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 existing methods are not capable of providing accurate estimation of MI with low-variance when the MI is large. We argue that estimating gradients of MI is more appealing for representation learning than directly estimating MI due to the difficulty of estimating MI. Therefore, we propose the Mutual Information Gradient Estimator (MIGE) for representation learning based on score estimation of implicit distributions. It exhibits a tight and smooth gradient estimation of MI in the high-dimensional and large-MI setting. 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 remarkable 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 mutual information, 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 mutual information. 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 mutual information 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 the Information Bottleneck method (Tishby et al., 2000; Alemi et al., 2017), where mutual information 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 of the data 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, mutual information 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, mutual information is challenging to analytically compute or to estimate from samples. A variety of mutual information 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 mutual information 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 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 from the product of the marginals. In detail, MINE employs a lower-bound to the mutual information based on the Donsker-Varadhan representation of the KL-divergence, while JSD follows the formulation of f-GAN KL-divergence. In general these estimates 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 estimates 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, however, 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 discussion about the bounds of MI and their relationship refers to Poole et al. (2019).

As discussed in the last paragraph, existing estimators first approximate MI and then use those approximations to optimize the associated parameters. In practice, we do not care about MI estimation and only care about computing gradients of MI during optimization. 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 the true MI. Estimating gradients of MI than estimating MI may be better approaches for MI optimization. To this end, to the best of our knowledge, we firstly propose the gradient estimator of MI 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 gradient estimation of MI 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 expensive and complex computation. 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).

In summary, the contributions of this paper are 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 4.

- We present a gradient estimation solution to the unsupervised representation learning based on InfoMax. It remarkably improves the performance of deep information models.

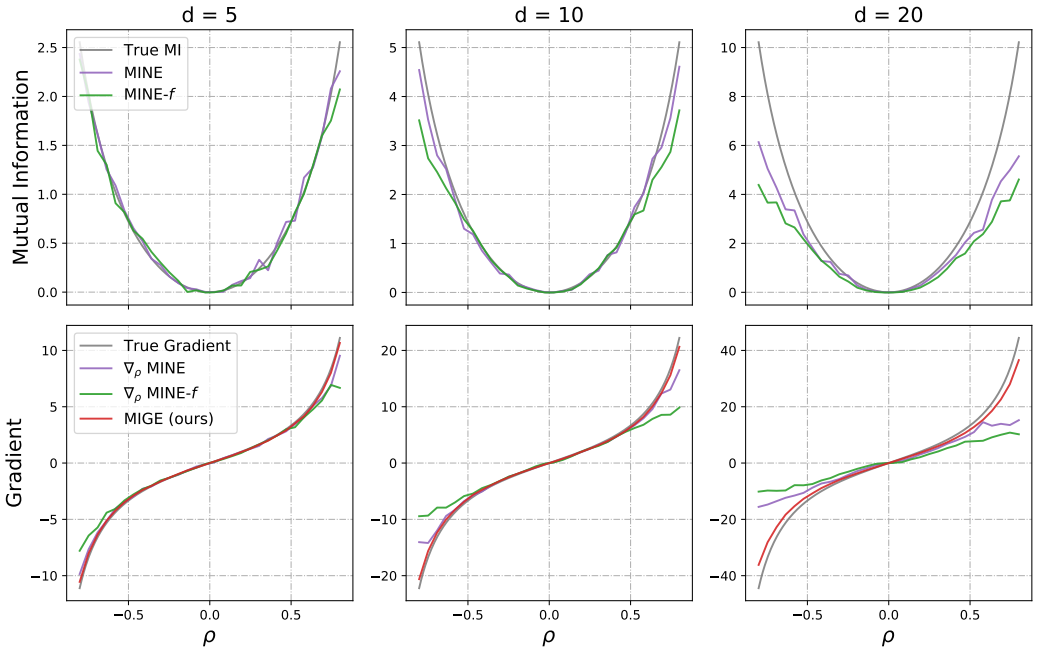

Figure 1: Estimation performance of MINE, MINE-f and MIGE. **Top:** True MI and corresponding estimation of MINE and MINE-f. **Bottom:** True gradient and corresponding estimation of MINE, MINE-f and MIGE. Our approach MIGE only appears in bottom figures since it gives gradient estimation directly. As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

- We present a gradient estimator of the Information Bottleneck method with MIGE in a continuous setting. Experimental results have indicated that our method outperforms variational bottleneck methods and MINE information bottleneck methods.

2 Background

In this section we briefly introduce score estimation and information bottleneck.

2.1 Score Estimation

Score estimation of implicit distributions has been widely explored in the past few years. Methods for score estimation of implicit distributions usually follow two lines. The first line of work is to directly estimate the score of implicit distributions, such as the sliced score matching (Song et al., 2019). The second line of work tries to estimate the score of the logarithmic density of implicit distributions, such as the Stein gradient estimator (Li & Turner, 2017; Shi et al., 2018).

Score matching (Hyvärinen, 2005) is a widely used method for estimating unnormalized statistical models, which minimizes the Fisher divergence between the estimated parameterized distribution and the implicit data distribution. However, score matching requires computing the diagonal elements of the Hessian of the logarithmic density function. It is known that the computation of the Hessian trace is expensive (Martens et al., 2012), since it requires multiple forward and backward propagation, proportional to the data dimension. In order to overcome this disadvantages, Song et al. (2019) proposes the sliced score matching method for estimating scores for implicit distributions. Compared with the original score matching and its variants, this estimator can be applied to deep models of high-dimensional data, while remaining easy to implement in modern automatic differentiation frameworks. However, in order to maximize MI between a pair of random variables, sliced score matching requires internal loop optimization to minimize Fisher divergence.

Stein gradient estimator (Li & Turner, 2017; Shi et al., 2018) is an alternative method of score estimation for the logarithmic density of an implicit distribution, whose core idea 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 $\boldsymbol{h}(\boldsymbol{x}) = [h_1(\boldsymbol{x}), h_2(\boldsymbol{x}), \ldots, h_{d'}(\boldsymbol{x})]^{\mathrm{T}}$ is a smooth vector function. Further, the boundary conditions on $\boldsymbol{h}$ is

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

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

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

Here $\boldsymbol{h}$ is called as the Stein class of $q(\boldsymbol{x})$ if Steins identity (3) holds. Monte Carlo estimates of expectation in Equation (3) builds the connection between $\nabla_{\mathbf{x}} \log q(\mathbf{x})$ and the samples from $q(x)$ in Steins identity. 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 being approximated by the Nyström method. 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 ith component of the gradient is $g_i(\mathbf{x}) = \nabla_{x_i} \log q(\mathbf{x})$. We assume $g_1, \ldots, g_d \in L^2(\mathcal{X}, q)$. And $\{\psi_j\}_{j\geq 1}$ denotes an orthonormal basis of $L^2(\mathcal{X}, q)$. So we can expand $g_i(\mathbf{x})$ into the following spectral series: $g_i(\mathbf{x}) = \sum_{j=1}^{\infty} \beta_{ij}\psi_j(\mathbf{x})$.

We can estimate the coefficients β_{ij} by Steins identity. And 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:

$$\hat{g}_i(\mathbf{x}) = \sum_{j=1}^{J} \hat{\beta}_{ij}\hat{\psi}_j(\mathbf{x}), \qquad \hat{\beta}_{ij} = -\frac{1}{M}\sum_{m=1}^{M} \nabla_{x_i}\hat{\psi}_j\left(\mathbf{x}^m\right), \tag{4}$$

where $\hat{\psi}_j(\mathbf{x})$ is the Nyström approximation of $\psi_j(\mathbf{x})$.

2.2 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). IB is first introduced by Tishby et al. (1999) as a method of seeking a representation that weighed the sufficiency for the target and the complexity of the representation. 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) $\mathbf{z}$ 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 achieved by maximizing the information between $\mathbf{z}$ and $\mathbf{z}$.
2) $\mathbf{z}$ is minimal. It can be known that there are many representations satisfying the point 1). Therefore for streamlining, $\mathbf{z}$ is required to contain the smallest information among all sufficient representations.

Since mutual information quantifies the dependence between two random variables, IB introduces it to characterize the above two characteristics. The first characteristic above can be represented by $I(\mathbf{z}; \mathbf{y}) = I(\mathbf{z}; \mathbf{x})$. In detail, we implement this by maximizing the $I(\mathbf{z}; \mathbf{y})$. And the second characteristic above indicates that $I(\mathbf{z}; \mathbf{x})$ should be smallest among all possible representations. More specifically, the IB applies a natural constraint to implement the second point, namely $I(\mathbf{z}, \mathbf{x}) \leq C$ (Witsenhausen & Wyner (1975)), where c is the information constraint.

Based on the goal of IB, the objective function is written as follows:

$$\max I(\mathbf{z}; \mathbf{y}), \quad \text{s.t. } I(\mathbf{z}; \mathbf{x}) \leq c. \tag{5}$$

Equivalently, by introducing a Lagrangian multiplier β, the IB method can maximize the following objective function:

$$G_{IB} = I(\mathbf{z}; \mathbf{y}) - \beta I(\mathbf{z}; \mathbf{x}). \tag{6}$$

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:

$$L_{IB} = H(\mathbf{y}|\mathbf{z}) + \beta I(\mathbf{z}; \mathbf{x}), \tag{7}$$

where $\beta \geq 0$ plays a role in trading off the sufficiency and minimality. Note that the above formulas omit the parameters for simplicity.

3 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 ψ 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 assume obtain to a representation $\mathbf{z} = E_\psi(\mathbf{x})$, which has some desirable properties for specific tasks.

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. encoder parameters ψ. There is a close relationship between mutual information and entropy, which is following:

$$I_\psi(\mathbf{x}; \mathbf{z}) = H(\mathbf{x}) + H_\psi(\mathbf{z}) - H_\psi(\mathbf{x}, \mathbf{z}), \tag{8}$$

Here $H(\mathbf{x})$ is data entropy and not relevant to ψ. The optimization of $I_\psi(\mathbf{x}, \mathbf{z})$ with parameters ψ can neglect the entry $H(\mathbf{x})$. And we decompose the gradient of the entropy of $q_\psi(\mathbf{z})$ and $q_\psi(\mathbf{x}, \mathbf{z})$ as (see Appendix A):

$$\nabla_\psi H(\mathbf{z}) = -\nabla_\psi \mathbb{E}_{q_\psi(\mathbf{z})}[\log q(\mathbf{z})], \quad \nabla_\psi H(\mathbf{x}, \mathbf{z}) = -\nabla_\psi \mathbb{E}_{q_\psi(\mathbf{x},\mathbf{z})}[\log q(\mathbf{x}, \mathbf{z})]. \tag{9}$$

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

$$\nabla_\psi I_\psi(\mathbf{x}; \mathbf{z}) = -\nabla_\psi \mathbb{E}_{q_\psi(\mathbf{z})}[\log q(\mathbf{z})] + \nabla_\psi \mathbb{E}_{q_\psi(\mathbf{x},\mathbf{z})}[\log q(\mathbf{x}, \mathbf{z})]. \tag{10}$$

However, this equation is intractable since an expectation w.r.t $q_\psi(\mathbf{z})$ is directly not differentiable w.r.t ψ. 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, called data reparameterization trick. 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:

$$\mathbf{z} = E_\psi(\mathbf{x}) \quad \text{with} \quad \mathbf{x} \sim p(\mathbf{x}), \tag{11}$$

where the data empirical distribution $p(\mathbf{x})$ is independent of encoder parameters ψ. 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 ψ.

Relying on data reparameterization trick, we can represent the gradient of MI w.r.t. encoder parameters ψ in Equation 10 as follows:

$$\begin{aligned}\nabla_\psi I_\psi(\mathbf{x}; \mathbf{z}) = &-\mathbb{E}_{q(\mathbf{x})}[\nabla_\mathbf{z} \log q(E_\psi(\mathbf{x}))\nabla_\psi E_\psi(\mathbf{x})] \\ &+ \mathbb{E}_{q(\mathbf{x})}[\nabla_{(\mathbf{x},\mathbf{z})} \log q(\mathbf{x}, E_\psi(\mathbf{x}))\nabla_\psi(\mathbf{x}, E_\psi(\mathbf{x}))],\end{aligned} \tag{12}$$

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 $q_{\boldsymbol{\psi}}(\mathbf{E}_\psi(\mathbf{x}))$ (Shi et al., 2018; Song et al., 2019). The samples form the joint distribution $p_\psi(\mathbf{x}, \mathbf{z})$ 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 $p_\psi(\mathbf{x}, E_\psi(\mathbf{x}))$. $\nabla_\psi E_\psi(\mathbf{x})$ and $\nabla_\psi(\mathbf{x}, E_\psi(\mathbf{x}))$ is 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 map into final representation, $E_\psi(\mathbf{x}) = f_\psi \circ C_\psi(\mathbf{x})$. The gradient estimator of MI between $\mathbf{h}$ and $\mathbf{z}$ is represent by the data reparameterization trick as follows:

$$\begin{aligned}\nabla_\psi I_\psi(\mathbf{h}; \mathbf{z}) &= \nabla_\psi H_\psi(\mathbf{h}) + \nabla_\psi H_\psi(\mathbf{z}) - \nabla_\psi H_\psi(\mathbf{h}, \mathbf{z}) \quad (13)\\ &= -\mathbb{E}_{q(\mathbf{x})}[\nabla_{\mathbf{z}} \log q(E_\psi(\mathbf{x}))\nabla_\psi E_\psi(\mathbf{x})] - \mathbb{E}_{q(\mathbf{x})}[\nabla_{\mathbf{h}} \log q(C_\psi(\mathbf{x}))\nabla_\psi C_\psi(\mathbf{x})]\\ &+ \mathbb{E}_{q(\mathbf{x})}[\nabla_{(\mathbf{h},\mathbf{z})} \log q(C_\psi(\mathbf{x}), E_\psi(\mathbf{x}))\nabla_\psi(C_\psi \mathbf{x}, E_\psi(\mathbf{x}))]. \quad (14)\end{aligned}$$

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):

$$\nabla_\psi H_\psi(\mathbf{z}|\mathbf{x}) = -\mathbb{E}_{p(\mathbf{x})}[\mathbb{E}_{p(\boldsymbol{\epsilon})}[\nabla_{(\mathbf{z}|\mathbf{x})} \log q(E_\psi(\mathbf{x}, \boldsymbol{\epsilon})|\mathbf{x})\nabla_\psi E_\psi(\mathbf{x}, \boldsymbol{\epsilon})]], \quad (15)$$

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 Equation (15), the gradient estimator of MI between input and encoder output can be represented as following:

$$\begin{aligned}\nabla_\psi I_\psi(\mathbf{x}; \mathbf{z}) &= \nabla_\psi H_\psi(\mathbf{z}) - \nabla_\psi H_\psi(\mathbf{z}|\mathbf{x}) \quad (16)\\ &= -\mathbb{E}_{p(\mathbf{x})p(\boldsymbol{\epsilon})}[\nabla_{\mathbf{z}}[\log p(E_\psi(\mathbf{x}, \boldsymbol{\epsilon}))]\nabla_\psi E_\psi(\mathbf{x}, \boldsymbol{\epsilon})]\\ &+ \mathbb{E}_{p(\mathbf{x})}[\mathbb{E}_{p(\boldsymbol{\epsilon})}[\nabla_{(\mathbf{z}|\mathbf{x})} \log q(E_\psi(\mathbf{x}, \boldsymbol{\epsilon})|\mathbf{x})\nabla_\psi E_\psi(\mathbf{x}, \boldsymbol{\epsilon})]]. \quad (17)\end{aligned}$$

In practical MI optimization, we can construct MIGE of the full dataset based on minibatche Monte Carlo estimates. The bias of MIGE depends on score estimation of implicit distributions. We prefer Spectral Stein Gradient Estimator (SSGE) for score estimation. And the error bound of SSGE is proved in Shi et al. (2018).

Scalable Spectral Stein Gradient Estimator It is worth noting that the estimation of the $\nabla_{(\mathbf{x},\mathbf{z})} \log q(\mathbf{x}, E_\psi(\mathbf{x}))$ by SSGE has a large computational complexity in high dimensional input spaces, such as text, images and videos. To alleviate this problem, we introduce random projection (RP) (Bingham & Mannila, 2001) to reduce the dimension of $\mathbf{x}$. 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,

$$X^{RP}_{k\times N} = R_{k\times d} X_{d\times N}. \quad (18)$$

After the RP, original distances between two original data vectors can be approximated by scaled Euclidean distance of these vectors in reduced spaces:

$$\|x_1 - x_2\| \approx \sqrt{d/k}\, \|Rx_1 - Rx_2\|, \quad (19)$$

where x_1 and x_2 denote the two data vectors in the original large dimensional space. It is obvious that the computation of RP is simple. By replacing the input of the SSGE with a projection obtained by random projection, we can derive a Scalable Spectral Stein Gradient Estimator, which is an efficient high-dimensional score function estimator.

4 EXPERIMENTS

To demonstrate the performance on gradient estimation, we evaluate our method MIGE in both a toy experiment and real-world tasks. In the toy experiment, we compare our method with two

Table 1: InfoMax on CIFAR-10 and CIFAR-100. JSD and infoNCE refer to the variational lower bound, and the PM refers to Prior Matching which is proposed by Hjelm et al. (2019) . The result of DIM (JSD + PM) is cited from 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 (ours)	**57.95**	**57.09**	**53.75**	**29.86**	**27.91**	**25.84**

baselines on analyzable problems and find that the gradient curve estimated by our method is far superior to other methods in terms of smoothness and accuracy. Furthermore, we deploy MIGE to the InfoMax principle and the Information Bottleneck respectively, namely replacing the original mutual information estimation term with MIGE. We find that MIGE achieves higher and more stable classification accuracy in CIFAR-10, CIFAR-100, and MNIST datasets, indicating that it has good performance in downstream tasks. In our experiments, we use the Stein gradient estimator (Shi et al., 2018) to estimate the score function term.

4.1 Toy Experiment

We evaluate our method MIGE in the toy experiment for the MI gradient estimation task, comparing to two high performing baselines, including MINE (Belghazi et al., 2018) and MINE-f (Nowozin et al., 2016; Belghazi et al., 2018). In the toy experiment, 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:

$$corr(\mathbf{x}_i, \mathbf{y}_i) = \delta_{ij}\rho, \quad \rho \in (-1, 1), \tag{20}$$

where δ_{ij} is Kronecker's delta and ρ is correlation coefficient. Since MI is invariant to smooth transformations of random variables $\mathbf{x}, \mathbf{y}$, we only consider standardized Gaussian for marginal distribution $p(\mathbf{x})$ and $p(\mathbf{y})$. MI between random variables $\mathbf{x}, \mathbf{y}$ is only relative to the correlation coefficient ρ. The gradient of MI w.r.t ρ has the analytic 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 ρ can be computed by backpropagation implemented in frameworks PyTorch.

Fig.1 presents our experiment 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 remarkable improvement over MINE and MINE-f when estimating MI gradient between twenty-dimensional random variables $\mathbf{x}, \mathbf{y}$. It provides a tighter and smoother gradient estimate of MI in a high-dimension and large-MI setting.

4.2 Deep InfoMax Experiment

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.

Table 2: 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%**

Follow Hjelm et al. (2019), we test Deep InfoMax(DIM) on image datasets CIFAR-10 and CIFAR-100 to evaluate our MI Gradient estimator MIGE. CIFAR-10 and CIFAR-100 each consists of 32×32 colored images, with 50,000 training images and 10,000 testing images. We adopt the same encoder architecture used in Hjelm et al. (2019), which uses a deep convolutional GAN (DCGAN, Radford et al. (2015)) consisting of 3 convolutional layers and 2 fully connected layer. The same empirical setup is used. Follow Hjelm et al. (2019), we choose images classification as the downstream task, then evaluate our representation in terms of the accuracy of transfer learning classification, that is, freezing the weights of the encoder and training a small fully-connected neural network classifier using the representation as the input.

As shown in Table 1, our proposed MIGE outperforms all the competitive models. Besides the numerical improvements, it is notable that our model have the less accuracy decrease across layers, whereas those methods based on variational lower bounds shrink a lot. The results indicate that, compared to variational lower bound methods, our approach MIGE gives much more favorable gradient direction, and demonstrates more power in controlling information flows without vast loss. 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.

4.3 Information Bottleneck

To overcome the intractability of MI in the continuous setting, Alemi et al. (2017) present a variational approximation to the information bottleneck, which adopts deep neural network encoder to produce a conditional multivariate normal distribution, called Deep Variational Bottleneck (DVB). Rencently, DVB is exploited to restricted 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.

Here, we demonstrate an implementation of the IB objective on permutation invariant MNIST using MIGE. We compare MIGE-IB with DVB and MINE-IB in the case of the same model structure. And most of the empirical setup is the same as DVB Alemi et al. (2017), but a little bit different. In our experiment, we use initial learning rate of 10_4 for Adam optimizer, and exponential decay, decaying the learning rate by a factor of 0.96 every 2 epochs. And the threshold of score function's Stein gradient estimator is set as 0.94. Our results can be seen in Table 2 and it manifests that our proposed MIGE-IB outperforms DVB and MINE-IB.

5 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.

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.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709, 2005.

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.

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.

James Martens, Ilya Sutskever, and Kevin Swersky. Estimating the hessian by back-propagating curvature. *arXiv preprint arXiv:1206.6464*, 2012.

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.

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

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.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

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. In *The 37th annual Allerton Conference on Communication, Control, and Computing*, pp. 368–377, 1999.

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

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.

H Witsenhausen and A Wyner. A conditional entropy bound for a pair of discrete random variables. *IEEE Transactions on Information Theory*, 21(5):493–501, 1975.

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

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 ψ. We can decompose the gradient of the entropy of $q_\psi(\mathbf{z})$ as:

$$\nabla_\psi H(z) = -\nabla_\psi \mathbb{E}_{q_\psi(\mathbf{z})}[\log q(\mathbf{z})] - \mathbb{E}_{q(\mathbf{z})}[\nabla_\psi \log q_\psi(\mathbf{z})], \tag{21}$$

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

$$\mathbb{E}_{q(\mathbf{z})}[\nabla_\psi \log q_\psi(\mathbf{z})] = \mathbb{E}_{q(\mathbf{z})}[\nabla_\psi q_\psi(\mathbf{z}) \times \frac{1}{q(\mathbf{z})}] = \int \nabla_\psi q_\psi(\mathbf{z}) d\mathbf{z} = \nabla_\psi \int q_\psi(\mathbf{z}) d\mathbf{z} = 0. \tag{22}$$

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

$$\nabla_\psi H(z) = -\nabla_\psi \mathbb{E}_{q_\psi(\mathbf{z})}[\log q(\mathbf{z})]. \tag{23}$$

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 ψ. The auxiliary variable $\boldsymbol{\epsilon}$ introduces randomness to the encoder. First, we decompose the gradients of Conditional Entropy as following:

$$\begin{aligned}
\nabla_\psi H\left(\mathbf{z}|\mathbf{x}\right) &= -\nabla_\psi \int p_\psi(\mathbf{z}, \mathbf{x}) \log p_\psi(\mathbf{z}|\mathbf{x}) dz dx \\
&= -\mathbb{E}_{p(\mathbf{x})}[\nabla_\psi \int p_\psi(\mathbf{z}|\mathbf{x}) \log p_\psi(\mathbf{z}|\mathbf{x}) dz] \\
&= -\mathbb{E}_{p(\mathbf{x})}[\nabla_\psi \mathbb{E}_{p_\psi(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{z}|\mathbf{x})] + \int p(\mathbf{z}|\mathbf{x}) \nabla_\psi \log p_\psi(\mathbf{z}|\mathbf{x}) dh] \\
&= -\mathbb{E}_{p(\mathbf{x})}[\nabla_\psi \mathbb{E}_{p_\psi(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{z}|\mathbf{x})] + \int \nabla_\psi p_\psi(\mathbf{z}|\mathbf{x}) dh] \\
&= -\mathbb{E}_{p(\mathbf{x})}[\nabla_\psi \mathbb{E}_{p_\psi(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{z}|\mathbf{x})] - \nabla_\psi \int p_\psi(\mathbf{h}, \mathbf{x}) dh dx] \\
&= -\mathbb{E}_{p(\mathbf{x})}[\nabla_\psi \mathbb{E}_{p_\psi(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{z}|\mathbf{x})]].
\end{aligned} \tag{24}$$

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 Equation (24),

$$H_\psi(\mathbf{z}|\mathbf{x}) = -\mathbb{E}_{p(\mathbf{x})}[\mathbb{E}_{p(\boldsymbol{\epsilon})}[\nabla_{(\mathbf{z}|\mathbf{x})} \log q(E_\psi(\mathbf{x}, \boldsymbol{\epsilon})|\mathbf{x}) \nabla_\psi E_\psi(\mathbf{x}, \boldsymbol{\epsilon})]]. \tag{25}$$