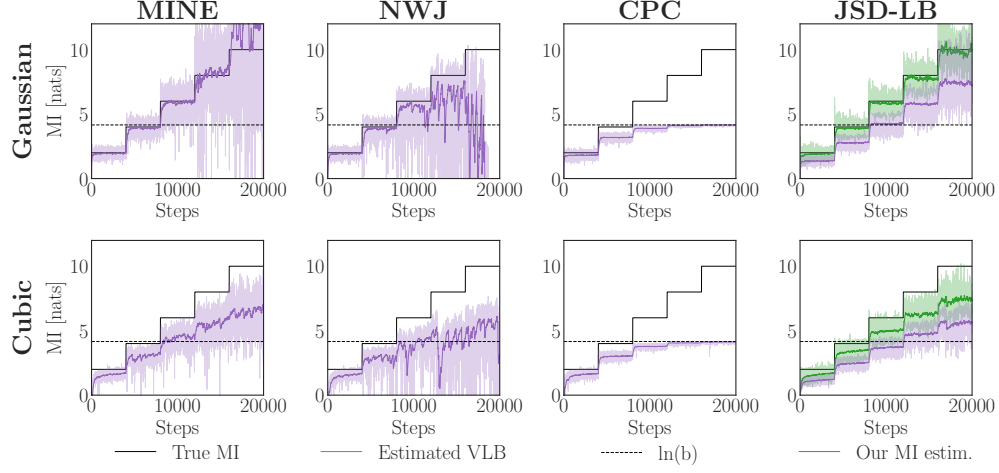


Figure 3: Staircase MI estimation comparison for $d = 5$ and batch size $b = 64$. The Gaussian case is reported in the top row, while the cubic case is shown in the bottom row.

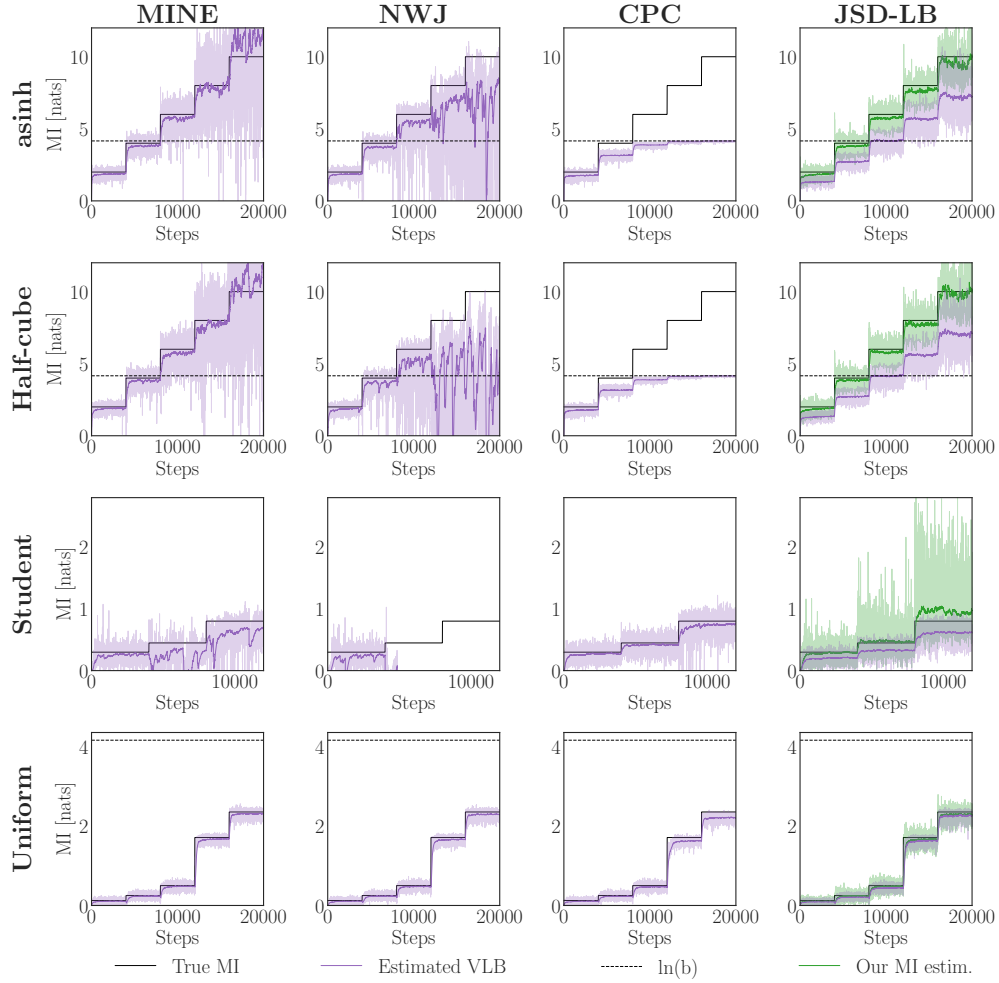


Figure 4: Staircase MI estimation comparison with batch size $b = 64$ for the asinh ($d = 5$), half-cube ($d = 5$), Student ($d = 5$), and uniform distributions.

5.2 Complex Gaussian and non-Gaussian distributions

We next evaluate our lower bound in continuous domains where MI is estimated from samples using neural networks. Our goal is to assess how tight the bound remains when used as a variational objective and how well it compares to other neural MI VLB under complex data distributions.

Setup and metrics We follow the protocol of Letizia *et al.* [12], who benchmark MI estimators in scenarios where the true MI is known analytically. For each estimator, we report the value of the lower bound objective and the ground truth MI. For our method, we additionally report the estimated MI obtained using the procedure described in Section 4.3 and equivalent to GAN-DIME [12].

All estimators are implemented using a fully connected discriminator with input dimension $2d$, two hidden layers of 256 ReLU units, and one scalar output. Training is performed for 4000 steps using the Adam optimizer and batch size $N = 64$, matching the architecture and hyperparameters from prior work for comparability. As the function Ξ is strictly increasing, maximizing Ξ ($\log 2 - \mathcal{L}_{CE}$) is equivalent to minimizing the cross-entropy loss \mathcal{L}_{CE} . Therefore, the approximation of Ξ is not used during optimization. Implementation details² are available in Appendix E.2.

We compare our approach against other differentiable variational lower bounds that, unlike two-step estimators, can be directly optimized as objective functions for mutual information maximization:

1. **MINE** [3], based on the Donsker-Varadhan dual representation of the KLD;
2. **NWJ** [16], based on another KLD dual representation, which is unbiased but less tight compared to MINE;
3. **CPC** [18], based on contrastive predictive coding with the InfoNCE lower bound;
4. **JSD-LB** (*ours*), based on I_{CE} from (27) with the estimated expected cross-entropy loss.

Gaussian benchmarks In the **Gaussian** setting, we define two random variables $U \sim \mathcal{N}(0, I_d)$ and $N \sim \mathcal{N}(0, I_d)$, sampled independently, and construct $V = \rho U + \sqrt{1 - \rho^2} N$, where ρ is the correlation coefficient. In the **cubic** case, we apply a nonlinear transformation $v \mapsto v^3$ to V . This transformation introduces non-Gaussian marginals while keeping the same mutual information. In both cases, the MI is known: $I[U; V] = -\frac{d}{2} \log(1 - \rho^2)$.

Figure 3 presents our results. In the **Gaussian** setting, our JSD-LB consistently provides a tight and stable lower bound on the true MI, outperforming MINE, NWJ, and CPC, especially in high-MI regimes. While MINE and NWJ exhibit high variance, JSD-LB and CPC remain more stable. In particular, MINE’s instability can lead to significant overestimation of its lower bound, sometimes even exceeding the true MI, whereas CPC remains limited by the contrastive bound $\log(b)$. In contrast, the estimated JSD-LB remains below the true MI, as expected for a valid lower bound, yet closely tracks it. Furthermore, our derived two-step MI estimator from (28), shown in green in Figures 3, 4 and equivalent to GAN-DIME [12], surpasses other variational lower bound estimators. In the **cubic** setting, JSD-LB provides a tight lower bound, while maintaining low-variance estimates compared to MINE and NWJ. Moreover, our two-step MI estimator achieves the most accurate MI estimation, outperforming all other VLBs.

Complex Gaussian and non-Gaussian benchmarks We also replicate the more challenging benchmark from Letizia *et al.* [12], following recommendations from Czyż *et al.* [5] using complex Gaussian transformations (**half-cube** and **asinh**, Fig 4) and non-Gaussian distributions (**Student** and **uniform** distributions, Fig 4). Across these settings (see details in Appendix E.2), our lower bound and its derived MI estimations remain competitive, with lower variance VLBs, confirming that the proposed JSD-based objective generalizes beyond Gaussian settings.

5.3 Variational Information Bottleneck Experiments

To assess the practical utility of our JSD-LB in representation learning, we evaluate it within the Information Bottleneck (IB) framework [1, 19], which aims to learn compact yet predictive representations T from inputs U for targets V , by maximizing $I[T; V] - I[T; U]$. We reproduce the

²<https://github.com/ReubenDo/JSDlowerbound>

Table 1: Generalization performance (%) on MNIST dataset. Performance is evaluated by the mean classification accuracy on the MNIST test set after training on the MNIST training set.

Method	VIB [1]	NIB [9]	squared-VIB [21]	squared-NIB [21]	DisenIB [19]	JSD-LB
Testing	97.6	97.2	96.2	93.3	98.2	98.8

Table 2: Adversarial robustness performance (%) on MNIST dataset.

	Method	VIB [1]	NIB [9]	squared-VIB [21]	squared-NIB [21]	DisenIB [19]	JSD-LB
Training	$\epsilon = 0.1$	74.1	75.2	42.1	61.3	94.3	99.5
	$\epsilon = 0.2$	19.1	21.8	8.7	24.1	81.5	96.0
	$\epsilon = 0.3$	3.5	3.2	5.9	9.3	68.4	91.4
Testing	$\epsilon = 0.1$	73.4	75.2	42.7	62.0	90.2	94.6
	$\epsilon = 0.2$	20.8	23.6	9.2	24.5	80.0	89.6
	$\epsilon = 0.3$	4.2	3.4	5.9	9.9	67.8	86.1

Table 3: Distinguishing in- and out-of-distribution test data for MNIST image classification (%). \uparrow (resp., \downarrow) indicates that a larger (resp., lower) value is better.

Method	VIB [1]	NIB [9]	squared-VIB [21]	squared-NIB [21]	DisenIB [19]	JSD-LB
TPR95 \downarrow	27.4	34.4	49.9	47.5	0.0	0.00
AUROC \uparrow	94.6	94.2	86.6	85.6	99.4	100.0
AUPR In \uparrow	94.8	95.2	83.5	83.3	99.6	99.9
AUPR Out \uparrow	93.7	91.8	83.2	83.3	98.9	99.9
Detection Err \downarrow	11.5	11.9	20.0	15.0	1.7	0.1

MNIST experiments of Pan *et al.* [19], where U denotes an image and V its label, using the same architecture and setup. In our setting (JSD-LB), $I[T; V]$ is optimized by maximizing the JSD via the cross-entropy loss. Following Pan *et al.* [19], we assess generalization, adversarial robustness, and out-of-distribution (OOD) detection. Robustness is evaluated by computing the average classification accuracy on adversarially perturbed inputs, where pixel-wise perturbation are obtained using a one-step gradient-based attack with perturbation magnitudes $\epsilon \in \{0.1, 0.2, 0.3\}$. OOD detection is tested on synthetic Gaussian noise samples, using standard OOD detection metrics including True Positive Rate at 95%(TPR95), Area Under the ROC Curve (AUROC), Area Under the Precision-Recall Curve (AUPR) and Detection Error.

Our results presented in Tables 1, 2, and 3 show that replacing the MI estimator with JSD-LB leads to state-of-the-art performance in terms of generalization, adversarial robustness, and out-of-distribution robustness. Specifically, it outperforms other Variational Information Bottleneck objectives [1, 9, 21] and the Disentangled Information Bottleneck approach [19]. This set of experiments shows that maximizing JSD via the cross-entropy loss has practical value for representation learning.

6 Conclusion

In this work, we addressed a fundamental gap in the theory of representation learning: the lack of formal guarantees linking Jensen–Shannon divergence (JSD), widely used in practice, to Kullback–Leibler divergence (KLD), the basis of MI. We derived a new and optimal lower bound on KLD as a function of JSD, and showed how this leads to a provable lower bound on MI when comparing joint and marginal distributions. Furthermore, we demonstrated that minimizing the cross-entropy loss of a binary classifier distinguishing between joint and marginal pairs increases a lower bound on the MI, unifying existing discriminator-based objectives in representation learning with an information-theoretic interpretation. Through extensive experiments in both discrete and continuous domains, we showed that our bound is tight and also performs competitively as a variational objective and as an estimator in practice. Moreover, we demonstrate its usefulness in representation learning within the Information Bottleneck (IB) framework, reaching state-of-the-art performance. Altogether, this work provides a rigorous theoretical foundation and a practical method for maximizing mutual information through discriminative representation learning.

Acknowledgment

We thank the anonymous reviewers for their insightful comments and helpful suggestions, which improved the quality of this paper. We would also like to thank Yury Polyanskiy for pointing us to the joint range. R.D. received a Marie Skłodowska-Curie grant No 101154248 (project: SafeREG). This work was performed using HPC resources from GENCI-IDRIS (Grant 2024-SafeREG, 2025-SafeREG). The research leading to these results has received funding from the program "Investissements d'avenir" ANR-10-IAIHU-06. W.W. is supported by the National Institutes of Health (P41EB028741 and R01EB032387).

References

- [1] Alexander A. Alemi and Ian Fischer and Joshua V. Dillon and Kevin Murphy. 2017. Deep Variational Information Bottleneck. *In: International Conference on Learning Representations*.
- [2] Barber, David, & Agakov, Felix. 2004. The IM algorithm: a variational approach to information maximization. *Advances in neural information processing systems*, **16**(320), 201.
- [3] Belghazi, Mohamed Ishmael, Baratin, Aristide, Rajeshwar, Sai, Ozair, Sherjil, Bengio, Yoshua, Courville, Aaron, & Hjelm, Devon. 2018. Mutual information neural estimation. *Pages 531–540 of: International conference on machine learning*. PMLR.
- [4] Chen, Xi, Duan, Yan, Houthoofd, Rein, Schulman, John, Sutskever, Ilya, & Abbeel, Pieter. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, **29**.
- [5] Czyż, Paweł, Grabowski, Frederic, Vogt, Julia, Beerenwinkel, Niko, & Marx, Alexander. 2023. Beyond normal: On the evaluation of mutual information estimators. *Advances in Neural Information Processing Systems*, **36**, 16957–16990.
- [6] Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, & Bengio, Yoshua. 2014. Generative adversarial nets. *Advances in neural information processing systems*, **27**.
- [7] Harremoës, Peter, & Vajda, Igor. 2011. On pairs of f -divergences and their joint range. *IEEE Transactions on Information Theory*, **57**(6), 3230–3235.
- [8] Hjelm, R Devon, Fedorov, Alex, Lavoie-Marchildon, Samuel, Grewal, Karan, Bachman, Phil, Trischler, Adam, & Bengio, Yoshua. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.
- [9] Kolchinsky, Artemy, Tracey, Brendan D, & Wolpert, David H. 2019. Nonlinear information bottleneck. *Entropy*, **21**(12), 1181.
- [10] Kraskov, Alexander, Stögbauer, Harald, & Grassberger, Peter. 2004. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, **69**(6), 066138.
- [11] Kwak, Nojun, & Choi, Chong-Ho. 2002. Input feature selection by mutual information based on Parzen window. *IEEE transactions on pattern analysis and machine intelligence*, **24**(12), 1667–1671.
- [12] Letizia, Nunzio Alexandro, Novello, Nicola, & Tonello, Andrea M. 2024. Mutual Information Estimation via f -Divergence and Data Derangements. *In: The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- [13] Liao, Ruizhi, Moyer, Daniel, Golland, Polina, & Wells, William M. 2020. DEMI: Discriminative Estimator of Mutual Information. *arXiv preprint arXiv:2010.01766*.
- [14] McAllester, David, & Stratos, Karl. 2020. Formal limitations on the measurement of mutual information. *Pages 875–884 of: International Conference on Artificial Intelligence and Statistics*. PMLR.

- [15] Moon, Young-II, Rajagopalan, Balaji, & Lall, Upmanu. 1995. Estimation of mutual information using kernel density estimators. *Physical Review E*, **52**(3), 2318.
- [16] Nguyen, XuanLong, Wainwright, Martin J, & Jordan, Michael I. 2010. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, **56**(11), 5847–5861.
- [17] Nowozin, Sebastian, Cseke, Botond, & Tomioka, Ryota. 2016. f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, **29**.
- [18] Oord, Aaron van den, Li, Yazhe, & Vinyals, Oriol. 2018. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748*.
- [19] Pan, Ziqi, Niu, Li, Zhang, Jianfu, & Zhang, Liqing. 2021. Disentangled information bottleneck. *Pages 9285–9293 of: Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35.
- [20] Poole, Ben, Ozair, Sherjil, Van Den Oord, Aaron, Alemi, Alex, & Tucker, George. 2019. On Variational Bounds of Mutual Information. *Pages 5171–5180 of: International conference on machine learning*. PMLR.
- [21] Rodríguez Gálvez, Borja, Thobaben, Ragnar, & Skoglund, Mikael. 2020. The convex information bottleneck lagrangian. *Entropy*, **22**(1), 98.
- [22] Shalev, Yuval, Painsky, Amichai, & Ben-Gal, Irad. 2022. Neural joint entropy estimation. *IEEE Transactions on Neural Networks and Learning Systems*, **35**(4), 5488–5500.
- [23] Shannon, Matt. 2020. Properties of f-divergences and f-GAN training. *arXiv preprint arXiv:2009.00757*.
- [24] Song, Jiaming, & Ermon, Stefano. 2019. Understanding the Limitations of Variational Mutual Information Estimators. *arXiv preprint arXiv:1910.06222*.
- [25] Tishby, Naftali, Pereira, Fernando C., & Bialek, William. 1999. The information bottleneck method. *Pages 368–377 of: Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*.
- [26] Tsai, Yao-Hung Hubert, Zhao, Han, Yamada, Makoto, Morency, Louis-Philippe, & Salakhutdinov, Russ R. 2020. Neural methods for point-wise dependency estimation. *Advances in Neural Information Processing Systems*, **33**, 62–72.
- [27] Yury Polyanskiy, Yihong Wu. 2024. *Information Theory From Coding to Learning*. Cambridge University Press. <https://people.lids.mit.edu/yp/homepage/data/itbook-export.pdf>.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The two main contributions of the paper are both included: a) a new, optimal and computationally tractable lower bound on Kullback-Leibler divergence as a function of Jensen-Shannon divergence; b) a proof that maximizing the cross-entropy loss of a discriminator leads to an increase in a lower bound on the mutual information.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: The manuscript acknowledges that the derived Ξ function does not admit a closed form and must be approximated numerically. Furthermore, we clearly state in the main manuscript that the proof of Theorem 4.1 relies on a conjectured functional property, supported by extensive numerical evidence

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: All the assumptions are included in the statements of Theorems, and Corollaries (that are cross-referenced). All the complete proofs are reported in Appendix B and C.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: A link to the code is available. In addition, the complete neural architectures are specified in Appendix E.2, together with the optimizer type, hyperparameters, and the hardware details. Finally, we report the other VLB objectives in Appendix A.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the code along with a README file detailing how to run the experiments. The first set of experiments does not require any external data, as all experiments are conducted using samples drawn from known probability distributions. For the second set, the widely used MNIST dataset is used.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: They are written in Appendix E.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: In line with prior work on mutual information estimators, our "staircase" plots display MI estimates over training steps. The shaded regions around the curves visually convey the variance of the estimators, providing an intuitive understanding of their stability and statistical behavior during optimization.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Details are given in Appendix E.2.5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: There are no potential harms caused by the research process and no potential harmful consequences.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Our work is foundational research. Mutual information estimation and Jensen-Shannon divergence are used for various machine learning algorithms and there are no direct paths to negative applications.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite all datasets used in our experiments, as well as the original papers from which we adapt code implementations (mainly *f*-DIME) of state-of-the-art mutual information estimators.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We fully disclose the code and provide a readme file to an anonymized repository, describing how to use it.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [No]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [No]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification: No large language models were used as part of the core methodology or technical contribution of this research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Appendix: Related MI estimators

In this section, we present some existing methods and their formulas for estimating MI. As mentioned in the “Related work” Section 2, these methods can be categorized as differentiable variational lower bounds and two-step (optimization then estimation) methods.

A.1 Other differentiable Variational Lower Bounds (VLBs)

In this subsection, we review some of the main Variational Lower Bounds (VLBs) on MI. These differentiable bounds can be used directly as minimization criteria in the context of representation learning.

MINE [3] The mutual information neural estimator (MINE) is a VLB on MI derived from the Donsker-Varadhan dual representation of the KLD [20] :

$$I[U; V] = \sup_{\theta} \left\{ I_{\text{MINE}}(\theta) \doteq \mathbb{E}_{(u,v) \sim p_{UV}(u,v)} [T_{\theta}(u, v)] - \log \left(\mathbb{E}_{(u,v) \sim p_U(u)p_V(v)} \left[e^{T_{\theta}(u,v)} \right] \right) \right\}, \quad (30)$$

where θ are the parameters of the functions $T_{\theta} : \mathcal{U} \times \mathcal{V} \mapsto \mathbb{R}$ which are defined using a deep neural network. A key limitation of MINE is that it considers the logarithm of the expectation, which leads to biased estimation when approximated using data samples.

NWJ [16] Another VLB based on the KLD dual representation:

$$I[U; V] = \sup_{\theta} \left\{ I_{\text{NWJ}}(\theta) \doteq \mathbb{E}_{(u,v) \sim p_{UV}(u,v)} [T_{\theta}(u, v)] - \mathbb{E}_{(u,v) \sim p_U(u)p_V(v)} \left[e^{T_{\theta}(u,v)-1} \right] \right\}. \quad (31)$$

This VLB is not as tight as MINE, i.e. $I_{\text{NWJ}}(\theta) \leq I_{\text{MINE}}(\theta) \leq I[U; V]$ [14], but its estimation is unbiased.

CPC [18] This VLB, based on contrastive predictive coding (CPC), is defined as

$$I_{\text{CPC}}(\theta) = \mathbb{E}_{(u,v) \sim p_{UV;b}(u,v)} \left[\frac{1}{N} \sum_{i=1}^b \log \left(\frac{e^{T_{\theta}(u_i, v_i)}}{\frac{1}{b} \sum_{j=1}^b e^{T_{\theta}(u_i, v_j)}} \right) \right], \quad (32)$$

where b is the batch size and $p_{UV;b}$ denotes the joint distribution of b i.i.d. samples drawn from p_{UV} . While CPC estimates have typically low variances, they are upper bounded by $\log b$, which introduces bias and restricts its ability to estimate high mutual information.

A.2 Two-step MI estimators

Alternatively, another class of estimators has been proposed. Unlike the previously presented VLB, their estimation is two-step. First, a neural network is trained to approximate a function that is later used for estimating the MI. These methods are, therefore, not as easily applied to representation learning, as that would involve solving bilevel (nested) optimization problems.

SMILE [24] Since MINE suffers from high-variance estimations, Song & Ermon [24] proposed to clip the outputs of the neural network T_{θ} , which estimates the log-density ratio, when estimating the partition function:

$$I_{\text{SMILE}}(\theta) \doteq \mathbb{E}_{(u,v) \sim p_{UV}(u,v)} [T_{\theta}(u, v)] - \log \left(\mathbb{E}_{(u,v) \sim p_U(u)p_V(v)} \left[\text{clip} \left(e^{T_{\theta}(u,v)}, e^{-\tau}, e^{\tau} \right) \right] \right) \quad (33)$$

where the parameter $\tau \geq 0$ is a hyperparameter and the clip function is defined as follows:

$$\text{clip}(v, l, u) = \max(\min(v, u), l). \quad (34)$$

The SMILE estimate I_{SMILE} is not a VLB but converges to the I_{MINE} VLB when $\tau \rightarrow \infty$.

In practice, the implementation of SMILE is significantly different from MINE. Indeed, Song & Ermon [24] proposed to train the neural network T_{θ} , by maximizing a Jensen-Shannon MI estimator from [17], as in the Deep InfoMax procedure [8]. SMILE is, therefore, a two-step estimator.

NJEE [22] The Neural Joint Entropy Estimator (NJEE) [22] estimates the MI between discrete random variables by modeling entropy estimation as a sequence of classification tasks. Let U_m denote the m -th component of a discrete random vector $U \in \mathbb{R}^d$, and define U^k as the subvector containing the first k components of U . Let V be another random variable. Given a batch of b samples, NJEE estimates the entropy of U by training a series of classifiers. Specifically, for each m , two neural networks $G_{\theta_m}(U_m | U^{m-1})$ and $G_{\theta_m}(U_m | V, U^{m-1})$ are trained to predict U_m respectively from its preceding components U^{m-1} , and from both V and U^{m-1} , using cross-entropy as the loss function. Once trained, the NJEE mutual information estimator is computed as:

$$I_{\text{NJEE}}(U; V) = \hat{H}_b(U_1) + \sum_{m=2}^d \text{CE}(G_{\theta_m}(U_m | U^{m-1})) - \sum_{m=1}^d \text{CE}(G_{\theta_m}(U_m | V, U^{m-1})), \quad (35)$$

where $\hat{H}_b(U_1)$ is the estimated marginal entropy of the first component. This two-stage procedure is designed for discrete random variables and does not directly extend to the continuous case.

f -DIME [12] is an f -divergence based two step method that decouples the optimization of a deep neural network T_θ and the estimation of the MI. Let f denote a convex lower semicontinuous function f that satisfies $f(1) = 0$. The Fenchel conjugate of f is denoted f^* . Specifically, the network \hat{T} is trained to maximize the following quantity:

$$\mathcal{J}_f(T) = \mathbb{E}_{(u,v) \sim p_{UV}(u,v)} [T(u, v) - f^*(T(x, \sigma(y)))] \quad (36)$$

and the MI is then estimated as:

$$I_{f\text{-DIME}} = \mathbb{E}_{(u,v) \sim p_{UV}(u,v)} \left[\log \left((f^*)'(\hat{T}(u, v)) \right) \right]. \quad (37)$$

In practice, Letizia *et al.* [12] proposed to use three f functions that are related to the Kullback–Leibler divergence, GAN, and Hellinger distance squared:

$$\begin{aligned} f_{KL} : u &\rightarrow u \log(u) \\ f_{GAN} : u &\rightarrow u \log u - (u+1) \log(u+1) + \log 4 \\ f_{HD} : u &\rightarrow (\sqrt{u} - 1)^2. \end{aligned}$$

Experimentally, they found that f_{GAN} achieves the best performance. Importantly, the f_{GAN} is closely related to the f function associated with the Jensen-Shannon Divergence defined as:

$$f_{JSD} : u \rightarrow \frac{1}{2} \left(u \log u - (u+1) \log\left(\frac{u+1}{2}\right) \right). \quad (38)$$

B Appendix: Proofs of theorems and derivations

B.1 Proof of Theorem 4.1

We begin by deriving the expression for the JSD between two Bernoulli distributions.

Lemma B.1. Let $p = B(\mu)$ and $q = B(\nu)$ be two Bernoulli distributions with parameters μ and ν . Then, the Jensen-Shannon Divergence between p and q is given by:

$$\begin{aligned} D_{\text{JS}}[B(\mu) || B(\nu)] &= \log 2 + \frac{1}{2} \left[\mu \log \frac{\mu}{\mu + \nu} + \nu \log \frac{\nu}{\mu + \nu} \right] \\ &\quad + \frac{1}{2} \left[(1 - \mu) \log \frac{1 - \mu}{2 - \mu - \nu} + (1 - \nu) \log \frac{1 - \nu}{2 - \mu - \nu} \right]. \quad (39) \end{aligned}$$

Proof. The Kullback–Leibler divergence between two Bernoulli distributions p and q with parameters μ and ν is given by:

$$D_{\text{KL}}[p || q] = \mu \log \frac{\mu}{\nu} + (1 - \mu) \log \frac{1 - \mu}{1 - \nu}. \quad (40)$$

Given that the mixture $\frac{1}{2}(p + q)$ is a Bernoulli distribution with parameter $m = \frac{p+q}{2}$, the JSD is defined as:

$$\begin{aligned} D_{\text{JS}}[B(\mu) \parallel B(\nu)] &= \frac{1}{2} D_{\text{KL}}[B(\mu) \parallel B(m)] + \frac{1}{2} D_{\text{KL}}[B(\nu) \parallel B(m)] \\ &= \frac{1}{2} \left[\mu \log \frac{2\mu}{\mu + \nu} + (1 - \mu) \log \frac{2(1 - \mu)}{2 - \mu - \nu} \right] \\ &\quad + \frac{1}{2} \left[\nu \log \frac{2\nu}{\mu + \nu} + (1 - \nu) \log \frac{2(1 - \nu)}{2 - \mu - \nu} \right] \\ &= \log 2 + \frac{1}{2} \left[\mu \log \frac{\mu}{\mu + \nu} + \nu \log \frac{\nu}{\mu + \nu} \right] \\ &\quad + \frac{1}{2} \left[(1 - \mu) \log \frac{1 - \mu}{2 - \mu - \nu} + (1 - \nu) \log \frac{1 - \nu}{2 - \mu - \nu} \right]. \end{aligned}$$

□

Lemma B.2. The function Ξ^{-1} defined by

$$\begin{aligned} \Xi^{-1} : \mathbb{R}_+ &\longrightarrow [0, \log 2] \\ y &\longmapsto \log 2 - \frac{1}{2} [(1 + e^{-y}) \log(1 + e^{-y}) + ye^{-y}], \end{aligned} \quad (41)$$

is C^∞ , strictly increasing, and satisfies

$$\Xi^{-1}(0) = 0, \quad \lim_{y \rightarrow +\infty} \Xi^{-1}(y) = \log 2. \quad (42)$$

Hence, Ξ^{-1} is a bijection from \mathbb{R}_+ onto $[0, \log 2)$.

Proof. The function Ξ^{-1} is obtained by finite compositions, additions and multiplication of elementary C^∞ functions. Therefore, $\Xi^{-1} \in C^\infty(\mathbb{R}_+)$.

Differentiating, we obtain

$$(\Xi^{-1})'(y) = -\frac{1}{2} \left[-e^{-y} \log(1 + e^{-y}) - (1 + e^{-y}) \frac{e^{-y}}{1 + e^{-y}} + e^{-y} - ye^{-y} \right] \quad (43)$$

$$= \frac{1}{2} e^{-y} [\log(1 + e^y)]. \quad (44)$$

Since $e^{-y} > 0$ and $\log(1 + e^y) > 0$ for all $y > 0$, it follows that $(\Xi^{-1})'(y) > 0$. Hence, Ξ^{-1} is strictly increasing on \mathbb{R}_+ .

Moreover, as $y \rightarrow +\infty$, $e^{-y} \rightarrow 0$ and thus $\lim_{y \rightarrow +\infty} \Xi^{-1}(y) = \log 2$. At $y = 0$, we have

$$\Xi^{-1}(0) = \log 2 - \frac{1}{2} [(1 + 1) \log(2) + 0] = 0.$$

By continuity and strict monotonicity, the image of Ξ^{-1} is therefore $[0, \log 2)$, establishing bijectivity. □

Lemma B.3. The function Ξ , defined as the inverse of Ξ^{-1} in Eq. (41), is C^∞ and strictly increasing.

Proof. Since Ξ^{-1} is C^∞ with strictly positive derivative on \mathbb{R}_+ , the inverse function theorem implies that Ξ is C^∞ and strictly increasing on $[0, \log 2)$. □

We will now prove the main result of this work.

We denote Ω the lower triangle $\Omega = \{(\mu, \nu) \in [0, 1] \times (0, 1] \mid \nu \leq \mu\}$.

Lemma B.4. Let the mapping ϕ be defined as on the lower triangle:

$$\phi : (\mu, \nu) \in \Omega \mapsto (j(\mu, \nu), k(\mu, \nu)) \quad (45)$$

where j and k are defined as follows:

$$j(\mu, \nu) \doteq D_{\text{JS}}[B(\mu) \parallel B(\nu)] \quad \text{and} \quad k(\mu, \nu) \doteq D_{\text{KL}}[B(\mu) \parallel B(\nu)]. \quad (46)$$

Then, the Jacobian J of the mapping ϕ is given by:

$$J(\mu, \nu) = \begin{bmatrix} \frac{\partial j}{\partial \mu} & \frac{\partial j}{\partial \nu} \\ \frac{\partial k}{\partial \mu} & \frac{\partial k}{\partial \nu} \end{bmatrix} (\mu, \nu) = \begin{bmatrix} \frac{1}{2} [\mathbb{L}(\mu) - \mathbb{L}(m)] & \frac{1}{2} [\mathbb{L}(\nu) - \mathbb{L}(m)] \\ \mathbb{L}(\mu) - \mathbb{L}(\nu) & -\left[\frac{\mu}{\nu} - \frac{1-\mu}{1-\nu}\right] \end{bmatrix} \quad (47)$$

where \mathbb{L} is the logit function.

Proof. Using Lemma B.1, the partial derivative $\frac{\partial j}{\partial \mu}$ is given by:

$$\begin{aligned}
\frac{\partial j}{\partial \mu}(\mu, \nu) &= \frac{1}{2} \left[\mu \left(\frac{1}{\mu} - \frac{1}{\mu + \nu} \right) + \log \frac{\mu}{\mu + \nu} - \frac{\nu}{\mu + \nu} \right] \\
&\quad + \frac{1}{2} \left[(1 - \mu) \left(\frac{1}{2 - \mu - \nu} - \frac{1}{1 - \mu} \right) - \log \frac{1 - \mu}{2 - \mu - \nu} + \frac{1 - \nu}{2 - \mu - \nu} \right] \\
&= \frac{1}{2} \left[\log \frac{\mu}{\mu + \nu} - \log \frac{1 - \mu}{2 - \mu - \nu} + 1 - \frac{\mu}{\mu + \nu} - \frac{\nu}{\mu + \nu} + \frac{1 - \mu}{2 - \mu - \nu} - 1 + \frac{1 - \nu}{2 - \mu - \nu} \right] \\
&= \frac{1}{2} \left[\log \frac{\mu}{\mu + \nu} - \log \frac{1 - \mu}{2 - \mu - \nu} \right] \\
&= \frac{1}{2} \left[\log \frac{\mu}{1 - \mu} - \log \frac{m}{1 - m} \right] \\
&= \frac{1}{2} [\mathbb{L}(\mu) - \mathbb{L}(m)] .
\end{aligned}$$

By symmetry of the JSD:

$$\frac{\partial j}{\partial \nu}(\mu, \nu) = \frac{1}{2} [\mathbb{L}(\nu) - \mathbb{L}(m)] . \quad (48)$$

Finally, using Eq (40):

$$\frac{\partial k}{\partial \mu}(\mu, \nu) = \log \frac{\mu(1 - \nu)}{\nu(1 - \mu)} \quad (49)$$

$$= \mathbb{L}(\mu) - \mathbb{L}(\nu) , \quad (50)$$

and

$$\frac{\partial k}{\partial \nu}(\mu, \nu) = - \left[\frac{\mu}{\nu} - \frac{1 - \mu}{1 - \nu} \right] . \quad (51)$$

□

Conjecture B.5. *The determinant of the Jacobian J defined in Eq.(47) is negative on the interior $\Omega^\circ = \{(\mu, \nu) \in (0, 1) \times (0, 1) \mid \nu < \mu\}$ of Ω , i.e.:*

$$\forall (\mu, \nu) \in \Omega^\circ, \quad \det(J(\mu, \nu)) < 0 . \quad (52)$$

While we have not found an analytical proof of the conjecture, it can be verified numerically, as shown in Figure 5. The negative of the determinant of the Jacobian is always positive.

Theorem B.6. *The optimal lower bound on Kullback-Leibler divergence established by the Jensen-Shannon divergence is:*

$$\Xi(D_{\text{JS}}[p \parallel q]) \leq D_{\text{KL}}[p \parallel q] \quad (53)$$

where $\Xi : [0, \log 2] \mapsto \mathbb{R}^+$ is a strictly increasing function defined implicitly by its inverse,

$$\Xi^{-1} : y \mapsto D_{\text{JS}}[B(1) \parallel B(\exp(-y))], \quad (54)$$

where $B(\exp(-y))$ is a Bernoulli distribution with parameter $\exp(-y) \in (1/2, 1]$. Ξ^{-1} has a closed-form expression related to Eq. (17).

Proof. Following the Theorem 3.1 by Harremoës & Vajda [7], the joint range between the JSD and the KLD can be fully characterized by restricting p, q to be pairs of Bernoulli distributions compatible with JSD and KLD,

$$x = D_{\text{JS}}[B(\mu) \parallel B(\nu)] \quad \text{and} \quad y = D_{\text{KL}}[B(\mu) \parallel B(\nu)] , \quad (55)$$

for $(\mu, \nu) \in [0, 1] \times (0, 1) \cup \{(0, 0), (1, 1)\}$, where $B(\theta)$ is a Bernoulli distribution with parameter θ .

We are therefore interested in the image of the mapping ϕ defined as:

$$\phi : (\mu, \nu) \mapsto (D_{\text{JS}}[B(\mu) \parallel B(\nu)], D_{\text{KL}}[B(\mu) \parallel B(\nu)]) \quad (56)$$

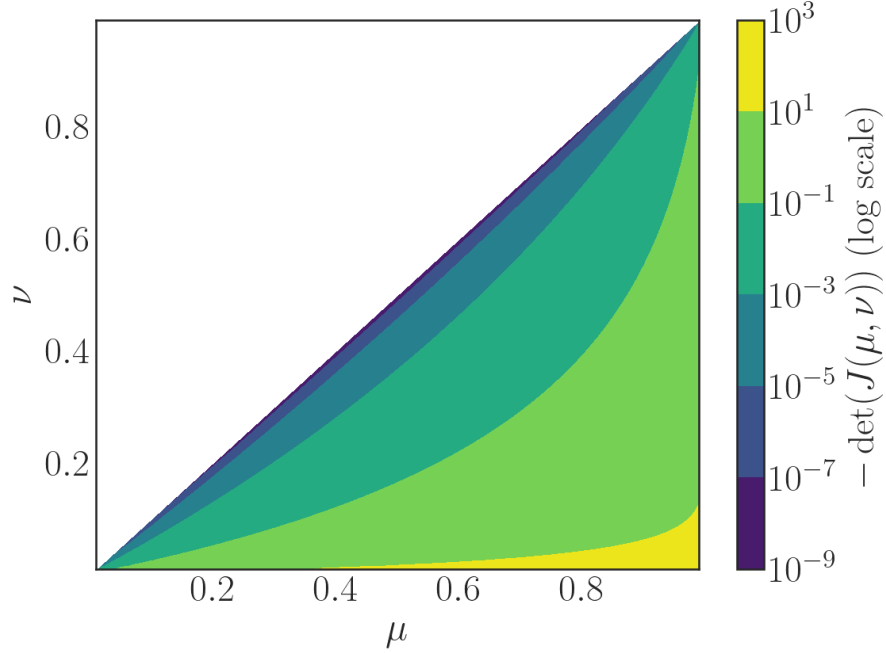


Figure 5: Visualization of the negative of the determinant of the Jacobian. We can observe that it is constantly positive, meaning that the determinant of the Jacobian is always negative.

from the unit square $[0, 1] \times (0, 1) \cup \{(0, 0), (1, 1)\}$ into \mathbb{R}^2 .

Exploiting the symmetry of the divergences:

$$D_{\text{KL}}[B(\mu) \parallel B(\nu)] = D_{\text{KL}}[B(1 - \mu) \parallel B(1 - \nu)], \quad (57)$$

$$D_{\text{JS}}[B(\mu) \parallel B(\nu)] = D_{\text{JS}}[B(1 - \mu) \parallel B(1 - \nu)], \quad (58)$$

we restrict our analysis to the image $\phi(\Omega)$ of the lower triangle $\Omega = \{(\mu, \nu) \in [0, 1] \times (0, 1) \mid \nu \leq \mu\}$.

Using Conjecture B.5, the determinant of the Jacobian is negative for all inner points in Ω . According to the open mapping theorem, the inner points are therefore mapped into interior points of the range. To determine the lower envelope of the image, we examine the image of the boundary $\delta\Omega = (S_1, S_2, S_3)$. Given that, S_1 and S_2 respectively map to a single point $(0, 0)$ and a horizontal line $y = \infty$, S_3 defines the lower envelope of the range and is, therefore, our bound.

B.2 Proof of Eq. (26)

The variational representation [17, 27] of can be expressed as:

$$I_{\text{JS}}[U; V] = D_{\text{JS}}[p_{UV} \parallel p_U \otimes p_V] = \frac{1}{2} \max_t \left[\mathbb{E}_{p_{UV}} [t] - \mathbb{E}_{p_U \otimes p_V} [-\log(2 - \exp(t))] \right] \quad (59)$$

where $t : \mathcal{U} \times \mathcal{V} \rightarrow (-\infty, \log 2)$ is a variational function.

Now, define the following reparameterization,

$$t(u, v) \doteq \log(2q_\theta(z = 1 \mid u, v)), \quad (60)$$

where $q_\theta(z = 1 \mid u, v)$ is a neural network.

Substituting (60) into (59), we obtain (25):

$$I_{\text{JS}}[U; V] = \frac{1}{2} \max_t \left[\mathbb{E}_{p_{UV}} [\log(2q_\theta(z=1|u, v))] - \mathbb{E}_{p_U \otimes p_V} [-\log(2 - 2q_\theta(z=1|u, v))] \right] \quad (61)$$

$$= \frac{1}{2} \max_\theta \left[\log 4 + \mathbb{E}_{p_{UV}} [\log q_\theta(z=1|u, v)] + \mathbb{E}_{p_U \otimes p_V} [\log(1 - q_\theta(z=1|u, v))] \right] \quad (62)$$

$$(63)$$

The expected cross-entropy (CE) under our classifier model in Eq. (21) is defined as:

$$\mathcal{L}_{CE} \doteq \mathbb{E}_{(u,v,z) \sim p_{UVZ}} [-\log q_\theta(z|u, v)] = \mathbb{E}_{z \sim p_Z} \left[\mathbb{E}_{(u,v) \sim p_{UV|Z}} [-\log q_\theta(z|u, v)] \right] \quad (64)$$

or, expanding the expectation over z ,

$$\mathcal{L}_{CE} = \frac{1}{2} \mathbb{E}_{(u,v) \sim p_{UV}} [-\log q_\theta(z=1|u, v)] + \frac{1}{2} \mathbb{E}_{(u,v) \sim p_U \otimes p_V} [-\log q_\theta(z=0|u, v)] . \quad (65)$$

With Eqn. (62) we obtain

$$I_{\text{JS}}[U; V] \geq \log 2 - \min_\theta \mathcal{L}_{CE}(\theta) . \quad (66)$$

□

B.3 Alternative to Section 4.2 to obtain (26) and gap quantification

Here, we revisit the relationship between Jensen-Shannon divergence and the cross-entropy training loss. In our main manuscript, the relationship was derived using properties of f -divergences. The following derivation is more discriminator-oriented, and does not use f -divergence properties directly.

Alternative proof to (26). Recall the classifier model of Equation (21). We use a standard binary classification setup to estimate the JSD between the joint distribution p_{UV} and the product of its marginals $p_U \otimes p_V$. Consider the mixture model:

$$(U, V) | Z = 1 \sim p_{UV}(u, v), \quad (U, V) | Z = 0 \sim p_U \otimes p_V, \quad \text{with } Z \sim B(1/2). \quad (67)$$

This defines a joint distribution over (u, v, z) , denoted $\tilde{p}(u, v, z)$, with marginal over (u, v) :

$$\tilde{p}(u, v) = m(u, v) \doteq \frac{1}{2} p_{UV}(u, v) + \frac{1}{2} p_U(u) p_V(v). \quad (68)$$

Next we study the MI among (U, V) and Z under that model,

$$I[(U, V); Z] \doteq D_{\text{KL}}[\tilde{p}_{UVZ} || p_{UV} \otimes p_Z] \quad (69)$$

$$= \mathbb{E}_{(u,v,z) \sim \tilde{p}_{UVZ}} \left[\log \frac{\tilde{p}_{UVZ}(u, v, z)}{\tilde{p}_{UV}(u, v) p_Z(z)} \right] \quad (70)$$

$$= \mathbb{E}_{z \sim p_Z} \left[\mathbb{E}_{(u,v) \sim \tilde{p}_{UV|Z}} \left[\log \frac{\tilde{p}_{UV|Z}}{m(u, v)} \right] \right] \quad (71)$$

$$(\text{expanding the expectation over } z) \quad (72)$$

$$= \frac{1}{2} \mathbb{E}_{(u,v) \sim p_{UV}} \left[\log \frac{p_{UV}(u, v)}{m(u, v)} \right] + \frac{1}{2} \mathbb{E}_{(u,v) \sim p_U \otimes p_V} \left[\log \frac{p_U \otimes p_V(u, v)}{m(u, v)} \right] . \quad (73)$$

Therefore:

$$I[(U, V); Z] = D_{\text{JS}}[p_{UV} || p_U \otimes p_V] = I_{\text{JS}}[U; V] . \quad (74)$$

From this, we observe that, in the classifier setup above, the MI between (U, V) and Z equals the Jensen-Shannon divergence between U and V . A similar property holds for general JS divergences.

MI also characterizes the information loss by conditioning. In this setup,

$$I[(U, V); Z] = H[Z] - H[Z|U, V] = \log 2 - H[Z|U, V] , \quad (75)$$

where the conditional entropy is defined as

$$H[Z|U, V] \doteq \mathbb{E}_{(u,v,z) \sim \tilde{p}_{UVZ}} [-\log \tilde{p}_{Z|UV}(z|u, v)] . \quad (76)$$

Combining Equations (74) and (75) yields

$$I_{JS}[U; V] = \log 2 - H[Z|U, V] . \quad (77)$$

Suppose we define a variational distribution $q_\theta(z|u, v)$; then

$$H[Z|U, V] = \mathbb{E}_{(u,v,z) \sim \tilde{p}_{UVZ}} [-\log \tilde{p}_{Z|UV}(z|u, v) + \log q_\theta(z|u, v)] + \mathbb{E}_{(u,v,z) \sim \tilde{p}_{UVZ}} [-\log q_\theta(z|u, v)] \quad (78)$$

$$= \mathbb{E}_{(u,v) \sim \tilde{p}_{UV}} \left[\mathbb{E}_{z \sim \tilde{p}_{Z|UV}} [-\log \tilde{p}_{Z|UV}(u, v|z) + \log q_\theta(z|u, v)] \right] + \mathbb{E}_{(u,v,z) \sim \tilde{p}_{UVZ}} [-\log q_\theta(z|u, v)] . \quad (79)$$

Therefore,

$$H[Z|UV] = - \underbrace{\mathbb{E}_{(u,v) \sim \tilde{p}_{UV}} [\text{D}_{\text{KL}}[\tilde{p}_{Z|UV}(u, v) \parallel q_\theta(z|u, v)]]}_{\doteq \delta} + \underbrace{\mathbb{E}_{(u,v,z) \sim \tilde{p}_{UVZ}} [-\log q_\theta(z|u, v)]}_{=\mathcal{L}_{CE}} . \quad (80)$$

The first term δ on the right-hand side of (80) is the expected KLD between the true posterior $p_{Z|UV}$ and the model posterior q_θ , which is always non-negative. The second term corresponds to the expected cross-entropy loss of the classifier q_θ .

Using the non-negativity of the KL divergence, we obtain the inequality:

$$H[Z|UV] \geq \mathcal{L}_{CE}, \quad (81)$$

where \mathcal{L}_{CE} denotes the expected cross-entropy loss. Equality holds if and only if the model distribution $q_\theta(z | u, v)$ is equal to the true posterior $p_{Z|UV}(z | u, v)$ almost everywhere.

Combining this with Equation (77) yields

$$I_{JS}[U; V] \geq \log 2 - \mathcal{L}_{CE} . \quad (82)$$

□

Thanks to this new derivation, we can explicitly quantify the gap between the true JSD lower bound $\Xi(I_{JS}[U; V])$ and the one obtained from the cross-entropy $I_{CE}(\theta)$ as follows:

Lemma B.7. *Let q_θ denote the discriminator, and let $\mathcal{L}_{CE}(\theta)$ be its expected binary cross-entropy. The gap between our new variational lower bound $I_{CE}(\theta)$ and the optimal lower bound $\Xi(I_{JS}[U; V])$ is given by*

$$\Xi(I_{JS}[U; V]) - I_{CE}(\theta) \approx \delta \cdot \Xi'(I_{CE}(\theta)) , \quad (83)$$

where δ denotes the expected Kullback–Leibler divergence between the true and approximate posteriors:

$$\mathbb{E}_{(u,v) \sim \tilde{p}_{UV}} [\text{D}_{\text{KL}}[\tilde{p}_{Z|UV}(u, v) \parallel q_\theta(z | u, v)]] . \quad (84)$$

Proof. Combining Equations (77) and (80), we obtain

$$I_{JS}[U; V] = \log(2) - \mathcal{L}_{CE}(\theta) + \delta . \quad (85)$$

Hence, the gap Δ_θ between the optimal JSD lower bound and our bound $I_{CE}(\theta)$ is

$$\Delta_\theta = \Xi(I_{JS}[U; V]) - \Xi(I_{CE}(\theta)) \quad (86)$$

$$= \Xi(\log(2) - \mathcal{L}_{CE}(\theta) + \delta) - \Xi(\log(2) - \mathcal{L}_{CE}(\theta)) . \quad (87)$$

Applying a first-order Taylor expansion around I_{CE} leads to:

$$\Delta_\theta \approx \delta \cdot \Xi'(I_{CE}(\theta)) . \quad (88)$$

□

B.4 Proof of Eq. (28)

We derive here a method for approximating the MI by way of the trained classifier, building on the setup of Section 4.2 and B.3.

Under with the joint model of Eq. (68) (originally (21)),

$$\tilde{p}_{UV|Z}(u, v|z) = \frac{\tilde{p}_{ZUV}(z, u, v)}{p_Z(z)} \quad \text{and} \quad \tilde{p}_{Z|UV}(z|u, v) = \frac{\tilde{p}_{ZUV}(z, u, v)}{\tilde{p}_{UV}(u, v)}. \quad (89)$$

Moreover, since $Z \sim B(1/2)$, we have $p(z = 1) = p(z = 0) = \frac{1}{2}$. Therefore,

$$\log \frac{\tilde{p}_{Z|UV}(z = 1|u, v)}{\tilde{p}_{Z|UV}(z = 0|u, v)} = \log \frac{\tilde{p}_{ZUV}(z = 1, u, v)}{\tilde{p}_{ZUV}(z = 0, u, v)} = \log \frac{\tilde{p}_{UV|Z}(u, v|z = 1)}{\tilde{p}_{UV|Z}(u, v|z = 0)} = \log \frac{p_{UV}(u, v)}{(p_U \otimes p_V)(u, v)}. \quad (90)$$

Using the definition of KLD,

$$D_{\text{KL}}[p_{UV} \parallel p_U \otimes p_V] \doteq \mathbb{E}_{(u,v) \sim p_{UV}} \left[\log \frac{p_{UV}(u, v)}{(p_U \otimes p_V)(u, v)} \right] = \mathbb{E}_{(u,v) \sim p_{UV}} \left[\log \frac{\tilde{p}_{Z|UV}(z = 1|u, v)}{\tilde{p}_{Z|UV}(z = 0|u, v)} \right], \quad (91)$$

and finally,

$$I[U; V] \doteq D_{\text{KL}}[p_{UV} \parallel p_U \otimes p_V] = \mathbb{E}_{(u,v) \sim p_{UV}} [\mathbb{L}(\tilde{p}_{Z|UV}(z = 1|u, v))] \quad (92)$$

where \mathbb{L} is the logit transform. Thus we can estimate the MI of p_{UV} by first training a classifier to approximate the true posterior, $p_{Z|UV}$, and after that use it to approximate the expectation by averaging over the ‘true’ data.

C Appendix: Approximation of the function Ξ

In this section, we derive Equation (17), which is used to define the inverse Ξ^{-1} . This allows the use of numerical solvers to approximate Ξ . Finally, we present the approximation of the function Ξ using the Logit transform.

C.1 Derivation of Equation (17)

Let’s recall (17):

$$D_{\text{JS}}[B(1) \parallel B(\nu)] = \log 2 - \frac{1}{2} \left[\log(1 + \nu) - \nu \log \left(\frac{\nu}{1 + \nu} \right) \right] \quad (93)$$

Proof. Using Lemma (B.1):

$$D_{\text{JS}}[B(1) \parallel B(\nu)] = \log 2 + \frac{1}{2} \left[\log \frac{1}{1 + \nu} + \nu \frac{\nu}{1 + \nu} + (1 - \nu) \log \frac{1 - \nu}{1 - \nu} \right] \quad (94)$$

$$= \log 2 - \frac{1}{2} \left[\log(1 + \nu) - \nu \log \left(\frac{\nu}{1 + \nu} \right) \right] \quad (95)$$

□

C.2 Python code for implementing Ξ

```

from scipy.optimize import root_scalar
import numpy as np

# Define the inverse function Xi(x)
def inv_xi(y):
    z = np.exp(-y)
    term1 = np.log(1 + z)
    term2 = z * np.log(z / (1 + z))
    return np.log(2) - 0.5 * (term1 - term2)

# Define the function Xi(x)
def xi(x):
    result = root_scalar(lambda y: inv_xi(y) - x, bracket=[0, 100],
        method='brentq')
    return result.root if result.converged else None

```

Algorithm 1: Algorithmic implementation of the Ξ function as the inverse of its known inverse Ξ^{-1}

C.3 Proposed differentiable approximation of Ξ

To enable the use of our bound within a differentiable objective, we introduce a smooth, differentiable approximation of the Ξ function. This approximation is derived from the Taylor expansion of the KL divergence in terms of the JSD. The lower bound is characterized by the following parametric curve:

$$x = D_{\text{JS}}[B(1) \parallel B(\nu)] = \Xi^{-1}(-\log \nu) \quad (96)$$

$$y = D_{\text{KL}}[B(1) \parallel B(\nu)] = -\log \nu, \quad (97)$$

where $\Xi : [0, \log 2) \mapsto \mathbb{R}^+$ is a strictly increasing function implicitly defined by its inverse:

$$\Xi^{-1} : y \mapsto \log 2 - \frac{1}{2} \left[(1 + e^{-y}) \log(1 + e^{-y}) + ye^{-y} \right]. \quad (98)$$

Hence, we can express:

$$x = \Xi^{-1}(y). \quad (99)$$

Let $\mathbb{L} : x \mapsto \log \frac{x}{1-x}$ denote the Logit function. We then observe that:

$$\mathbb{L} \left(\frac{1}{2} \left(\frac{x}{\log 2} + 1 \right) \right) = \log \left(\frac{\frac{x}{\log 2} + 1}{1 - \frac{x}{\log 2}} \right) \quad (100)$$

$$= \log \left(\frac{\frac{\Xi^{-1}(y)}{\log 2} + 1}{1 - \frac{\Xi^{-1}(y)}{\log 2}} \right) \quad (101)$$

$$= \log \left(\frac{1 - \frac{1}{2\log 2} [(1 + e^{-y}) \log(1 + e^{-y}) + ye^{-y}]}{\frac{1}{2\log 2} [(1 + e^{-y}) \log(1 + e^{-y}) + ye^{-y}]} \right) \quad (102)$$

$$= \log \left(\frac{4\log 2}{(1 + e^{-y}) \log(1 + e^{-y}) + ye^{-y}} - 1 \right) \quad (103)$$

The first-order Taylor expansion of the right-hand side of Eq. (103) around $y = 0$ is simply $y \mapsto y$. This motivates approximating Ξ using a scaled logit function. Through a grid search, we found that scaling the logit by a factor of 1.15 yields the lowest median approximation error over $x \in (0, 1)$:

$$\Xi(x) \approx 1.15 \cdot \mathbb{L} \left(\frac{1}{2} \left(\frac{x}{\log 2} + 1 \right) \right). \quad (104)$$

As shown in Figure 6, our proposed approximation closely matches the numerically estimated inverse obtained via a solver. Importantly, this formulation is fully differentiable and can be directly integrated into a variational lower bound (VLB) objective in representation learning.

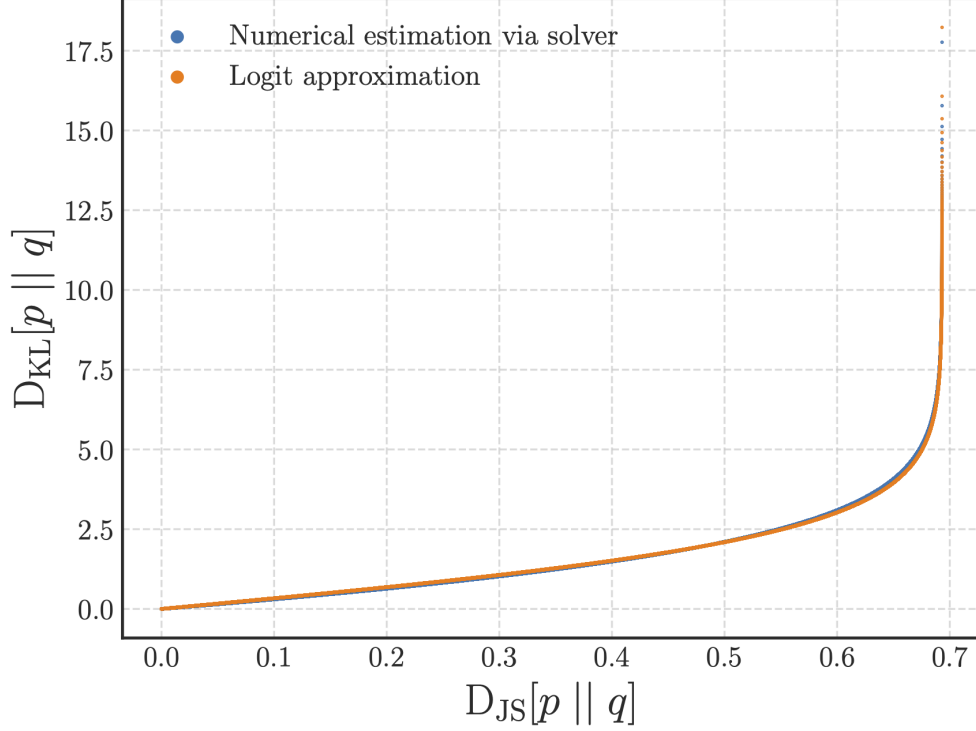


Figure 6: Comparison between the numerical evaluation of Ξ using Algorithm 1 and the proposed differentiable approximation from Eq. (104).

D Appendix: Tightness Analysis in the Mutual Information Setting

While our JSD–KLD lower bound in Theorem 4.1 is the tightest possible (i.e., equality is achieved for a family of distributions) between these two divergences for arbitrary distributions, one might expect a tighter bound when restricted to the mutual information setting. i.e., when comparing a joint distribution to the product of its marginals. Remarkably, as illustrated in Figure 2 of the main text, discrete joint–marginal pairs can lie close to our general bound. In this section, we analytically show that there exists an infinite family of distributions in the mutual information setting that exactly lie on our lower bound.

Lemma D.1. *Let (U, V) be a pair of uniform categorical variables of dimension $k \in \mathbb{N}^+$ that are fully dependent (i.e., $U = V$). Then:*

$$I[U; V] = \Xi(I_{\text{JS}}[U; V]) . \quad (105)$$

Proof. Let $k \in \mathbb{N}^+$, and assume U and V are fully dependent (i.e., $U = V$) uniform categorical variables of dimension k . Their joint and marginal distributions are given by:

$$P_U = P_V = \frac{1}{k} \mathbf{1}_k , \quad P_{UV} = \frac{1}{k} I_k ,$$

and consequently,

$$P_U \otimes P_V = \frac{1}{k^2} \mathbf{1}_{(k,k)} .$$

Let $M = \frac{1}{2} (P_U \otimes P_V + P_{UV})$ denote the mixture distribution. Then $M_{i,j} = \frac{1/k + 1/k^2}{2}$ if $i = j$, and $M_{i,j} = \frac{1}{2k^2}$ otherwise.

We now compute $I[U; V]$ and $I_{\text{JS}}[U; V]$:

$$I[U; V] = \sum_{i=1}^k \frac{1}{k} \log \frac{1/k}{1/k^2} = \log k . \quad (106)$$

$$I_{\text{JS}}[U; V] = \frac{1}{2} \left(\sum_{i=1}^k \frac{1}{k} \log \frac{1/k}{\frac{1}{2}(1/k + 1/k^2)} + \sum_{i=1}^k \frac{1}{k^2} \log \frac{1/k^2}{\frac{1}{2}(1/k + 1/k^2)} + \sum_{i=1}^k \sum_{\substack{j=1 \\ j \neq i}}^k \frac{1}{k^2} \log \frac{1/k^2}{1/(2k^2)} \right) \quad (107)$$

$$= \frac{1}{2} \left(\log \frac{2}{1+1/k} + \frac{1}{k} \log \frac{2/k}{1+1/k} + \frac{k-1}{k} \log 2 \right) \quad (108)$$

$$= \log 2 - \frac{1}{2} \left[\log(1 + 1/k) - \frac{1}{k} \log \frac{1/k}{1+1/k} \right] \quad (109)$$

$$= \Xi^{-1}(\log k). \quad (110)$$

Combining Equations (106) and (110), we obtain

$$I_{\text{JS}}[U; V] = \Xi^{-1}(I[U; V]), \quad (111)$$

which completes the proof. \square

E Appendix: Experimental details

E.1 Discrete case experiments

In this section, we present a synthetic experiment designed to empirically validate the tightness of our proposed lower bound between mutual information (MI) and the Jensen-Shannon divergence (JSD), as introduced in Theorem 4.1. Unlike general divergences between arbitrary distributions, here we specialize the analysis to the mutual information setting, which corresponds to comparing a joint distribution to the product of its marginals.

We consider a family of discrete joint distributions $P_{UV}^{(\alpha)} \in \mathbb{R}^{k \times k}$ parameterized by a dependence factor $\alpha \in [0, 1]$, which interpolates between the independent and perfectly dependent cases. Specifically, we fix the marginals to be uniform categorical distributions: $P_U = P_V = \frac{1}{k} \mathbf{1}_k$, where k denotes the number of categories. The joint distribution is constructed as a convex combination of the independent and deterministic cases:

$$P_{UV}^{(\alpha)} = (1 - \alpha) P_U \otimes P_V + \alpha \cdot \text{diag}(P_U), \quad (112)$$

where $\text{diag}(P_U)$ denotes the diagonal matrix with P_U on its diagonal, ensuring perfect dependence along the diagonal. This construction ensures that:

- When $\alpha = 0$, U and V are independent.
- When $\alpha = 1$, $U = V$ with probability one.

For each α , we compute the mutual information $I(U; V) = D_{\text{KL}}[P_{UV}^{(\alpha)} \parallel P_U \otimes P_V]$ and the Jensen-Shannon divergence $D_{\text{JS}}[P_{UV}^{(\alpha)} \parallel P_U \otimes P_V]$, and compare them to the lower bound $\Xi(D_{\text{JS}}[P_{UV}^{(\alpha)} \parallel P_U \otimes P_V])$, where Ξ is defined in Equation (19).

E.2 Complex Gaussian and non-Gaussian distributions experiments

In this section, we provide further details on the experimental setup described in Section 5.2 and present supplementary results for the linear and cubic Gaussian experiments.

E.2.1 Experimental setup

The neural network architectures employed in our experiments follow the joint architecture, as in [12]. We do not use the deranged architecture from [12], as it is incompatible with the CPC objective.

Joint architecture. Following the architecture described in [12], we employ a fully connected feed-forward neural network where the input layer has dimensionality $2d$, corresponding to the concatenation of two d -dimensional samples. The network comprises two hidden layers, each with 256 neurons and ReLU activations, and a single scalar output neuron. During training, the network

receives as input b^2 pairs (u, v) per iteration, where these pairs are formed by taking all combinations of samples u and v independently drawn from the joint distribution $p_{(u,v)}$.

Staircase In the Gaussian setting, we define two random variables $U \sim \mathcal{N}(0, I_d)$ and $N \sim \mathcal{N}(0, I_d)$, sampled independently, and construct $V = \rho U + \sqrt{1 - \rho^2} N$, where ρ is the correlation coefficient.

To explore settings with non-Gaussian marginals while preserving the same mutual information, we apply nonlinear transformations to V . In the cubic case, we use the mapping $v \mapsto v^3$, which preserves the dependency structure but introduces non-Gaussian marginals. Additionally, we benchmark two alternative transformations: the inverse hyperbolic sine (asinh), which shortens the tails of the distribution, and the half-cube transformation, which stretches them. These mappings allow us to evaluate the robustness of our method under controlled deviations from Gaussianity.

During the training procedure, every 4k iterations, the target value of the MI is increased by 2 nats, for 5 times, obtaining a target staircase with 5 steps. The change in target MI is obtained by increasing ρ , which affects the true MI according to $I[U; V] = -\frac{d}{2} \log(1 - \rho^2)$.

Training procedure. Each neural estimator is trained using Adam optimizer, with learning rate 0.002, $\beta_1 = 0.9$, $\beta_2 = 0.999$. The batch size is initially set to $b = 64$.

E.2.2 Analysis for different values of batch size b

Our VLB is robust to changes in batch size b . In Figures 7, 8, and 9, we observe that MINE and NWJ exhibit high variance, even for large batch sizes (e.g., $b = 128$), while JSD-LB and CPC remain more stable. Notably, MINE’s instability can cause substantial overestimation of its lower bound, occasionally exceeding the true MI. CPC, on the other hand, provides low-variance estimates but is upper bounded by the contrastive limit $\log(b)$, which results in higher bias in high-MI regimes.

E.2.3 Analysis for different values of dimension d

We additionally perform an analysis for different values of dimension d . We report the achieved bias, variance, and mean squared error (MSE) corresponding to the four settings $d \in \{5, 10, 20, 100\}$ in Table 4. We experimentally found that MINE produces estimate with high variance, NWJ is unstable in high dimension, and CPC is bounded by $\ln(b)$. When repurposed for two-step MI estimation, our method surpasses other variational lower bound estimators across dimensions. In contrast, while the bias of JSD-LB increases with dimension d , its variance remains stable, making it a reliable surrogate for MI maximization. Finally, when repurposed for two-step MI estimation, our method surpasses other variational lower bound estimators across dimensions.

E.2.4 Comparison of our two-step estimator with other two-step estimators

While our primary contribution is a novel variational lower bound (JSD-LB) that can be directly used for MI maximization, we additionally report results for our two-step estimator (Our MI estim.) alongside other two-step approaches to provide better context. Comparisons with SMILE, KL-DIME, and GAN-DIME are presented in Tables 5 and 6. The results show that our derived two-step MI estimator, which is equivalent to GAN-DIME up to a linear transformation, achieves competitive performance.

E.2.5 Computational time analysis

A critical characteristic is the computation time. The computational time analysis is developed on a server with CPU “Intel Xeon Platinum 8468 48-Core Processor” and an NVIDIA GPU H100 and reported in Table 7. As in [12], we found that the influence of the MI lower bound objective functions on the algorithm’s time requirements is minor.

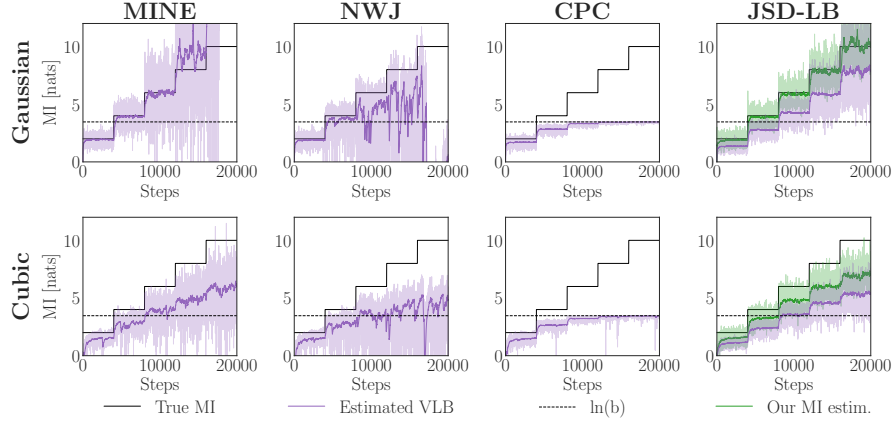


Figure 7: Staircase MI estimation comparison for $d = 5$ and batch size $b = 32$. The Gaussian case is reported in the top row, while the cubic case is shown in the bottom row.

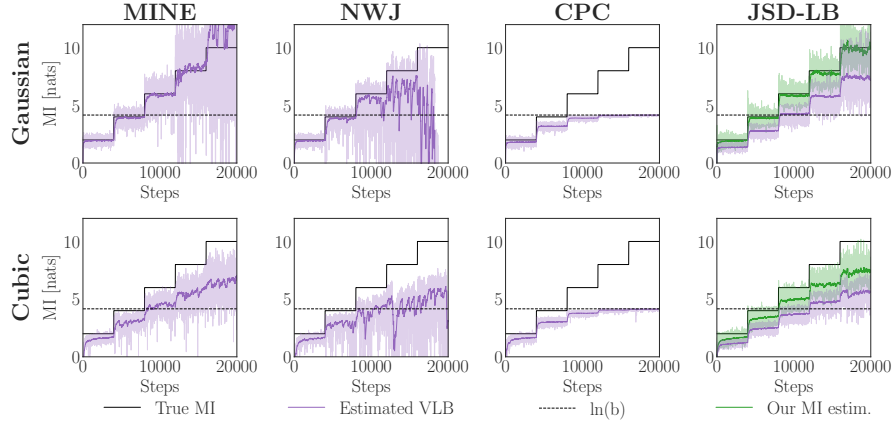


Figure 8: Staircase MI estimation comparison for $d = 5$ and batch size $b = 64$. The Gaussian case is reported in the top row, while the cubic case is shown in the bottom row.

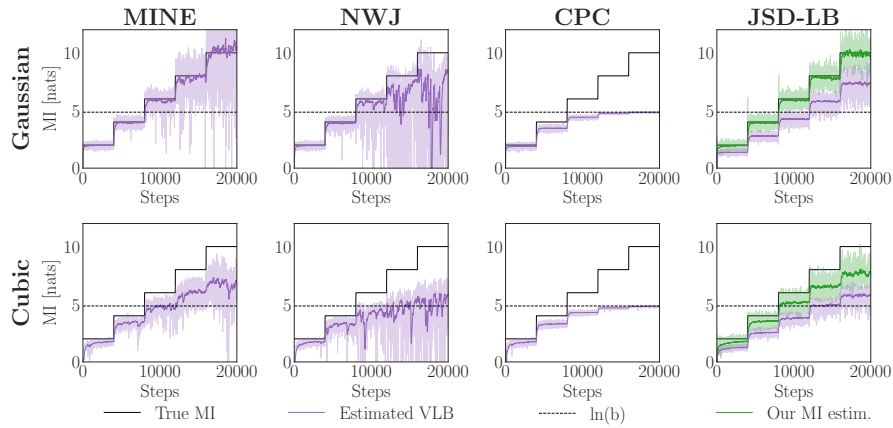


Figure 9: Staircase MI estimation comparison for $d = 5$ and batch size $b = 128$. The Gaussian case is reported in the top row, while the cubic case is shown in the bottom row.

Table 4: Bias, variance, and mean squared error (MSE) of the MI estimators using the joint architecture, when $b = 128$, for the Gaussian setting with varying dimensionality $d \in \{5, 10, 20, 100\}$.

MI	$d = 5$					$d = 10$				
	2	4	6	8	10	2	4	6	8	10
a) Bias										
MINE	0.07	0.11	0.12	0.05	-1.58	0.12	0.19	0.32	0.12	-1.26
NWJ	0.09	0.19	0.8	1.77	∞	0.15	0.28	0.86	∞	∞
CPC	0.2	0.81	2.11	3.89	5.85	0.25	0.91	2.21	3.94	5.86
JSD-LB	0.64	1.22	1.75	2.22	2.62	0.77	1.55	2.33	3.07	3.75
Our MI estim.	0.1	0.09	0.12	0.19	0.1	0.17	0.21	0.28	0.35	0.29
b) Variance										
MINE	0.05	0.14	0.41	2.05	12.23	0.06	0.15	0.49	1.63	8.87
NWJ	0.07	0.23	4.09	12.83	∞	0.09	0.23	1.49	∞	∞
CPC	0.04	0.03	0.01	0.0	0.0	0.04	0.03	0.01	0.0	0.0
JSD-LB	0.03	0.06	0.12	0.23	0.72	0.03	0.06	0.11	0.19	0.35
Our MI estim.	0.07	0.13	0.22	0.36	0.93	0.08	0.15	0.25	0.44	0.73
c) Mean Square Error										
MINE	0.06	0.15	0.43	2.05	14.72	0.08	0.19	0.59	1.64	10.46
NWJ	0.07	0.27	4.74	15.98	∞	0.11	0.31	2.23	∞	∞
CPC	0.08	0.69	4.45	15.15	34.23	0.1	0.87	4.89	15.5	34.36
JSD-LB	0.45	1.55	3.17	5.16	7.57	0.63	2.47	5.56	9.62	14.43
Our MI estim.	0.08	0.13	0.23	0.39	0.94	0.11	0.2	0.33	0.56	0.82
MI	$d = 20$					$d = 100$				
	2	4	6	8	10	2	4	6	8	10
a) Bias										
MINE	0.19	0.38	0.8	0.82	0.7	1.0	1.36	2.61	3.24	4.0
NWJ	0.23	0.48	∞	∞	∞	1.13	1.67	3.05	4.1	∞
CPC	0.31	1.0	2.3	4.0	5.89	0.99	1.39	2.59	4.18	5.98
JSD-LB	0.88	1.8	2.77	3.73	4.67	1.42	2.36	3.54	4.79	6.06
Our MI estim.	0.26	0.39	0.52	0.64	0.65	1.1	1.22	1.67	2.22	2.83
b) Variance										
MINE	0.08	0.19	0.65	1.62	5.74	0.19	0.21	0.99	1.15	1.75
NWJ	0.12	0.27	∞	∞	∞	0.14	0.17	1.19	1.81	∞
CPC	0.06	0.04	0.02	0.01	0.0	0.18	0.04	0.03	0.01	0.01
JSD-LB	0.04	0.06	0.1	0.18	0.29	0.06	0.04	0.06	0.11	0.16
Our MI estim.	0.12	0.17	0.28	0.44	0.73	0.17	0.14	0.19	0.27	0.41
c) Mean Square Error										
MINE	0.12	0.34	1.29	2.29	6.23	1.18	2.05	7.81	11.62	17.76
NWJ	0.17	0.5	∞	∞	∞	1.41	2.95	10.46	18.61	∞
CPC	0.15	1.04	5.32	15.97	34.65	1.15	1.96	6.72	17.47	35.83
JSD-LB	0.82	3.28	7.75	14.12	22.06	2.07	5.6	12.61	23.05	36.89
Our MI estim.	0.19	0.32	0.55	0.86	1.15	1.39	1.62	2.99	5.18	8.41

Table 5: Bias, Variance, and Mean Square Error of SMILE, KL-DIME, GAN-DIME, and Our MI estimator across different b , and MI values for $d = 5$.

MI	$b = 64$					$b = 256$				
	2	4	6	8	10	2	4	6	8	10
a) Bias										
SMILE	-0.23	-0.59	-0.73	-0.68	-0.65	-0.28	-0.65	-0.81	-0.78	-0.68
KL-DIME	0.09	0.12	0.22	0.44	0.9	0.05	0.05	0.09	0.19	0.41
GAN-DIME	0.1	0.1	0.13	0.24	0.09	0.06	0.05	0.07	0.12	0.01
Our MI estim.	0.1	0.09	0.12	0.19	0.1	0.06	0.05	0.07	0.12	0.04
b) Variance										
SMILE	0.08	0.13	0.22	0.33	0.72	0.03	0.07	0.1	0.18	0.3
KL-DIME	0.06	0.07	0.08	0.11	0.17	0.03	0.02	0.02	0.03	0.05
GAN-DIME	0.07	0.13	0.22	0.37	0.92	0.03	0.06	0.11	0.18	0.4
Our MI estim.	0.07	0.13	0.22	0.36	0.93	0.03	0.06	0.1	0.2	0.43
c) Mean Square Error										
SMILE	0.13	0.48	0.74	0.8	1.14	0.11	0.49	0.75	0.8	0.77
KL-DIME	0.07	0.08	0.13	0.3	0.97	0.03	0.02	0.03	0.07	0.22
GAN-DIME	0.08	0.14	0.23	0.43	0.92	0.04	0.07	0.11	0.2	0.4
Our MI estim.	0.08	0.13	0.23	0.39	0.94	0.04	0.07	0.11	0.22	0.43

Table 6: Bias, Variance, and Mean Square Error of SMILE, KL-DIME, GAN-DIME, and Our MI estimator across different b , and MI values for $d = 20$.

MI	$b = 64$					$b = 256$				
	2	4	6	8	10	2	4	6	8	10
a) Bias										
SMILE	-0.08	-0.24	-0.25	-0.24	-0.16	-0.17	-0.38	-0.43	-0.39	-0.28
KL-DIME	0.24	0.36	0.62	1.0	1.57	0.16	0.2	0.34	0.58	0.92
GAN-DIME	0.26	0.38	0.54	0.67	0.72	0.19	0.26	0.35	0.48	0.5
Our MI estim.	0.26	0.39	0.52	0.64	0.65	0.19	0.26	0.36	0.47	0.53
b) Variance										
SMILE	0.12	0.16	0.27	0.46	0.77	0.06	0.07	0.12	0.2	0.32
KL-DIME	0.1	0.11	0.15	0.19	0.22	0.05	0.03	0.04	0.05	0.07
GAN-DIME	0.12	0.17	0.26	0.41	0.73	0.06	0.08	0.12	0.2	0.36
Our MI estim.	0.12	0.17	0.28	0.44	0.73	0.06	0.08	0.12	0.2	0.32
c) Mean Square Error										
SMILE	0.12	0.22	0.33	0.52	0.79	0.08	0.21	0.3	0.35	0.41
KL-DIME	0.16	0.24	0.53	1.19	2.69	0.07	0.07	0.16	0.38	0.91
GAN-DIME	0.19	0.32	0.55	0.86	1.25	0.09	0.15	0.24	0.43	0.61
Our MI estim.	0.19	0.32	0.55	0.86	1.15	0.09	0.14	0.25	0.42	0.6

Table 7: Comparison of the time requirements (in minutes) to complete the 5-step staircase MI ($d = 5$) over the batch size.

	$b = 32$	$b = 64$	$b = 128$	$b = 256$	$b = 512$	$b = 1024$
MINE	0.54	0.46	0.52	1.0	2.67	9.07
NWJ	0.39	0.4	0.46	0.94	2.61	9.02
CPC	0.39	0.37	0.42	0.91	2.59	8.99
JSD-LB	0.45	0.48	0.51	0.96	2.62	9.05