The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)

Improved Mutual Information Estimation

Youssef Mroueh, Igor Melnyk, Pierre Dognin, Jarret Ross, Tom Sercu *

IBM Research AI

Abstract

We propose to estimate the KL divergence using a relaxed likelihood ratio estimation in a Reproducing Kernel Hilbert space. We show that the dual of our ratio estimator for KL in the particular case of Mutual Information estimation corresponds to a lower bound on the MI that is related to the so called Donsker Varadhan lower bound. In this dual form, MI is estimated via learning a witness function discriminating between the joint density and the product of marginal, as well as an auxiliary scalar variable that enforces a normalization constraint on the likelihood ratio. By extending the function space to neural networks, we propose an efficient neural MI estimator, and validate its performance on synthetic examples, showing advantage over the existing baselines. We demonstrate its strength in large-scale self-supervised representation learning through MI maximization.

1 Introduction

Mutual information (MI) is an ubiquitous measure of dependency between a pair of random variables, and is one of the corner stones of information theory. In machine learning, the information maximization principle for learning representation from unlabeled data through self-supervision (Bell and Sejnowski 1995) motivated the development of many MI estimators and applications (Hjelm et al. 2019; Noroozi and Favaro 2016; Kolesnikov, Zhai, and Beyer 2019; Doersch, Gupta, and Efros 2015; Oord, Li, and Vinyals 2018; Hu et al. 2017). The information bottleneck (Tishby, Pereira, and Bialek 1999; Kolchinsky, Tracey, and Wolpert 2017) is another principle that triggered recent interest in mutual information estimation. MI is also used to understand the information flow in neural networks, in learning clusters (Krause, Perona, and Gomes 2010) and in regularizing the training of Generative Adversarial Networks (GANs) (Chen et al. 2016).

In many of these machine learning applications and other scientific fields, one has to estimate MI given samples from the joint distribution of high dimensional random variables. Since MI is defined as the Kullback-Leibler (KL) divergence between the joint distribution and the product of marginals, one can leverage non parametric estimators of f-divergences (Nguyen, Wainwright, and Jordan 2008; Nowozin, Cseke, and Tomioka 2016; Sriperumbudur et al. 2009). Specifically of interest to us is the Donsker-Varadhan (DV) representation of the KL divergence (Donsker and Varadhan 1976) that was used recently with neural networks estimators (Belghazi et al. 2018; Poole et al. 2019). Other approaches to estimating the MI are through finding lower bounds using variational Bayesian methods (Alemi et al. 2017, 2018; Barber and Agakov 2003; Blei, Kucukelbir, and McAuliffe 2017), through geometric methods like binning (Kraskov, Stögbauer, and Grassberger 2004a), k-nearest neighbors (Kraskov, Stögbauer, and Grassberger 2004b), kernel density (Kandasamy et al. 2015; Han et al. 2017), ensemble estimation (Moon, Sricharan, and Hero 2017), jackknife approach (Zeng, Xia, and Tong 2018), Gaussian copula (Singh and Póczos 2017), to name a few.

In this paper, we propose a new estimator of MI that can be used in direct MI maximization or as a regularizer, thanks to its unbiased gradients. Our starting point is the Donsker-Varadhan (**DV**) lower bound of the KL divergence that we represent equivalently via a joint optimization that we call **$\boldsymbol{\eta}$-DV** on a witness function f and an auxiliary variable η in Section 2. In Section 3, we show that when the witness function f is learned in a Reproducing Kernel Hilbert Space (RKHS) the **$\boldsymbol{\eta}$-DV** problem is jointly convex in both f and η. The dual of this problem sheds the light on this estimator as a constrained ratio estimation where η plays the role of a Lagrange multiplier that ensures proper normalization of the likelihood ratio. We also show how the witness function can be estimated as a neural network akin to (Belghazi et al. 2018). We specify our estimator for MI in Section 4, and show how it compares to alternatives in the literature (Nguyen, Wainwright, and Jordan 2008; Belghazi et al. 2018; Poole et al. 2019). The experiments are presented in Section 5. On synthetic data, we validate our estimators by estimating MI on Gaussian variables and by regularizing GAN training as in (Chen et al. 2016). On real data, we explore our estimator in deep MI maximization for learning representation from unlabeled data. Figure 1 shows an overview of all the bounds and related MI estimators discussed in this paper.

2 Lower Bounds on KL Divergences and MI

Consider two probability distributions $\mathbb{P}$ and $\mathbb{Q}$, where $\mathbb{P}$ is absolutely continuous w.r.t. $\mathbb{Q}$. Let p and q be their respective

[*] Tom Sercu is now with FAIR.

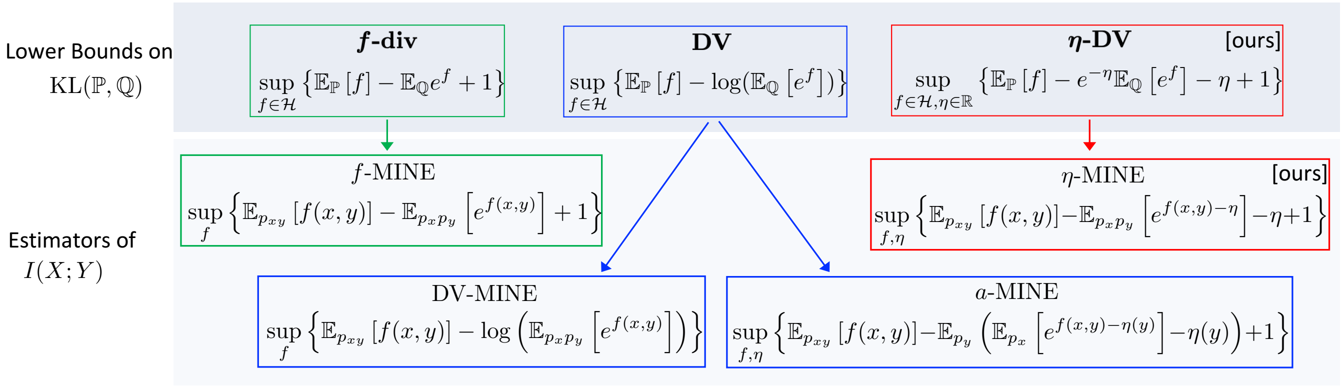
Lower Bounds on
$\mathrm{KL}(\mathbb{P},\mathbb{Q})$
f-div
$\sup_{f\in\mathcal{H}} \left\{\mathbb{E}_{\mathbb{P}}\left[f\right] - \mathbb{E}_{\mathbb{Q}} e^{f} + 1\right\}$
DV
$\sup_{f\in\mathcal{H}} \left\{\mathbb{E}_{\mathbb{P}}\left[f\right] - \log(\mathbb{E}_{\mathbb{Q}}\left[e^{f}\right])\right\}$
η-DV
[ours]
$\sup_{f\in\mathcal{H},\eta\in\mathbb{R}} \left\{\mathbb{E}_{\mathbb{P}}\left[f\right] - e^{-\eta}\mathbb{E}_{\mathbb{Q}}\left[e^{f}\right] - \eta + 1\right\}$
Estimators of
$I(X;Y)$
f-MINE
$\sup_{f} \left\{\mathbb{E}_{p_{xy}}\left[f(x,y)\right] - \mathbb{E}_{p_x p_y}\left[e^{f(x,y)}\right] + 1\right\}$
η-MINE
[ours]
$\sup_{f,\eta} \left\{\mathbb{E}_{p_{xy}}\left[f(x,y)\right] - \mathbb{E}_{p_x p_y}\left[e^{f(x,y)-\eta}\right] - \eta + 1\right\}$
DV-MINE
$\sup_{f} \left\{\mathbb{E}_{p_{xy}}\left[f(x,y)\right] - \log\left(\mathbb{E}_{p_x p_y}\left[e^{f(x,y)}\right]\right)\right\}$
a-MINE
$\sup_{f,\eta} \left\{\mathbb{E}_{p_{xy}}\left[f(x,y)\right] - \mathbb{E}_{p_y}\left(\mathbb{E}_{p_x}\left[e^{f(x,y)-\eta(y)}\right] - \eta(y)\right) + 1\right\}$

Figure 1: Overview of the paper. Top row shows several KL divergence lower bounds for two probability distributions $\mathbb{P}$ and $\mathbb{Q}$. By substituting $\mathbb{P} = p_{xy}$, $\mathbb{Q} = p_x p_y$, and defining f as a neural network, we obtain corresponding MI estimators. $\eta - DV$ and η-MINE are the proposed bound and its derived estimator.

densities defined on $\mathcal{X} \subset \mathbb{R}^d$. Their KL divergence is defined as $\mathrm{KL}(\mathbb{P},\mathbb{Q}) = \mathbb{E}_{x\sim\mathbb{P}} \log\left(\frac{p(x)}{q(x)}\right) = \int p(x)\log\left(\frac{p(x)}{q(x)}\right)dx$. We are interested in the MI between two random variables X, Y where X is defined on $\mathcal{X} \subset \mathbb{R}^{d_x}$, and Y on $\mathcal{Y} \subset \mathbb{R}^{d_y}$. Let p_{xy} be their joint densities and p_x, p_y the marginals of X and Y respectively. The MI is defined as follows:

$$I(X;Y) = \mathrm{KL}(p_{xy}, p_x p_y), \tag{1}$$

which is the KL divergence between the joint density and the product of marginals. Non-parametric estimation of MI from samples is an important problem in science and machine learning. In what follows, we review variational lower bounds on KL to enable such estimation.

Variational Characterization of KL divergence. Let $\mathscr{H}$ be any function space mapping $\mathcal{X}$ to $\mathbb{R}$. The first variational characterization of the KL divergence goes back to (Donsker and Varadhan 1976):

$$\begin{aligned}\mathrm{KL}(\mathbb{P},\mathbb{Q}) \geq D^{\mathscr{H}}_{\mathrm{DV}}(\mathbb{P},\mathbb{Q}) = \sup\{&\mathbb{E}_{x\sim\mathbb{P}} f(x) \\ &- \log(\mathbb{E}_{x\sim\mathbb{Q}} e^{f(x)}) : f \in \mathscr{H}\},\end{aligned} \tag{2}$$

where the equality holds if and only if $f^* = \log(p/q) \in \mathscr{H}$. We refer to this bound as the ***DV** bound*.

The second variational representation was introduced in (Nguyen, Wainwright, and Jordan 2008; Nowozin, Cseke, and Tomioka 2016) and derived through convex duality to be finally stated as follows:

$$\begin{aligned}\mathrm{KL}(\mathbb{P},\mathbb{Q}) \geq D^{\mathscr{H}}_{f}(\mathbb{P},\mathbb{Q}) = 1 + \sup\{&\mathbb{E}_{x\sim\mathbb{P}} f(x) \\ &- \mathbb{E}_{x\sim\mathbb{Q}} e^{f(x)} : f \in \mathscr{H}\},\end{aligned} \tag{3}$$

with equality iif $f^* = \log(p/q) \in \mathscr{H}$. We call this bound the ***f*-div** bound (as in f-divergence).

From Eq. 2 and Eq. 3 we see that the variational bounds are attempting to estimate the log-likelihood ratio $f^* = \log(p/q)$, and the tightness of the bound depends on the representation power of $\mathscr{H}$. In order to compare these two lower bounds, observe that $\log(t) \leq t - 1, t > 0$. Therefore $\log\left(\mathbb{E}_{x\sim\mathbb{Q}} e^{f(x)}\right) \leq \mathbb{E}_{x\sim\mathbb{Q}} e^{f(x)} - 1$, which means that for any function space $\mathscr{H}$ we have:

$$\mathrm{KL}(\mathbb{P},\mathbb{Q}) \geq D^{\mathscr{H}}_{\mathrm{DV}}(\mathbb{P},\mathbb{Q}) \geq D^{\mathscr{H}}_{f}(\mathbb{P},\mathbb{Q}), \tag{4}$$

from which we conclude that the **DV** bound is tighter than the ***f*-div** bound.

Now, given samples $\{x_i, i = 1 \ldots N, x_i \sim \mathbb{P}\}$, $\{y_i, i = 1 \ldots N, y_i \sim \mathbb{Q}\}$, estimating the KL divergence can be done by computing the variational bound from Monte-Carlo simulation. Specifically, for the **DV** bound we have the following estimator:

$$\widehat{D}^{\mathscr{H}}_{\mathrm{DV}}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathscr{H}} \frac{1}{N}\sum_{i=1}^{N} f(x_i) - \log\left(\frac{1}{N}\sum_{i=1}^{N} e^{f(y_i)}\right). \tag{5}$$

The Mutual Information Neural Estimator (MINE) (Belghazi et al. 2018) considered Eq. 5 with the hypothesis class $\mathscr{H}$ being a neural network. For the ***f*-div** bound we have a similar estimator (for which a convex version with an RKHS $\mathscr{H}$ was introduced by (Nguyen, Wainwright, and Jordan 2008)):

$$\widehat{D}^{\mathscr{H}}_{f}(\mathbb{P},\mathbb{Q}) = 1 + \sup_{f\in\mathscr{H}} \frac{1}{N}\sum_{i=1}^{N} f(x_i) - \frac{1}{N}\sum_{i=1}^{N} e^{f(y_i)}. \tag{6}$$

While **DV** bound is tighter than the ***f*-div** bound, in order to learn the function f using stochastic gradient optimization, ***f*-div** is a better fit because the cost function is linear in the expectation, whereas in the **DV** bound, a log non-linearity is applied to the expectation. This non-linearity introduces biases in the mini-batch estimation of the cost function as noted in (Belghazi et al. 2018). In the following, we show how to alleviate this problem and remove the non-linearity at the price of an additional auxiliary variable that will enable better estimation of the **DV** bound.

An η-trick for the DV Bound. We start with the following elementary Lemma, that gives a variational characterization of the log. All proofs are given in the Appendix A.

Lemma 1. *Let* $x > 0$, *we have:* $\log(x) = \min_\eta e^{-\eta}x + \eta - 1$.

Using Lemma 1 we can now linearize the log in the **DV** bound, Eq. 5.

Lemma 2 (η-Donsker-Varadhan)**.** *Let $\mathscr{H}$ be any function space mapping $\mathcal{X}$ to $\mathbb{R}$:*

$$\mathrm{KL}(\mathbb{P},\mathbb{Q}) \geq D^{\mathscr{H}}_{\eta\text{-}DV}(\mathbb{P},\mathbb{Q}) = -\inf\{L(f,\eta){:}f{\in}\mathscr{H},\eta{\in}\mathbb{R}\} \quad (7)$$

$$L(f,\eta) = e^{-\eta}\mathbb{E}_{x\sim\mathbb{Q}}e^{f(x)} - \mathbb{E}_{x\sim\mathbb{P}}f(x) + \eta - 1, \quad (8)$$

We refer to this bound as ***η-DV*** *bound. Note that for $\eta = 0$ we recover the* ***f-div*** *bound.*

Using Lemma 2, we can now rewrite the estimator for the **DV** bound in Eq. 5 as follows:

$$\begin{aligned}\widehat{D}^{\mathscr{H}}_{\mathrm{DV}}(\mathbb{P},\mathbb{Q}) &= -\inf_{f,\eta}\hat{L}(f,\eta)\\ &= -\inf_{f,\eta} e^{-\eta}\frac{1}{N}\sum_{i=1}^{N}e^{f(y_i)} - \sum_{i=1}^{N}f(x_i) + \eta - 1, \quad (9)\end{aligned}$$

which enables unbiased stochastic gradient optimization of the function f. We note that similar variational tricks of nonlinearities have been devised for $g(\eta) = \sqrt{\eta}$ in (Argyriou, Evgeniou, and Pontil 2008; Bach, Jenatton, and Mairal 2011).

3 What Do KALE Learn?

KALE from (Gretton, Sutherland, and Jitkrittum 2019) refers to KL Approximate Lower-bound Estimators introduced earlier (**DV**, ***η*-DV**, and ***f*-div**). In this section we show that KALE estimates, whose witness functions are estimated in RKHS, learn likelihood ratio estimates r in the maximum mean discrepancy sense. Considering the dual of the KL lower bounds optimized in RKHS , we establish that likelihood ratio r appears naturally as the dual variable of the witness function f. See Figure 2 for an illustration.

For simplicity, we consider RKHS with a finite dimensional feature map, i.e., $\mathscr{H} = \{f|f(x) = \langle w, \Phi(x)\rangle, \Phi : \mathcal{X} \to \mathbb{R}^m, w \in \mathbb{R}^m\}$. Now for $f \in \mathscr{H}$, the loss given in Eq. 8 for ***η*-DV** can be rewritten as follows:

$$\begin{aligned}L(f,\eta) &\triangleq L(w,\eta)\\ &= e^{-\eta}\mathbb{E}_{x\sim\mathbb{Q}}e^{\langle w,\Phi(x)\rangle} - \langle w, \mathbb{E}_{x\sim\mathbb{P}}\Phi(x)\rangle + \eta - 1.\end{aligned}$$

Following (Nguyen, Wainwright, and Jordan 2008), we consider the following regularized loss $\mathscr{L}(w,\eta) = L(w,\eta) + \Omega(w)$, and the corresponding sample-based formulation $\hat{\mathscr{L}}(w,\eta) = \hat{L}(w,\eta) + \Omega(w)$, where

$$\hat{L}(w,\eta) = \frac{e^{-\eta}}{N}\sum_{i=1}^{N}e^{\langle w,\Phi(y_i)\rangle} - \left\langle w, \frac{1}{N}\sum_{i=1}^{N}\Phi(x_i)\right\rangle + \eta - 1,$$

and $\Omega(w)$ is a convex regularizer, e.g., $\Omega(w) = \frac{\lambda}{2}\left\|w\right\|_2^2$. The ***η*-DV** primal is defined as:

$$\boldsymbol{\eta}\textbf{-DV}\text{–P} : \min_{w,\eta} L(w,\eta) + \Omega(w), \quad (10)$$

Let $(w^*,\eta^*) \in \arg\min_{w,\eta}\hat{L}(w,\eta) + \Omega(w)$ be the empirical estimator. The KALE estimate is: $\widehat{D}^{\mathscr{H}}_{\mathrm{DV}}(\mathbb{P},\mathbb{Q}) = -\hat{L}(w^*,\eta^*)$. In the following, we show that $\mathscr{L}(w,\eta)$ is jointly convex in (w,η) and derive its dual, which will shed light on the nature of the likelihood ratio estimate and the role of η.

Convex Estimate in RKHS. In Lemma 3 we first establish the convexity of the ***η*-DV** loss function.

Lemma 3. *$\mathscr{L}$ and $\hat{\mathscr{L}}$ are jointly convex in w and η.*

***η*-DV Dual is a Constrained Likelihood Ratio Estimation.** Given a ratio estimate $\hat{r}$ of p/q, the KL divergence is computed as $\mathbb{E}_{\mathbb{Q}}\hat{r}\log(\hat{r})$ (Mohamed and Lakshminarayanan 2016), which can be easily estimated using samples from $\mathbb{Q}$.

In Theorem 1, proven in Appendix A, we show that the dual problem (denoted D), corresponding to the primal minimization problem (denoted P) of ***η*-DV**, reduces to a constrained likelihood ratio estimation problem (denoted C).

Theorem 1. *Let $\Omega^*(.)$ be the Fenchel conjugate of $\Omega(.)$. The* ***η*-DV** *bound restricted to an RKHS amounts to the following regularized convex minimization problem:*
$P = -\left(\min_{w,\eta} L(w,\eta) + \Omega(w)\right)$, with its dual form (D):

$$\min_{r:\mathcal{X}\to\mathbb{R}^+}\max_{\eta}\mathbb{E}_{\mathbb{Q}}r\log(r) + (\eta-1)\left(\mathbb{E}_{\mathbb{Q}}r - 1\right) + \Omega^*\left(\Delta(r)\right),$$

$$\textit{where } \Delta(r) = \mathbb{E}_{x\sim\mathbb{P}}\Phi(x) - \mathbb{E}_{y\sim\mathbb{Q}}r(y)\Phi(y). \quad (11)$$

Noticing that $\eta - 1$ plays the role of Lagrangian multiplier, it is equivalent to the likelihood ratio estimation problem (C):

$$\min_{r>0}\mathbb{E}_{\mathbb{Q}}r(y)\log\left(r(y)\right) + \Omega^*\left(\mathbb{E}_{x\sim\mathbb{P}}\Phi(x) - \mathbb{E}_{y\sim\mathbb{Q}}r(y)\Phi(y)\right)$$

$$\textit{such that } \mathbb{E}_{y\sim\mathbb{Q}}r(y) = 1. \quad (12)$$

Therefore, we have $P = D = C$. Let (w^*,η^*) be an optimizer of P. Let r^* be an optimizer of D, then the KL estimate is:

$$D^{\mathscr{H}}_{DV}(\mathbb{P},\mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}r^*\log(r^*) = -L(w^*,\eta^*),$$

and the KALE witness function $f^*(x') = \langle w^*, \Phi(x')\rangle$ is

$$f^*(x') = \langle\mathbb{E}_{x\sim\mathbb{P}}\Phi(x) - \mathbb{E}_{y\sim\mathbb{Q}}r^*(y)\Phi(y), \Phi(x')\rangle\,.$$

The regularizer $\Omega(w) = \frac{\lambda}{2}\left\|w\right\|^2$ can now be given the following interpretation based on the results of Theorem 1. Recall the definition of the MMD (Gretton et al. 2012) between distributions using an RKHS:

$$\mathrm{MMD}_{\Phi}(\mathbb{P},\mathbb{Q}) = \left\|\mathbb{E}_{x\sim\mathbb{P}}\Phi(x) - \mathbb{E}_{y\sim\mathbb{Q}}\Phi(y)\right\|. \quad (13)$$

Replacing the Fenchel conjugate $\Omega^*(.)$ by its expression in Theorem 1, we see that ***η*-DV** is equivalent to the following dual (***η*-DV**–D):

$$\min_{r>0}\mathbb{E}_{\mathbb{Q}}r\log(r) + (\eta-1)\left(\mathbb{E}_{\mathbb{Q}}r - 1\right) + \frac{1}{2\lambda}\mathrm{MMD}^2_{\Phi}(rq,p) \quad (14)$$

or as a constrained ratio estimation problem (***η*-DV**–C):

$$\min_{r>0}\mathbb{E}_{\mathbb{Q}}r\log(r) + \frac{1}{2\lambda}\mathrm{MMD}^2_{\Phi}(rq,p) \;\text{ s.t. } \mathbb{E}_{y\sim\mathbb{Q}}r(y) = 1. \quad (15)$$

Hence, it is clear now that the ***η*-DV** optimization problem is equivalent to the constrained likelihood ratio estimation problem r given in Eq. 15, where the ratio is estimated using the MMD distance in the RKHS between rq and p. It is also easy to see that p/q is a feasible point and for the universal feature map $\mathrm{MMD}_{\Phi}(rq,p) = 0$ iif $r = p/q$, therefore, for a universal kernel, p/q is optimal and we recover the KL divergence for $r = p/q$. When comparing ***η*-DV**–D (Eq. 14) and ***η*-DV**–C (Eq. 15), we see that $\eta - 1$ plays the role of a

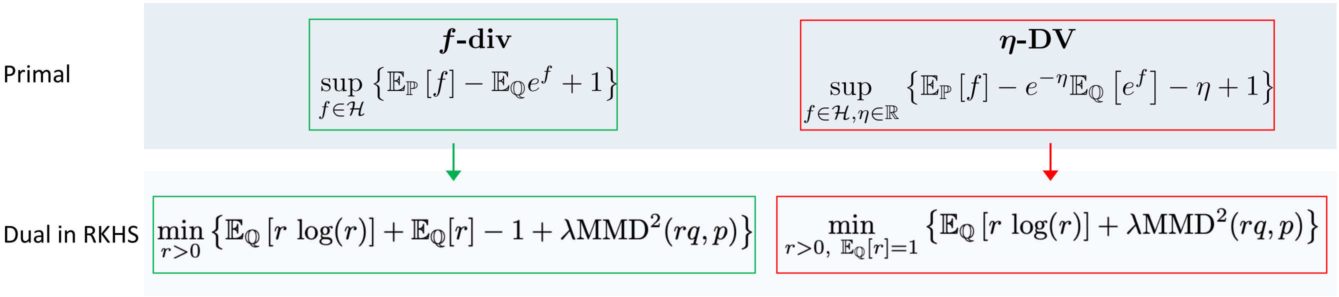
Primal
$\boldsymbol{f}$-div
$\sup_{f\in\mathcal{H}} \{\mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{Q}} e^{f} + 1\}$
$\boldsymbol{\eta}$-DV
$\sup_{f\in\mathcal{H},\eta\in\mathbb{R}} \{\mathbb{E}_{\mathbb{P}}[f] - e^{-\eta}\mathbb{E}_{\mathbb{Q}}[e^{f}] - \eta + 1\}$
Dual in RKHS
$\min_{r>0} \{\mathbb{E}_{\mathbb{Q}}[r \log(r)] + \mathbb{E}_{\mathbb{Q}}[r] - 1 + \lambda \mathrm{MMD}^2(rq, p)\}$
$\min_{r>0,\ \mathbb{E}_{\mathbb{Q}}[r]=1} \{\mathbb{E}_{\mathbb{Q}}[r \log(r)] + \lambda \mathrm{MMD}^2(rq, p)\}$

Figure 2: Comparison of dual formations of $\eta-DV$ and $f-div$ in RKHS. Here, r is an estimator for density ratio p/q and $\mathrm{MMD}_{\Phi}(\mathbb{P},\mathbb{Q}) = \|\mathbb{E}_{x\sim\mathbb{P}}\Phi(x) - \mathbb{E}_{y\sim\mathbb{Q}}\Phi(y)\|$. Both duals have a relative entropy term, but $f-div$ bound does not impose a normalization constraint on the ratio, which biases the estimate, while $\eta-DV$ uses η as Lagrangian multiplier to impose the constraint $\mathbb{E}_{\mathbb{Q}}[r]=1$ and ensure density normalization.

Lagrangian that ensures that rq is indeed a normalized distribution. In practice, we solve the primal problem (P), while the dual problem (D) and its equivalent constrained form (C) explains why this formulation estimates the KL divergence and reveals η as a **Lagrangian** multiplier, enforcing a normalization constraint. Let r^* be the solution, then:

$$D^{\mathscr{H}}_{\mathrm{DV}}(\mathbb{P},\mathbb{Q}) = \mathbb{E}_{\mathbb{Q}} r^* \log(r^*) = -L(w^*, \eta^*). \qquad (16)$$

KALE's witness function f^* in Theorem 1 shares similarity with the witness function with the MMD. It is MMD witness function of a rescaled distribution $\mathbb{Q}$ with likelihood ratio r^*. Ratio estimation via MMD matching for covariate shift appeared in (Gretton et al. 2009; Sugiyama et al. 2008).

Comparison to $\boldsymbol{f}$-div. We also restrict $\boldsymbol{f}$**-div** bound to an RKHS, which is equivalent to the following ratio estimation (follows from the proof of Theorem 1 by eliminating max on η, and setting $\eta = 0$), and is consistent with the results of (Nguyen, Wainwright, and Jordan 2008) (see Eq. 51 therein):

$$\boldsymbol{f}\textbf{-div}: \min_{r>0} \mathbb{E}_{\mathbb{Q}} r \log(r) + (1 - \mathbb{E}_{\mathbb{Q}} r) + \frac{1}{2\lambda}\mathrm{MMD}^2_{\Phi}(rq, p).$$

Let r^* be the optimum, then the KL divergence can be estimated as follows:

$$D^{\mathscr{H}}_{f}(\mathbb{P},\mathbb{Q}) = \mathbb{E}_{\mathbb{Q}} r^* \log(r^*) + (1 - \mathbb{E}_{\mathbb{Q}} r^*). \qquad (17)$$

Comparing $D^{\mathscr{H}}_{\mathrm{DV}}(\mathbb{P},\mathbb{Q})$ and $D^{\mathscr{H}}_{f}(\mathbb{P},\mathbb{Q})$ we see that they both have a relative entropy term but the $\boldsymbol{f}$**-div** bound does not impose the normalization constraint on the ratio, which biases the estimate.

Empirical Estimation. Note that in Theorem 1, if we replace the loss L by its empirical counterpart $\hat{L}$ from Eq. 9, the equivalent Dual takes the following form:

$$\min_{r_i>0} \frac{1}{N}\sum_{i=1}^{N} r_i \log(r_i) + \Omega^*(\frac{1}{N}\sum_{i=1}^{N}\Phi(x_i) - \frac{1}{N}\sum_{i=1}^{N} r_i \Phi(y_i))$$

$$\text{subject to: } \frac{1}{N}\sum_{i=1}^{N} r_i = 1, \qquad (18)$$

and the KALE is given by:

$$\widehat{D}^{\mathscr{H}}_{\mathrm{DV}}(\mathbb{P},\mathbb{Q}) = \frac{1}{N}\sum_{i=1}^{N} r^*_i \log(r^*_i) = -\hat{L}(w^*, \eta^*).$$

From RKHS to Neural Estimation. One shortcoming of the RKHS approach is that the method depends on the choice of feature map Φ. We propose to learn Φ as a deep neural network, as in MINE (Belghazi et al. 2018). Given samples x_i from $\mathbb{P}$, y_i from $\mathbb{Q}$, the KL estimation becomes:

$$\widehat{D}^{\mathrm{NN}}_{\mathrm{DV}}(\mathbb{P},\mathbb{Q}) = -\left(\min_{\Phi} \hat{L}_{\Phi}(w,\eta)\right), \qquad (19)$$

which can be solved using BCD on (w, η, Φ). Note that if $\Phi(\cdot)$ is known and fixed, then optimization problem in Eq. 19 becomes convex in η and w. We refer to this bound as $\boldsymbol{\eta}$**-DV**–convex.

```
Algorithm 1 $\eta$-MINE (Stochastic BCD )
Inputs: $X, Y$ dataset $X \in \mathbb{R}^{N\times d_x}, Y \in \mathbb{R}^{N\times d_y}$, such that $(x_i = X_{i,.}, y_i = Y_{i,.}) \sim p_{xy}$
Hyperparameters: $\alpha_\eta, \alpha_\theta$ (learning rates), $n_c$ (number of critic updates)
Initialize $\eta$, $\theta$ parameter of the neural network $f_\theta$
for $i = 1 \ldots$ Maxiter do
  for $j = 1 \ldots n_c$ do
    Fetch a minibatch of size $N$ $(x_i, y_i) \sim p_{xy}$
    Fetch a minibatch of size $N$ $(x_i, \tilde{y}_i) \sim p_x p_y$
    Evaluate $\hat{L}(f_\theta, \eta)$
    Stochastic Gradient step on $\theta$: $\theta \leftarrow \theta - \alpha \frac{\partial \hat{L}(f_\theta,\eta)}{\partial \theta}$
  end for
  Update $\eta$: $\eta \leftarrow \eta - \alpha_\eta \frac{\partial \hat{L}(f_\theta,\eta)}{\partial \eta}$
end for
Output: $f_\theta, \eta,$ $\hat{I}^{\mathscr{H}}_{\eta\text{-DV}}(X,Y) = \frac{1}{N}\sum_{i=1}^{N} f_\theta(x_i,y_i) - e^{-\eta}\frac{1}{N}\sum_{i=1}^{N} e^{f_\theta(x_i,\tilde{y}_i)} - \eta + 1$
```

Observations about what Neural KL/MI Estimators Learn: 1) *Ratio estimation via Feature Matching.* KL divergence estimation with variational bounds boils down to a ratio estimation using a form of MMD matching. 2) *Choice of Feature/Architecture.* The choice of feature space or architecture of the network introduces bias in ratio estimation; also observed in (Tschannen et al. 2019). 3) *Ratio Normalization.* $\boldsymbol{\eta}$**-DV** bound introduces a normalization constraint on the ratio ensuring a better estimate.

MI Estimator (bound)	**Loss to minimize $\hat{L}$**	**Constraints**
DV-MINE (**DV**)	$\log\left(\frac{1}{N}\sum_{i=1}^{N} e^{f(x_i,\tilde{y}_i)}\right) - \frac{1}{N}\sum_{i=1}^{N} f(x_i,y_i)$	f is a DNN
Unbiased DV-MINE	$\frac{\frac{1}{N}\sum_{i=1}^{N} e^{f(x_i,\tilde{y}_i)}}{m} - \frac{1}{N}\sum_{i=1}^{N} f(x_i,y_i)$	f is a DNN, m running avg. of $\frac{1}{N}\sum_{i=1}^{N} e^{f(x_i,\tilde{y}_i)}$
f-MINE ($\boldsymbol{f}$-div)	$\frac{1}{N}\sum_{i=1}^{N} e^{f(x_i,\tilde{y}_i)} - \frac{1}{N}\sum_{i=1}^{N} f(x_i,y_i) - 1$	f is in RKHS (convex) or f is a DNN
a-MINE (**DV**)	$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{e^{f(x_i,\tilde{y}_i)}}{a(\tilde{y}_i)} + \log(a(\tilde{y}_i))\right) - \frac{1}{N}\sum_{i=1}^{N} f(x_i,y_i) - 1$	a is a DNN, $a > 0$ f is a DNN
InfoNCE (**-**)	$\frac{1}{N}\sum_{i=1}^{N}\left(\log\frac{1}{N}\sum_{j=1}^{N} e^{f(x_i,\tilde{y}_j)} - f(x_i,y_i)\right)$	f is a DNN
η-MINE (ours:$\boldsymbol{\eta}$-**DV**)	$e^{-\eta}\frac{1}{N}\sum_{i=1}^{N} e^{f(x_i,\tilde{y}_i)} - \sum_{i=1}^{N} f(x_i,y_i) + \eta - 1$	f is in RKHS or f is a DNN; $\eta \in \mathbb{R}$
η-MINE-convex (ours:$\boldsymbol{\eta}$-**DV**)	$e^{-\eta}\frac{1}{N}\sum_{i=1}^{N} e^{\langle w,\Phi(x_i,\tilde{y}_i)\rangle} - \sum_{i=1}^{N}\langle w,\Phi(x_i,y_i)\rangle + \eta - 1$	$\Phi(\cdot)$ is fixed $w \in \mathbb{R}^{\text{dim}(\Phi)}$, $\eta \in \mathbb{R}$

Table 1: Given iid samples from marginals $x_i \sim p_x$ and $\tilde{y}_i \sim p_y$ and samples from the joint $(x_i, y_i) \sim p_{xy}$, we list some MI estimators, corresponding variational bounds, associated losses, and constraints on the function space. MI estimators include biased and unbiased DV-MINE from DV (Belghazi et al. 2018), f-MINE from f-div (Nguyen, Wainwright, and Jordan 2008; Nowozin, Cseke, and Tomioka 2016), InfoNCE (Oord, Li, and Vinyals 2018), a-MINE from DV (Poole et al. 2019) and η-MINE, η-MINE-convex from $\eta$$-DV$(ours).

4 η-DV Mutual Information Estimation

We now specialize the KL estimators given in Section 3 for the task of MI estimation. Given a function space $\mathscr{H}$ defined on $\mathcal{X} \times \mathcal{Y}$,

$$I(X;Y) \geq I_{\text{DV}}^{\mathscr{H}}(X;Y) \geq I_{f\text{-div}}^{\mathscr{H}}(X;Y)$$
$$= \sup_{f\in\mathscr{H}} \mathbb{E}_{p_{xy}} f(x,y) - \mathbb{E}_{p_x}\mathbb{E}_{p_y} e^{f(x,y)}, \qquad (20)$$

where $I_{\text{DV}}^{\mathscr{H}} = \sup_{f\in\mathscr{H}} \mathbb{E}_{p_{xy}} f(x,y) - \log\left(\mathbb{E}_{p_x}\mathbb{E}_{p_y} e^{f(x,y)}\right)$.

Equivalently, with the η-trick we can rewrite

$$I_{\text{DV}}^{\mathscr{H}} = -\inf_{f\in\mathscr{H},\eta} e^{-\eta}\mathbb{E}_{p_x p_y} e^{f(x,y)} - \mathbb{E}_{p_{xy}} f(x,y) + \eta - 1.$$

Now, given iid samples from marginals $x_i \sim p_x$, and $\tilde{y}_i \sim p_y$, for $i = 1 \ldots N$, and samples from the joint $(x_i, y_i) \sim p_{xy}$, for $i = 1 \ldots N$, we can estimate the MI as follows $\hat{D}_{\text{DV}}^{\mathscr{H}}(p_{x,y}, p_x p_y)$ given in the expression below:

$$-\inf_{f\in\mathscr{H},\eta} e^{-\eta}\frac{1}{N}\sum_{i=1}^{N} e^{f(x_i,\tilde{y}_i)} - \frac{1}{N}\sum_{i=1}^{N} f(x_i,y_i) + \eta - 1.$$

When $\mathscr{H}$ is an RKHS, η can be seen as a Lagrangian ensuring that the ratio estimation $r(x,y)$ of $\frac{p_{xy}}{p_x p_y}$ is normalized to one when integrated on $p_x p_y$, i.e., η is a Lagrangian associated with the constraint $\int_{\mathcal{X}\times\mathcal{Y}} r(x,y) p_x(x) p_y(y) dx dy = 1$. Table 1 is a review of other variational bounds for MI based on $\boldsymbol{f}$-div, **DV**, and $\boldsymbol{\eta}$-**DV** bounds; a-MINE from (Poole et al. 2019) is discussed next.

We establish in Appendix C that the **DV** bound can be made tighter and we land on a-MINE estimator from (Poole et al. 2019) that we refer to as $I_{a\text{-DV}}^{\mathscr{H}}(X;Y)$. We then derive the following hierarchy of lower bounds: $I(X;Y) \geq I_{a\text{-DV}}^{\mathscr{H}}(X;Y) \geq I_{\eta\text{-DV}}^{\mathscr{H}}(X;Y) \geq I_{f\text{-div}}^{\mathscr{H}}(X;Y)$. We discuss in Appendix C that while this may suggest a tighter bound, than $\boldsymbol{\eta}$-**DV**, it is prone to higher estimation errors since a-MINE estimates a function $a(y)$ as it can be seen in Table 1, while $\boldsymbol{\eta}$-**DV** estimates a scalar.

Sample Complexity. We discuss in Appendix B the sample complexity of η-MINE and show that by a reduction of the DV bound to $\boldsymbol{f}$-div bound, the convergence results from (Nguyen, Wainwright, and Jordan 2008) apply and we have a sample complexity of $O(1/\sqrt{N})$.

Algorithm 1 outlines steps of η-MINE for MI estimation.

5 Experiments

Proposed η-MINE MI estimator is compared to existing baselines on few applications using synthetic and real data.

MI estimation. We compared different MI estimators on three synthetically generated Gaussian datasets [5K training and 1K testing samples]. Each evaluated MI estimator was run 10 times and average performance (± standard deviation) is shown at the top of Fig. 3. Clearly, MI estimation in high dimensions is a challenging task, where the estimation error for all methods increases as the data dimensionality grows [red line shows true MI value]. Nevertheless, the proposed η-MINE achieves on average more accurate results compared to existing baselines. Moreover, the convex formulation of η-MINE has overall a better performance and fast convergence rate. This estimator has a linear witness function defined as $f(\cdot) = \langle w, \Phi(\cdot)\rangle$ using pre-trained fixed feature map $\Phi(\cdot)$ from regular η-MINE. In the experiments that follow, we compare the proposed estimator η-MINE to DV-MINE, as a main baseline approach.

MI-regularized GAN. We investigate GAN training improvements with MI, especially its diminishing mode collapse, as addressed in (Belghazi et al. 2018) in Section 5.1. (Belghazi et al. 2018) uses a 25-Gaussian dataset to show improvements on GAN clustering by using MI objective for

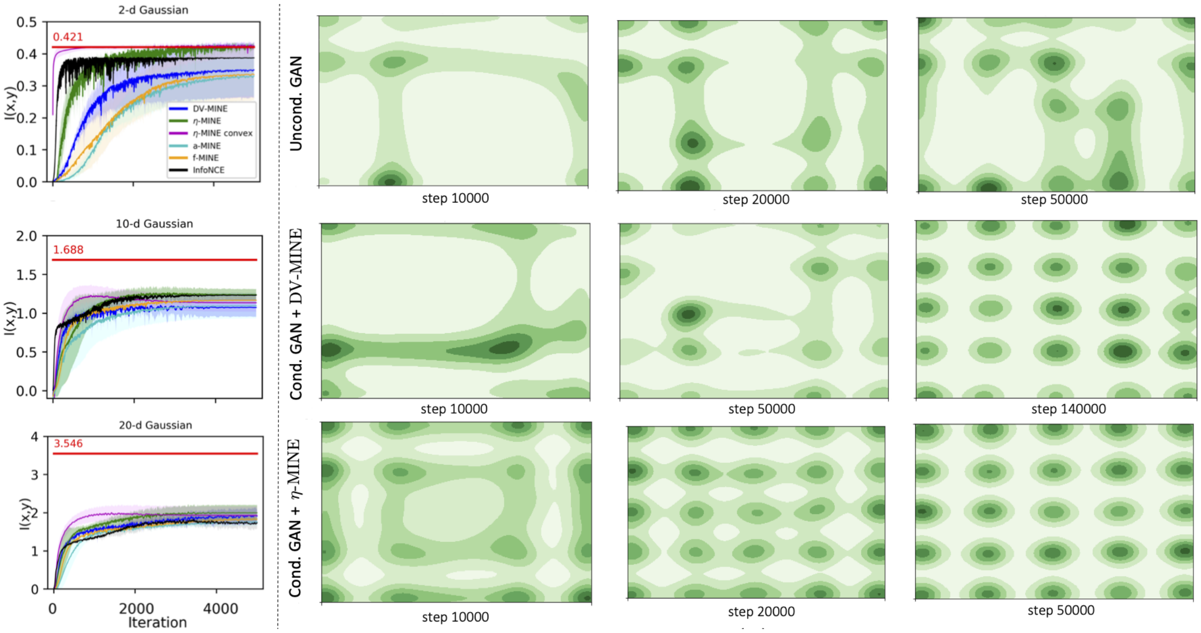

Figure 3: Performance of different MI estimators on synthetic Gaussian data. Left: *MI estimation*. Data was sampled from 2-, 10- and 20-dimensional Gaussian distributions with random means and random symmetric, positive-definite covariance matrices (i.e., random dependencies, which is a difficult scenario). As we increase data complexity, difference between estimators decreases, although we observed that η-MINE (or its convex extension) on average performed better than baseline methods, converging to the true MI [red line] faster. Right: *MI for GAN regularization*. Top row: unconditional GAN baseline fails at capturing all 25 modes in the Gaussian mixture. Middle row: MI-regularized conditional GAN using DV-MINE (Belghazi et al. 2018) converges after 140K steps of the generator. We found this estimator to be sensitive to the hyper-parameters and unstable. Bottom row: MI-regularized conditional GAN using η-MINE; the model converges in 50K steps.

regularization. As in InfoGAN (Chen et al. 2016), the conditional generator G is supplied with random codes c along with noise z; we maximize the mutual information between $I(G(z,c),c)$ using η-MINE estimators. In Fig. 3, we establish that η-MINE recovers all modes within fewer steps than DV-MINE and with a more stable training.

Self-supervised: Deep InfoMax. In unsupervised or self-supervised learning, the objective is to build a model without relying on labeled data but using an auxiliary task to learn informative features that can be useful for various downstream tasks. Here, we evaluate the effectiveness of the proposed η-MINE estimator for unsupervised feature learning using the recently proposed Deep InfoMax method from (Hjelm et al. 2019). For feature representation we used an encoder similar to DCGAN (Radford, Metz, and Chintala 2015), shown in Fig. 4.a and evaluated results on CIFAR10 and STL10 datasets (STL10 images were scaled down to match CIFAR10 resolution).

The encoder is trained by maximizing MI $I(x',y')$ between features from one shallow layer l ($x' = E_l(x)$) and a deeper layer k ($y' = E_k(y)$). We examined different layer combinations and found that the encoder composed of only the first two convolutional layers give the best performance on the downstream tasks. As shown in Fig. 4.b the encoder features are passed through additional trainable neural network layers, whose job is to build a classifier $f(x',y')$, discriminating cases when x' are y' are coming from the same image and cases when x' are y' are unrelated. Finally, we attach a linear layer to the pre-trained and now fixed encoder (see Fig. 4.c) to perform supervised training. Table 2 presents the results for two MI estimators: η-MINE and DV-MINE, whose loss functions are listed in Table 1. As can be seen, η-MINE-based pre-training performs competitively with DV-MINE, achieving overall better results on both datasets, showing practical benefits of the proposed approach.

Self-Supervised: Jigsaws with MI. The self-supervision *Jigsaw* pretext task (Noroozi and Favaro 2016) aims at solving an image jigsaw puzzle by predicting a scrambling per-

	Test					
	C10			S10		
Train	DV	η	Sup.	DV	η	Sup.
C10	**77.5**	74.8	84.2	55.1	**56.4**	-
S10	67.5	**68.3**	-	61.8	**63.3**	61.3

Table 2: Classification accuracy (top1) results on CIFAR10 (C10) and STL10 (S10) for unsupervised pre-training task with DV-MINE and η-MINE using encoder in Fig. 4.b. For reference we list results for supervised (CE) training using full encoder in Fig. 4.a. η-MINE-based pre-training achieves better results, outperforming supervised model on S10.

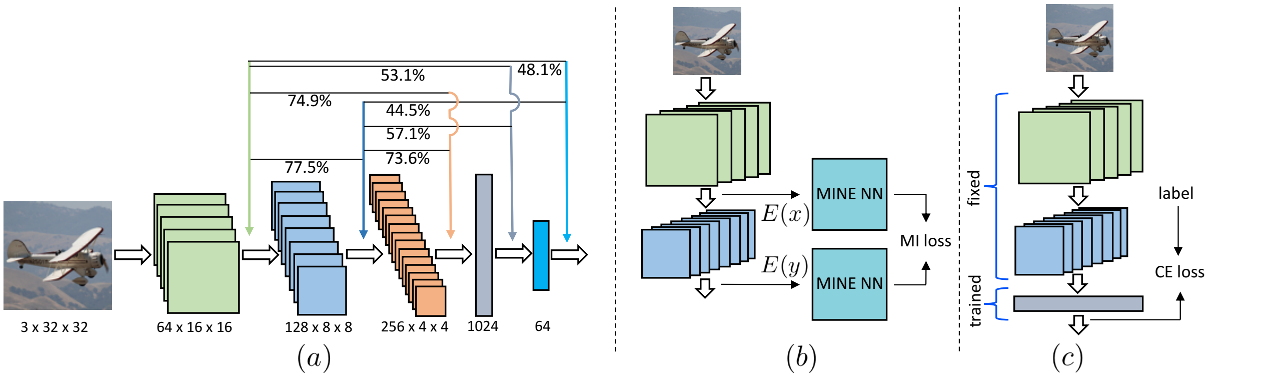

Figure 4: (a) Encoder architecture and classification accuracy (top1) on CIFAR-10 dataset for different pre-trained encoders. Each number represents test accuracy of the system trained by maximizing MI between features from layers pointed by the corresponding arrows. Interestingly, the highest accuracy was obtained by pre-training encoder composed of just the first two convolutional layers (see (b) and (c) for details of this process). (b) Model pre-training by maximizing MI between features from different layers [additionally transformed by a neural net, aimed at constructing witness function $f(\cdot)$]. (c) After pre-training, we fix encoder and attach a trainable linear layer to perform supervised tasks.

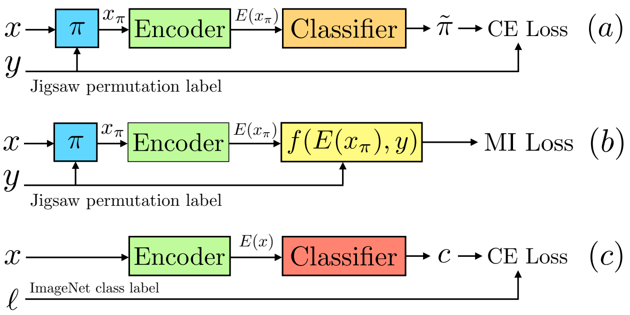

Figure 5: (a) Jigsaw CE training. (b) Jigsaw MI training. (c) ImageNet Classification CE training.

	DV-MINE	η-MINE	CE
top1	8.5 ± 1.3	**11.0** ± **1.1**	12.9 ± 0.3
top5	20.0 ± 2.5	**24.1** ± **2.2**	28.1 ± 0.6
top10	27.8 ± 3.1	**32.7** ± **2.2**	37.7 ± 0.6

Table 3: ImageNet (10% subset) classification accuracies (in %). DV-MINE and η-MINE use fixed Encoder from MI training. CE uses a CE Jigsaw Encoder. Means ± std. deviations over 8 models from different initialization are reported.

mutation π. From image X with $x = \{x_1 \dots x_9\}$ 3×3 jigsaw patches, and permutation $y = \pi_i : [1,9] \rightarrow [1,9]$, scrambled patches $x_{\pi_i} = \{x_{\pi_i(1)} \dots x_{\pi_i(9)}\}$ generate the puzzle. Each patch is fed to encoder E which must learn meaningful representations so a classifier C_J can solve the puzzle by predicting the scrambling order π_i (Noroozi and Favaro 2016) (see Fig. 5.a). While this standard formulation relies on a CE-based classification of the permutation, we propose to use MI Jigsaw, where an encoder E is trained to maximize $I(E(x_{\pi_i}); \pi_i)$ by using MI estimators DV- and η-MINE, as seen in Fig. 5.b. A patch preprocessing similar to (Kolesnikov, Zhai, and Beyer 2019) avoids shortcuts based on color aberration, patch edge matching, etc.; for details of our implementation, see Appendix E. All models are built on a 10% subset of ImageNet (128K train., 5K val., 1K classes) as proposed by (Kolesnikov, Zhai, and Beyer 2019). This is a larger set than Tiny ImageNet (200 classes) used in many publications. E is a ResNet50 for all our experiments. In our ImageNet target classification task, E from CE and MI Jigsaw trainings are frozen (at 200 epochs) and followed by linear classifier C (Fig. 5.c); an adequate setup for comparing encoders as argued by (Kolesnikov, Zhai, and Beyer 2019).

Table 3 reports best accuracies for all models on target task for Cs trained for exactly 200 epochs. For all results, E is trained from Jigsaw task (CE or MI) and frozen, with *only* C trained as in (Kolesnikov, Zhai, and Beyer 2019). DV- and η-MINE share the same f architecture. η-MINE gives better accuracy performance compared to DV-MINE. CE model trained from a Jigsaw-supervised encoder E provides an upper-bound for supervised performance for E from the Jigsaw task. Despite CE being better than η- and DV-MINE, η-MINE does a respectable job at learning a representation space for the target task, better than DV-MINE. Details of the Jigsaw experiments are given in Appendix E.

6 Conclusion

In this paper, we introduced a new lower bound on the KL divergence and demonstrated how it can improve the estimation of MI using neural networks. Theoretically, we proved that the dual of our η-DV formulation reduces to a constrained likelihood ratio estimation. In practice, the stability of η-MINE is due to its unbiased gradient. We tested our estimator η-MINE on synthetic data and applied it on various real-world tasks where MI can be used as a regularizer, or as an objective in self-supervised learning. We used η-MINE in unsupervised learning of representations through MI maximization and by solving Jigsaw puzzles.

References

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2017. Deep Variational Information Bottleneck. In *ICLR*.

Alemi, A. A.; Poole, B.; Fischer, I.; Dillon, J. V.; Saurous, R. A.; and Murphy, K. 2018. Fixing a Broken ELBO. In *International Conference on Machine Learning*.

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2008. Convex Multi-task Feature Learning. *Mach. Learn.* .

Bach, F.; Jenatton, R.; and Mairal, J. 2011. *Optimization with Sparsity-Inducing Penalties (Foundations and Trends(R) in Machine Learning)*. Hanover, MA, USA: Now Publishers Inc. ISBN 160198510X, 9781601985101.

Barber, D.; and Agakov, F. V. 2003. The IM Algorithm: A Variational Approach to Information Maximization. In *NIPS*.

Belghazi, M. I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, D. 2018. Mutual Information Neural Estimation. In *Proceedings of the 35th International Conference on Machine Learning*, 531–540.

Bell, A. J.; and Sejnowski, T. J. 1995. An Information-maximization Approach to Blind Separation and Blind Deconvolution. *Neural Comput.* .

Blei, D. M.; Kucukelbir, A.; and McAuliffe, J. D. 2017. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association* abs/1601.00670.

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

Doersch, C.; Gupta, A.; and Efros, A. A. 2015. Unsupervised Visual Representation Learning by Context Prediction. *2015 IEEE International Conference on Computer Vision (ICCV)* .

Donsker, M.; and Varadhan, S. 1976. Asymptotic evaluation of certain Markov process expectations for large time. *Communications on Pure and Applied Mathematics* .

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A Kernel Two-sample Test. *JMLR* .

Gretton, A.; Smola, A.; Huang, J.; Schmittfull, M.; Borgwardt, K.; and Schölkopf, B. 2009. Covariate Shift by Kernel Mean Matching. In *Dataset Shift in Machine Learning*. MIT press.

Gretton, A.; Sutherland, D.; and Jitkrittum, W. 2019. Interpretable comparison of distributions and models. In *NIPS*.

Han, Y.; Jiao, J.; Weissman, T.; and Wu, Y. 2017. Optimal rates of entropy estimation over Lipschitz balls. *arXiv preprint arXiv:1711.02141* .

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2019. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*.

Hu, W.; Miyato, T.; Tokui, S.; Matsumoto, E.; and Sugiyama, M. 2017. Learning discrete representations via information maximizing self-augmented training. In *Proceedings of the 34th International Conference on Machine Learning*.

Kandasamy, K.; Krishnamurthy, A.; Poczos, B.; Wasserman, L.; et al. 2015. Nonparametric von mises estimators for entropies, divergences and mutual informations. In *Advances in Neural Information Processing Systems*, 397–405.

Kolchinsky, A.; Tracey, B. D.; and Wolpert, D. H. 2017. Nonlinear information bottleneck. *arXiv preprint arXiv:1705.02436* .

Kolesnikov, A.; Zhai, X.; and Beyer, L. 2019. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1920–1929.

Kraskov, A.; Stögbauer, H.; and Grassberger, P. 2004a. Estimating mutual information. *Physical review. E, Statistical, nonlinear, and soft matter physics* .

Kraskov, A.; Stögbauer, H.; and Grassberger, P. 2004b. Estimating mutual information. *Physical review E* 69(6).

Krause, A.; Perona, P.; and Gomes, R. G. 2010. Discriminative Clustering by Regularized Information Maximization. In *Advances in Neural Information Processing Systems 23*.

Mohamed, S.; and Lakshminarayanan, B. 2016. Learning in Implicit Generative Models. *arXiv:1610.03483* .

Moon, K. R.; Sricharan, K.; and Hero, A. O. 2017. Ensemble estimation of mutual information. In *IEEE International Symposium on Information Theory*, 3030–3034.

Nguyen, X.; Wainwright, M. J.; and Jordan, M. I. 2008. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *NIPS*.

Noroozi, M.; and Favaro, P. 2016. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. *Lecture Notes in Computer Science* 69–84.

Nowozin, S.; Cseke, B.; and Tomioka, R. 2016. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization. In *NIPS*.

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

Poole, B.; Ozair, S.; Van Den Oord, A.; Alemi, A.; and Tucker, G. 2019. On variational bounds of mutual information. In *International Conference on Machine Learning*, 5171–5180. PMLR.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434* .

Singh, S.; and Póczos, B. 2017. Nonparanormal information estimation. In *Proceedings of the International Conference on Machine Learning*, 3210–3219.

Sriperumbudur, B. K.; Fukumizu, K.; Gretton, A.; Scholkopf, B.; and Lanckriet, G. R. G. 2009. On integral probability metrics, ϕ-divergences and binary classification. In *arXiv preprint arXiv:0901.2698.*

Sugiyama, M.; Nakajima, S.; Kashima, H.; Buenau, P. V.; and Kawanabe, M. 2008. Direct Importance Estimation with

Model Selection and Its Application to Covariate Shift Adaptation. In *Advances in Neural Information Processing Systems 20*, 1433–1440. NeurIPS.

Tishby, N.; Pereira, F. C.; and Bialek, W. 1999. The Information Bottleneck Method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, 368–377.

Tschannen, M.; Djolonga, J.; Rubenstein, P. K.; Gelly, S.; and Lucic, M. 2019. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625* .

Zeng, X.; Xia, Y.; and Tong, H. 2018. Jackknife approach to the estimation of mutual information. *Proceedings of the National Academy of Sciences* 115(40): 9956–9961.