Information bottleneck through variational glasses

Slava Voloshynovskiy[*]
svolos@unige.ch

Mouad Kondah[*]
Mouad.Kondah@etu.unige.ch

Shideh Rezaeifar[*]
Shideh.Rezaeifar@unige.ch

Olga Taran[*]
Olga.Taran@unige.ch

Taras Holotyak[*]
Taras.Holotyak@unige.ch

Danilo Jimenez Rezende[†]
danilor@google.com

1 Abstract

Information bottleneck (IB) principle [1] has become an important element in information-theoretic analysis of deep models. Many state-of-the-art generative models of both Variational Autoencoder (VAE) [2, 3] and Generative Adversarial Networks (GAN) [4] families use various bounds on mutual information terms to introduce certain regularization constraints [5, 6, 7, 8, 9, 10]. Accordingly, the main difference between these models consists in add regularization constraints and targeted objectives.

In this work, we will consider the IB framework for three classes of models that include supervised, unsupervised and adversarial generative models. We will apply a variational decomposition leading a common structure and allowing easily establish connections between these models and analyze underlying assumptions.

Based on these results, we focus our analysis on unsupervised setup and reconsider the VAE family. In particular, we present a new interpretation of VAE family based on the IB framework using a direct decomposition of mutual information terms and show some interesting connections to existing methods such as VAE [2, 3], $\beta-$VAE [11], AAE [12], InfoVAE [5] and VAE/GAN [13]. Instead of adding regularization constraints to an evidence lower bound (ELBO) [2, 3], which itself is a lower bound, we show that many known methods can be considered as a product of variational decomposition of mutual information terms in the IB framework. The proposed decomposition might also contribute to the interpretability of generative models of both VAE and GAN families and create a new insights to a generative compression [14, 15, 16, 17]. It can also be of interest for the analysis of novelty detection based on one-class classifiers [18] with the IB based discriminators.

Notations: We will denote a joint generative distribution as $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{z}) p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$, whereas marginal $p_{\boldsymbol{\theta}}(\mathbf{z})$ is interpreted as a targeted distribution of latent space and marginal $p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{z})}\left[p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right] = \int_{\mathbf{z}} p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) p_{\boldsymbol{\theta}}(\mathbf{z}) d\mathbf{z}$ as a generated data distribution with a generative model described by $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$. A joint data distribution $q_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{z}) = p_{\mathcal{D}}(\mathbf{x}) q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, where $p_{\mathcal{D}}(\mathbf{x})$ denotes an empirical data distribution and $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ is an inference or encoding model and marginal $q_{\boldsymbol{\phi}}(\mathbf{z})$ denotes a "true" or "aggregated" distribution of latent space data.

2 Information bottleneck for different models

In this section, we consider the IB framework and summarize some known results for supervised and unsupervised models. Having introduced a common base, we will also extend these results to generative adversarial models. Along this analysis, we will introduce several interesting bounds that will be used to develop a proposed bounded IB auto-encoding.

[*] Department of Computer Science, University of Geneva, Carouge 1227, Switzerland
[†] DeepMind

4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

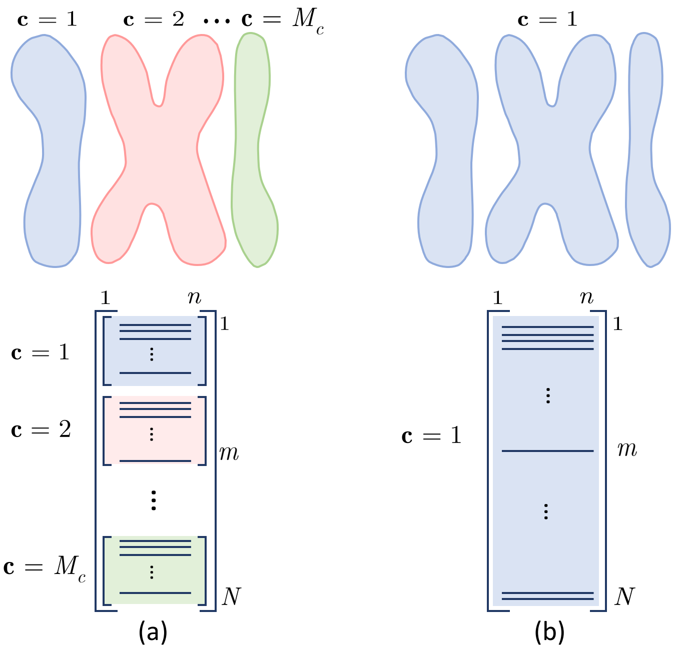

Figure 1: Labeling for supervised (a) and unsupervised (b) models. All data points are shown as sequences of dimension n at the bottom part of plot.

2.1 Information bottleneck for supervised models

We consider a true joint distribution $p(\mathbf{c}, \mathbf{x})$ from which the training set $\{\mathbf{x}_m, \mathbf{c}_m\}_{m=1}^N$ is sampled from, where each data sample is $\mathbf{x} \in \mathbb{R}^n$, n denotes the dimensionality of data and N stands for the number of training samples. We will use $\mathbf{c} \in \mathcal{M}$, with $\mathcal{M} = \{1, \cdots, M_c\}$, to denote a class label. We use a vector notation for $\mathbf{c}$ to highlight that each label can be encoded according to some representation. The number of classes is denoted as M_c. The labeling of N sequences into M_c classes is shown in Figure 1,a. It should be noted that many sequences might be assigned to the same class according to a set of chosen common features. We use different colors to reflect this labeling. At the same time, one can consider a "binning" organization principle shown in the bottom part of Figure 1,a, where N training sequences are allocated into M_c bins representing M_c classes.

The supervised IB framework is considered based on Figure 2,a. A sample $\mathbf{x}$ from a class $\mathbf{c}$ is generated by a mapping $p(\mathbf{x}, \mathbf{c}) = p(\mathbf{c})p(\mathbf{x}|\mathbf{c})$. The supervised IB can be formulated according to [1] as:

$$\min_{\boldsymbol{\phi}: I(\mathbf{Z};\mathbf{C}) \geq I_c} I_{\boldsymbol{\phi}}(\mathbf{X}; \mathbf{Z}). \tag{1}$$

The supervised IB framework assumes an existence of a parametrized probabilistic mapping $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ with a controllable set of parameters $\boldsymbol{\phi}$, where $\mathbf{z}$ is considered to be a latent or bottleneck representation with dimensionality and statistical properties different of those of $\mathbf{x}$. It is assumed that three concerned vectors form a Markov chain $\mathbf{C} \to \mathbf{X} \to \mathbf{Z}$ and the objective is to find such a mapping $\boldsymbol{\phi}$, when $\mathbf{z}$ is a minimal sufficient statistic for task $\mathbf{c}$. The term $I_{\boldsymbol{\phi}}(\mathbf{X}; \mathbf{Z})$ denotes the mutual information between $\mathbf{X}$ and $\mathbf{Z}$ considering the above parametric mapping and $I(\mathbf{Z}; \mathbf{C})$ corresponds to the mutual information between $\mathbf{Z}$ and $\mathbf{C}$.

The main idea behind the supervised IB (1) consists in a search of parameters $\boldsymbol{\phi}$ that ensures the preservation of the information I_c about the class $\mathbf{c}$ in the latent or bottleneck representation $\mathbf{z}$, while filtering out all irrelevant information from $\mathbf{x}$ that corresponds to the minimisation of $I_{\boldsymbol{\phi}}(\mathbf{X}; \mathbf{Z})$ over $\boldsymbol{\phi}$. It should be pointed out that the minimization of mutual information can be obtained in different ways that include but are not limited to dimensionality reduction, compression that might include both clustering and quantization, additional of noise or sparsification of $\mathbf{z}$. All these techniques are well known and often used in practical deep net mappers implementing $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$.

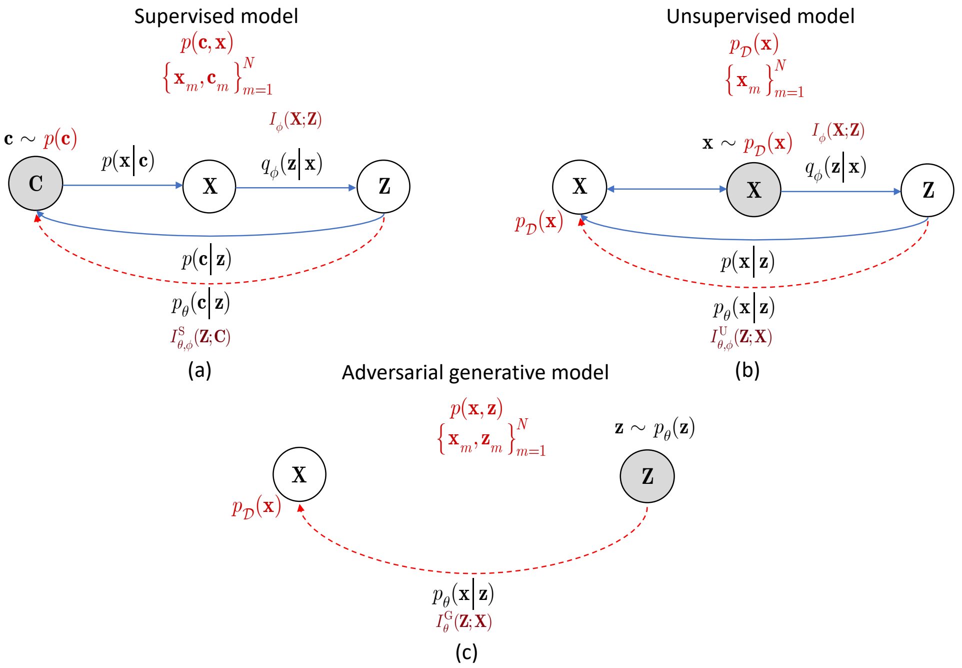
Supervised model
$p(\mathbf{c},\mathbf{x})$
$\{\mathbf{x}_m,\mathbf{c}_m\}_{m=1}^N$
$I_\phi(\mathbf{X};\mathbf{Z})$
$\mathbf{c} \sim p(\mathbf{c})$
$p(\mathbf{x}|\mathbf{c})$
$q_\phi(\mathbf{z}|\mathbf{x})$
C
X
Z
$p(\mathbf{c}|\mathbf{z})$
$p_\theta(\mathbf{c}|\mathbf{z})$
$I^{\mathrm{S}}_{\theta,\phi}(\mathbf{Z};\mathbf{C})$
(a)
Unsupervised model
$p_{\mathcal{D}}(\mathbf{x})$
$\{\mathbf{x}_m\}_{m=1}^N$
$\mathbf{x} \sim p_{\mathcal{D}}(\mathbf{x})$
$I_\phi(\mathbf{X};\mathbf{Z})$
$q_\phi(\mathbf{z}|\mathbf{x})$
X
X
Z
$p_{\mathcal{D}}(\mathbf{x})$
$p(\mathbf{x}|\mathbf{z})$
$p_\theta(\mathbf{x}|\mathbf{z})$
$I^{\mathrm{U}}_{\theta,\phi}(\mathbf{Z};\mathbf{X})$
(b)
Adversarial generative model
$p(\mathbf{x},\mathbf{z})$
$\{\mathbf{x}_m,\mathbf{z}_m\}_{m=1}^N$
$\mathbf{z} \sim p_\theta(\mathbf{z})$
X
Z
$p_{\mathcal{D}}(\mathbf{x})$
$p_\theta(\mathbf{x}|\mathbf{z})$
$I^{\mathrm{G}}_\theta(\mathbf{Z};\mathbf{X})$
(c)

Figure 2: Information bottleneck models: (a) supervised, (b) unsupervised and (c) adversarial generative ones.

Tishby *et. al.* [19] also proposed the Langrangian of IB optimization (1) defined as:

$$\mathcal{L}^{\mathrm{S}}(\boldsymbol{\phi}) = I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z}) - \beta I(\mathbf{Z};\mathbf{C}), \tag{2}$$

where S stands for the supervised setup and β is a regularization parameter corresponding to I_c that leads to an optimization formulation:

$$\hat{\boldsymbol{\phi}} = \operatorname*{argmin}_{\boldsymbol{\phi}} \mathcal{L}^{\mathrm{S}}(\boldsymbol{\phi}). \tag{3}$$

In the following part, we will consider both terms of mutual information in (2) and establish some useful bounds on them.

2.1.1 Decomposition of the first term

The first mutual information term $I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z})$ in (2) is defined as:

$$\begin{aligned} I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z}) &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z},\mathbf{x})}\left[\log \frac{q_{\boldsymbol{\phi}}(\mathbf{z},\mathbf{x})}{q_{\boldsymbol{\phi}}(\mathbf{z})p_{\mathcal{D}}(\mathbf{x})}\right] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z},\mathbf{x})}\left[\log \frac{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}{q_{\boldsymbol{\phi}}(\mathbf{z})}\right] \\ &= H_{\boldsymbol{\phi}}(\mathbf{Z}) - H_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X}), \end{aligned} \tag{4}$$

where $p_{\mathcal{D}}(\mathbf{x})$ denotes the data distribution and $H_{\boldsymbol{\phi}}(\mathbf{Z}) = -\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z})}\left[\log q_{\boldsymbol{\phi}}(\mathbf{z})\right]$ denotes the entropy of distribution $q_{\boldsymbol{\phi}}(\mathbf{z}) = \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right]$ and $H_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X}) = -\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z},\mathbf{x})}\left[\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right]$ denotes the conditional entropy defined by $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$. In (4), we used the decomposition of the joint distribution $q_{\boldsymbol{\phi}}(\mathbf{z},\mathbf{x}) = q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})p_{\mathcal{D}}(\mathbf{x})$. At the moment, we will not address technical details of computing $q_{\boldsymbol{\phi}}(\mathbf{z})$ and focus on them along the unsupervised setup analysis.

2.1.2 Decomposition of the second term

The second mutual information term $I(\mathbf{Z};\mathbf{C})$ in (2) can be defined via $p(\mathbf{c}|\mathbf{z})$ as:

$$I(\mathbf{Z};\mathbf{C}) = \mathbb{E}_{p(\mathbf{c},\mathbf{z})}\left[\log \frac{p(\mathbf{c},\mathbf{z})}{p(\mathbf{c})p(\mathbf{z})}\right] = \mathbb{E}_{p(\mathbf{c},\mathbf{z})}\left[\log \frac{p(\mathbf{c}|\mathbf{z})}{p(\mathbf{c})}\right]. \tag{5}$$

We show in Appendix A, that this mutual information can be lower bounded by $I(\mathbf{Z};\mathbf{C}) \geq I^{\mathrm{S}}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{C})$, where:

$$\begin{aligned} I^{\mathrm{S}}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{C}) &\triangleq -\mathbb{E}_{p(\mathbf{c})}\left[\log p(\mathbf{c})\right] + \mathbb{E}_{p(\mathbf{c},\mathbf{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{z})\right]\right] \\ &= H(\mathbf{C}) - H_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{C}|\mathbf{Z}), \end{aligned} \tag{6}$$

with $H(\mathbf{C}) = -\mathbb{E}_{p(\mathbf{c})}\left[\log p(\mathbf{c})\right]$ and $H_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{C}|\mathbf{Z}) = -\mathbb{E}_{p(\mathbf{c},\mathbf{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{z})\right]\right]$.

Therefore, the corresponding IB Lagrangian is redefined as:

$$\mathcal{L}^{\mathrm{S}}(\boldsymbol{\phi},\boldsymbol{\theta}) = I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z}) - \beta I^{\mathrm{S}}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{C}), \tag{7}$$

that leads to the optimization problem:

$$(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\phi}}) = \underset{\boldsymbol{\theta},\boldsymbol{\phi}}{\operatorname{argmin}}\, \mathcal{L}^{\mathrm{S}}(\boldsymbol{\phi},\boldsymbol{\theta}). \tag{8}$$

Remark: since $H(\mathbf{C})$ in (6) is constant and does not depend on the parameters $\boldsymbol{\theta},\boldsymbol{\phi}$, the supervised IB Lagrangian (7) can be rewritten in yet another commonly know form of supervised IB:

$$\mathcal{L}^{\mathrm{S}}(\boldsymbol{\phi},\boldsymbol{\theta}) \propto I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z}) + \beta H_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{C}|\mathbf{Z}). \tag{9}$$

In turns, it can be considered as finding a trade-off between the reduction of mutual information between $\mathbf{X}$ and $\mathbf{Z}$ according to the first term and the prediction accuracy of class $\mathbf{c}$ based on $\mathbf{z}$ according to the second term.

2.2 Information bottleneck for unsupervised models

In the case of unsupervised setup, the data samples are not labelled by the classes $\mathbf{c}$. We will consider a true data distribution $p_{\mathcal{D}}(\mathbf{x})$ from which the training set $\{\mathbf{x}_m\}_{m=1}^{N}$ is sampled from. The data samples can be considered as belonging to a common class with the same label $\mathbf{c}=1$ as shown in Figure 1,b. Each sequence $\mathbf{x}$ is indexed by its proper index m. It means that the mapping between m and $\mathbf{x}$ is unique $m \leftrightarrow \mathbf{x}$ in contrast to the supervised setup, where knowing $\mathbf{c}$ does not automatically imply that one knows a sample $\mathbf{x}$ but rather a set or bin to which it belongs to.

Alternatively, one can interpret the unsupervised setup as the supervised one with $M_c = N$ classes, i.e., when each class is represented by just one sequence as shown in Figure 1b. Therefore, by the direct analogy with the supervised setup, one can replace each class $\mathbf{c}$ by its proper representative sequence $\mathbf{x}$ as depicted in Figure 2,b. Therefore, the generative process can be considered to start directly from $\mathbf{x}$ as shown by a gray circle.

Thus, the unsupervised IB can be considered as a "compression" of $\mathbf{x}$ to $\mathbf{z}$ via the parametrized mapping $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ leading to a bottleneck representation $\mathbf{z}$ yet preserving a certain level of information I_x in $\mathbf{z}$ about $\mathbf{x}$. Accordingly, the unsupervised IB problem can be formulated as:

$$\min_{\boldsymbol{\phi}: I(\mathbf{Z};\mathbf{X}) \geq I_x} I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z}), \tag{10}$$

and in the Lagrangian formulation as a minimization of:

$$\mathcal{L}^{\mathrm{U}}(\boldsymbol{\phi}) = I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z}) - \beta I(\mathbf{Z};\mathbf{X}), \tag{11}$$

where we use the same β as for the supervised setup for the sake of simplicity and U denotes the unsupervised case.

In the following sections, we will consider decompositions of both mutual information terms.

2.2.1 Decomposition of the first term

The first term $I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z})$ in (11) can be defined similarly to the supervised case (4) using entropies. The conditional entropy $H_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X})$ is computable, since $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ is defined. However, the entropy $H_{\boldsymbol{\phi}}(\mathbf{Z}) = -\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z})}\left[\log q_{\boldsymbol{\phi}}(\mathbf{z})\right]$ requires computation of marginal distribution $q_{\boldsymbol{\phi}}(\mathbf{z}) = \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right]$ that might be a computationally expensive task in practice. Therefore, we will proceed with a variational approximation of $q_{\boldsymbol{\phi}}(\mathbf{z})$ by a distribution $p_{\boldsymbol{\theta}}(\mathbf{z})$[3]:

$$\begin{aligned} I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z}) &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z},\mathbf{x})}\left[\log\frac{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z})}{q_{\boldsymbol{\phi}}(\mathbf{z})p_{\mathcal{D}}(\mathbf{x})}\right] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z},\mathbf{x})}\left[\log\frac{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}{q_{\boldsymbol{\phi}}(\mathbf{z})}\frac{p_{\boldsymbol{\theta}}(\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{z})}\right]\\ &= \underbrace{\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X}=\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{z})\right)\right]}_{\mathrm{A}} - \underbrace{D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z})\|p_{\boldsymbol{\theta}}(\mathbf{z})\right)}_{\mathrm{B}}, \end{aligned} \tag{12}$$

where the term (A) denotes the KL-divergence $\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X}=\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{z})\right)\right] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z},\mathbf{x})}\left[\log\frac{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{z})}\right] = \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log\frac{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{z})}\right]\right]$ and the term (B) denotes the KL-divergence $D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z})\|p_{\boldsymbol{\theta}}(\mathbf{z})\right) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z},\mathbf{x})}\left[\log\frac{q_{\boldsymbol{\phi}}(\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{z})}\right] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z})}\left[\log\frac{q_{\boldsymbol{\phi}}(\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{z})}\right]$.

2.2.2 Decomposition of the second term

The second mutual information term $I(\mathbf{Z};\mathbf{X})$ in (11) is defined as:

$$I(\mathbf{Z};\mathbf{X}) = \mathbb{E}_{p(\mathbf{z},\mathbf{x})}\left[\log\frac{p(\mathbf{x}|\mathbf{z})}{p_{\mathcal{D}}(\mathbf{x})}\right]. \tag{13}$$

To find a variational approximation to the unknown $p(\mathbf{x}|\mathbf{z})$, one can proceed in the same way as with the supervised model. However, one can also directly obtain a variational lower bound on $I(\mathbf{Z};\mathbf{X})$ by assuming $\mathbf{c} \equiv \mathbf{x}$ in (6). This leads to $I(\mathbf{Z};\mathbf{X}) \geq I^{\mathrm{U}}_{\boldsymbol{\theta},\phi}(\mathbf{Z};\mathbf{X})$, where:

$$\begin{aligned} I^{\mathrm{U}}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{X}) &\triangleq -\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\log p_{\mathcal{D}}(\mathbf{x})\right] + \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right]\right]\\ &= H_{\mathcal{D}}(\mathbf{X}) - H_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{X}|\mathbf{Z}), \end{aligned} \tag{14}$$

with $H_{\mathcal{D}}(\mathbf{X}) = -\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\log p_{\mathcal{D}}(\mathbf{x})\right]$ and $H_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{X}|\mathbf{Z}) = -\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right]\right]$. Therefore, the corresponding IB Lagrangian is defined as:

$$\mathcal{L}^{\mathrm{U}}(\boldsymbol{\phi},\boldsymbol{\theta}) = I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z}) - \beta I^{\mathrm{U}}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{X}), \tag{15}$$

thus leading to the minimization problem:

$$(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\phi}}) = \underset{\boldsymbol{\theta},\boldsymbol{\phi}}{\operatorname{argmin}}\,\mathcal{L}^{\mathrm{U}}(\boldsymbol{\phi},\boldsymbol{\theta}). \tag{16}$$

In should be pointed out that similarly to the supervised case (9), the term $H_{\mathcal{D}}(\mathbf{X})$ in (14) does not depend on the encoder and decoder parameters $\boldsymbol{\phi},\boldsymbol{\theta}$ and can be skipped from the further consideration, if one is only concerned about the reconstruction task.

Nevertheless, the same model can also be considered for a generative task, which will also be considered below, when a trained encoder-decoder pair or just a sole decoder can be used for the generation of new samples from the latent space distribution. For these reasons, it is of interest to ensure that newly generated samples closely follow the statistics of original data. That is why one can also consider a decomposition of (14) as:

$$\begin{aligned} I^{\mathrm{U}}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{X}) &= \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{z})}\left[\log\frac{p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})}{p_{\mathcal{D}}(\mathbf{x})}\right]\right]\\ &= \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{z})}\left[\log\frac{p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})}{p_{\mathcal{D}}(\mathbf{x})}\frac{p_{\boldsymbol{\theta}}(\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{x})}\right]\right]\\ &= -\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x})\right] - \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\log\frac{p_{\mathcal{D}}(\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{x})}\right] + \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{z})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right]\right]\\ &= H(p_{\mathcal{D}}(\mathbf{x});p_{\boldsymbol{\theta}}(\mathbf{x})) - D_{\mathrm{KL}}\left(p_{\mathcal{D}}(\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{x})\right) + \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{z})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right]\right]. \end{aligned} \tag{17}$$

[3] Technically, the same factorization can be applied to the supervised counterpart (4). However, since in practice it is rarely of interest to generate labels $\mathbf{c}$ from $\mathbf{z}$, we only consider it in the scope of unsupervised generative and compression models.

where $H(p_{\mathcal{D}}(\mathbf{x}); p_{\boldsymbol{\theta}}(\mathbf{x})) = -\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x})\right]$ denotes a cross-entropy. Since $H(p_{\mathcal{D}}(\mathbf{x}); p_{\boldsymbol{\theta}}(\mathbf{x})) \geq 0$, one can lower bound (17) as $I^{U}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{X}) \geq I^{\mathrm{U}_L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{X})$, where[4]:

$$I^{\mathrm{U}_L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{X}) \triangleq \underbrace{\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right]\right]}_{\mathrm{C}} - \underbrace{D_{\mathrm{KL}}\left(p_{\mathcal{D}}(\mathbf{x}) \| p_{\boldsymbol{\theta}}(\mathbf{x})\right)}_{\mathrm{D}}. \tag{18}$$

Remark: The term (D) in (18) can be implemented based on the density ratio estimation [20] that will be addressed below. The term (C) can be defined explicitly using Gaussian or Laplacian priors. In the Laplacian case, one can define $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) \propto \exp(-\lambda\|\mathbf{x} - g_{\boldsymbol{\theta}}(\mathbf{z})\|_1)$ with a scale parameter λ, which leads to ℓ_1-norm, and $g_{\boldsymbol{\theta}}(\mathbf{z})$ denotes the decoder. It also corresponds to the model $\mathbf{x} = g_{\boldsymbol{\theta}}(\mathbf{z}) + \mathbf{e}_x$, where $\mathbf{e}_x$ is a reconstruction error vector following the Laplacian pdf. Therefore, (18) reduces to:

$$I^{\mathrm{U}_L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{X}) = \underbrace{-\lambda\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\|\mathbf{x} - g_{\boldsymbol{\theta}}(\mathbf{z})\|_1\right)\right]}_{\mathrm{C}} - \underbrace{D_{\mathrm{KL}}\left(p_{\mathcal{D}}(\mathbf{x}) \| p_{\boldsymbol{\theta}}(\mathbf{x})\right)}_{\mathrm{D}}. \tag{19}$$

2.2.3 Comparison of supervised and unsupervised IB

Having considered the supervised and unsupervised IB formulations, it should be remarked several differences.

The main origin of these differences is in the entropy of classes $H(\mathbf{C})$ and entropy of data $H_{\mathcal{D}}(\mathbf{X})$, i.e., $H_{\mathcal{D}}(\mathbf{X}) \gg H(\mathbf{C})$. The supervised IB describing the classification task only needs to ensure that the latent space data $\mathbf{Z}$, representing the sufficient statistics for $\mathbf{C}$, should preserve just $\log_2(M_c)$ bits to uniquely encode and recognize each class. In the unsupervised setup, the IB suggests to compress $\mathbf{X}$ to the encoding representation $\mathbf{Z}$ such that each sequence $\mathbf{X}$ is uniquely decodable or identifiable from $\mathbf{Z}$. It means that the entropy of latent space should correspond to the entropy of observation space, i.e., it should encode at least $\log_2(N)$ bits to uniquely distinguish all N sequences, unless some tolerance is allowed in terms of reconstruction error[5].

Naturally, this difference also leads to different encoding strategies. In the supervised setup, all common information within the same labeled class is "compressed" or disregarded and only the "differences" between the classes are encoded. With the increase of the number of classes, the differences might be minor that could be a potential source of vulnerability to adversarial attacks. An "informed" attacker knowing how these features are selected, that can be learned having an access to the same training data, might change only several of them to achieve a flipping between the classes. In contrast, the entropy of latent data for the unsupervised setup should be considerably higher than those for the supervised setup.

Finally, the nature of encoding is also different. In the unsupervised encoding, the classes are encoded to satisfy the reconstruction on average, i.e., the sequences close in the observation space might be close or even collude in the latent space, and the features of data contributing the most to the chosen metric of fidelity are preserved while less significant features are compressed or disregarded. As pointed above, all features that are irrelevant to a given classification task will be disregarded in the supervised setup. Using different re-labeling, new class-relevant features will be extracted while class irrelevant information will be filtered out. In the unsupervised case, there is no labeling and the encoding solely depends on statistics of data.

2.3 A link to generative adversarial models

The generative adversarial models can be considered as in Figure 2c, i.e., the latent representation $\mathbf{z}$ of these models is not derived from the input of the network. Instead, it is assumed that the randomly assigned pairs $\{\mathbf{x}_m, \mathbf{z}_m\}_{m=1}^{N}$ are generated from $p_{\mathcal{D}}(\mathbf{x})$ and $p_{\boldsymbol{\theta}}(\mathbf{z})$.

Hence, the samples $\mathbf{z}$ are not produced by mapping $p_{\mathcal{D}}(\mathbf{x})$ via $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ but directly from $\mathbf{z} \sim p_{\boldsymbol{\theta}}(\mathbf{z})$ and thus the term $I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z}) = 0$. Therefore, the unsupervised setup (15) reduces to the minimization

[4] The cross-entropy computation requires knowledge of model $p_{\boldsymbol{\theta}}(\mathbf{x})$, whereas the KL-divergence is based on the ratio of two distributions and can be computed without an explicit knowledge of distributions but only from the training samples. For this reason, we proceed further with the KL-term.

[5] The total number of samples in the training set is upper limited by $2^{nH(X)}$ under the i.i.d. assumption, whereas the training set is assumed to contain only N sequences.

of:

$$\hat{\boldsymbol{\theta}} = \min_{\boldsymbol{\theta}} \mathcal{L}^{\mathrm{G}}(\boldsymbol{\theta}), \tag{20}$$

where $\mathcal{L}^{\mathrm{G}}(\boldsymbol{\theta}) = -\beta I^{\mathrm{G}}_{\boldsymbol{\theta}}(\mathbf{Z};\mathbf{X})$ and:

$$I^{\mathrm{G}}_{\boldsymbol{\theta}}(\mathbf{Z};\mathbf{X}) \triangleq \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{z})}\left[\log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})}{p_{\mathcal{D}}(\mathbf{x})}\right]\right], \tag{21}$$

corresponds $I^{\mathrm{U}}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{X})$ in (15) due to the fact that the sole link between $\mathbf{Z}$ and $\mathbf{X}$ is via $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ and the latent vectors are generated from $p_{\boldsymbol{\theta}}(\mathbf{z})$ and there is no dependence on $\boldsymbol{\phi}$.

Equivalently, the minimization problem (20) can be reformulated as:

$$\hat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta}} I^{\mathrm{G}}_{\boldsymbol{\theta}}(\mathbf{Z};\mathbf{X}). \tag{22}$$

Accordingly, using the factorization with respect to the marginal distribution of generated data $p_{\boldsymbol{\theta}}(\mathbf{x})$ similarly to the unsupervised case (17), one can define $I^{\mathrm{G}}_{\boldsymbol{\theta}}(\mathbf{Z};\mathbf{X})$ as:

$$\begin{aligned}
I^{\mathrm{G}}_{\boldsymbol{\theta}}(\mathbf{Z};\mathbf{X}) &\triangleq \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{z})}\left[\log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})}{p_{\mathcal{D}}(\mathbf{x})}\right]\right] \\
&= \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{z})}\left[\log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})}{p_{\mathcal{D}}(\mathbf{x})}\frac{p_{\boldsymbol{\theta}}(\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{x})}\right]\right] \\
&= -\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x})\right] - \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\log \frac{p_{\mathcal{D}}(\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{x})}\right] + \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{z})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right]\right] \\
&= H(p_{\mathcal{D}}(\mathbf{x}); p_{\boldsymbol{\theta}}(\mathbf{x})) - D_{\mathrm{KL}}\left(p_{\mathcal{D}}(\mathbf{x}) \| p_{\boldsymbol{\theta}}(\mathbf{x})\right) + \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{z})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right]\right].
\end{aligned} \tag{23}$$

Since $H(p_{\mathcal{D}}(\mathbf{x}); p_{\boldsymbol{\theta}}(\mathbf{x})) \geq 0$, one can lower bound (23) as $I^{\mathrm{G}}_{\boldsymbol{\theta}}(\mathbf{Z};\mathbf{X}) \geq I^{\mathrm{G}_L}_{\boldsymbol{\theta}}(\mathbf{Z};\mathbf{X})$ where:

$$I^{\mathrm{G}_L}_{\boldsymbol{\theta}}(\mathbf{Z};\mathbf{X}) \triangleq -D_{\mathrm{KL}}\left(p_{\mathcal{D}}(\mathbf{x}) \| p_{\boldsymbol{\theta}}(\mathbf{x})\right) + \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{z})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right]\right]. \tag{24}$$

Similarly to (19), one can further develop (24) using $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) \propto \exp(-\lambda\|\mathbf{x} - g_{\boldsymbol{\theta}}(\mathbf{z})\|_1)$ with a scale parameter λ that results in:

$$\hat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta}} I^{\mathrm{G}_L}_{\boldsymbol{\theta}}(\mathbf{Z};\mathbf{X}) = \min_{\boldsymbol{\theta}} D_{\mathrm{KL}}\left(p_{\mathcal{D}}(\mathbf{x}) \| p_{\boldsymbol{\theta}}(\mathbf{x})\right) + \lambda \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{z})}\left[\|\mathbf{x} - g_{\boldsymbol{\theta}}(\mathbf{z})\|_1)\right]\right]. \tag{25}$$

Remark : Vanilla GANs use only an approximation to the first term for the generator optimization. However, GANs might face a mode collapse and the likelihood term can at least theoretically regularize it.

3 Bounded information bottleneck AE formulation

Having considered the unsupervised and adversarial generative models, we can proceed with the formulation of a new auto-encoding framework. More particularly, we will use the results (12) and (18) to propose a new type of unsupervised auto-encoder that combines the elements of VAE and GAN families and is built on the IB principle. We will refer to this auto-encoder as a *bounded information bottleneck AE* (BIB-AE) and link it to the VAE family of auto-encoders, generative compression and one-class classification. It should also be pointed out that the BIB-AE framework is rather considered as a conceptual generalization then as practical implementation. However, we will comment how to implement the BIB-AE components in practice using known techniques of KL-divergence approximation.

The BIB-AE Lagrangian is based on (15) and is defined as:

$$\mathcal{L}_{\mathrm{BIB-AE}}(\boldsymbol{\theta},\boldsymbol{\phi}) = I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z}) - \beta I^{\mathrm{U}_L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{X}), \tag{26}$$

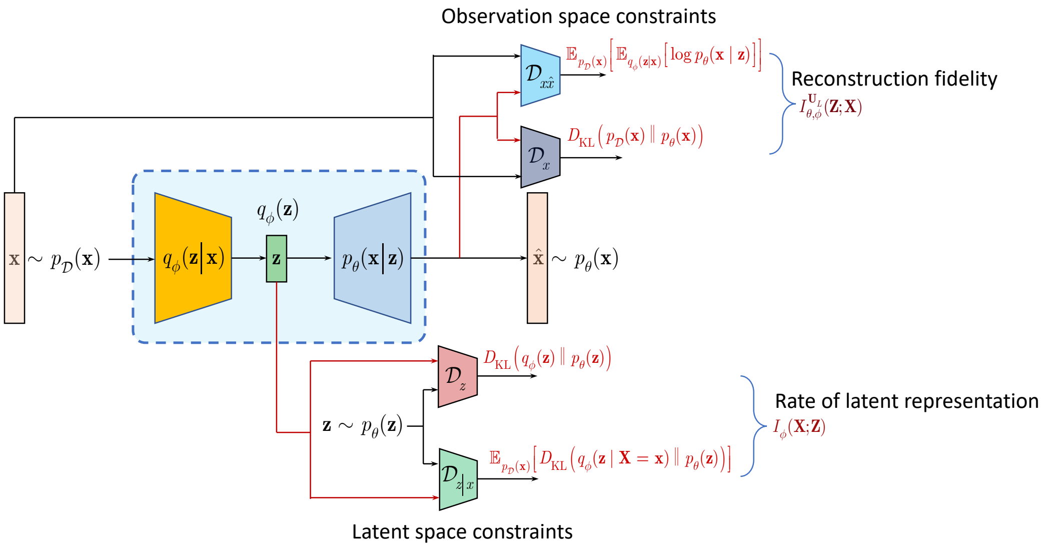
Observation space constraints
$\mathcal{D}_{x\hat{x}}$
$\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right]\right]$
Reconstruction fidelity
$I^{\mathrm{U}_L}_{\theta,\phi}(\mathbf{Z};\mathbf{X})$
$\mathcal{D}_x$
$D_{\mathrm{KL}}\left(p_{\mathcal{D}}(\mathbf{x}) \parallel p_\theta(\mathbf{x})\right)$
$\mathbf{x} \sim p_{\mathcal{D}}(\mathbf{x})$
$q_\phi(\mathbf{z}|\mathbf{x})$
$q_\phi(\mathbf{z})$
$\mathbf{z}$
$p_\theta(\mathbf{x}|\mathbf{z})$
$\hat{\mathbf{x}} \sim p_\theta(\mathbf{x})$
$\mathcal{D}_z$
$D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}) \parallel p_\theta(\mathbf{z})\right)$
$\mathbf{z} \sim p_\theta(\mathbf{z})$
Rate of latent representation
$I_\phi(\mathbf{X};\mathbf{Z})$
$\mathcal{D}_{z|x}$
$\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[D_{\mathrm{KL}}\left(q_\phi(\mathbf{z} \mid \mathbf{X} = \mathbf{x}) \parallel p_\theta(\mathbf{z})\right)\right]$
Latent space constraints

Figure 3: Generalized diagram of BIB-AE.

where $I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z})$ and $I^{\mathrm{U}_L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{X})$ correspond to (12) and (18) that we summarize below for the convenience of analysis:

$$I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z}) = \underbrace{\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X}=\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{z})\right)\right]}_{\mathrm{A}} - \underbrace{D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z})\|p_{\boldsymbol{\theta}}(\mathbf{z})\right)}_{\mathrm{B}}, \tag{27}$$

$$I^{\mathrm{U}_L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{X}) = \underbrace{\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right]\right]}_{\mathrm{C}} - \underbrace{D_{\mathrm{KL}}\left(p_{\mathcal{D}}(\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{x})\right)}_{\mathrm{D}}. \tag{28}$$

The BIB-AE parameters are found according to the following minimization problem:

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}) = \underset{\boldsymbol{\theta},\boldsymbol{\phi}}{\operatorname{argmin}}\, \mathcal{L}_{\mathrm{BIB-AE}}(\boldsymbol{\theta}, \boldsymbol{\phi}). \tag{29}$$

The diagram explaining the BIB-AE setup is shown in Figure 3. The reconstruction fidelity is ensured jointly by the terms (C) and (D), while the minimization of mutual information between $\mathbf{X}$ and $\mathbf{Z}$ is guided by the targeted distribution of the latent space $p_{\boldsymbol{\theta}}(\mathbf{z})$ according to the terms (A) and (B). The "stochasticity" of the encoder will determine to which extend the mappings of data points from the observation space will "overlap" in the latent space yet satisfying the correspondence between the marginal posterior and the prior.

More particularly, as shown in Figure 4, the data distribution $p_{\mathcal{D}}(\mathbf{x})$ is mapped to the latent space marginal distribution $q_{\boldsymbol{\phi}}(\mathbf{z})$ via the stochastic mapping $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$. According to the variational approach, the targeted distribution of latent space is $p_{\boldsymbol{\theta}}(\mathbf{z})$ and the encoder tries to optimize the parameters of encoder $\boldsymbol{\phi}$ according to (29) to meet both the constraints on the latent space and the reconstruction fidelity by satisfying the targeted $\beta I^{\mathrm{U}_L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{X})$. One can imagine several forms of stochastic encoding: (i) $\mathbf{z} = f_{\boldsymbol{\phi}}(\mathbf{x}) + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ follows the distribution defying the properties of conditional distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, (ii) $\mathbf{z} = f_{\boldsymbol{\phi}}(\mathbf{x} + \boldsymbol{\epsilon})$ or (iii) $\mathbf{z} = f_{\boldsymbol{\phi}}([\mathbf{x}, \boldsymbol{\epsilon}])$. However, in practice depending on a chosen way of computing KL-divergence, one might be interested in a tractable density. In this case, the encoding of the first type is used as for example in the VAE family. Disregarding a particular form of randomness injecting mechanism, the green circles in the latent space of Figure 4 denote the resulting stochastic mappings of each point from the observable space.

4 Connections to the prior art AEs

4.1 Generative models of VAE family

VAE [2, 3] Lagrangian is defined as:

$$\mathcal{L}_{\mathrm{VAE}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X}=\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{z}))\right] - \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right]\right], \tag{30}$$

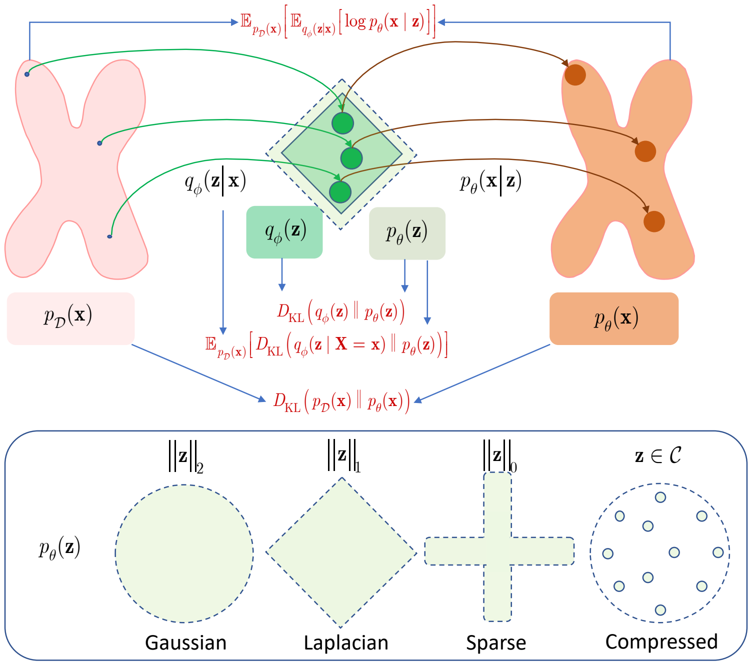
$\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right]\right]$
$q_\phi(\mathbf{z}|\mathbf{x})$
$p_\theta(\mathbf{x}|\mathbf{z})$
$q_\phi(\mathbf{z})$
$p_\theta(\mathbf{z})$
$p_{\mathcal{D}}(\mathbf{x})$
$p_\theta(\mathbf{x})$
$D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}) \parallel p_\theta(\mathbf{z})\right)$
$\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[D_{\mathrm{KL}}\left(q_\phi(\mathbf{z} \mid \mathbf{X}=\mathbf{x}) \parallel p_\theta(\mathbf{z})\right)\right]$
$D_{\mathrm{KL}}\left(p_{\mathcal{D}}(\mathbf{x}) \parallel p_\theta(\mathbf{x})\right)$
$\|\mathbf{z}\|_2$
$\|\mathbf{z}\|_1$
$\|\mathbf{z}\|_0$
$\mathbf{z} \in \mathcal{C}$
$p_\theta(\mathbf{z})$
Gaussian
Laplacian
Sparse
Compressed

Figure 4: BIB-AE as a stochastic mapping. Possible targeted priors of latent space are shown in the bottom of figure.

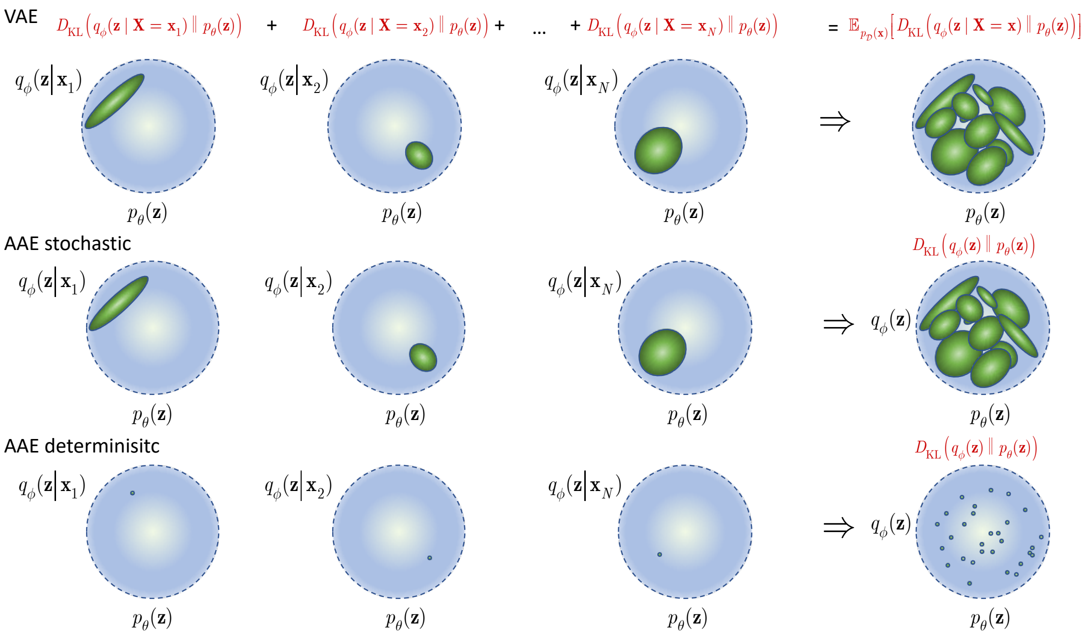
VAE
$D_{\mathrm{KL}}\left(q_\phi(\mathbf{z} \mid \mathbf{X}=\mathbf{x}_1) \parallel p_\theta(\mathbf{z})\right)$ +
$D_{\mathrm{KL}}\left(q_\phi(\mathbf{z} \mid \mathbf{X}=\mathbf{x}_2) \parallel p_\theta(\mathbf{z})\right)$ +
...
+ $D_{\mathrm{KL}}\left(q_\phi(\mathbf{z} \mid \mathbf{X}=\mathbf{x}_N) \parallel p_\theta(\mathbf{z})\right)$
= $\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[D_{\mathrm{KL}}\left(q_\phi(\mathbf{z} \mid \mathbf{X}=\mathbf{x}) \parallel p_\theta(\mathbf{z})\right)\right]$
$q_\phi(\mathbf{z}|\mathbf{x}_1)$
$q_\phi(\mathbf{z}|\mathbf{x}_2)$
$q_\phi(\mathbf{z}|\mathbf{x}_N)$
$\Rightarrow$
$p_\theta(\mathbf{z})$
AAE stochastic
$D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}) \parallel p_\theta(\mathbf{z})\right)$
$q_\phi(\mathbf{z}|\mathbf{x}_1)$
$q_\phi(\mathbf{z}|\mathbf{x}_2)$
$q_\phi(\mathbf{z}|\mathbf{x}_N)$
$\Rightarrow$ $q_\phi(\mathbf{z})$
$p_\theta(\mathbf{z})$
AAE determinisitc
$D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}) \parallel p_\theta(\mathbf{z})\right)$
$q_\phi(\mathbf{z}|\mathbf{x}_1)$
$q_\phi(\mathbf{z}|\mathbf{x}_2)$
$q_\phi(\mathbf{z}|\mathbf{x}_N)$
$\Rightarrow$ $q_\phi(\mathbf{z})$
$p_\theta(\mathbf{z})$

Figure 5: Schematic visualization of latent space for VAE and AAE.

and contains only 2 terms (A) and (C) in (26) with $\beta = 1$. It can be shown that the VAE is based on an upper bound on $I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z}) \leq I^{U}_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z}) = \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X}=\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{z})\right)\right]$, since $D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z})\|p_{\boldsymbol{\theta}}(\mathbf{z})\right) \geq 0$. Similarly, since $D_{\mathrm{KL}}\left(p_{\mathcal{D}}(\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{x})\right) \geq 0$, and denoting $I^{\mathrm{VAE}}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{X}) = \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right]\right]$, one obtains $I^{\mathrm{VAE}}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{X}) \geq I^{\mathrm{U}_L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{X})$.

The VAE encoder can be considered as a stochastic mapping with a particular form of parametrization [2] $\mathbf{z} = \boldsymbol{\mu}(\mathbf{x}) + \boldsymbol{\sigma}(\mathbf{x}) \odot \boldsymbol{\epsilon}$, where $\boldsymbol{\mu}(\mathbf{x})$ and $\boldsymbol{\sigma}(\mathbf{x})$ are outputs of the network $f_{\boldsymbol{\phi}}(\mathbf{x})$ and $\boldsymbol{\epsilon}$ is assumed to be a zero mean unit variance vector, i.e., $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$, and $\odot$ denotes element wise product. As a result, the conditional distribution of $\mathbf{Z}$ given an input variable $\mathbf{X}$ follows a Gaussian distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}(\mathbf{x}), \mathrm{diag}(\boldsymbol{\sigma}(\mathbf{x})))$. The VAE also assumes a prior on the latent space to be $p_{\boldsymbol{\theta}}(\mathbf{z}) = \mathcal{N}(\mathbf{0},\mathbf{I})$. Under these conditions the KL-term (A) can be computed analytically.

It should be pointed out that the VAE encoder maps a point from the observation space into a probabilistic output of Gaussian cloud with mean $\boldsymbol{\mu}(\mathbf{x})$ and "ellipsoid" orientation determined by the diagonal covariance matrix diag$(\boldsymbol{\sigma}(\mathbf{x}))$. This is schematically shown in a form of green ellipsoids for different samples $\mathbf{x}_i$, $i = 1, \cdots, N$, in the latent space according to Figure 5. Moreover, since the targeted marginal prior is $p_{\boldsymbol{\theta}}(\mathbf{z}) = \mathcal{N}(\mathbf{0},\mathbf{I})$, and the KL term for all mappings of $\mathbf{x}$'s via $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ should match this $p_{\boldsymbol{\theta}}(\mathbf{z})$ in $\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{z})\right)\right]$, the encoder optimized in such a way will target to make the mean of all mappings close to zero and whiten the ellipsoids.

Without a special guidance, these mappings will converge to the zero mean unit variance Gaussian marginal shape under asymptomatically many input mappings. Obviously, there is a little control on this process but the final goal of stochastic minimization of the upper bound on the mutual information $I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z})$ considered as a "compression" is achieved according to the IB framework.

β-VAE [11] is linked to (26) in the same way as the VAE but with a varying relaxation parameter β:

$$\mathcal{L}_{\beta-\mathrm{VAE}}(\boldsymbol{\theta},\boldsymbol{\phi}) = \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X}=\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{z}))\right] - \beta\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right]\right]. \quad (31)$$

The main advantage of β-VAE over VAE is a possibility to relax the described above stochastic "compression" via mapping everything to a big Gaussian "heap" by applying the relaxation parameter β that might give more preference to the reconstruction cost. By increasing β, one might achieve a sort of "disentangliation", yet weakly controllable by one global parameter, by allowing Gaussian clouds in the latent space to be far away from each other by less satisfying the KL-term constraint to fit the marginally Gaussian distribution. The semantically similar inputs might be mapped closer thus creating a sort of clusters that might be interpreted as a disentangled representation. Surely, it is only an interpretation of such a relaxed stochastic mapping and the process of "semantic clustering" highly depends on statistics of data. It seems to be quite difficult to achieve a semantically meaningful encoding and interepretability of the latent space without either at least some weak supervision or specially constructed latent space.

AAE [12] can be defined according to the equivalent Lagrangian cost:

$$\mathcal{L}_{\mathrm{AAE}}(\boldsymbol{\theta},\boldsymbol{\phi}) = D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z})\|p_{\boldsymbol{\theta}}(\mathbf{z})) - \beta\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right]\right], \quad (32)$$

where we do not explicitly consider the technical details of KL-divergence approximation and computation whereas one can use adversarial discriminator for this purpose or the maximum mean discrepancy (MMD) [21] based discriminator.

It should be pointed out that (32) contains the term (C) which origin can be explained in the same way as for the VAE. Despite of the fact that the term (B) indeed appears in (32) with the opposite sign, it cannot be interpreted either as an upper bound on $I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z})$ similarly to the VAE or as a lower bound. The goal of AAE is to minimize the reconstruction loss or to maximize the log-likelihood by ensuring that the latent space marginal distribution $q_{\boldsymbol{\phi}}(\mathbf{z})$ matches the prior $p_{\boldsymbol{\theta}}(\mathbf{z})$. The latter corresponds to the minimization of $D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z})\|p_{\boldsymbol{\theta}}(\mathbf{z})\right)$.

It is interesting to point out that the original AAE paper considers as a potential encoding all options that include: a *deterministic encoding*, i.e., $\mathbf{z} = f_{\boldsymbol{\phi}}(\mathbf{x})$, as well as the considered in section 3 *stochastic encodings*. A nice flexibility of AAE comes from a possibility to match the observed marginal distribution $q_{\boldsymbol{\phi}}(\mathbf{z})$ to a desired targeted distribution $p_{\boldsymbol{\theta}}(\mathbf{z})$ without the need to have explicitly defined distributions in contrast to the VAE.

An actual implementation of AAE is based on the deterministic encoding. We can imagine this sort of mapping by considering Figure 5. A point of the observation space is mapped just to one point in the latent space. Under the deterministic encoder the mutual information $I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z}) = H_{\boldsymbol{\phi}}(\mathbf{Z})$ since $H_{\boldsymbol{\phi}}(\mathbf{Z}|\mathbf{X}) = 0$[6].

That is why the ability to compress the observation space to the latent space or to generate from the latent space comes from the relationship between the entropy of observation space distribution $p_{\mathcal{D}}(\mathbf{x})$ and targeted latent space distribution $p_{\boldsymbol{\theta}}(\mathbf{z})$. If the entropy of the observation space is large, i.e., the data are on a complex distributed manifold with a large variance, and the latent space is characterized by a small variance, many samples from the observation space will be mapped very closely to meet the KL-term constraint on the marginal latent space distribution. Naturally, it is a form of "deterministic" compression leading to the reduction of entropy by a "collusion" of many samples from the observation space in the latent space. It should be noticed that in this case, the centroids typically used in quantization based compression are not even used. At the same time, the "continuity" of latent space filling is determined by the randomness of $p_{\mathcal{D}}(\mathbf{x})$ with respect to $p_{\boldsymbol{\theta}}(\mathbf{z})$. If for some reason $p_{\boldsymbol{\theta}}(\mathbf{z})$ is chosen to be relatively "broad", it is not excluded that one might observe some "holes" in the latent space as a result of such a mapping.

Nevertheless, as shown in Figure 4, one can impose any constraint on $p_{\boldsymbol{\theta}}(\mathbf{z})$ like Gaussian, Laplacian or even sparsifying prior. Moreover, one can predefine some centroids or clusters and target that the closest samples in the observation space to be mapped into the same centroids. In this sense, the AAE can also implement a form of deterministic compression by clustering.

At the same time, one can relax the quantization requirement to map an input to exactly one closest centroid and instead to envision some relaxation within the allowed KL-term. These options are not directly implemented in the AAE but can be envisioned. We mention and consider them in view of a link to InfoVAE and generative compression that will be addressed in the next section.

InfoVAE [5] consists of 3 terms obtained by adding the regularisation term $I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z})$ to an alternative form of the VAE. Since this original way of deriving InfoVAE is not straightforward and does not naturally comes from the IB framework, we will show that the InfoVAE has its BIB-AE counterpart with the terms (A), (B) and (C) and can be defined according to the Lagrangian[7]:

$$\begin{aligned}\mathcal{L}_{\text{InfoVAE}}(\boldsymbol{\theta},\boldsymbol{\phi}) = \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[D_{\text{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X}=\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{z})\right)\right] - D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z})\|p_{\boldsymbol{\theta}}(\mathbf{z})) \\ - \beta\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right]\right].\end{aligned} \tag{33}$$

In fact, the terms (A) and (B) correspond to $I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z})$ while the term (C) corresponds to $I^{\text{VAE}}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{X})$. Besides, it should also be pointed out that in the original paper [5] the above three terms have not been used jointly in the reported simulations. Instead, the original InfoVAE uses 2 terms depending on the VAE form, i.e., the terms (A) and (C), or the terms (B) and (C), i.e., the AAE form.

The InfoVAE can also be considered as yet another form of compression by the minimization of $I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z})$. Since it contains both KL-terms (A) and (B) in $I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z})$, the encoder can minimize $I_{\boldsymbol{\phi}}(\mathbf{X};\mathbf{Z})$ by seeking an equality between the terms (A) and (B) since both terms are non-negative. One can consider the presence of term (B) with the regularization parameter β as a regularization of VAE term (A). As a result, it will relax the condition to map all conditional distributions to one Gaussian heap how it is done in the VAE case.

Having considered all these connections, it should be pointed out that the interpretability of the latent space in all considered methods is a quite complex task unless special supervised constraints are imposed how it was finally suggested in a semi-supervised AAE framework. For this reason, we will also consider other possibilities of controllable latent space encoding and generation using generative compression. However, it should be noted that the initial goal of this type of encoding has different roots and requires the selection of optimal distribution to meet a rate-distortion trade-off.

GANs [4]: not pretending to consider the whole GAN family, we can mention that the IB considered for the generative adversarial models in section 2.3 makes it possible to link GAN with BIB-AE. Considering the generation from the targeted latent space distribution $p_{\boldsymbol{\theta}}(\mathbf{z})$ via the generator $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$

[6] One can use a variational decomposition $H_{\boldsymbol{\phi}}(\mathbf{Z}) = -\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z})}\left[\log q_{\boldsymbol{\phi}}(\mathbf{z})\frac{p_{\boldsymbol{\theta}}(\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{z})}\right] = H(q_{\boldsymbol{\phi}}(\mathbf{z}); p_{\boldsymbol{\theta}}(\mathbf{z})) - D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z})\|p_{\boldsymbol{\theta}}(\mathbf{z}))$. Thus, if one wants to reduce the entropy of latent space to the entropy $H_{\boldsymbol{\theta}}(\mathbf{Z})$ of targeted distribution $p_{\boldsymbol{\theta}}(\mathbf{z})$, one should ensure that the encoder targets $q_{\boldsymbol{\phi}}(\mathbf{z}) \to p_{\boldsymbol{\theta}}(\mathbf{z})$ leading to $H(q_{\boldsymbol{\phi}}(\mathbf{z}); p_{\boldsymbol{\theta}}(\mathbf{z})) \to H_{\boldsymbol{\theta}}(\mathbf{Z})$. Therefore, the term (B) in the AAE follows from the minimization of $D_{\text{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z})\|p_{\boldsymbol{\theta}}(\mathbf{z})\right)$.

[7] The original InvoVAE contains different multipliers in front of KL-terms.

one uses (24) that corresponds to the terms (D) and (C) in (28), respectively. Therefore, the BIB-AE is linked to GANs via the IB framework.

It should be remarked that the original GAN does not include the likelihood term (C). However, according to the BIB-AE analysis, this regularizer naturally follows from the IB framework. It is interesting to mention that Rosca *et. al.* [22] have considered this option as a potential solution to the GAN mode collapse problem.

VAE/GAN [13]: an option to use jointly the VAE represented by term (A) and (C) and the GAN represented by term (D) was envisioned in VAE/GAN model. An equivalent VAE/GAN Lagrangian is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{VAE/GAN}}(\boldsymbol{\theta},\boldsymbol{\phi}) = \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[D_{\text{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{X}=\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{z}))\right] - \beta\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right]\right] \\ + \beta D_{\text{KL}}\left(p_{\mathcal{D}}(\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{x})\right).\end{aligned} \tag{34}$$

In the original paper, the log-likelihood term was replaced by a special metric in the latent space[8].

In conclusion, many existing variations of VAE and GAN families can be considered directly from the BIB-AE framework perspectives. The main difference between these approaches, where either the VAE based on ELBO or GAN are taken as a basis and then some regularization terms are added, and the proposed one is in a fact that we proceed directly with the IB formulation and impose the corresponding bounds on the mutual information components of the IB.

Extending the same methodology, next we consider a compression formulation of IB from the Shannon's rate-distortion perspectives and link it with generative models.

4.2 Shannon's rate-distortion and generative compression AEs

In the previous analysis, the targeted latent space distribution was assumed to be any manifold specified by $p_{\boldsymbol{\theta}}(\mathbf{z})$. However, if one wants additionally to have a latent space with a bounded rate below the entropy $H_{\boldsymbol{\theta}}(\mathbf{Z}) = -\mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{z})}[\log p_{\boldsymbol{\theta}}(\mathbf{z})]$, i.e., targeting some compression, yet providing the best reconstruction and possibly generation from the latent space samples, it is of interest to link the considered analysis to the Shannon's rate-distortion theory.

Since the latent space of compression AE should be limited to some rate R_Q, we will assume that the latent space consists of a codebook $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_L\}$, containing the codewords $\mathbf{c}_i \in \mathbb{R}^{n_z}$ of dimension n_z with probabilities $\{p_j\}_{j=1}^{L}$ such that $R_Q = -\sum_{j=1}^{L} p_j \log p_j$. The codewords of $\mathcal{C}$ can be considered as realizations or centroids generated from $p_{\boldsymbol{\theta}}(\mathbf{z})$ that makes it conceptually similar to the AAE. This is conceptually shown in Figure 4 as "compressed" latent space.

At the same time, an essential simplification comes from the fact that the encoder is deterministic and maps the input to one of the above centroids. This can be achieved by a vector quantizer $\hat{\mathbf{z}} = Q(f_{\boldsymbol{\phi}}(\mathbf{x})) := \text{argmin}_{1\leq j\leq L} ||f_{\boldsymbol{\phi}}(\mathbf{x}) - \mathbf{c}_j||_2$, where $f_{\boldsymbol{\phi}}(\mathbf{x})$ denotes a deterministic encoder and $Q(.)$ a vector quantizer (VQ). Hence, the distribution of the quantized latent space is $p_{\boldsymbol{\theta}}(\hat{\mathbf{z}}) = \sum_{j=1}^{L} p_j \boldsymbol{\delta}(\hat{\mathbf{z}} - \mathbf{c}_j)$ that defines the rate R_Q.

Shannon's rate-distortion [23] can be expressed as a special case of (26) with (27) and (28):

$$\mathcal{L}_{\text{Shannon}-\text{AE}}(\boldsymbol{\theta},\boldsymbol{\phi}) = I_{\boldsymbol{\phi}}^{Q}(\mathbf{X};\hat{\mathbf{Z}}) - \beta I_{\boldsymbol{\theta},\boldsymbol{\phi}}^{\text{U}_L}(\hat{\mathbf{Z}};\mathbf{X}). \tag{35}$$

It is easy to show that $I_{\boldsymbol{\phi}}(\mathbf{X};\hat{\mathbf{Z}}) = I_{\boldsymbol{\phi}}^{Q}(\mathbf{X};\hat{\mathbf{Z}}) = H_{\boldsymbol{\phi}}(\hat{\mathbf{Z}})$ due to the deterministic encoding with quantization, while $I_{\boldsymbol{\theta},\boldsymbol{\phi}}^{\text{U}_L}(\hat{\mathbf{Z}};\mathbf{X})$ is reduced to the term (C) that under the deterministic decoding further reduces to $\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\hat{\mathbf{z}})\right]$. This term corresponds to the reconstruction distortion that is often expressed as the ℓ_2-norm that in turns corresponds to the Shannon's lower bound on rate-distortion function. Therefore, the classical compression schemes satisfy the trade-off between the rate $I_{\boldsymbol{\phi}}^{Q}(\mathbf{X};\hat{\mathbf{Z}}) = R_Q$ and distortion $\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\log(-p_{\boldsymbol{\theta}}(\mathbf{x}|\hat{\mathbf{z}}))\right] = D$. Finally, the latent space distribution $p_{\boldsymbol{\theta}}(\hat{\mathbf{z}})$ is optimized to ensure the achievability of rate-distortion limit. This is a fundamental difference with the AAE, where the latent space distribution is chosen in advance for the technical reasons.

It is important to note that the Shannon's rate distortion framework in the considered interpretation is closely linked with the AAE, when the targeted distribution latent space is represented by the

[8] One can use both encoded-reconstructed samples and samples generated from $p_{\boldsymbol{\theta}}(\mathbf{z})$ in the third term for the adversarial discrimination.

compression codebook Figure 5. The difference is in the practical implementation. The VQ implementation assumes a hard assignment of the input to one of centroids[9], whereas the AAE proceeds with the optimization of the KL-term to fit the targeted latent space distribution. It means that some deviation from the centroids is still possible. However, in both cases to proceed with the generation from the latent compressed space, one needs to ensure a proper randomness. Otherwise, the space of reconstructed signals will correspond to the number of centroids in the latent space. For this reason, we will consider a generative compression and link it with the IB framework.

Generative compression [14, 15, 16, 17] can be considered as an "extension" of Shannon's rate distortion with the Lagrangian:

$$\mathcal{L}_{\mathrm{GC-AE}}(\boldsymbol{\theta},\boldsymbol{\phi}) = I^{Q}_{\boldsymbol{\phi}}(\mathbf{X};\hat{\mathbf{Z}}) - \beta I^{\mathrm{U}_L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\hat{\mathbf{Z}}+\mathbf{U};\mathbf{X}). \tag{36}$$

The first term is the same as in the classical compression setup, while the second one contains a stochastic component achieved by the addition of the permutation $\mathbf{U} \sim p_{\mathbf{u}}(\mathbf{u})$ to the centroids[10]. At the same time, it contains both equivalent terms (C) and (D) in (28). In practice, the KL-divergence is lower bounded by f-divergence that is implemented in a form of adversarial loss based on a density ratio estimation [24, 25] or its Wasserstein's counterpart [26]. In the original generative compression papers, the origin of the f-divergence term interpreted as a perceptual loss was only explained from the heuristic point of view to make highly compressed fragments of images under a low compression rate to look more naturally but not necessarily to be close to the original fragments. However, we can trace the origin of this term as an outcome of the IB factorization.

4.3 Novelty detection AEs

The novelty detection problem aims at detecting outliers with respect to some manifold represented by the training data set. It assumed that similarly to the unsupervised setup, the training set consisting of N samples is given. One can use different techniques to measure the relevance of a test sample to the training set or even to train a one class classifier for this purpose.

Alternatively, one can consider a novelty detection problem from the position of unsupervised IB framework in the BIB-AE formulation. It is interesting to note that [18] proposed the architecture similar to the BIB-AE presented in Figure 3 and trained with the terms (B), (C) and (D) for the novelty detection. The AE trained in this way might use several metrics such as output of term (D) to detect outliers. This also corresponds to a one-class classification problem. Therefore, the mechanism of novelty detection can be seen from the perspective of using the BIB-AE architecture.

5 Conclusions

In this paper, we considered the IB for several practical tasks covering supervised, unsupervised, generative adversarial, generative compressive and novelty detection models. We show that the IB for all these models reduces to four terms in the Lagrangian cost. We call this formulation as BIB-AE. This formulation is closely linked with many models ranging from the VAE to VAE/GAN.

Besides this remarkable similarity, we note that this connection is seen via the IB framework with application of variational approach to the decomposition of mutual information terms in contrast to the VAE family that is based on various attempts to regularize the ELBO. As a result, the interpretability of obtained results and connection between methods leads to different conclusions.

Along the same line, we consider the new framework of generative compression in a close link to the IB framework whereas the original works on the generative compression considered it from the "perceptual" perspectives by adding the regularizer similar to the ELBO.

Finally, we also show that the novelty detection problem in the recent interpretability of AE encoding with the adversarial loss can be linked to the BIB-AE interpretation. Altogether the performed analysis gives new insights on the connections between different problems and methods and creates an interesting basis for the interpretability of the latent space.

Acknowledgement

The research was supported by the SNF project No. 200021-182063. The authors are thankful to Behrooz Razeghi for his feedback and fruitful discussions.

[9] The multiple assignments are also possible that is know as *soft-encoding*.

[10] We use another variable $\mathbf{u}$ for the randomization of centroids to reflect a fact that it is assigned to the decoder part in contrast to the randomization based on the encoder randomization using $\boldsymbol{\epsilon}$.

Appendix A

In this part, we derive a lower bound on $I(\mathbf{Z};\mathbf{C})$. According to the definition (5), this mutual information can be further decomposed as:

$$\begin{aligned} I(\mathbf{Z};\mathbf{C}) &= \mathbb{E}_{p(\mathbf{c},\mathbf{z})}\left[\log\frac{p(\mathbf{c}|\mathbf{z})}{p(\mathbf{c})}\right] \\ &= -\mathbb{E}_{p(\mathbf{c})}\left[\log p(\mathbf{c})\right] + \mathbb{E}_{p(\mathbf{c},\mathbf{z})}\left[\log p(\mathbf{c}|\mathbf{z})\right]. \end{aligned} \tag{37}$$

The first term of this decomposition corresponds to the entropy of classes $H(\mathbf{C}) = -\mathbb{E}_{p(\mathbf{c})}\left[\log p(\mathbf{c})\right]$.

We consider the second term since the transition probability $p(\mathbf{c}|\mathbf{z})$ is unknown. At the same time, it can be written as:

$$\mathbb{E}_{p(\mathbf{c},\mathbf{z})}\left[\log p(\mathbf{c}|\mathbf{z})\right] = \int_{\mathbf{c}}\int_{\mathbf{z}} p(\mathbf{c},\mathbf{z})\log p(\mathbf{c}|\mathbf{z})\,\mathrm{d}\mathbf{c}\,\mathrm{d}\mathbf{z}. \tag{38}$$

The expectation is with respect to the joint distribution $p(\mathbf{c},\mathbf{z})$ that can also be defined via the marginalization $p(\mathbf{c},\mathbf{z}) = \int_{\mathbf{x}} p(\mathbf{c},\mathbf{z},\mathbf{x})d\mathbf{x} = \int_{\mathbf{x}} p(\mathbf{c},\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\mathrm{d}\mathbf{x}$. Therefore, combing these results, one can obtain:

$$\begin{aligned} \mathbb{E}_{p(\mathbf{c},\mathbf{z})}\left[\log p(\mathbf{c}|\mathbf{z})\right] &= \int_{\mathbf{c}}\int_{\mathbf{z}}\int_{\mathbf{x}} p(\mathbf{c},\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\log p(\mathbf{c}|\mathbf{z})\,\mathrm{d}\mathbf{c}\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{x} \\ &= \mathbb{E}_{p(\mathbf{c},\mathbf{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p(\mathbf{c}|\mathbf{z})\right]\right]. \end{aligned} \tag{39}$$

To overcome the problem of unknown $p(\mathbf{c}|\mathbf{z})$, we will apply a variational distribution $p_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{z})$ parametrized via a set of parameters $\boldsymbol{\theta}$ to approximate $p(\mathbf{c}|\mathbf{z})$. This can be considered as a bypass network and formulated as:

$$\begin{aligned} \mathbb{E}_{p(\mathbf{c},\mathbf{z})}\left[\log p(\mathbf{c}|\mathbf{z})\right] &= \mathbb{E}_{p(\mathbf{c},\mathbf{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p(\mathbf{c}|\mathbf{z})\frac{p_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{z})}\right]\right] \\ &= \mathbb{E}_{p(\mathbf{c},\mathbf{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{z})\right]\right] + \mathbb{E}_{p(\mathbf{c},\mathbf{z})}\left[\log\frac{p(\mathbf{c}|\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{z})}\right], \end{aligned} \tag{40}$$

where in the second term we used the expectation defined in (39).

At the same time, we can re-write $p(\mathbf{c},\mathbf{z}) = p(\mathbf{z})p(\mathbf{c}|\mathbf{z})$ that leads to[11]:

$$\begin{aligned} \mathbb{E}_{p(\mathbf{c},\mathbf{z})}\left[\log\frac{p(\mathbf{c}|\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{z})}\right] &= \mathbb{E}_{p(\mathbf{z})}\left[\mathbb{E}_{p(\mathbf{c}|\mathbf{z})}\left[\log\frac{p(\mathbf{c}|\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{z})}\right]\right] \\ &= \mathbb{E}_{p(\mathbf{z})}\left[D_{\mathrm{KL}}\left(p(\mathbf{c}|\mathbf{Z}=\mathbf{z})\|p_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{Z}=\mathbf{z})\right)\right] = D_{\mathrm{KL}}\left(p(\mathbf{c}|\mathbf{z})\|p_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{z})\right). \end{aligned} \tag{41}$$

Since the KL-divergence $D_{\mathrm{KL}}\left(p(\mathbf{c}|\mathbf{z})\|p_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{z})\right) \geq 0$, we can lower bound (40) as:

$$\mathbb{E}_{p(\mathbf{c},\mathbf{z})}\left[\log p(\mathbf{c}|\mathbf{z})\right] \geq \mathbb{E}_{p(\mathbf{c},\mathbf{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{z})\right]\right]. \tag{42}$$

Therefore, the mutual information (37) can be lower bounded as $I(\mathbf{Z};\mathbf{C}) \geq I^{\mathrm{S}}_{\boldsymbol{\theta},\phi}(\mathbf{Z};\mathbf{C})$, where we define a lower bound as:

$$\begin{aligned} I^{\mathrm{S}}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{Z};\mathbf{C}) &\triangleq H(\mathbf{C}) + \mathbb{E}_{p(\mathbf{c},\mathbf{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{z})\right]\right] \\ &= H(\mathbf{C}) - H_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{C}|\mathbf{Z}), \end{aligned} \tag{43}$$

where $H_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{C}|\mathbf{Z}) = -\mathbb{E}_{p(\mathbf{c},\mathbf{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{c}|\mathbf{z})\right]\right]$.

References

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

[2] D.P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114*, 2014.

[11] One can note that $p(\mathbf{z}) = q_{\boldsymbol{\phi}}(\mathbf{z})$.

[3] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

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

[5] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017.

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

[7] Shengjia Zhao, Jiaming Song, and Stefano Ermon. The information autoencoding family: A lagrangian perspective on latent variable generative models. *arXiv preprint arXiv:1806.06514*, 2018.

[8] Alex Alemi, Ben Poole, Ian Fischer, Josh Dillon, Rif A Saurus, and Kevin Murphy. An information-theoretic analysis of deep latent-variable models. 2018.

[9] Alexander A Alemi, Ben Poole, Ian Fischer, Joshua V Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. *arXiv preprint arXiv:1711.00464*, 2017.

[10] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

[11] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2(5):6, 2017.

[12] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[13] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.

[14] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. *arXiv preprint arXiv:1804.02958*, 2018.

[15] Shibani Santurkar, David Budden, and Nir Shavit. Generative compression. In *2018 Picture Coding Symposium (PCS)*, pages 258–262. IEEE, 2018.

[16] Michael Tschannen, Eirikur Agustsson, and Mario Lucic. Deep generative models for distribution-preserving lossy compression. In *Advances in Neural Information Processing Systems*, pages 5929–5940, 2018.

[17] Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. *arXiv preprint arXiv:1901.07821*, 2019.

[18] Stanislav Pidhorskyi, Ranya Almohsen, and Gianfranco Doretto. Generative probabilistic novelty detection with adversarial autoencoders. In *Advances in Neural Information Processing Systems*, pages 6822–6833, 2018.

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

[20] I. Goodfellow et al. Generative adversarial nets. *arXiv:1406.2661*, 2014.

[21] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[22] Mihaela Rosca, Balaji Lakshminarayanan, David Warde-Farley, and Shakir Mohamed. Variational approaches for auto-encoding generative adversarial networks. *arXiv preprint arXiv:1706.04987*, 2017.

[23] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

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

[25] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

[26] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.