# Combating the Instability of Mutual Information-based Losses via Regularization


## Abstract

Notable progress has been made in numerous fields of machine learning based on neural network-driven mutual information (MI) bounds. However, utilizing the conventional MI-based losses is often challenging due to their practical and mathematical limitations. In this work, we first identify the symptoms behind their instability: (1) the neural network not converging even after the loss seemed to converge, and (2) saturating neural network outputs causing the loss to diverge. We mitigate both issues by adding a novel regularization term to the existing losses. We theoretically and experimentally demonstrate that added regularization stabilizes training. Finally, we present a novel benchmark that evaluates MI-based losses on both the MI estimation power and its capability on the downstream tasks, closely following the pre-existing supervised and contrastive learning settings. We evaluate six different MI-based losses and their regularized counterparts on multiple benchmarks to show that our approach is simple yet effective.


## 1 INTRODUCTION

Identifying a relationship between two variables of interest is one of the key problems in mathematics, statistics, and machine learning [Goodfellow et al., 2014, Ren et al., 2015, He et al., 2016, Vaswani et al., 2017]. One of the fundamental approaches is information theory-based measurement, namely the measure of mutual information (MI). Due to its mathematical soundness and the rise of deep learning, many have designed differentiable MI-based losses for neural networks. Some utilize the MI-based losses to bridge the gap between latent variables and representations in generative adversarial networks [Nowozin et al., 2016, Chen et al., 2016, Belghazi et al., 2018, Oord et al., 2018, Hjelm et al., 2019], where others introduce MI-based methodologies identifying the relationship between input, output, and hidden variables [Tishby and Zaslavsky, 2015, Shwartz-Ziv and Tishby, 2017, Saxe et al., 2019]. Furthermore, recent self-supervised losses use contrastive losses, where its origin can be traced back to MI-based losses [Cheng et al., 2020, Hénaff, 2020, Chuang et al., 2020].

https://www.overleaf.com/project/621070228159485cf786a1f3

Although many have shown computational tractability and usefulness of MI-based losses, others still struggle with their instability during optimization. Contrastive learning literature with MI-based losses such as Chen et al. [2020], He et al. [2020] use huge batch sizes to reduce the variance of losses. Bardes et al. [2021] adds a regularization term to the neural network embeddings to stabilize the training. McAllester and Stratos [2018] and Song and Ermon [2020] 'further provide theoretical limitations of variational MI estimators, arguing that the limited batch size induces a MI estimation variance too large to handle. We argue that mitigating the variance of MI-based losses is critical for stabilizing training, where it is well known that more stable optimization of neural networks yields better predictive performance on the downstream tasks [Rothfuss et al., 2019, Bear and Cushman, 2020, Chavdarova et al., 2019, Richter et al., 2020, Zeng et al., 2020, Colombo et al., 2021].

In this paper, we concentrate on identifying the cause behind the instability of MI-based losses and propose a simple yet effective regularization method that can be applied to various MI-based losses. We start by analyzing the behaviors of two MI estimators; the MI Neural Estimator (MINE) loss [Belghazi et al., 2018] and Nguyen-Wainwright-Jordan loss (NWJ) loss [Nguyen et al., 2010]. We identify two distinctive behaviors that induce instability during training, drifting and exploding neural network outputs. Based on these observations, we design two novel dual representations of the KL-divergence called Regularized Donsker-Varadhan representation (ReDV) and Regularized NWJ representation (ReNWJ). We show theoretically and experimentally

*Submitted to the 38th Conference on Uncertainty in Artificial Intelligence* (UAI 2022). **To be used for reviewing only**.

that adding our regularizer term suppresses two behaviors of drifting and exploding, avoiding instability during training. Finally, we design a novel benchmark that bridges the gap between variational MI estimators and real-world tasks, whereas previous works either do not directly show the MI estimation performance or evaluate only on toy problems. We reformulate both the supervised and the contrastive learning problem [Chen et al., 2020, He et al., 2020, Khosla et al., 2020] as MI estimation problems and show that our regularization yields better performance on both perspectives, downstream task and MI estimation performance.

## 2 BACKGROUND & RELATED WORKS

**Definition of MI** The mutual information between two random variables $X$ and $Y$ is defined as

$$\begin{aligned} I(X,Y) &= D_{\mathrm{KL}}(\mathbb{P}_{XY}||\mathbb{P}_X \otimes \mathbb{P}_Y) \\ &= \mathbb{E}_{\mathbb{P}_{XY}}(\log \frac{d\mathbb{P}_{XY}}{d\mathbb{P}_{X\otimes Y}}) \end{aligned} \tag{1}$$

where $\mathbb{P}_{XY}$ and $\mathbb{P}_X \otimes \mathbb{P}_Y$ are the joint distribution and the product of the marginal distributions, respectively. $D_{\mathrm{KL}}$ is the Kullback-Leibler (KL) divergence. Without loss of generality, we consider $\mathbb{P}_{XY}$ and $\mathbb{P}_X \otimes \mathbb{P}_Y$ as being distributions on a compact domain $\Omega \subset \mathbb{R}^d$.

**MI through dual representation of $D_{\mathrm{KL}}$** We first introduce two dual representations of $D_{\mathrm{KL}}$, as MI is defined using $D_{\mathrm{KL}}$. The most widely known is the Donsker-Varadhan representation $D_{\mathrm{DV}}$ [Donsker and Varadhan, 1975]. For given two distribution $\mathbb{P}$ and $\mathbb{Q}$ on some compact domain $\Omega \subset \mathbb{R}^d$,

$$D_{\mathrm{DV}}(X,Y) := \sup_{T:\Omega\to\mathbb{R}} \mathbb{E}_{\mathbb{P}}(T) - \log(\mathbb{E}_{\mathbb{Q}}(e^T)), \tag{2}$$

where both the expectations $\mathbb{E}_{\mathbb{P}}(T)$ and $\mathbb{E}_{\mathbb{Q}}(e^T)$ are finite. If we substitute $\mathbb{P}$ and $\mathbb{Q}$ into $\mathbb{P}_{XY}$ and $\mathbb{P}_X \otimes \mathbb{P}_Y$, $D_{\mathrm{DV}}$ yields the definition of MI. The optimal $T^* = \log \frac{d\mathbb{P}}{d\mathbb{Q}} + C$, where $C \in \mathbb{R}$ can be any constant.

In contrast to $D_{\mathrm{DV}}$, the Nguyen-Wainwright-Jordan representation $D_{\mathrm{NWJ}}$ [Nguyen et al., 2010] is induced by the convex conjugate known as Fenchel's inequality [Hiriart-Urruty and Lemaréchal, 2004]:

$$D_{\mathrm{NWJ}}(X,Y) := \sup_{T:\Omega\to\mathbb{R}} \mathbb{E}_{\mathbb{P}}(T) - \mathbb{E}_{\mathbb{Q}}(e^{T-1}) \tag{3}$$

The optimal $T^* = \log \frac{d\mathbb{P}}{d\mathbb{Q}} + 1$ is unique unlike the optimal $T^*$ of $D_{\mathrm{DV}}$ due to its self-normalizing property [Belghazi et al., 2018]. However, $D_{\mathrm{DV}}$ guarantees tighter lower bounds than $D_{\mathrm{NWJ}}$ [Ruderman et al., 2012, Polyanskiy and Wu, 2014]. These two representations provide the theoretical soundness for numerous variational MI bounds.

**Variational MI estimation** With the increasing success of neural networks, several neural network-driven variational bounds of MI are proposed. They are widely employed, such as contrastive learning [Oord et al., 2018, He et al., 2020, Chen et al., 2020] or generative adversarial training [Belghazi et al., 2018, Nowozin et al., 2016]. Variational bounds of MI commonly focus on estimating $T^*$ via a neural network $T_\theta : \Omega \to \mathbb{R}$, called the statistics network [Belghazi et al., 2018], which outputs a single real value given the input sample pairs.

$I_{\mathrm{MINE}}$ [Belghazi et al., 2018] directly maximize $D_{\mathrm{DV}}$ as the objective function by feeding the samples $(x, y)$ of $\mathbb{P}_{XY}$ and $\mathbb{P}_X \otimes \mathbb{P}_Y$ into $T_\theta$:

$$\begin{aligned} &I_{\mathrm{MINE}}(X,Y) := \\ &\quad \mathbb{E}_{\mathbb{P}^{(n)}_{XY}}(T_\theta(x,y)) - \log(\mathbb{E}_{\mathbb{P}^{(n)}_X \otimes \hat{\mathbb{P}}^{(n)}_Y}(e^{T_\theta(x,y)})), \end{aligned} \tag{4}$$

where $\mathbb{P}^{(n)}$ is the empirical distribution associated to $n$ i.i.d. samples for given distribution $\mathbb{P}$. Belghazi et al. [2018] also utilizes moving averages of mini-batches to reduce the MI estimation variance caused by the limited batch size.

$I_{\mathrm{InfoNCE}}$ [Oord et al., 2018] is also commonly used due to its stability and decent performance:

$$I_{\mathrm{InfoNCE}}(X,Y) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{T_\theta(x_i,y_i)}}{\frac{1}{N}\sum_j^N e^{T_\theta(x_i,y_j)}} \tag{5}$$

where the $N$ samples $(x_i, y_i)_{i=1}^N$ are drawn from $\mathbb{P}_{XY}$, which becomes equivalent to using the Softmax function with the negative log loss. $I_{\mathrm{InfoNCE}}$ is also equivalent to $I_{\mathrm{MINE}}$ up to a constant, but upper bounded by $\log N$, hence not able to estimate large MI values [Oord et al., 2018].

Poole et al. [2019] introduced $I_{\mathrm{TUBA}}$, a unified lower bound, by expanding $D_{\mathrm{NWJ}}$ [Barber and Agakov, 2004, Nguyen et al., 2010].

$$\begin{aligned} &I_{\mathrm{NWJ}}(X,Y) := \\ &\quad \mathbb{E}_{\mathbb{P}^{(n)}_{XY}}(T_\theta(x,y)) - \mathbb{E}_{\mathbb{P}^{(n)}_X \otimes \mathbb{P}^{(n)}_Y}(e^{T_\theta(x,y)-1}), \end{aligned} \tag{6}$$

$$\begin{aligned} &I_{\mathrm{TUBA}}(X,Y) := \mathbb{E}_{\mathbb{P}^{(n)}_{XY}}(T_\theta(x,y)) \\ &- \mathbb{E}_{\mathbb{P}^{(n)}_Y}\left(\mathbb{E}_{\mathbb{P}^{(n)}_X}(e^{T_\theta(x,y)})/a(y) + \log(a(y)) - 1\right), \end{aligned} \tag{7}$$

where $a(y)$ is the variational parameter. However, unlike $I_{\mathrm{MINE}}$ or $I_{\mathrm{InfoNCE}}$, directly using the exponential term often causes numerical instability. Even if $T_\theta$ outputs a moderately sized value, $e^{T_\theta}$ can easily exceed the floating-point range.

To avoid this problem, Poole et al. [2019] introduce $D_{\mathrm{NWJ}}$-based lower bound $I_{\mathrm{JS}}$ by using a softplus-activated neural network as $T_\theta$,

$$\begin{aligned} I_{\mathrm{JS}}(X,Y) &:= 1 + \mathbb{E}_{\mathbb{P}^{(n)}_{XY}}(T_\theta(x,y)) \\ &\quad - \mathbb{E}_{\mathbb{P}^{(n)}_Y \otimes \mathbb{P}^{(n)}_X}((e^{T_\theta(x,y)})). \end{aligned} \tag{8}$$

**Variance problem of MI estimators** Despite the variety of bounds proposed, many still suffer from the bias-variance trade-off [Poole et al., 2019]. McAllester and Stratos [2018] and Song and Ermon [2020] prove that the $I_{\text{MINE}}$ estimator must have a batch size proportional to the exponential of true MI to control the variance of the estimation.

Many bounds try to mitigate this problem by reducing the variance of low-biased estimators, such as by interpolating with a low variance bound [Poole et al., 2019] or dropping the formal theoretical guarantees [McAllester and Stratos, 2018]. Song and Ermon [2020] proposed $I_{\text{SMILE}}$ to clip the range of $T_\theta$ trained with $I_{\text{MINE}}$, sacrificing the estimation quality to reduce the variance.

$$I_{\text{SMILE}}(X,Y) := \mathbb{E}_{\mathbb{P}^{(n)}_{XY}}(T_\theta(x,y)) - \log(\mathbb{E}_{\mathbb{P}^{(n)}_X \otimes \hat{\mathbb{P}}^{(n)}_Y}(\text{clip}(e^{T_\theta(x,y)}, e^{-\tau}, e^{\tau})), \quad (9)$$

where $\text{clip}(v,l,u) = \max(\min(v,u),l)$ for $v,u,l \in \mathbb{R}$.

**Practical usages of MI** MI-based losses are often applied in generative modeling, such as for better mode coverage [Belghazi et al., 2018] or learning disentangled representations without supervision [Chen et al., 2016, Ojha et al., 2020, Li et al., 2021, Jeon et al., 2021]. Representation learning employs MI-based losses [Tian et al., 2020b, Hjelm et al., 2019, Tschannen et al., 2020, Cheng et al., 2020, Wu et al., 2020, Wen et al., 2020, Boudiaf et al., 2020, Tian et al., 2020a, Li et al., 2020] to yield feature extractors that reflect its downstream tasks well. We emphasize that these approaches can be further utilized to measure the performance of MI estimators.

**Comparing between MI estimators** Toy datasets such as correlated multivariate Gaussian distributions has been widely accepted for the evaluation of MI estimation [Belghazi et al., 2018, Poole et al., 2019, Song and Ermon, 2020, Cheng et al., 2020, Lin et al., 2019]. However, we emphasize that using synthetic data as a definitive benchmark will end up in a disparity with real-world tasks. There have been some approaches that compared different MI estimators on generative modeling [Belghazi et al., 2018, Hjelm et al., 2019] or representational learning [Tian et al., 2020b]. However, finding the ideal MI for each downstream task is not trivial, making it impossible to directly assess the MI estimation quality. Moreover, Tschannen et al. [2020] and Tian et al. [2020b] showed the gap between MI estimation quality and downstream performance on specific tasks. Hence, it is crucial to evaluate both perspectives. The closest work to our benchmark is the consistency test of Song and Ermon [2020] using CIFAR-10 [Krizhevsky et al., 2009] and MNIST [LeCun et al., 1998]. However, the test only offered to assess the ratio of two separate MI estimations, making it difficult to separately measure the quality of each estimation.

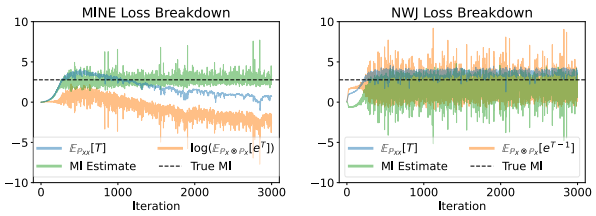


Figure 1: Training $T_\theta$ using $I_{\text{MINE}}$ and $I_{\text{NWJ}}$ with batch size 100 for 3000 iterations. We breakdown the MI loss into two components. We split $I_{\text{MINE}}$ into first term $\mathbb{E}_{\mathbb{P}_{XX}}(T)$ and second term $\log \mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_X}(e^T)$. Similarly, we split $I_{\text{NWJ}}$ into first term $\mathbb{E}_{\mathbb{P}_{XX}}(T)$ and second term $\mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_X}(e^{T-1})$.

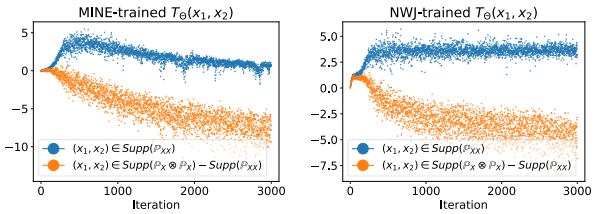


Figure 2: Training $T_\theta$ using $I_{\text{MINE}}$ and $I_{\text{NWJ}}$ with batch size 100 for 3000 iterations. We observe the statistics network outputs $T_\theta(x_1, x_2)$, where we split the outputs into two: $(x_1, x_2) \in Supp(\mathbb{P}_{XX})$ and $(x'_1, x'_2) \in Supp(\mathbb{P}_X \otimes \mathbb{P}_X) \setminus Supp(\mathbb{P}_{XX})$.

## 3 INSTABILITY OF MI BOUNDS

To demonstrate and analyze the instability of variational MI bounds, we design a synthetic problem with the One-hot dataset. We then solve the task via $I_{\text{MINE}}$ and $I_{\text{NWJ}}$, which are the losses derived from the two most commonly used representations of KL-divergence, $D_{\text{DV}}$ and $D_{\text{NWJ}}$, respectively. Both losses consist of two terms, each derived from the statistics of joint distribution $\mathbb{E}_{\mathbb{P}_{XY}}$ and the product of marginal distributions $\mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Y}$. Hence, to observe the behavior of each loss during training, we plot the two terms separately. Also, to observe how each distribution differ by the statistics network outputs $T_\theta(x,y)$, we plot each output from $(x,y) \sim Supp(\mathbb{P}_{XY})$ and $(x,y) \sim Supp(\mathbb{P}_X \otimes \mathbb{P}_Y) \setminus Supp(\mathbb{P}_{XY})$, where we denote the support of $\mathbb{P}$ as $Supp(\mathbb{P})$. Support is the set of values that the random variable can take [Taboga, 2021].

**One-hot Dataset** We design a one-hot discrete problem with uniform distribution $X \sim U(1,N)$ to estimating $I(X,X) = \log N$ for a given integer $N$. This task is intentionally created to easily discern samples $(x,x) \sim \mathbb{P}_{XX}$ from $(x,x) \sim \mathbb{P}_X \otimes \mathbb{P}_X$, so that we can directly observe its network outputs $T_\theta(x,x)$.

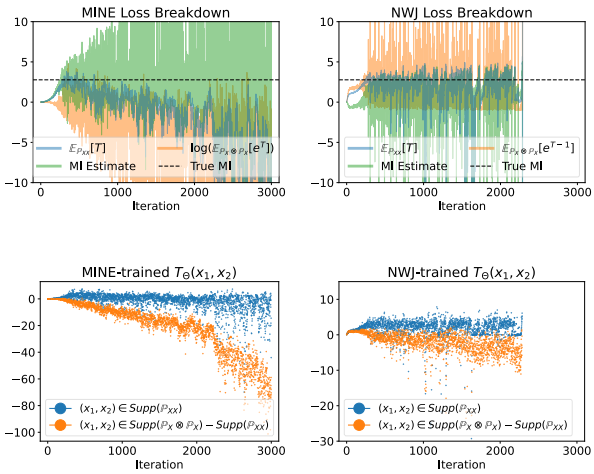


Figure 3: Training $T_\theta$ using $I_{\text{MINE}}$ and $I_{\text{NWJ}}$ with a reduced batch size of 32 for 3000 iterations. MI estimate diverges for both cases. Also, $I_{\text{NWJ}}$ incurs exploding $T_\theta$ outputs, hence the empty plot after 23k iterations.

**Seemingly Stable Case** We first observe the behaviors of the statistics network $T_\theta$ when the losses are seemingly stable, producing a successful MI estimate. Fig. 1 shows the MI estimates and the two terms that construct each MI estimate per batch. We observe that the first and the second term estimates of $I_{\text{MINE}}$, unlike $I_{\text{NWJ}}$, drifting in parallel even after the MI estimate converge. This is due to the free constant term $C$ in the optimal $T^*$ of $D_{\text{DV}}$, where the self-normalizing $D_{\text{NWJ}}$ avoids this problem. This drifting phenomenon implies that $T_\theta$ is not stable even after the loss seems to be converged, as shown in Fig. 2. Also, the plot demonstrates how $T_\theta$ is trained; it isolates the samples $(x, y) \sim \mathbb{P}_{XY}$ from the samples $(x, y) \sim \mathbb{P}_X \otimes \mathbb{P}_Y$.

**Unstable Case** We also demonstrate the behaviors of $T_\theta$ when the losses get unstable in Fig. 3. We reduce the batch size to make the optimization unstable, where this behavior is often reported in multiple works [Oord et al., 2018, He et al., 2020, Chen et al., 2020]. However, even though the losses seem unstable, $T_\theta$ successfully discerns the samples before the outputs explode. We believe that this is because of how $T_\theta$ is optimized during training. The statistics network outputs $T_\theta(x_1, x_2)$ of $(x_1, x_2) \in Supp(\mathbb{P}_{XX})$ gets increased by the first term but occasionally decreased by the second term. However, $T_\theta(x_1', x_2')$ of $(x_1', x_2') \in Supp(\mathbb{P}_X \otimes \mathbb{P}_X) \setminus Supp(\mathbb{P}_{XX})$ gets decreased whatsoever, as $(x_1', x_2')$ is used only for the second term. This makes the second term more unstable and motivates us to regularize it for better numerical stability during optimization.

To summarize, we suspect the instability of variational bounds comes from two reasons. Firstly, the statistics network did not converge even after the loss seemingly converged. We argue that this is due to the unnormalized constant term in the optimal $T^*$ of $D_{\text{DV}}$, where $D_{\text{NWJ}}$ successfully avoids via self-normalization. Secondly, the loss gets unstable as $T_\theta(x_1', x_2')$ endlessly decrease due to the second term. This observation is also consistent with the theoretical findings of Song and Ermon [2020], McAllester and Stratos [2018], where they show that large variance of the second term leads to failed MI estimation. We claim that the outputs have to be regularized in some form to avoid the instability.

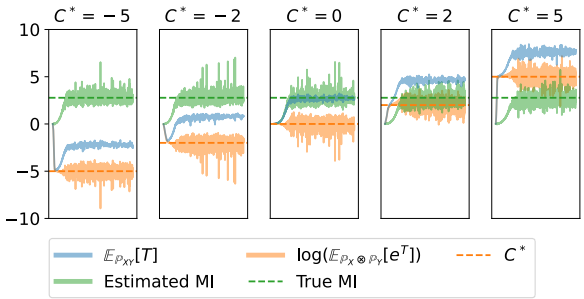


Figure 4: Training $T_\theta$ with batch size 100 for 1500 iterations using $I_{\text{ReMINE}}$ with different $C^*$ (orange dotted line).

## 4 STABILIZING THE MI BOUNDS

In this section, we introduce two novel regularized representations and its corresponding losses to tackle the instability during optimization. We show both theoretically and experimentally that adding regularization mitigates the unstable behavior of the statistics network $T_\theta$. We also describe a simple windowing method that can sidestep the batch size limitation problem of the MI estimation problem. We defer all the proofs to the Appendix.

**Regularized representations** We stabilize the two existing representations $D_{\text{DV}}$ and $D_{\text{NWJ}}$ by regularizing the second term. We introduce two novel representations: Regularized DV ($D_{\text{ReDV}}$) and Regularized NWJ ($D_{\text{ReNWJ}}$),

$$D_{\text{ReDV}}(X, Y) := \sup_{T:\Omega\to\mathbb{R}} \mathbb{E}_{\mathbb{P}}(T) - \log(\mathbb{E}_{\mathbb{Q}}(e^T)) - d(\log(\mathbb{E}_{\mathbb{Q}}(e^T)), C^*), \tag{10}$$

$$D_{\text{ReNWJ}}(X, Y) := \sup_{T:\Omega\to\mathbb{R}} \mathbb{E}_{\mathbb{P}}(T) - \mathbb{E}_{\mathbb{Q}}(e^{T-1})) - d(\mathbb{E}_{\mathbb{Q}}(e^{T-1})), 1), \tag{11}$$

where $C^* \in \mathbb{R}$ is any constant and $d(*, *)$ is a distance function on $\mathbb{R}$.

**Theorem 1.** *$D_{ReDV}$ and $D_{ReNWJ}$ is a dual representation for $D_{KL}$ such that*

$$D_{KL}(\mathbb{P}||\mathbb{Q}) = D_{ReDV}(X, Y), \tag{12}$$

$$D_{KL}(\mathbb{P}||\mathbb{Q}) = D_{ReNWJ}(X, Y). \tag{13}$$

We emphasize that both representations are not MI-specific but dual representations of $D_{\mathrm{KL}}$, which can be easily extended to numerous variational MI bounds based on $D_{\mathrm{DV}}$ and $D_{\mathrm{NWJ}}$. Especially, the newly added regularizer grants $D_{\mathrm{ReDV}}$ the normalizing property, effectively solving the drifting problem of $D_{\mathrm{DV}}$.

**Regularizing $I_{\mathrm{MINE}}$ and $I_{\mathrm{NWJ}}$** Based on $D_{\mathrm{ReDV}}$ and $D_{\mathrm{ReNWJ}}$, we propose a novel neural network-driven variational MI bound $I_{\mathrm{ReMINE}}$ and $I_{\mathrm{ReNWJ}}$ by choosing the Euclidean distance $d(x,y) = (x-y)^2$ and the log-Euclidean distance $d(x,y) = (\log x - \log y)^2$, respectively.

$$\begin{aligned} I_{\mathrm{ReMINE}}(X,Y) := {} & \mathbb{E}_{\mathbb{P}_{XY}^{(n)}}(T_\theta(x,y)) \\ & - \log(\mathbb{E}_{\mathbb{P}_X^{(n)} \otimes \mathbb{P}_Y^{(n)}}(e^{T_\theta(x,y)})) \\ & - \lambda(\log(\mathbb{E}_{\mathbb{P}_X^{(n)} \otimes \mathbb{P}_Y^{(n)}}(e^{T_\theta(x,y)})) - C^*)^2, \end{aligned} \tag{14}$$

$$\begin{aligned} I_{\mathrm{ReNWJ}}(X,Y) := {} & \mathbb{E}_{\mathbb{P}_{XY}^{(n)}}(T_\theta(x,y)) \\ & - \mathbb{E}_{\mathbb{P}_X^{(n)} \otimes \mathbb{P}_Y^{(n)}}(e^{T_\theta(x,y)-1}) \\ & - \lambda(\log(\mathbb{E}_{\mathbb{P}_X^{(n)} \otimes \mathbb{P}_Y^{(n)}}(e^{T_\theta(x,y)-1})))^2, \end{aligned} \tag{15}$$

where $C^* \in \mathbb{R}$ is any constant and $\lambda$ is a hyperparameter that controls the degree of regularization. We can also easily regularize other losses such as $I_{\mathrm{InfoNCE}}$, $I_{\mathrm{SMILE}}$, $I_{\mathrm{TUBA}}$, and $I_{\mathrm{JS}}$ in a plug-and-play manner. See Table 2 for more details on its regularized counterparts.

**Solving the drifting problem** Due to the self-regularizing nature of $D_{\mathrm{NWJ}}$, we must fix $C^* = 1$ for $I_{\mathrm{ReNWJ}}$. We also set $C^* = 0$ for $I_{\mathrm{ReMINE}}$ on future experiments, but to demonstrate the ability of the regularizer term to stop the drifting, we experiment with various $C^*$ in Fig. 4. Comparing to $I_{\mathrm{MINE}}$ in Fig. 1, we can observe that $I_{\mathrm{ReMINE}}$ successfully solves the drifting problem by regularizing the second term to have a single value.

**Solving the explosion problem** We previously observed the instability of $I_{\mathrm{MINE}}$ and $I_{\mathrm{NWJ}}$ when using a small batch in Fig. 3. We apply the same setting to $I_{\mathrm{ReMINE}}$ and $I_{\mathrm{ReNWJ}}$ to observe if the regularizer mitigates the instability problem. Both regularized losses successfully avoid the explosion problem and limit the statistics network outputs $T_\theta(x_1, x_2)$ within a certain boundary. As discussed in Section 3, the second term was the culprit of the variance in MI estimation. The newly added term directly regularizes it to stabilize training, giving the statistics network $T_\theta$ additional hints for the second term to converge to a specific value $C^*$ successfully. Furthermore, we empirically found that our regularization works well with $I_{\mathrm{SMILE}}$'s strategy of clipping $T_\theta$. Gradient zeros out for the original $I_{\mathrm{SMILE}}$ if $T_\theta(x,y)$ exceeds a certain threshold. This behavior makes $T_\theta$ act as if it were frozen, failing to further optimize during training. However, with the regularizer term, we can clip $T_\theta(x,y)$ only on first and second term, i.e., on the original loss. Now, clipping filters out the noisy gradients while the gradients calculated from the regularizer avoid freezing $T_\theta$ entirely.

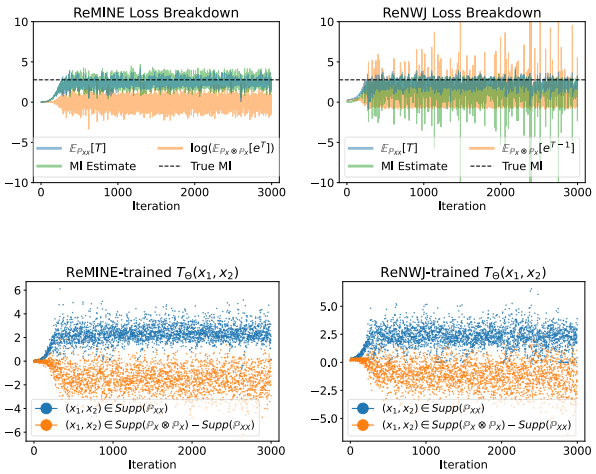


Figure 5: Training $T_\theta$ using the regularized counterparts, $I_{\mathrm{ReMINE}}$ and $I_{\mathrm{ReNWJ}}$, with the same small batch settings from Fig. 3. Regularization effectively mitigates both instability symptoms, shifting and exploding.

**Mathematical properties of $I_{\mathrm{ReMINE}}$ and $I_{\mathrm{ReNWJ}}$** Following Belghazi et al. [2018], we show the soundness of $I_{\mathrm{ReMINE}}$ and $I_{\mathrm{ReNWJ}}$ in two perspectives, strong consistency and sample complexity. These properties relate to whether the trained $T_\theta$ can be sufficiently similar to the optimal $T^*$.

**Theorem 2.** *$I_{ReMINE}$ and $I_{ReNWJ}$ are strongly consistent.*

For the two losses, we also provide the mathematical bound on the number of samples required for the empirical MI estimation at a given accuracy and with high confidence. Similar to Belghazi et al. [2018], let $T_\theta$ satisfy $L$-Lipschitz with respect to the parameter $\theta$ such that $|\theta| < K$ and $d$ is dimension of the parameter space of $T_\theta$.

**Theorem 3.** *Assume that $T_\theta$ is bounded above by $M$. Let $k$ be the number of sample means. Given any $\epsilon, \delta$ of the desired accuracy and confidence parameters, we have*

$$\mathcal{P}(|I_{ReMINE}(X;Y) - I(X,Y)| \leq \epsilon) \geq 1 - \delta, \tag{16}$$

*whenever the number $n$ of samples satisfies*

$$n \geq \frac{d\log(24KL\sqrt{d}/\epsilon) + 2dM + log(2/\delta)}{\epsilon^2 k/(2M^2)}. \tag{17}$$

**Theorem 4.** *Assume that $1 \leq |T_\theta| < M$ and $d(x,1) \leq |x-1|$. Let $k$ be the number of sample means. Given any $\epsilon, \delta$ of the desired accuracy and confidence parameters, we have*

$$\mathcal{P}(|I_{ReNWJ}(X;Y) - I(X,Y)| \leq \epsilon) \geq 1 - \delta, \tag{18}$$

*whenever the number $n$ of samples satisfies*

$$n \geq \frac{d\log(24KL\sqrt{d}/\epsilon) + 2dM + \log(2/\delta)}{\epsilon^2 k/(2M^2)}. \quad (19)$$

**Drifting may lead to noisy MI estimate** We prove that the variance of the second term on the empirical distributions is affected by the constant term $C^*$.

**Theorem 5.** *Let $Q^{(n)}$ be the empirical distributions of $n$ i.i.d. samples from $\mathbb{Q}$. For the optimal $T_1 = \log \frac{dp}{dq} + C_1$ and $T_2 = \log \frac{dp}{dq} + C_2$ where $C_1 \geq C_2$,*

$$Var_{\mathbb{Q}}(\mathbb{E}_{\mathbb{Q}^{(n)}}(e^{T_1})) \geq Var_{\mathbb{Q}}(\mathbb{E}_{\mathbb{Q}^{(n)}}(e^{T_2})) \quad (20)$$

This implies that unregulated $C^*$ may lead to worse MI estimation quality, as the source of the estimate variance are mainly due to the second term.

**Increasing the effective sample size for MI estimation** The drifting problem caused by the unnormalized constant term $C^*$ raises more issues when estimating MI. Poole et al. [2019] use a simple macro-averaging technique, i.e., averaging the estimated MI from each batch. We can also consider a slight modification to the technique, where we call it the micro-averaging technique, by saving all the statistics network outputs $T_\theta(x, y)$ for each batch and producing a single estimate based on all the outputs. However, we proved that both averaging techniques yield wrong final estimates for biased estimators like $I_{\text{MINE}}$ [Belghazi et al., 2018], $I_{\text{SMILE}}$ [Song and Ermon, 2020], $I_{\text{CLUB}}$ [Cheng et al., 2020], and $I_{\text{InfoNCE}}$ [Oord et al., 2018].

**Theorem 6.** *(Estimation bias caused by drifting) Both macro- and micro-averaging strategies produce a biased MI estimate when the drifting problem occurs.*

To the contrary, self-normalizing or regularized MI estimators have the upper hand in this perspective. By utilizing all the samples from multiple batches, they can effectively sidestep the batch size limitation problem [McAllester and Stratos, 2018, Song and Ermon, 2020].

# 5 EXPERIMENTS

## 5.1 MI ESTIMATION VS. DOWNSTREAM TASK PERFORMANCE

**Benchmark Design** To measure the performance of the MI estimators, one must design the target task to have the ground truth MI. This constraint led previous works to evaluate the estimators only on artificial toy problems [Belghazi et al., 2018, Poole et al., 2019], where its connection to actual problems is fairly limited. We design the two types of MI estimation tasks with de facto image datasets to improve the existing benchmarks to reflect on the real-world tasks. We defer all the proofs to the Appendix.

**Theorem 7.** *(Supervised learning) Given a dataset $D = (X, Y)$ where $X$ is an sample, $Y$ is the label for $X$, and $H(Y)$ is the entropy of the label set, $I(X, Y) = H(Y)$.*

Similarly, the true MI between images from the same class is also tractable based on the same assumption.

**Theorem 8.** *(Contrastive learning) Consider the dataset $D = (X, Y)$. Let $X_1$ be a sample drawn from the dataset and $X_2$ be another sample drawn from the subset with the same label $Y$ to $X_1$. Then, $I(X_1, X_2) = I(X_1, Y) = I(X_2, Y) = H(Y)$.*

Note that we assume statistical dependence between the image $X$ and label $Y$ from the point of view of information bottleneck [Tishby and Zaslavsky, 2015]. We derive the theorems above based on the assumption, where $Y$ implicitly determines $X$.

Based on the above theorems, we use the two MI estimation problems as benchmarks that evaluate the performance of estimators. We intentionally design Theorem 7 and Theorem 8 to mimic the existing tasks closely, namely, supervised and contrastive learning. For Theorem 7, we can set the statistics network $T_\theta(X, Y) = f_\theta(X) \cdot o(Y)$ where $f_\theta(X)$ is the logits obtained from feeding the image $X$ to the classification neural networks and $o(Y)$ is the one-hot representation of the label $Y$. If we use the InfoNCE estimator, this formulation becomes identical to solving the classification problem using negative log loss with the Softmax function, hence the name being the supervised learning benchmark (SLB). Similarly, for Theorem 8, we can set $T_\theta(X_1, X_2) = f_\theta(X_1) \cdot f_\theta(X_2)$ and use the InfoNCE estimator to yield a commonly used contrastive loss [Oord et al., 2018, Chen et al., 2020].

Due to the strict assumption of statistical dependence, the theorems above cannot be used on standard datasets like ImageNet dataset [Deng et al., 2009], as its samples often violate the single-label assumption. However, we can still empirically compare the MI estimators by the relative size of their final MI estimation.

**Evaluation** To verify the performance of MI estimators, we perform our benchmark tasks on the CIFAR10 and CIFAR100 dataset [Krizhevsky et al., 2009]. As both CIFAR10 and CIFAR100 have a uniform label distribution, ideal MI is $\log 10$ and $\log 100$, respectively. In addition, to check whether this MI estimate task is actually helpful for downstream tasks, we evaluate each estimator on both dimensions: MI estimation and test set accuracy. Similar to the existing settings in the contrastive learning literature [Chen et al., 2020, He et al., 2020], we design the test accuracy of CLB by defining the label estimate $\hat{y}$ of each test set sample $x_{\text{Test}}$ to be the label of $x = \text{argmax}_{x \in X_{\text{Train}}} f(x) \cdot f(x_{\text{Test}})$ of the train dataset $X_{\text{Train}}$. Similarly, for SLB, we chose $\hat{y} = \text{argmax}_y f(x_{\text{Test}}) \cdot o(y)$ where $o(y)$ is the one-hot encod-

| Task | | Loss | MI Estimation | | Test Accuracy | |
|---|---|---|---|---|---|---|
| | | | Original | Regularized | Original | Regularized |
| Supervised Learning Benchmark | CIFAR-10 | MINE | **2.300 ± 0.003** | 2.298 ± 0.005 | 0.850 ± 0.009 | **0.856 ± 0.004** |
| | | SMILE | 2.297 ± 0.009 | **2.300 ± 0.003** | **0.854 ± 0.008** | 0.853 ± 0.009 |
| | | InfoNCE | 2.301 ± 0.002 | **2.302 ± 0.001** | 0.845 ± 0.006 | 0.845 ± 0.005 |
| | | NWJ | **2.297 ± 0.009** | 2.294 ± 0.013 | 0.859 ± 0.003 | **0.862 ± 0.004** |
| | | TUBA | 2.297 ± 0.008 | **2.300 ± 0.003** | **0.862 ± 0.008** | 0.859 ± 0.003 |
| | | JS | 1.944 ± 0.039 | **2.000 ± 0.049** | 0.838 ± 0.012 | **0.842 ± 0.004** |
| | CIFAR-100 | MINE | 4.597 ± 0.011 | **4.603 ± 0.001** | 0.610 ± 0.007 | 0.610 ± 0.006 |
| | | SMILE | 4.595 ± 0.015 | **4.602 ± 0.002** | 0.601 ± 0.015 | **0.606 ± 0.007** |
| | | InfoNCE | 4.594 ± 0.017 | **4.599 ± 0.005** | 0.589 ± 0.010 | **0.593 ± 0.005** |
| | | NWJ | 4.572 ± 0.055 | **4.586 ± 0.034** | 0.558 ± 0.042 | **0.599 ± 0.009** |
| | | TUBA | 4.495 ± 0.207 | **4.603 ± 0.002** | 0.543 ± 0.055 | **0.611 ± 0.007** |
| | | JS | 4.088 ± 0.430 | **4.240 ± 0.116** | 0.591 ± 0.026 | **0.598 ± 0.010** |
| Contrastive Learning Benchmark | CIFAR-10 | MINE | 2.233 ± 0.674 | **2.240 ± 0.657** | 0.812 ± 0.026 | **0.823 ± 0.012** |
| | | SMILE | 0.000 ± 0.000 | **2.065 ± 0.842** | 0.100 ± 0.001 | **0.830 ± 0.008** |
| | | InfoNCE | 1.705 ± 0.462 | **1.739 ± 0.431** | **0.830 ± 0.008** | 0.826 ± 0.006 |
| | | NWJ | 0.000 ± 0.000 | **1.910 ± 0.662** | 0.100 ± 0.000 | **0.831 ± 0.005** |
| | | TUBA | 0.000 ± 0.000 | **1.358 ± 0.590** | 0.100 ± 0.000 | **0.830 ± 0.009** |
| | | JS | 1.552 ± 0.485 | **1.556 ± 0.546** | **0.837 ± 0.003** | 0.832 ± 0.009 |
| | CIFAR-100 | MINE | **4.634 ± 0.186** | 4.563 ± 0.162 | 0.522 ± 0.026 | **0.540 ± 0.020** |
| | | SMILE | 0.000 ± 0.000 | **4.677 ± 0.162** | 0.012 ± 0.003 | **0.585 ± 0.007** |
| | | InfoNCE | 4.112 ± 0.147 | **4.115 ± 0.145** | 0.576 ± 0.019 | **0.585 ± 0.014** |
| | | NWJ | 0.000 ± 0.000 | **4.065 ± 0.255** | 0.010 ± 0.000 | **0.521 ± 0.025** |
| | | TUBA | 0.000 ± 0.000 | **2.731 ± 0.786** | 0.010 ± 0.000 | **0.490 ± 0.023** |
| | | JS | 3.253 ± 0.368 | **3.393 ± 0.124** | 0.451 ± 0.020 | **0.463 ± 0.031** |

Table 1: Our supervised and contrastive learning benchmark results. We provide the 95% confidence interval of 5 runs for both MI estimation and test accuracy, where we clip the negative MI estimations to 0. We compare the performance of original and regularized loss. **Bold text** and **blue text** indicates the better performance with overlapping and non-overlapping confidence interval, respectively.

| Loss | Loss settings | Regularizer settings |
|---|---|---|
| $I_{\text{MINE}}$ | No grad. moving average | Euclidean dist. |
| $I_{\text{SMILE}}$ | Clipping ($\tau = 10$) | Euclidean dist. |
| $I_{\text{InfoNCE}}$ | - | Euclidean dist. |
| $I_{\text{NWJ}}$ | - | Log-Euclidean dist., Clipping ($\tau = 10$) |
| $I_{\text{TUBA}}$ | $a(y) = 1$ | Log-Euclidean dist., Clipping ($\tau = 10$) |
| $I_{\text{JS}}$ | Estimate with $I_{\text{NWJ}}$ | Euclidean dist. |

Table 2: List of MI estimators with its hyperparameters

ing of $y$. We ran the same experiment 5 times with different seeds to yield a 95% confidence interval.

### 5.2 COMPARISON WITH OUR BENCHMARK

To demonstrate the effectiveness of our novel regularization term, we regularize the two representations, $D_{\text{DV}}$ and $D_{\text{NWJ}}$. We test three realizations for each representation, $I_{\text{MINE}}$ [Belghazi et al., 2018], $I_{\text{SMILE}}$ [Song and Ermon, 2020], and $I_{\text{InfoNCE}}$ [Oord et al., 2018] for $D_{\text{DV}}$, $I_{\text{NWJ}}$ [Nguyen et al., 2010], $I_{\text{TUBA}}$ [Poole et al., 2019], and $I_{\text{JS}}$ [Hjelm et al., 2019] for $D_{\text{NWJ}}$. We compare the original losses with its regularized counterparts, a total of $6 \times 2 = 12$. We do not apply averaging scheme on any of the losses, and choose the regularization weight $\lambda \in \{0.1, 0.01, 0.001\}$ that shows the best MI estimation results. See Table 2 for more details.

We observe in Table 1 that additional regularization generally induces better performance on both the MI estimation task and the downstream task (test accuracy). Hence, adding the regularizer to a pre-existing supervised or contrastive learning loss seems to be a viable option to increase the performance further. Even when the performance of the regularized loss slightly degrades, its negative impact is minimal. This implies that even for the case where the regularizer is not greatly helpful, it does not greatly hinder optimization. Especially, it is intriguing that many losses, $I_{\text{MINE}}$, $I_{\text{ReMINE}}$, and $I_{\text{ReTUBA}}$, are better than $I_{\text{InfoNCE}}$ in SLB, which is used as the de facto standard in classification. Also, $I_{\text{SMILE}}$, $I_{\text{NWJ}}$, and $I_{\text{TUBA}}$ fail to converge in CLB, where simply adding a regularization term solves the issue altogether, yielding a competitive or even better performance than all the other losses. Given the fact that numerous contrastive learning literature suffers from instability [Caron et al., 2021, Bardes et al., 2021, Chen et al., 2020, He et al., 2020, Bardes et al., 2021], we emphasize that adding our regularization term can be a simple yet effective method to stabilize training.

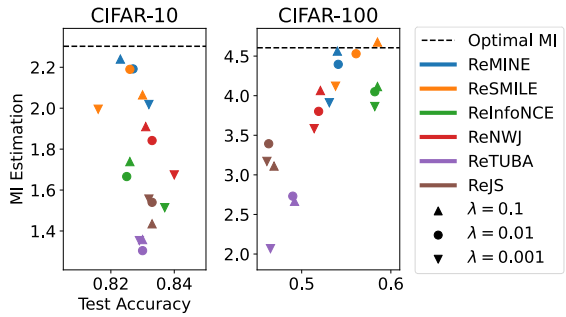


Figure 6: Ablation study on different $\lambda$s with CLB CIFAR-10 and CIFAR-100.

Additionally, to observe the impact of regularization strength $\lambda$, we plot the benchmark performance for each $\lambda$ in Fig. 6. We compare the losses on CLB as experimental results suggest that CLB is a more difficult task than SLB, showing significant performance differences between various losses. On CIFAR-10, $\lambda$ acts as a trade-off parameter between test accuracy and MI estimation quality. Performance trade-off has also been reported in other literature, where better MI estimation does not necessarily deliver better downstream performance [Tschannen et al., 2020, Tian et al., 2020b]. However, compared to CIFAR-100, test accuracy differences are minimal, where MI differences are apparent. $I_{\text{ReMINE}}$ and $I_{\text{ReSMILE}}$ show excellent MI estimation quality in CIFAR-10 compared to other losses. In contrast, test accuracy and MI estimation quality align well in the CIFAR-100 case. $I_{\text{ReSMILE}}$ shows good overall performance, albeit its sensitivity towards regularization strength. $I_{\text{ReInfoNCE}}$, on the other hand, shows stable performance in the downstream task, sacrificing the MI estimation quality. This result is further supported by the prominence of $I_{\text{InfoNCE}}$ in the contrastive learning domain. It is yet unclear where the difference between CIFAR-10 and CIFAR-100 comes from, whether it is due to the difference in the level of difficulty of the dataset or the batch size used throughout the training. We leave further analysis as future work.

### 5.3 COMPARISON WITH THE STANDARD TOY PROBLEM

We provide the quality of MI-based losses on the 20D Correlated Gaussian task [Belghazi et al., 2018, Poole et al., 2019] where the true MI is increased 5 times during optimization in Fig. 7. This experiment demonstrates how stable the MI-based losses estimate MI in a dynamically changing environment. We apply the same settings from Table 2, where we fix the regularization strength $\lambda = 1.0$ for all the losses. With the exception of $I_{\text{InfoNCE}}$, regularized losses show clear superiority over the original losses. Regularization facilitates $I_{\text{MINE}}$ and $I_{\text{SMILE}}$ to avoid the instability which is mentioned in Section 3. Also, regularization greatly

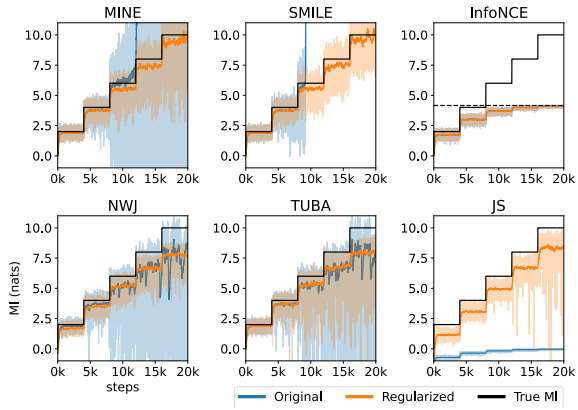


Figure 7: Estimation performance on 20-D Gaussian. The estimated MI (light) and the smoothed estimation with exponential moving average (dark) are plotted for each methods with its regularized counterparts. Black line represents the true MI. Dotted line shows the bound of $I_{\text{InfoNCE}}$ due to the limited batch size of 64.

enhances the MI estimation quality of $I_{\text{JS}}$ and lessens the variance of both $I_{\text{NWJ}}$ and $I_{\text{TUBA}}$.

## 6 CONCLUSION

In this paper, we identify the two symptoms behind the instability: The statistics network weights were not converging even after the loss seemed to converge, and the statistics network outputs from the product of marginal distribution explode during training. We propose a novel regularization term to mitigate the instability during training by adding to various existing MI-based losses. We theoretically and experimentally demonstrate that the added regularizer directly alleviates the two instability symptoms. Finally, we present a benchmark that evaluates both the MI estimation power and its capability on the downstream tasks by imitating the supervised or contrastive learning settings. We compare six different losses and their regularized counterparts on various benchmarks to show the method's effectiveness and broad applicability.

## LIMITATIONS AND FUTURE WORKS

We suspect that the instability of MI estimators can also be related to the collapse problem [Bardes et al., 2021, Caron et al., 2021]. Further loss-based approaches to combat this problem by regularizing the network outputs may be helpful. We expect that extending our methods to various contrastive learning losses may yield fruitful results for self-supervised learning. Also, we did not use larger image datasets due to the limited computational budget. Future works can investigate other domains such as text or audio.

## References


David Barber and Felix V Agakov. Information maximization in noisy channels: A variational approach. In *Advances in Neural Information Processing Systems*, pages 201–208, 2004.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *CoRR*, abs/2105.04906, 2021.

Adam Bear and Fiery Cushman. Loss functions modulate the optimal bias-variance trade-off. In Stephanie Denison, Michael Mack, Yang Xu, and Blair C. Armstrong, editors, *Proceedings of the 42th Annual Meeting of the Cognitive Science Society - Developing a Mind: Learning in Humans, Animals, and Machines, CogSci 2020, virtual, July 29 - August 1, 2020*. cognitivesciencesociety.org, 2020. URL https://cogsci.mindmodeling.org/2020/papers/0087/index.html.

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

Malik Boudiaf, Jérôme Rony, Imtiaz Masud Ziko, Eric Granger, Marco Pedersoli, Pablo Piantanida, and Ismail Ben Ayed. A unifying mutual information view of metric learning: Cross-entropy vs. pairwise losses. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI*, volume 12351 of *Lecture Notes in Computer Science*, pages 548–564. Springer, 2020. doi: 10.1007/978-3-030-58539-6_33. URL https://doi.org/10.1007/978-3-030-58539-6_33.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 391–401, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/58a2fc6ed39fd083f55d4182bf88826d-Abstract.html.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020. URL http://proceedings.mlr.press/v119/chen20j.html.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. CLUB: A contrastive log-ratio upper bound of mutual information. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1779–1788. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/cheng20b.html.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/63c3ddcc7b23daa1e42dc41f9a44a873-Abstract.html.

Pierre Colombo, Pablo Piantanida, and Chloé Clavel. A novel estimator of mutual information for learning to disentangle textual representations. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6539–6550. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.511. URL ahttps://doi.org/10.18653/v1/2021.acl-long.511.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*

*(CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735. IEEE, 2020. doi: 10.1109/CVPR42600.2020.00975. URL https://doi.org/10.1109/CVPR42600.2020.00975.

Olivier J. Hénaff. Data-efficient image recognition with contrastive predictive coding. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4182–4192. PMLR, 2020. URL http://proceedings.mlr.press/v119/henaff20a.html.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.

R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bklr3j0cKX.

Insu Jeon, Wonkwang Lee, Myeongjang Pyeon, and Gunhee Kim. Ib-gan: Disengangled representation learning with information bottleneck generative adversarial networks. 2021.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations. *CoRR*, abs/2005.04966, 2020. URL https://arxiv.org/abs/2005.04966.

Wei Li, Zhixuan Liang, Julian Neuman, Jinlin Chen, and Xiaohui Cui. Multi-generator GAN learning disconnected manifolds with mutual information. *Knowl. Based Syst.*, 212:106513, 2021. doi: 10.1016/j.knosys.2020.106513. URL https://doi.org/10.1016/j.knosys.2020.106513.

Xiao Lin, Indranil Sur, Samuel A Nastase, Ajay Divakaran, Uri Hasson, and Mohamed R Amer. Data-efficient mutual information neural estimator. *arXiv preprint arXiv:1905.03319*, 2019.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. *arXiv preprint arXiv:1811.04251*, 2018.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

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

Utkarsh Ojha, Krishna Kumar Singh, Cho-Jui Hsieh, and Yong Jae Lee. Elastic-infogan: Unsupervised disentangled representation learning in class-imbalanced data. In

Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d1e39c9bda5c80ac3d8ea9d658163967-Abstract.html.

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

Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6 (2012-2016):7, 2014.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180, 2019.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

Lorenz Richter, Ayman Boustati, Nikolas Nüsken, Francisco J. R. Ruiz, and Ömer Deniz Akyildiz. Vargrad: A low-variance gradient estimator for variational inference. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/9c22c0b51b3202246463e986c7e205df-Abstract.html.

Jonas Rothfuss, Dennis Lee, Ignasi Clavera, Tamim Asfour, and Pieter Abbeel. Promp: Proximal meta-policy search. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=SkxXCi0qFX.

Avraham Ruderman, Mark D. Reid, Dario García-García, and James Petterson. Tighter variational representations of f-divergences via restriction to probability measures. *ArXiv*, abs/1206.4664, 2012.

Andrew M Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D Tracey, and David D Cox. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020, 2019.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1x62TNtDS.

Marco Taboga. *Support of a random variable*. Kindle Direct Publishing, 2021. URL https://www.statlect.com/glossary/support-of-a-random-variable.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, volume 12356 of *Lecture Notes in Computer Science*, pages 776–794. Springer, 2020a. doi: 10.1007/978-3-030-58621-8_45. URL https://doi.org/10.1007/978-3-030-58621-8_45.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/4c2e5eaae9152079b9e95845750bb9ab-Abstract.html.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2015.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rkxoh24FPH.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Liangjian Wen, Yiji Zhou, Lirong He, Mingyuan Zhou, and Zenglin Xu. Mutual information gradient estimation

for representation learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=ByxaUgrFvH.

Mike Wu, Chengxu Zhuang, Milan Mosse, Daniel Yamins, and Noah D. Goodman. On mutual information in contrastive learning for visual representations. *CoRR*, abs/2005.13149, 2020. URL https://arxiv.org/abs/2005.13149.

Fanyu Zeng, Chen Wang, and Weitian Wang. Visual navigation with asynchronous proximal policy optimization in artificial agents. *J. Robotics*, 2020:8702962:1–8702962:7, 2020. doi: 10.1155/2020/8702962. URL https://doi.org/10.1155/2020/8702962.