# MUTUAL INFORMATION GRADIENT ESTIMATION FOR REPRESENTATION LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Mutual Information (MI) plays an important role in representation learning. However, MI is unfortunately intractable in continuous and high-dimensional settings. Recent advances establish tractable and scalable MI estimators to discover useful representation. However, most of the existing methods are not capable of providing an accurate estimation of MI with low-variance when the MI is large. We argue that directly estimating the gradients of MI is more appealing for representation learning than estimating MI in itself. To this end, we propose the Mutual Information Gradient Estimator (MIGE) for representation learning based on the score estimation of implicit distributions. MIGE exhibits a tight and smooth gradient estimation of MI in the high-dimensional and large-MI settings. We expand the applications of MIGE in both unsupervised learning of deep representations based on InfoMax and the Information Bottleneck method. Experimental results have indicated the significant performance improvement in learning useful representation.

## 1 INTRODUCTION

Mutual information (MI) is an appealing metric widely used in information theory and machine learning to quantify the amount of shared information between a pair of random variables. Specifically, given a pair of random variables $\mathbf{x}, \mathbf{y}$, the MI, denoted by $I(\mathbf{x}; \mathbf{y})$, is defined as

$$I(\mathbf{x}; \mathbf{y}) = \mathbb{E}_{p(\mathbf{x},\mathbf{y})} \left[ \log \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} \right], \tag{1}$$

where $\mathbb{E}$ is the expectation over the given distribution. Since MI is invariant to invertible and smooth transformations, it can capture non-linear statistical dependencies between variables (Kinney & Atwal, 2014). These appealing properties make it act as a fundamental measure of true dependence. Therefore, MI has found applications in a wide range of machine learning tasks, including feature selection (Kwak & Choi, 2002; Fleuret, 2004; Peng et al., 2005), clustering (Müller et al., 2012; Ver Steeg & Galstyan, 2015), and causality (Butte & Kohane, 1999). It has also been pervasively used in science, such as biomedical sciences (Maes et al., 1997), computational biology (Krishnaswamy et al., 2014), and computational neuroscience (Palmer et al., 2015).

Recently, there has been a revival of methods in unsupervised representation learning based on MI. A seminal work is the InfoMax principle (Linsker, 1988), where given an input instance $x$, the goal of the InfoMax principle is to learn a representation $E_\psi(x)$ by maximizing the MI between the input and its representation. A growing set of recent works have demonstrated promising empirical performance in unsupervised representation learning via MI maximization (Krause et al., 2010; Hu et al., 2017; Alemi et al., 2018b; Oord et al., 2018; Hjelm et al., 2019). Another closely related work is the Information Bottleneck method (Tishby et al., 2000; Alemi et al., 2017), where MI is used to limit the contents of representations. Specifically, the representations are learned by extracting task-related information from the original data while being constrained to discard parts that are irrelevant to the task. Several recent works have also suggested that by controlling the amount of information between learned representations and the original data, one can tune desired characteristics of trained models such as generalization error (Tishby & Zaslavsky, 2015; Vera et al., 2018), robustness (Alemi et al., 2017), and detection of out-of-distribution data (Alemi et al., 2018a).

Despite playing a pivotal role across a variety of domains, MI is notoriously intractable. Exact computation is only tractable for discrete variables, or for a limited family of problems where the

probability distributions are known. For more general problems, MI is challenging to analytically compute or estimate from samples. A variety of MI estimators have been developed over the years, including likelihood-ratio estimators (Suzuki et al., 2008), binning (Fraser & Swinney, 1986; Darbellay & Vajda, 1999; Shwartz-Ziv & Tishby, 2017), k-nearest neighbors (Kozachenko & Leonenko, 1987; Kraskov et al., 2004; Pérez-Cruz, 2008; Singh & Póczos, 2016), and kernel density estimators (Moon et al., 1995; Kwak & Choi, 2002; Kandasamy et al., 2015). However, few of these mutual information estimators scale well with dimension and sample size in machine learning problems (Gao et al., 2015).

In order to overcome the intractability of MI in the continuous and high-dimensional settings, Alemi et al. (2017) combines variational bounds of Barber & Agakov (2003) with neural networks for the estimation. However, the tractable density for the approximate distribution is required due to variational approximation. This limits its application to the general-purpose estimation, since the underlying distributions are often unknown. Alternatively, the Mutual Information Neural Estimation (MINE, Belghazi et al. (2018)) and the Jensen-Shannon MI estimator (JSD, Hjelm et al. (2019)) enable differentiable and tractable estimation of MI by training a discriminator to distinguish samples coming from the joint distribution or the product of the marginals. In detail, MINE employs a lower-bound to the MI based on the Donsker-Varadhan representation of the KL-divergence, and JSD follows the formulation of f-GAN KL-divergence. In general, these estimations are often noisy and can lead to unstable training due to their dependence on the discriminator used to estimate the bounds of mutual information. As pointed out by Poole et al. (2019), these unnormalized critic estimations of MI exhibit high variance and are challenging to tune for estimation. An alternative low-variance choice of MI estimator is Information Noise-Contrastive Estimation (InfoNCE, Oord et al. (2018)). It introduces the Noise-Contrastive Estimation with flexible critics parameterized by neural networks as a bound to approximate MI. Nonetheless, its estimation saturates at log of the batch size and suffers from high bias. Despite their modeling power, none of the estimators are capable of providing accurate estimation of MI with low variance when the MI is large and the batch size is small (Poole et al., 2019). As supported by the theoretical findings in McAllester & Statos (2018), any distribution-free high-confidence lower bound on entropy requires a sample size exponential in the size of the bound. More discussions about the bounds of MI and their relationship refers to Poole et al. (2019).

In summary, existing estimators first approximate MI and then use these approximations to optimize the associated parameters. For estimating MI based on any finite number of samples, there exists an infinite number of functions, with arbitrarily diverse gradients, that can perfectly approximate the true MI at these samples. However, these approximate functions can lead to unstable training and poor performance in optimization due to gradients discrepancy between approximate estimation and true MI. Estimating gradients of MI rather than estimating MI may be a better approach for MI optimization. To this end, to the best of our knowledge, we firstly propose the Mutual Information Gradient Estimator (MIGE) in representation learning. In detail, we estimate the score function of an implicit distribution, $\nabla_{\mathbf{x}} \log q(\mathbf{x})$, to achieve a general-purpose MI gradient estimation for representation learning. In particular, to deal with high-dimensional inputs, such as text, images and videos, score function estimation via Spectral Stein Gradient Estimator (SSGE) (Shi et al., 2018) is computationally expensive and complex. We thus propose an efficient high-dimensional score function estimator to make SSGE scalable. To this end, we derive a new reparameterization trick for the representation distribution based on the lower-variance reparameterization trick proposed by Roeder et al. (2017).

We summarize the contributions of this paper as follows:

- We propose the Mutual Information Gradient Estimator (MIGE) for representation learning based on the score function estimation of implicit distributions. Compared with MINE and MINE-$f$, MIGE provides a tighter and smoother gradient estimation of MI in a high-dimensional and large-MI setting, as shown in Figure 1 of Section 5.
- We propose the Scalable SSGE to alleviate the exorbitant computational cost of SSGE in high-dimensional settings.
- To learn meaningful representations, we apply SSGE as gradient estimators for both InfoMax and Information Bottlenck, and have achieved improved performance than their corresponding competitors.

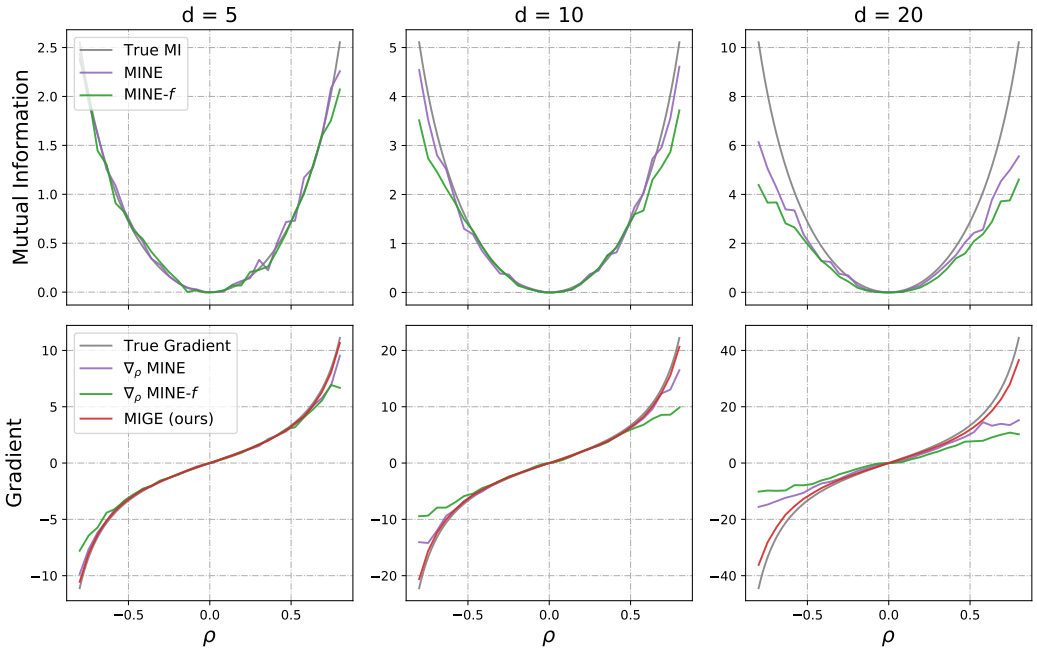


Figure 1: Estimation performance of MINE, MINE-$f$ and MIGE. **Top:** True MI and corresponding estimation of MINE and MINE-$f$. **Bottom:** True gradient and corresponding estimation of MINE, MINE-$f$ and MIGE. Our approach MIGE only appears in bottom figures since it directly gives gradient estimation. As we observe, MIGE gives more stable, smooth and accurate results.

## 2 Background: Spectral Stein Gradient Estimator

Score estimation of implicit distributions has been widely explored in the past few years (Song et al., 2019; Li & Turner, 2017; Shi et al., 2018). A promising method of score estimation is the Stein gradient estimator (Li & Turner, 2017; Shi et al., 2018), which is proposed for implicit distributions. It is inspired by generalized Steins identity (Gorham & Mackey, 2015; Liu & Wang, 2016) as follows.

**Steins identity.** Let $q(\mathbf{x})$ be a continuously differentiable (also called smooth) density supported on $\mathcal{X} \subseteq \mathbb{R}^d$, and $\boldsymbol{h}(\mathbf{x}) = [h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_{d'}(\mathbf{x})]^{\mathrm{T}}$ is a smooth vector function. Further, the boundary conditions on $\boldsymbol{h}$ is

$$q(\mathbf{x})\boldsymbol{h}(\mathbf{x}) = 0, \forall \mathbf{x} \in \partial\mathcal{X} \text{ if } \mathcal{X} \text{ is compact, or } \lim_{\mathbf{x}\to\infty} q(\mathbf{x})\boldsymbol{h}(\mathbf{x}) = 0 \text{ if } \mathcal{X} = \mathbb{R}^d. \tag{2}$$

Under this condition, the following identity can be easily checked using integration by parts, assuming mild zero boundary conditions on $\boldsymbol{h}$,

$$\mathbb{E}_q \left[\boldsymbol{h}(\mathbf{x})\nabla_{\mathbf{x}} \log q(\mathbf{x})^{\mathrm{T}} + \nabla_{\mathbf{x}}\boldsymbol{h}(\mathbf{x})\right] = \mathbf{0}. \tag{3}$$

Here $\boldsymbol{h}$ is called the Stein class of $q(\boldsymbol{x})$ if Steins identity Eq. (3) holds. Monte Carlo estimation of the expectation in Eq. (3) builds the connection between $\nabla_{\mathbf{x}} \log q(\mathbf{x})$ and the samples from $q(\mathbf{x})$ in Steins identity. For modeling implicit distributions, Motivated by Steins identity, Shi et al. (2018) proposed Spectral Stein Gradient Estimator (SSGE) for implicit distributions based on Stein's identity and a spectral decomposition of kernel operators where the eigenfunctions is approximated by the Nyström method. Below we briefly review SSGE. More details refer to Shi et al. (2018). Specifically, we denote the target gradient function to estimate by $\mathbf{g} : \mathcal{X} \to \mathbb{R}^d : \mathbf{g}(\mathbf{x}) = \nabla_{\mathbf{x}} \log q(\mathbf{x})$. The $i^{th}$ component of the gradient is $g_i(\mathbf{x}) = \nabla_{\mathbf{x}_i} \log q(\mathbf{x})$. We assume $g_1, \ldots, g_d \in L^2(\mathcal{X}, q)$. $\{\psi_j\}_{j\geq 1}$ denotes an orthonormal basis of $L^2(\mathcal{X}, q)$. We can expand $g_i(\mathbf{x})$ into the spectral series, i.e., $g_i(\mathbf{x}) = \sum_{j=1}^{\infty} \beta_{ij}\psi_j(\mathbf{x})$. The value of the $j^{th}$ eigenfunction $\psi_j$ at $\mathbf{x}$ can be approximated by

the Nyström method. Due to the orthonormality of eigenfunctions $\{\psi_j\}_{j\geq 1}$, there is a constraint under the probability measure q(.): $\int \psi_i(\mathbf{x})\psi_j(\mathbf{x})q(\mathbf{x})d\mathbf{x} = \delta_{ij}$, where $\delta_{ij} = \mathbb{1}[i = j]$. Based on this constraint, we can obtain the following equation for $\{\psi_j\}_{j\geq 1}$:

$$\int k(\mathbf{x}, \mathbf{y})\psi(\mathbf{y})q(\mathbf{y})d\mathbf{y} = \mu\psi(\mathbf{x}), \tag{4}$$

where $k(.)$ is a kernel function. The left side of the above equation can be approximated by the Monte Carlo estimate using i.i.d. samples $\mathbf{x}^1, ..., \mathbf{x}^M$ from $q(.)$ : $\frac{1}{M}\mathbf{K}\psi \approx \mu\psi$, where $\mathbf{K}$ is the Gram Matrix and $\psi = \left[\psi\left(\mathbf{x}^1\right), \ldots, \psi\left(\mathbf{x}^M\right)\right]^\top$. We can solve this eigenvalue problem by choose the $J$ largest eigenvalues $\lambda_1 \geq \cdots \geq \lambda_J$ for $\mathbf{K}$. $\boldsymbol{u}_j$ denotes the eigenvector of the Gram matrix. The approximation for $\{\psi_j\}_{j\geq 1}$ can be obtained combined with Eq. (4) as following: $\psi_j(\mathbf{x}) \approx \hat{\psi}_j(\mathbf{x}) = \frac{\sqrt{M}}{\lambda_j}\sum_{m=1}^{M} u_{jm}k\left(\mathbf{x}, \mathbf{x}^m\right)$

Furthermore, based on the orthonormality of $\{\psi_j\}_{j\geq 1}$, we can easily obtain $\beta_{ij} = -\mathbb{E}_q\nabla_{\mathbf{x}_i}\psi_j(\mathbf{x})$. By taking derivative both sides of Eq. (4), we can show that:

$$\mu_j\nabla_{\mathbf{x}_i}\psi_j(\mathbf{x}) = \nabla_{\mathbf{x}_i}\int k(\mathbf{x}, \mathbf{y})\psi_j(\mathbf{y})q(\mathbf{y})d\mathbf{y} = \int \nabla_{\mathbf{x}_i}k(\mathbf{x}, \mathbf{y})\psi_j(\mathbf{y})q(\mathbf{y})d\mathbf{y}. \tag{5}$$

Then we can estimate as following: $\hat{\nabla}_{\mathbf{x}_i}\psi_j(\mathbf{x}) \approx \frac{1}{\mu_j M}\sum_{m=1}^{M}\nabla_{\mathbf{x}_i}k\left(\mathbf{x}, \mathbf{x}^m\right)\psi_j\left(\mathbf{x}^m\right)$

Finally, by truncating the expansion to the first $J$ terms and plugging in the Nyström approximations of $\{\psi_j\}_{j\geq 1}$, we can get the score estimator:

$\hat{g}_i(\mathbf{x}) = \sum_{j=1}^{J}\hat{\beta}_{ij}\hat{\psi}_j(\mathbf{x}), \qquad \hat{\beta}_{ij} = -\frac{1}{M}\sum_{m=1}^{M}\nabla_{\mathbf{x}_i}\hat{\psi}_j\left(\mathbf{x}^m\right)$

## 3 Scalable Spectral Stein Gradient Estimator

In general, representation learning for large-scale datasets is usually costly in terms of storage and computation. For instance, the dimension of images in the STL-10 dataset is $96 \times 96 \times 3$ (i.e., the vector length is 27648). This makes it almost impossible to directly estimate the gradient of MI between the input and representation. To alleviate this problem, we introduce random projection (RP) (Bingham & Mannila, 2001) to reduce the dimension of $\mathbf{x}$.

We briefly review RP. More details refer to Bingham & Mannila (2001). RP projects the original $d$-dimensional data into a $k$-dimensional ($k << d$) subspace. Concretely, let matrix $X_{d\times N}$ denote the original set of N $d$-dimensional data, the projection of the original data $X^{RP}_{k\times N}$ is obtained by introducing a random matrix $R_{k\times d}$ whose columns have unit length, as follows (Bingham & Mannila, 2001), $X^{RP}_{k\times N} = R_{k\times d}X_{d\times N}$. After the RP, distance between two original data vectors can be approximated by scaled Euclidean distance of these vectors in reduced spaces:

$$\|\mathbf{x}_1 - \mathbf{x}_2\| \approx \sqrt{d/k}\,\|R\mathbf{x}_1 - R\mathbf{x}_2\|, \tag{6}$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ denote the two data vectors in the original large dimensional space.

Based on the principle of RP, we can derive a Salable Spectral Stein Gradient Estimator, which is an efficient high-dimensional score function estimator. One can show that the RBF kernel satisfies Steins identity (Liu & Wang, 2016). Shi et al. (2018) also shows that it is a promising choice for SSGE with a lower error bound. To reduce the computation of the kernel similarities of SSGE in high-dimensional settings, we replace the input of SSGE with a projections obtained by RP according to the approximation of Eq. (6) for the computation of the RBF kernel.

## 4 Mutual Information Gradient Estimator

As gradient estimation is a straightforward and effective method in optimization, we propose a gradient estimator for MI based on score estimation of implicit distributions, which is called Mutual Information Gradient estimator (MIGE). In this section, we focus on three most general cases of MI gradient estimation for representation learning, and derive the corresponding MI gradient estimator for these circumstances.

We outline the general setting of training an encoder to learn a representation. Let $\mathcal{X}$ and $\mathcal{Z}$ be the domain, and $E_\psi : \mathcal{X} \to \mathcal{Z}$ with parameters $\psi$ denotes a continuous and (almost everywhere) differentiable parametric function, which is usually a neural network, namely an encoder. $p(\mathbf{x})$ denotes the empirical distribution given the input data $\mathbf{x} \in \mathcal{X}$. We can obtain the representation of the input data through the encoder, $\mathbf{z} = E_\psi(\mathbf{x})$. $q_\psi(\mathbf{z})$ is defined as the marginal distribution induced by pushing samples from $p(\mathbf{x})$ through encoder $E_\psi(.)$ We also define $q_\psi(\mathbf{x}, \mathbf{z})$ as the joint distribution with $\mathbf{x}$ and $\mathbf{z}$, which is determined by encoder $E_\psi(.)$.

**Circumstance I**. Given that the encoder $E_\psi(.)$ is deterministic, our goal is to estimate the gradient of MI between input $\mathbf{x}$ and encoder output $\mathbf{z}$ w.r.t. the encoder parameters $\psi$. There is a close relationship between mutual information and entropy, which is as following:$I_\psi(\mathbf{x}; \mathbf{z}) = H(\mathbf{x}) + H_\psi(\mathbf{z}) - H_\psi(\mathbf{x}, \mathbf{z})$.Here $H(\mathbf{x})$ is data entropy and not relevant to $\psi$. The optimization of $I_\psi(\mathbf{x}, \mathbf{z})$ with parameters $\psi$ can neglect the entry $H(\mathbf{x})$. We decompose the gradient of the entropy of $q_\psi(\mathbf{z})$ and $q_\psi(\mathbf{x}, \mathbf{z})$ as (see Appendix A.1):

$$\nabla_\psi H(\mathbf{z}) = -\nabla_\psi \mathbb{E}_{q_\psi(\mathbf{z})}[\log q(\mathbf{z})], \quad \nabla_\psi H(\mathbf{x}, \mathbf{z}) = -\nabla_\psi \mathbb{E}_{q_\psi(\mathbf{x},\mathbf{z})}[\log q(\mathbf{x}, \mathbf{z})]. \tag{7}$$

Hence, we can represent the gradient of MI between input $\mathbf{x}$ and encoder output $\mathbf{z}$ w.r.t. encoder parameters $\psi$ as following:

$$\nabla_\psi I_\psi(\mathbf{x}; \mathbf{z}) = -\nabla_\psi \mathbb{E}_{q_\psi(\mathbf{z})}[\log q(\mathbf{z})] + \nabla_\psi \mathbb{E}_{q_\psi(\mathbf{x},\mathbf{z})}[\log q(\mathbf{x}, \mathbf{z})]. \tag{8}$$

However, this equation is intractable since an expectation w.r.t $q_\psi(\mathbf{z})$ is directly not differentiable w.r.t $\psi$. Roeder et al. (2017) proposed a general variant of the standard reparameterization trick for the variational evidence lower bound, which demonstrates lower-variance. To address above problem, we adapt this trick for MI gradient estimator in representation learning. Specifically, we can obtain the samples from the marginal distribution of $\mathbf{z}$ by pushing samples from the data empirical distribution $p(\mathbf{x})$ through $E_\psi(.)$ for representation learning. Hence we can reparameterize the representations variable $\mathbf{z} \sim q_\psi(\mathbf{z})$ using a differentiable transformation:$\mathbf{z} = E_\psi(\mathbf{x}) \quad \text{with} \quad \mathbf{x} \sim p(\mathbf{x})$, where the data empirical distribution $p(\mathbf{x})$ is independent of encoder parameters $\psi$. This reparameterization can rewrite an expectation w.r.t $q_\psi(\mathbf{z})$ and $q_\psi(\mathbf{x}, \mathbf{z})$ such that the Monte Carlo estimate of the expectation is differentiable w.r.t $\psi$.

Relying on this reparameterization trick, we can represent the gradient of MI w.r.t. encoder parameters $\psi$ in Eq. 8 as follows:

$$\begin{aligned}\nabla_\psi I_\psi(\mathbf{x}; \mathbf{z}) = &-\mathbb{E}_{q(\mathbf{x})}[\nabla_\mathbf{z} \log q(E_\psi(\mathbf{x}))\nabla_\psi E_\psi(\mathbf{x})] \\ &+ \mathbb{E}_{q(\mathbf{x})}[\nabla_{(\mathbf{x},\mathbf{z})} \log q(\mathbf{x}, E_\psi(\mathbf{x}))\nabla_\psi(\mathbf{x}, E_\psi(\mathbf{x}))],\end{aligned} \tag{9}$$

where the score function $\nabla_\mathbf{z} \log q_{\boldsymbol{\psi}}(E_\psi(\mathbf{x}))$ can be estimated based on i.i.d. samples from an implicit density $q_\psi(\mathbf{E}_\psi(\mathbf{x}))$ (Shi et al., 2018; Song et al., 2019). The samples form the joint distribution $q_\psi(\mathbf{x}, \mathbf{z})$ are produced as following: we sample observations from empirical distribution $p(\mathbf{x})$; then the corresponding samples of $\mathbf{z}$ is obtained through $E_\psi(.)$. Hence we can also estimate $\nabla_{(\mathbf{x},\mathbf{z})} \log q(\mathbf{x}, E_\psi(\mathbf{x}))$ based on i.i.d. samples from $q_\psi(\mathbf{x}, E_\psi(\mathbf{x}))$. $\nabla_\psi E_\psi(\mathbf{x})$ and $\nabla_\psi(\mathbf{x}, E_\psi(\mathbf{x}))$ are directly computed with $\mathbf{x}$.

**Circumstance II**. Assume that we encode the input to latent data space $\mathbf{h} = C_\psi(\mathbf{x})$ that reflects useful structure in the data. Next, we summarize this latent variable mapping into final representations, $E_\psi(\mathbf{x}) = f_\psi \circ C_\psi(\mathbf{x})$. The gradient estimator of MI between $\mathbf{h}$ and $\mathbf{z}$ is represented by the data reparameterization trick as follows:

$$\nabla_\psi I_\psi(\mathbf{h}; \mathbf{z}) = \nabla_\psi H_\psi(\mathbf{h}) + \nabla_\psi H_\psi(\mathbf{z}) - \nabla_\psi H_\psi(\mathbf{h}, \mathbf{z}) \tag{10}$$

$$\begin{aligned}&= -\mathbb{E}_{q(\mathbf{x})}[\nabla_\mathbf{z} \log q(E_\psi(\mathbf{x}))\nabla_\psi E_\psi(\mathbf{x})] - \mathbb{E}_{q(\mathbf{x})}[\nabla_\mathbf{h} \log q(C_\psi(\mathbf{x}))\nabla_\psi C_\psi(\mathbf{x})] \\ &+ \mathbb{E}_{q(\mathbf{x})}[\nabla_{(\mathbf{h},\mathbf{z})} \log q(C_\psi(\mathbf{x}), E_\psi(\mathbf{x}))\nabla_\psi(C_\psi\mathbf{x}, E_\psi(\mathbf{x}))].\end{aligned} \tag{11}$$

**Circumstance III**. Consider stochastic encoder function $E_\psi(., \boldsymbol{\epsilon})$ where $\boldsymbol{\epsilon}$ is an auxiliary variable with independent marginal $p(\boldsymbol{\epsilon})$. By utilizing data reparameterization trick and reparameterization trick, we can represent the gradient of the conditional entropy $H_\psi(\mathbf{z}|\mathbf{x})$ as following (see Appendix A):

$$\nabla_\psi H_\psi(\mathbf{z}|\mathbf{x}) = -\mathbb{E}_{p(\mathbf{x})}[\mathbb{E}_{p(\boldsymbol{\epsilon})}[\nabla_{(\mathbf{z}|\mathbf{x})} \log q(E_\psi(\mathbf{x}, \boldsymbol{\epsilon})|\mathbf{x})\nabla_\psi E_\psi(\mathbf{x}, \boldsymbol{\epsilon})]], \tag{12}$$

where the term $\nabla_{(\mathbf{z}|\mathbf{x})} \log q(E_\psi(\mathbf{x}, \boldsymbol{\epsilon})|\mathbf{x})$ can be easily estimated by score estimation.

Based on the condition entropy gradient estimation in Eq. (12), the gradient estimator of MI between input and encoder output can be represented as following:

$$\nabla_\psi I_\psi(\mathbf{x};\mathbf{z}) = \nabla_\psi H_\psi(\mathbf{z}) - \nabla_\psi H_\psi(\mathbf{z}|\mathbf{x}) \tag{13}$$

$$= -\mathbb{E}_{p(\mathbf{x})p(\boldsymbol{\epsilon})}[\nabla_{\mathbf{z}}[\log p(E_\psi(\mathbf{x},\boldsymbol{\epsilon}))]\nabla_\psi E_\psi(\mathbf{x},\boldsymbol{\epsilon})] + \mathbb{E}_{p(\mathbf{x})}[\mathbb{E}_{p(\boldsymbol{\epsilon})}[\nabla_{(\mathbf{z}|\mathbf{x})}\log q(E_\psi(\mathbf{x},\boldsymbol{\epsilon})|\mathbf{x})\nabla_\psi E_\psi(\mathbf{x},\boldsymbol{\epsilon})]]. \tag{14}$$

In practical MI optimization, we can construct MIGE of the full dataset based on mini-batch Monte Carlo estimates. We have provided an algorithm description for MIGE in Appendix A2.

## 5 Toy Experiment

Recently, MINE and MINE-$f$ enable effective computation of MI in the continuous and high-dimensional settings. To compare with MINE and MINE-$f$, we evaluate MIGE in the correlated Gaussian problem taken from (Belghazi et al., 2018).

**Experimental Settings**. We consider two random variables $\mathbf{x}$ and $\mathbf{y}$ ($\mathbf{x},\mathbf{y} \in \mathcal{R}^d$), coming from a $2d$-dimension multivariate Gaussian distribution. The component-wise correlation of $\mathbf{x}$ and $\mathbf{y}$ is defined as follows: $corr(\mathbf{x}_i,\mathbf{y}_i) = \delta_{ij}\rho, \quad \rho \in (-1,1)$, where $\delta_{ij}$ is Kronecker's delta and $\rho$ is the correlation coefficient. Since MI is invariant to smooth transformations of $\mathbf{x},\mathbf{y}$, we only consider standardized Gaussian for marginal distribution $p(\mathbf{x})$ and $p(\mathbf{y})$. The gradient of MI w.r.t $\rho$ has the analytical solution: $\nabla_\rho I(\mathbf{x};\mathbf{y}) = \frac{\rho d}{1-\rho^2}$. We apply MINE and MINE-$f$ to estimate MI of $\mathbf{x},\mathbf{y}$ by sampling from the correlated Gaussian distribution and its marginal distributions, and the corresponding gradient of MI w.r.t $\rho$ can be computed by backpropagation implemented in Pytorch.

**Results**. Fig.1 presents our experimental results in different dimensions $d = \{5, 10, 20\}$. In the case of low-dimensional ($d = 5$), all the estimators give promising estimation of MI and its gradient. However, the MI estimation of MINE and MINE-$f$ are unstable due to its relying on a discriminator to produce estimation of the bound on MI. Hence, as showed in Fig.1, corresponding estimation of MI and its gradient is unsmoothed. As the dimension $d$ and the absolute value of correlation coefficient $|\rho|$ increase, MINE and MINE-$f$ are apparently hard to reach the True MI, and their gradient estimation of MI is thus high biased. This phenomenon would be more significant in the case of high-dimensional or large MI. Contrastively, MIGE demonstrates the significant improvement over MINE and MINE-$f$ when estimating MI gradient between twenty-dimensional random variables $\mathbf{x},\mathbf{y}$. In this experiment, we compare our method with two baselines on an analyzable problem and find that the gradient curve estimated by our method is far superior to other methods in terms of smoothness and tightness in a high-dimensional and large-MI setting compared with MINE and MINE-$f$.

## 6 Applications

To demonstrate the performance in downstream tasks, we deploy MIGE to Deep InfoMax (Hjelm et al., 2019) and Information Bottleneck (Tishby et al., 2000) respectively, namely replacing the original MI estimators with MIGE. We find that MIGE achieves higher and more stable classification accuracy, which indicating its good gradient estimation performance in practical applications.

### 6.1 Deep InfoMax

Discovering useful representations from unlabeled data is one core problem for deep learning. Recently, a growing set of methods is explored to train deep neural network encoders by maximizing the mutual information between its input and output. A number of methods based on tractable variational lower bounds, such as JSD and infoNCE, have been proposed to improve the estimation of MI between high dimensional input/output pairs of deep neural networks (Hjelm et al., 2019). To compare with JSD and infoNCE, we expand the application of MIGE in unsupervised learning of deep representations based on the InfoMax principle.

**Experimental Settings** For consistent comparison, we follow the experiments of Deep InfoMax(DIM) to set the experimental setup as in Hjelm et al. (2019). The implementation of DIM is

Table 1: CIFAR-10 and CIFAR-100 classification accuracy (top 1) of downstream tasks compared with vanilla DIM. JSD and infoNCE are MI estimators, and PM denotes matching representations to a prior distribution (Hjelm et al., 2019).

| **Model** | **CIFAR-10** | | | **CIFAR-100** | | |
|---|---|---|---|---|---|---|
| | conv | fc(1024) | Y(64) | conv | fc(1024) | Y(64) |
| DIM (JSD) | 55.81% | 45.73% | 40.67% | 28.41% | 22.16% | 16.50% |
| DIM (JSD + PM) | 52.2% | 52.84% | 43.17% | 24.40% | 18.22% | 15.22% |
| DIM (infoNCE) | 51.82% | 42.81% | 37.79% | 24.60% | 16.54% | 12.96% |
| DIM (infoNCE + PM) | 56.77% | 49.42% | 42.68% | 25.51% | 20.15% | 15.35% |
| MIGE | **57.95%** | **57.09%** | **53.75%** | **29.86%** | **27.91%** | **25.84%** |

Table 2: STL-10 classification accuracy (top 1) of downstream tasks compared with vanilla DIM. The dimension of STL-10 images (27648) results in exorbitant computational cost. Random Projection (RP) is applied to reduce the dimension.

| **Model** | **STL-10** | | |
|---|---|---|---|
| | conv | fc(1024) | Y(64) |
| DIM (JSD) | 42.03% | 30.28% | 28.09% |
| DIM (infoNCE) | 43.13% | 35.80% | 34.44% |
| MIGE | unaffordable computational cost | | |
| MIGE + RP to 1024d | 49.08% | 40.09% | **38.95%** |
| MIGE + RP to 512d | 49.89% | **41.05%** | 38.56% |
| MIGE + RP to 256d | **49.91%** | 40.24% | 38.83% |

available from its authors[1]. We test Deep InfoMax (DIM) on image datasets CIFAR-10, CIFAR-100 and STL-10 to evaluate our MIGE. For the high-dimensional images in STL-10, directly applying SSGE is almost impossible since it results in exorbitant computational cost. Our proposed Scalable SSGE is applied, to reduce the dimension of images and achieve reasonable computational cost. As mentioned in Hjelm et al. (2019), non-linear classifier is chosen to evaluate our representation, After learning representation, we freeze the parameters of the encoder and train a non-linear classifier using the representation as the input. The same classifiers are used for all methods. Our baseline results are directly copied from Hjelm et al. (2019) or by running the code of author.

**Results** As shown in Table 1, MIGE outperforms all the competitive models in DIM experiments on CIFAR-10 and CIFAR-100. Besides the numerical improvements, it is notable that our model have the less accuracy decrease across layers than that of DIM(JSD) and DIM(infoNCE). The results indicate that, compared to variational lower bound methods, MIGE gives more favorable gradient direction, and demonstrates more power in controlling information flows without significant loss. Table 2 shows the result of DIM experiments on STL-10. With the aid of Random Projection, we could evaluate on bigger datasets, e.g., STL-10. Especially, when the dimension of images is reduced to 256, we can observe significant improvement over the baselines. Note that our proposed gradient estimator can also be extended to the multi-view setting(i.e., with local and global features) of DIM, it is beyond the scope of this paper. More discussions refer to Appendix A.3.

### 6.2 Information Bottleneck

Information Bottleneck (IB) has been widely applied to a variety of application domains, such as classification (Tishby & Zaslavsky, 2015; Alemi et al., 2017; Chalk et al., 2016; Kolchinsky et al., 2017), clustering (Slonim & Tishby, 2000), and coding theory and quantization (Zeitler et al., 2008; Courtade & Wesel, 2011). In particular, given the input variable $\mathbf{x}$ and the target variable $\mathbf{y}$, the goal of the IB is to learn a representation of $\mathbf{x}$ (denoted by the variable $\mathbf{z}$) that satisfies the following characteristics:

[1] https://github.com/rdevon/DIM

Table 3: Permutation-invariant MNIST misclassification rate. Datas except our model are cited from Belghazi et al. (2018)

| **Model** | **Misclass rate** |
|---|---|
| Baseline | 1.38% |
| Dropout | 1.34% |
| Confidence penalty | 1.36% |
| Label Smoothing | 1.4% |
| DVB | 1.13% |
| MINE-IB | 1.11% |
| MIGE-IB (ours) | **1.05%** |

1) $\mathbf{z}$ is sufficient for the target $\mathbf{y}$, that is, all information about target $\mathbf{y}$ contained in $\mathbf{x}$ should also be contained in $\mathbf{z}$. In optimization, it should be
2) $\mathbf{z}$ is minimal. In order not to contain irrelevant information that is not related to $\mathbf{y}$, $\mathbf{z}$ is required to contain the smallest information among all sufficient representations.

The objective function of IB is written as follows:

$$\max I(\mathbf{z};\mathbf{y}), \quad \text{s.t. } I(\mathbf{z};\mathbf{x}) \leq c. \tag{15}$$

Equivalently, by introducing a Lagrangian multiplier $\beta$, the IB method can maximize the following objective function: $G_{IB} = I(\mathbf{z};\mathbf{y}) - \beta I(\mathbf{z};\mathbf{x})$. Further, it is generally acknowledged that $I(\mathbf{z};\mathbf{y}) = H(\mathbf{y}) - H(\mathbf{y}|\mathbf{z})$, and $H(\mathbf{y})$ is constant. Hence we can also minimize the objective function of the following form:

$$L_{IB} = H(\mathbf{y}|\mathbf{z}) + \beta I(\mathbf{z};\mathbf{x}), \tag{16}$$

where $\beta \geq 0$ plays a role in trading off the sufficiency and minimality. Note that the above formulas omit the parameters for simplicity. To overcome the intractability of MI in the continuous and high-dimension setting, Alemi et al. (2017) presents a variational approximation to IB, which adopts deep neural network encoder to produce a conditional multivariate normal distribution, called Deep Variational Bottleneck (DVB). Rencently, DVB is exploited to restrict the capacity of discriminators in GANs (Peng et al., 2019). However, a tractable density is required for the approximate posterior in DVB due to their reliance on a variational approximation while MIGE does not.

To evaluate our method, we compare MIGE-IB with DVB and MINE-IB in IB application. We demonstrate an implementation of the IB objective on permutation invariant MNIST (LeCun, 1998) using MIGE.

**Experiment** For consistent comparison, we adopt the same architecture and empirical settings used in Alemi et al. (2017) except that the initial learning rate of 2e-4 is set for Adam optimizer, and exponential decay with decaying rate by a factor of 0.96 was set for every 2 epochs. The implementation of DVB is available from its authors[2]. Under these experimental settings, we use our MI Gradient Estimator to replace the MI estimator in DVB experiment. The threshold of score function's Stein gradient estimator is set as $0.94$. The threshold is the hyper-parameter of Spectral Stein Gradient Estimator (SSGE), and it is used to set the kernel bandwidth of RBF kernel. Our results can be seen in Table 3 and it manifests that our proposed MIGE-IB outperforms DVB and MINE-IB.

## 7 Conclusion

In this paper, we present a gradient estimator, called Mutual Information Gradient Estimator (MIGE), to avoid the various problems met in direct mutual information estimation. We manifest the effectiveness of gradient estimation of MI over direct MI estimation by applying it in unsupervised or supervised representation learning. Experimental results have indicated the remarkable improvement over MI estimation in the Deep InfoMax method and the Information Bottleneck method.

[2] `https://github.com/alexalemi/vib_demo`

## REFERENCES


Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *ICLR*, 2017.

Alexander A Alemi, Ian Fischer, and Joshua V Dillon. Uncertainty in the variational information bottleneck. *arXiv preprint arXiv:1807.00906*, 2018a.

Alexander A. Alemi, Ben Poole, Ian Fischer, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a broken ELBO. In *ICML*, 2018b.

David Barber and Felix V Agakov. The im algorithm: a variational approach to information maximization. In *Advances in neural information processing systems*, pp. None, 2003.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *ICML*, 2018.

Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, San Francisco, CA, USA*, pp. 245–250, 2001.

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

Matthew Chalk, Olivier Marre, and Gasper Tkacik. Relevant sparse codes with variational information bottleneck. In *Advances in Neural Information Processing Systems*, pp. 1957–1965, 2016.

Thomas A Courtade and Richard D Wesel. Multiterminal source coding with an entropy-based distortion measure. In *2011 IEEE International Symposium on Information Theory Proceedings*, pp. 2040–2044. IEEE, 2011.

Georges A Darbellay and Igor Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321, 1999.

François Fleuret. Fast binary feature selection with conditional mutual information. *Journal of Machine learning research*, 5(Nov):1531–1555, 2004.

Andrew M Fraser and Harry L Swinney. Independent coordinates for strange attractors from mutual information. *Physical review A*, 33(2):1134, 1986.

Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Efficient estimation of mutual information for strongly dependent variables. In *Artificial intelligence and statistics*, pp. 277–286, 2015.

Jackson Gorham and Lester Mackey. Measuring sample quality with stein's method. In *Advances in Neural Information Processing Systems*, pp. 226–234, 2015.

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 *ICLR*, 2019.

Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *ICML*, 2017.

Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabas Poczos, Larry Wasserman, et al. Nonparametric von mises estimators for entropies, divergences and mutual informations. In *Advances in Neural Information Processing Systems*, pp. 397–405, 2015.

Justin B Kinney and Gurinder S Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359, 2014.

Artemy Kolchinsky, Brendan D Tracey, and David H Wolpert. Nonlinear information bottleneck. *arXiv preprint arXiv:1705.02436*, 2017.

LF Kozachenko and Nikolai N Leonenko. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16, 1987.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

Andreas Krause, Pietro Perona, and Ryan G Gomes. Discriminative clustering by regularized information maximization. In *Advances in neural information processing systems*, 2010.

Smita Krishnaswamy, Matthew H Spitzer, Michael Mingueneau, Sean C Bendall, Oren Litvin, Erica Stone, Dana Peer, and Garry P Nolan. Conditional density-based analysis of t cell signaling in single-cell data. *Science*, 346(6213):1250689, 2014.

Nojun Kwak and Chong-Ho Choi. Input feature selection by mutual information based on parzen window. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12):1667–1671, 2002.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Yingzhen Li and Richard E Turner. Gradient estimators for implicit models. *arXiv preprint arXiv:1705.07107*, 2017.

Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in neural information processing systems*, pp. 2378–2386, 2016.

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

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

Young-Il Moon, Balaji Rajagopalan, and Upmanu Lall. Estimation of mutual information using kernel density estimators. *Physical Review E*, 52(3):2318, 1995.

Andreas C Müller, Sebastian Nowozin, and Christoph H Lampert. Information theoretic clustering using minimum spanning trees. In *Joint DAGM (German Association for Pattern Recognition) and OAGM Symposium*, pp. 205–215. Springer, 2012.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *NIPS*, 2018.

Stephanie E Palmer, Olivier Marre, Michael J Berry, and William Bialek. Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, 112(22):6908–6913, 2015.

Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (8):1226–1238, 2005.

Xue Bin Peng, Angjoo Kanazawa, Sam Toyer, Pieter Abbeel, and Sergey Levine. Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

Fernando Pérez-Cruz. Kullback-leibler divergence estimation of continuous distributions. In *2008 IEEE international symposium on information theory*, pp. 1666–1670. IEEE, 2008.

Ben Poole, Sherjil Ozair, Aäron van den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 5171–5180, 2019.

Geoffrey Roeder, Yuhuai Wu, and David K Duvenaud. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *Advances in Neural Information Processing Systems*, pp. 6925–6934, 2017.

Jiaxin Shi, Shengyang Sun, and Jun Zhu. A spectral approach to gradient estimation for implicit distributions. *arXiv preprint arXiv:1806.02925*, 2018.

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

Shashank Singh and Barnabás Póczos. Finite-sample analysis of fixed-k nearest neighbor density functional estimators. In *Advances in neural information processing systems*, pp. 1217–1225, 2016.

Noam Slonim and Naftali Tishby. Document clustering using word clusters via the information bottleneck method. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 208–215. ACM, 2000.

Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. *arXiv preprint arXiv:1905.07088*, 2019.

Taiji Suzuki, Masashi Sugiyama, Jun Sese, and Takafumi Kanamori. Approximating mutual information by maximum likelihood density ratio estimation. In *New challenges for feature selection in data mining and knowledge discovery*, pp. 5–20, 2008.

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

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

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *ArXiv*, abs/1907.13625, 2019.

Greg Ver Steeg and Aram Galstyan. Maximally informative hierarchical representations of high-dimensional data. In *Artificial Intelligence and Statistics*, pp. 1004–1012, 2015.

Matias Vera, Pablo Piantanida, and Leonardo Rey Vega. The role of the information bottleneck in representation learning. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 1580–1584. IEEE, 2018.

Georg Zeitler, Ralf Koetter, Gerhard Bauch, and Joerg Widmer. Design of network coding functions in multihop relay networks. In *2008 5th International Symposium on Turbo Codes and Related Topics*, pp. 249–254. IEEE, 2008.

# A APPENDIX

**Algorithm 1** MIGE (Circumstance I)

**1. Sampling**:
Draw $n$ samples from the data distribution $p(x)$, $n$ denotes mini-batch size,
then compute the corresponding output of the encoder
$(\mathbf{x}^{(1)}, \mathbf{z}^{(1)}), \cdots, (\mathbf{x}^{(n)}, \mathbf{z}^{(n)}) \sim q_\psi(\mathbf{x}, \mathbf{z})$
$\mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(n)} \sim q_\psi(\mathbf{z})$

**2. Estimate the score function**:
$\nabla_{\mathbf{z}} \log q_\psi(\mathbf{z}^{(i)}) \leftarrow \mathrm{SSGE}(\mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(n)})$
$\nabla_{(\mathbf{x},\mathbf{z})} \log q_\psi(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) \leftarrow \mathrm{SSGE}((\mathbf{x}^{(1)}, \mathbf{z}^{(1)}), \cdots, (\mathbf{x}^{(n)}, \mathbf{z}^{(n)}))$

**3. Estimate the entropy gradient**:
$\nabla_\psi H(\mathbf{z}) \leftarrow -\frac{1}{n}\sum_{i=1}^{n}\left[\nabla_\psi \mathbf{z}^{(i)} \nabla_{\mathbf{z}} \log q_\psi(\mathbf{z}^{(i)})\right]$
$\nabla_\psi H(\mathbf{x}, \mathbf{z}) \leftarrow -\frac{1}{n}\sum_{i=1}^{n}\left[\nabla_\psi (\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) \nabla_{(\mathbf{x},\mathbf{z})} \log q_\psi(\mathbf{x}^{(i)}, \mathbf{z}^{(i)})\right]$

**4. Estimate the MI gradient**:
$\nabla_\psi I(\mathbf{x}; \mathbf{z}) \leftarrow \nabla_\psi H(\mathbf{z}) - \nabla_\psi H(\mathbf{x}; \mathbf{z})$

## A.1 DERIVATION OF GRADIENT ESTIMATES FOR ENTROPY

**Unconditional Entropy** Given that the encoder $E_\psi(.)$ is deterministic, our goal is to optimize the entropy $\mathrm{H}(q) = -\mathbb{E}_q \log q$, where $q$ is short for the distribution $q_\psi(\mathbf{z})$ of the representation $\mathbf{z}$ w.r.t. its parameters $\psi$. We can decompose the gradient of the entropy of $q_\psi(\mathbf{z})$ as:

$$\nabla_\psi H(z) = -\nabla_\psi \mathbb{E}_{q_\psi(\mathbf{z})}[\log q(\mathbf{z})] - \mathbb{E}_{q(\mathbf{z})}[\nabla_\psi \log q_\psi(\mathbf{z})], \tag{17}$$

The second term on the right side of the equation can be calculated:

$$\mathbb{E}_{q(\mathbf{z})}[\nabla_\psi \log q_\psi(\mathbf{z})] = \mathbb{E}_{q(\mathbf{z})}[\nabla_\psi q_\psi(\mathbf{z}) \times \frac{1}{q(\mathbf{z})}] = \int \nabla_\psi q_\psi(\mathbf{z}) d\mathbf{z} = \nabla_\psi \int q_\psi(\mathbf{z}) d\mathbf{z} = 0. \tag{18}$$

Therefore, the gradient of the entropy of $q_\psi(\mathbf{z})$ becomes

$$\nabla_\psi H(z) = -\nabla_\psi \mathbb{E}_{q_\psi(\mathbf{z})}[\log q(\mathbf{z})]. \tag{19}$$

**Conditional Entropy** Consider nondeterministic encoder function $E_\psi(., \boldsymbol{\epsilon})$ where $\boldsymbol{\epsilon}$ is an auxiliary variable with independent marginal $p(\boldsymbol{\epsilon})$. The distribution $q_\psi(z|x)$ is determined by $\boldsymbol{\epsilon}$ and the encoder parameters $\psi$. The auxiliary variable $\boldsymbol{\epsilon}$ introduces randomness to the encoder. First, we decompose the gradients of Conditional Entropy as following:

$$\begin{aligned}
\nabla_\psi H\left(\mathbf{z}|\mathbf{x}\right) &= -\nabla_\psi \int p_\psi(\mathbf{z}, \mathbf{x}) \log p_\psi(\mathbf{z}|\mathbf{x}) dz dx \\
&= -\mathbb{E}_{p(\mathbf{x})}[\nabla_\psi \int p_\psi(\mathbf{z}|\mathbf{x}) \log p_\psi(\mathbf{z}|\mathbf{x}) dz] \\
&= -\mathbb{E}_{p(\mathbf{x})}[\nabla_\psi \mathbb{E}_{p_\psi(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{z}|\mathbf{x})] + \int p(\mathbf{z}|\mathbf{x}) \nabla_\psi \log p_\psi(\mathbf{z}|\mathbf{x}) dh] \\
&= -\mathbb{E}_{p(\mathbf{x})}[\nabla_\psi \mathbb{E}_{p_\psi(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{z}|\mathbf{x})] + \int \nabla_\psi p_\psi(\mathbf{z}|\mathbf{x}) dh] \\
&= -\mathbb{E}_{p(\mathbf{x})}[\nabla_\psi \mathbb{E}_{p_\psi(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{z}|\mathbf{x})] - \nabla_\psi \int p_\psi(\mathbf{h}, \mathbf{x}) dh dx] \\
&= -\mathbb{E}_{p(\mathbf{x})}[\nabla_\psi \mathbb{E}_{p_\psi(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{z}|\mathbf{x})]].
\end{aligned} \tag{20}$$

Note that $\mathbf{z} = E_\psi(\mathbf{x}, \boldsymbol{\epsilon})$, such that we can apply reparameterization trick to the gradient estimator of conditional entropy in Eq. (20),

$$H_\psi(\mathbf{z}|\mathbf{x}) = -\mathbb{E}_{p(\mathbf{x})}[\mathbb{E}_{p(\boldsymbol{\epsilon})}[\nabla_{(\mathbf{z}|\mathbf{x})} \log q(E_\psi(\mathbf{x}, \boldsymbol{\epsilon})|\mathbf{x}) \nabla_\psi E_\psi(\mathbf{x}, \boldsymbol{\epsilon})]]. \tag{21}$$

Table 4: CIFAR-10 and CIFAR-100 classification accuracy (top 1) of downstream tasks compared with vanilla DIM(L).

| **Model** | **CIFAR-10** | | | **CIFAR-100** | | |
|---|---|---|---|---|---|---|
| | conv | fc(1024) | Y(64) | conv | fc(1024) | Y(64) |
| DIM(L) (JSD) | 72.16% | 67.99% | 66.35% | 41.65% | 39.60% | 39.66% |
| DIM(L) (JSD + PM) | 73.25% | 73.62% | 66.96% | 48.13% | 45.92% | 39.6% |
| DIM(L) (infoNCE) | 75.05% | 70.68% | 69.24% | 44.11% | 42.97% | 42.74% |
| DIM(L) (infoNCE + PM) | 75.21% | 75.57% | 69.13% | 49.74% | 47.72% | 41.61% |
| MIGE | **59.72%** | **56.14%** | **54.01%** | **30.0%** | **28.96%** | **27.65%** |

### A.2 Algorithm Description

The algorithm description of our proposed MIGE is stated in Algorithm 1.

### A.3 Discussion on DIM(L)

DIM(L) (Hjelm et al., 2019) is the state-of-the-art unsupervised model for representaion learning, which maximizes the average MI between the high-level representation and local patches of the image, and achieve an even higher classification accuracy than supervised learning. As shown in Table 4, we apply MIGE into DIM(L) and surprisingly find there is a significant performance gap to DIM(L).

To our knowledge, the principle of DIM(L) is still unclear. Tschannen et al. (2019) argues that maximizing tighter bounds in DIM(L) can lead to worse results, and the success of these methods cannot be attributed to the properties of MI alone, and they strongly depend on the inductive bias in both the choice of feature extractor architectures and the parameterization of the employed MI estimators. For MIGE, we are investigating the behind reasons, e.g., to investigate the distributions of the patches.