UNIVERSITY OF AMSTERDAM

MSc ARTIFICIAL INTELLIGENCE
MASTER THESIS

Understanding Deep Representation Models Through Information Theory

---

by
MARCO FEDERICI
11413042

September 8, 2018

36 EC
07/01/2018 - 08/09/2018

Supervisor:
Karen Ullrich MSc
Co-Supervisor:
Dr Jakub M. Tomczak

Examiner:
Dr Zeynep Akata

UNIVERSITY OF AMSTERDAM

Faculty of Science

Abstract

Master of Science

Understanding Deep Representation Models Through Information Theory

by Marco FEDERICI

Recent literature has been proposing an increasing number of learning objectives for unsupervised representation learning. By approaching the problem of building a data representation first from a generative perspective, then from an information theoretic point of view, we present a novel generalized loss function that allows to better understand and compare different popular models. In order to facilitate a direct comparison, we introduce a refined approximation for some information theoretical quantities of interest that allows to measure and compare several aspects of the presented architectures. An empirical evaluation shows that the theory is consistent with the proposed estimations and gives additional insights regarding the strengths and weaknesses of various architectures in the literature. Additionally, this research suggests new directions of exploration that arise from the presented view.

Acknowledgements

I would like to thank my supervisors, Karen Ullrich and Jakub M. Tomczak, for their trust and feedback that granted me the freedom to explore my personal area of interest.

Thank you to all the fellow student of the AI master room that tolerated me during these last months. In particular, I express my gratitude to Jose Daniel Gallego for the insightful discussions, Iris Verweij, Heng Lin, Joop Pascha and Janosch Harber for the continuous support and feedback.

Un immenso grazie alla mia famiglia che non ha mai fatto mancare il supporto che mi ha portato a questo traguardo.

Contents

Abstract iii

Acknowledgements v

1 Introduction 1
- 1.1 Structure . . . 2
- 1.2 Contributions . . . 2
- 1.3 Related work . . . 2

2 Learning Objectives for Generative Models 5
- 2.1 Information and moment projection . . . 5
  - 2.1.1 Parametric models . . . 6
    - M-projection and maximum log-likelihood . . . 7
    - I-projection and density ratio estimation . . . 7
- 2.2 Modeling joint distributions . . . 8
  - 2.2.1 A parametric view . . . 9

3 Information Theoretical Quantities for Representation Learning 13
- 3.1 Encoding mutual information . . . 13
  - 3.1.1 Data distortion . . . 14
  - 3.1.2 Code rate . . . 15
  - 3.1.3 Computation . . . 15
- 3.2 Decoding mutual information . . . 15
  - 3.2.1 Code distortion . . . 16
  - 3.2.2 Data rate . . . 17
  - 3.2.3 Computation . . . 17

4 Learning Objectives for Unsupervised Representation Learning 19
- 4.1 A generalized learning objective . . . 19
- 4.2 Analysis of the different training objectives . . . 20
  - 4.2.1 Variational Autoencoder . . . 20
  - 4.2.2 β-Variational Autoencoder . . . 22
  - 4.2.3 Information Variational Autoencoder . . . 23
  - 4.2.4 Information Generative Adversarial Network . . . 24
  - 4.2.5 Cycle-Consistent Generative Adversarial Network . . . 26
- 4.3 Summary and comparison . . . 27

5 Experiments 29
- 5.1 Experimental setup . . . 29
  - 5.1.1 1D-Mixture . . . 30
  - 5.1.2 2D-Mixture . . . 31
- 5.2 Experiments and Results . . . 32
  - 5.2.1 Effect of the joint divergence . . . 32

5.2.2 Effect of the hyper-parameters on the generative performances 33
5.2.3 Effect of the hyper-parameters on the mutual information . . . 35
5.2.4 Comparing the different models . . . 37
5.3 Conclusions . . . 41

6 Discussion and Future Work **43**
6.1 Future work . . . 44

Bibliography **45**

A Information Theoretical Quantities **51**
A.1 Entropy and mutual information . . . 51
A.2 Cross-entropy . . . 53
A.3 Kullback-Leibler divergence . . . 53
A.4 Notes on differential entropy and cross-entropy . . . 54

B Interpolating the I-projection and M-projection **57**

C The Density-ratio Trick **59**
C.1 Estimating the Kullback-Leibler divergence . . . 59
C.2 Estimating the Jensen-Shannon divergence . . . 60
C.3 Modeling the approximate classifier . . . 61

D Modeling the Parametric Distributions **63**
D.1 The encoding distribution . . . 63
D.2 The decoding distribution . . . 63

E Estimation of the Information Theoretical Quantities **65**
E.1 Entropy . . . 65
E.2 Data distortion and code distortion . . . 65
E.3 Data rate and the code rate . . . 66
E.4 Kullback-Leibler divergence . . . 66
E.5 Mutual information . . . 67
E.6 Summary . . . 67

F Experimental Details **69**
F.1 The Smoothed Uniform Distribution . . . 69
F.2 Details on the experiments . . . 70
F.3 Details on the visualizations . . . 70
F.4 Additional visualizations . . . 72

List of Figures

2.1 Visualization of the characteristics of the information (in yellow) and moment projection (in blue) of a mixture model (in red), onto the set of Normal distributions. The characteristics of the parameters' configurations are visualized on the right side of the picture. 6
2.2 Visualization of the maximum likelihood interpretation of the Moment projection for two distinct parameterizations defined through the parameters spaces $\Theta \supseteq \Theta'$. In the reported example, $\hat{\theta}'_M$ represents parameter configuration in Θ' that achieves the maximum likelihood for the empirical distribution $\tilde{r}_{\mathcal{D}}(X)$. 7
2.3 Graphical visualization of the sets $\mathcal{C}(\tilde{r}_{\mathcal{D}}(X))$ and $\mathcal{C}(r(Z))$ of distributions with marginals $\tilde{r}_{\mathcal{D}}(X)$ and $r(Z)$ respectively. Since their parametric restriction $\mathcal{C}_\Theta(\tilde{r}_{\mathcal{D}}(X))$ and $\mathcal{C}_\Phi(r(Z))$ do not intersect, the probability distributions $q_\theta(X,Z)$ and $p_\phi(X,Z)$, identified through the minimization of a joint divergence, do not coincide. Note that in this scenario the choice of different divergences influences the parameter configuration of the minimum. 10

3.1 Graphical representation of the relationships between the information theoretical quantities defined by the encoding distribution, $q_\theta(X,Z)$, and its interaction with the decoding distribution, $p_\phi(X,Z)$. The coloring used for the different terms represents their tractability. A red color surrounds the terms that cannot be approximated directly, while the green box denotes the components that can be easily computed. The KL-divergence $KL(q_\theta(Z)||p(Z))$ is colored in yellow since we have access to a rough estimation obtained by using the density-ratio trick. 16
3.2 Graphical representation of the relationships between the information theoretical quantities defined by the decoding distribution, $p_\phi(X,Z)$, and its interaction with the encoding distribution, $q_\theta(X,Z)$. Analogously to the representation reported in Figure 3.1, the color scheme is used to represents the tractability of different quantities. 18

5.1 Visualization of the probability density of the data-generating distribution $r(X)$ (in red) and the empirical distribution $\tilde{r}_{\mathcal{D}}(X)$ (vertical orange lines) for the 1D-Mixture dataset. 30
5.2 Visualization of the data generating distribution $r(X)$ (in red) and the empirical distribution $\tilde{r}_{\mathcal{D}}(X)$ (in orange) for the 2D-Mixture experiments. 31

5.3 Visualization of joint encoding and decoding distributions $p_\phi(X, Z)$ and $q_\theta(X, Z)$ trained on the 1D-Mixture dataset for β-VAE, Info-VAE, Info-GAN and Cycle-GAN (one for each column). The two rows correspond to the choice of divergence used to match the marginal distributions (Kullback-Leibler divergence for the first row, Jensen-Shannon for the second one). Each plot presents an approximation of the joint distribution (in the center) together with the data marginals (on the top) and the latent code marginals (on the right). The data-generating distribution is represented trough a dashed red line. The plot shows that the for the β-VAE and Info-VAE models the decoding distribution completely covers the support of the encoding distribution. The Info-GAN architecture leads to the opposite scenario, while the Cycle-GAN loss results in a balance between the zero-avoiding and zero-forcing behavior. 33

5.4 Visualization of the data marginals obtained by training the same model using the VAE (on the left) and Info-GAN (on the right) learning objectives on the 2D-Mixture dataset. The plot compares the modes of the data-generating distribution $r(X)$ (in red) with an approximation of the density computed by sampling the data marginal $p_\phi(X)$. By observing the area covered by the samples, one may notice that the VAE model over-estimates the support of $r(X)$, while the Info-GAN model does not cover one of the three modes of the data-generating distribution. 34

5.5 Plot of the trade-off between $KL(q_\theta(Z)||p(Z))$ (in red) and $KL(q_\theta(X|Z)||p_\phi(X|Z))$ (in green) measured for β-VAE, Info-VAE and Cycle-GAN by varying the respective hyper-parameter. Each model has been trained 5 times with different instances of the 1D-Mixture distribution. The average values of divergence achieved by the different architectures after convergence are reported together with the estimation of the standard deviations. The picture shows that the trade-off of $KL(q_\theta(Z)||p(Z))$ and $KL(q_\theta(X|Z)||p_\phi(X|Z))$ measured empirically is consistent with the scaling coefficient determined theoretically as a function of their respective hyper-parameter (Table 4.2). Furthermore the reduced variance validates the effectiveness of the proposed estimations for $KL(q_\theta(Z)||p(Z))$ and $KL(q_\theta(X|Z)||p_\phi(X|Z))$. Note that the Info-VAE and Info-GAN have been trained using the Kullback-Leibler adversarial approximation. 34

5.6 Plot of the trade-off between two components of the joint KL divergence $KL(p_\phi(X)||q(X))$ (in blue) and $KL(p_\phi(Z|X)||q_\theta(Z|X))$ obtained by the Info-GAN and Cycle-GAN by varying the respective hyper-parameters. The mean and the variance of the measurements have been estimated by following the same procedure described in Figure 5.5. Note that since $KL(p_\phi(Z|X)||q_\theta(Z|X))$ can not be estimated directly, the plot reports its upper bound that is represented by the code distortion $D_Z(\theta, \phi)$ instead (dotted line in pink). For both the Cycle-GAN and Info-GAN architectures, the empirical measurements are consistent with the values of the coefficients reported in Table 4.2. . 35

5.7 Comparison of encoding (in blue) and decoding (in red) mutual informations for the β-VAE, Info-VAE, Info-GAN and Cycle-GAN architectures obtained by varying the respective parameters on the 1D-Mixture dataset. Each plot reports the lower bound (dashed blue line), the upper bound (dotted blue line) and the estimation for the encoding mutual information (solid line) together with the lower bound of the decoding mutual information (dashed red line). The empirical observations reported in the four graphs are coherent with the coefficients α_q and α_p determined in Chapter 4. The procedure used to determine the mean and the variance of the measurement is identical to the one reported in Figures 5.6 and 5.5. 36
5.8 Visualization of the correlation between the encoding mutual information estimation $I(q_\theta(X, Z))$ (x-axis) and the lower bound of the decoding mutual information $\underline{I}(p_\phi(X, Z))$ (y-axis) measured for the different models, hyper-parameters and choices of marginal divergences on 5 different instances of the 1D-Mixture dataset. The two mutual information exhibit a strong correlation even when the training objective is addressing only one of the two components. 36
5.9 Plot of the trade-off between code rate $R_Z(\theta)$ and data distortion $D_X(\theta, \phi)$ achieved by the β-VAE, Info-VAE and Cycle-GAN objectives by varying respective parameters. The curve defined by the Info-GAN objective is not visualized as the data distortion is orders of magnitudes bigger then the ones achieved by the other architectures. The difference in scale underlines that the Rate-Distortion interpretation (Alemi et al., 2018) is inadequate to evaluate the different aspects of the 4 learning objectives. 37
5.10 Visualization of the effect of the mode-collapse operated by the Info-GAN architecture. The plot visualizes both the estimations for the marginal $p_\phi(X)$ and $q_\theta(Z)$ (first and second columns from the left respectively) together with the data and the code reconstructions (last two columns on the right) obtained by encoding and decoding the dataset $\mathcal{D}$ and a uniform code grid G using the mean of the encoding $\mu_\theta(x)$ and decoding $\mu_\phi(z)$ distributions respectively. By observing that the decoding distribution does not model one of the modes of the real data distribution we can expect high values of data distortion $D_X(\theta, \phi)$. This is because the behavior of the encoder $q_\theta(Z|X)$ is un-constraint in the regions that are not modeled by $p_\phi(X)$. In fact, $q_\theta(Z|X)$ may map the observations that are not included in its support arbitrarily far from the prior (second figure from the left). At the same time, the decoder, $p_\phi(X|Z)$, is not necessarily consistent with the encoder in the areas that are unlikely to be sampled from $p(Z)$. As a consequence, the reconstruction for the data belonging to the missing modes lies arbitrarily far from the original observations (top-right mode in the third column) resulting in an increased data distortion $D_X(\theta, \phi)$. 38

5.11 Comparison between the estimation of the values of the Variational Lower Bound (ELBO), data distortion ($D_X(\theta,\phi)$), code rate ($R_Z(\theta)$) and code distortion ($D_Z(\theta,\phi)$). Each model has been train on 5 different instances of the 1D-Mixture dataset with their respective hyperparameter set to 1. The measures reported on the histograms refer the mean and standard deviation of the different quantities evaluated after convergence. 39
5.12 Comparison between the estimation of $KL(p_\phi(X)||q(X))$, $KL(q_\theta(Z)||p(Z))$ and $KL(q_\theta(X|Z)||p_\phi(X|Z))$ for the different models tested with $\lambda = 1$ ($\beta = 1$ for the β-VAE). The experimental setup is equivalent to the one used to produce Figure 5.11. 40
5.13 Visualization of the estimation for the values of encoding mutual information ($I(q_\theta(X,Z))$) and lower bound on the decoding mutual information ($\underline{I}(p_\phi(X,Z))$) obtained with the same experimental setup described in Figure 5.11. 40

A.1 Visualization of the entropy diagram for the random variables X and Z. The different areas correspond to the additive relationships that characterize the entropies, conditional entropies and mutual information. 52

B.1 γ-projections of the mixture distribution $p(X)$ obtained for different values of γ onto the set of Normal distributions. 57

F.1 Plot of the probability density function for the smoothed uniform distribution. The values of the parameters m, s and σ are graphically visualized on the picture. 69
F.2 Visualization of the encoding and decoding joint distributions induced by the Variational Autoencoder training objective $\mathcal{L}_{VAE}(\theta,\phi)$ on the 1D-Mixture dataset. 71
F.3 Visualization of the encoding and decoding joint distributions induced by the Variational Autoencoder training objective $\mathcal{L}_{VAE}(\theta,\phi)$ on the 2D-Mixture dataset. 72
F.4 Visualization of the generative and reconstruction performances for the Beta-VAE training objective. 73
F.5 Visualization of the generative and reconstruction performances for the Info-VAE training objective with the adversarial Jensen-Shannon divergence approximation. 73
F.6 Visualization of the generative and reconstruction performances for the Info-VAE training objective with the adversarial Kullback-Leibler divergence approximation. 74
F.7 Visualization of the generative and reconstruction performances for the Info-GAN training objective with the adversarial Jensen-Shannon divergence approximation. 74
F.8 Visualization of the generative and reconstruction performances for the Info-GAN training objective with the adversarial Kullback-Leibler divergence approximation. 75
F.9 Visualization of the generative and reconstruction performances for the Cycle-GAN training objective with the adversarial Jensen-Shannon divergence approximation. 75

F.10 Visualization of the generative and reconstruction performances for the Cycle-GAN training objective with the adversarial Kullback-Leibler divergence approximation. 76

List of Tables

4.1 Comparison between the coefficient introduced by the different models in the original formulation. The different columns report the coefficients for the code-rate $R_Z(\theta)$, data distortion $D_X(\theta, \phi)$, code distortion $D_Z(\theta, \phi)$, divergence between the data-marginal and the empirical distribution $D(p_\phi(X)||q(X))$ and divergence between the prior and the aggregated posterior $D(q_\theta(Z)||p(Z))$. 28

4.2 Comparison of the weighting of the components of the joint divergence (π_M and π_C) and the coefficients of the encoding and decoding mutual information (α_q and α_p) as a function of the model hyperparameters. The joint divergence $D(q_\theta(X, Z)||p_\phi(X, Z))$ is expressed in the form $\pi_M \, M(\theta, \phi) + \pi_C \, C(\theta, \phi)$. 28

Chapter 1

Introduction

Choosing an appropriate data representation has a significant impact on the performance of Machine Learning algorithms. For this reason, over the years, a lot of effort has been invested in designing models that automatically extract relevant features from the data. In particular, reducing the data dimensionality and creating fixed-size representations led to a number of successful applications in the areas of speech recognition (Hinton et al., 2012), object recognition (Krizhevsky, 2010) and natural language processing (Mikolov et al., 2013). Other than representing a helpful tool for data pre-processing, the use of latent feature representations allows to create abstractions that capture useful characteristics of the data. In fact, unsupervised representation learning has been successfully used to approach multiple classification problems at the same time (Passos et al., 2012), adapt data from different domains (Liu, Breuel, and Kautz, 2017) and recognize objects that have not been observed during training (Akata et al., 2013; Yu et al., 2017).

Deep generative models currently represent one of the most successful solutions for creating effective latent representations. In particular, Variational Autoencoders (Kingma and Welling, 2013) accomplished remarkable results in both textual (Miao, Yu, and Blunsom, 2016) and visual (Pu et al., 2016) domains by optimizing a variational lower bound of the data likelihood. On the other side of the spectrum, Generative Adversarial Networks (Goodfellow et al., 2014a) have been recently extended to include an explicit latent representation (Radford, Metz, and Chintala, 2015; Chen et al., 2017; Li et al., 2017), presenting a viable alternative for learning the data features in an unsupervised fashion. The introduction of new hybrid methodologies (Makhzani et al., 2014; Larsen et al., 2015) led to a wide range of learning objectives and evaluation methods that in their diversity obfuscate the common underlying principles of representation learning (Bengio, Courville, and Vincent, 2013; Huszár, 2018).

Because the learning objective represents one of the fundamental aspects of a model, in this thesis we want to take one step back and focus on the analysis and comparison of the different approaches in literature, trying to shed light in the world of unsupervised deep representation models through the use of information theoretical tools. The presented analysis will approach the problem of defining a learning goal in two steps: modeling the data-generating distribution through observations and the study of the amount of information that is represented in the latent codes. By defining a generalized loss function, this work allows to re-formulate the learning objectives of different state-of-the-art models to allow for a direct comparison between their goals and design choices.

1.1 Structure

This thesis aims to guide the reader through the definition of a unified learning objective in subsequent steps. A first part introduces the problem of creating a latent feature representation from a generative perspective. Secondly, an information theoretical analysis presents what the important characteristic of a useful representation are and how they can be estimated. The two perspectives are then unified through the definition of a loss function that takes both points of views into account and allows to re-interpret various well-known models in the literature.

- Chapter 2: Learning objectives for generative models
- Chapter 3: Information theoretical quantities for representation learning
- Chapter 4: Unified learning goal and literature analysis
- Chapter 5 Experiments
- Chapter 6 Discussion and future work

An introduction to the information theoretical quantities and properties that will used in this work is included in appendix A.

1.2 Contributions

The main goal of this work consist in defining a perspective that allows to interpret the different unsupervised representation learning objectives proposed in the literature. To this end, a theoretical analysis furnishes a novel interpretation that expresses the goal of learning a data representation as weighed combination of two components: i) the matching between two joint distributions and ii) a constraint on the mutual information between the data and its latent representation. Other then allowing a direct comparison between the diverse loss functions, the proposed view provides a new interpretation for some of the hyper-parameter included in the expression of the training objectives that is supported by empirical evidence.

The information theoretic analysis presented in this work introduces elements of novelty through the definition of a refined estimation for the mutual information that allows for a direct comparison between various models. Furthermore, this work suggests an intuitive interpretation for different quantities induced by modeling the joint distributions of data and codes. This is done by considering coding and a novel "reverse-coding" scheme that allow to relate several quantities considered in literature. Other than unifying multiple classes of learning objectives, the proposed interpretation opens up a number possible direction for future explorations that are discussed in the last section of this work.

1.3 Related work

In recent years, deep architectures received an increasing attention from the community in the area of representation learning. The majority of the successful unsupervised approaches are relying on the assumption that useful representations are a result of an effective generative model. This is because the ability to generate new observations is considered to be related to some form of understanding of the underlying generating process. In this context, Variational Autoencoders (VAEs) (Kingma and Welling, 2013) suggest a generative approach to model the joint occurrences of the data and their latent codes. The obtained representation has been shown to

be useful for semi-supervised tasks on small datasets (Kingma et al., 2014). However, the performance deteriorates for more complicated data distributions (Larsen et al., 2015; Zhao, Song, and Ermon, 2017b). One of the main reasons behind the limitation of VAEs is due to the restrictions imposed by the parametric representation (Rezende and Mohamed, 2016; Kingma et al., 2017; Tomczak and Welling, 2017; Chen et al., 2017). Nevertheless, recent work has also shown that the additional flexibility does not necessarily result in better latent representations (Makhzani and Frey, 2017; Huszár, 2018), which suggests that solely modeling the data distribution is an insufficient objective for learning meaningful representations.

Another prominent direction of exploration in the area of unsupervised deep learning considered the use of an adversarial procedure to approximately model the data distribution. Generative Adversarial Networks (GANs) (Goodfellow et al., 2014b) rapidly gained popularity thanks to the sharpness of their generated samples (Goodfellow, 2016; Ledig et al., 2017) often obtained at the price of a less stable training procedure (Salimans et al., 2016; Arjovsky, Chintala, and Bottou, 2017). Although vanilla GANs do not directly provide a latent representation for the observations, several approaches have been proposing strategies to learn a mapping from the data to the code space (Radford, Metz, and Chintala, 2015; Chen et al., 2017; Li et al., 2017; Belghazi et al., 2018), resulting in effective data representations.

Over the years, a number of studies have been trying to combine the stability of VAE architectures with the advantages of the adversarial training procedure (Makhzani et al., 2014; Zhao, Song, and Ermon, 2017a; Mescheder, Nowozin, and Geiger, 2017; Tiao, Bonilla, and Ramos, 2018), vastly increasing the number of possible options for creating a latent representation in unsupervised settings. Despite the growing number of learning objectives proposed in the literature, only a few work examined into the comparison between different approaches. In particular, Hu et al., 2017 suggests a unified view for VAEs and GANs based on two phases of the Wake-Sleep algorithm (Hinton et al., 1995) focusing on their generative aspect; Alemi et al., 2018 proposes to interpret some VAE-based models with rate-distortion theory; and Zhao and Ermon, 2017 introduces an interpretation that links different objectives in literature to a constraint optimization problem without focusing on the specific effect of the model hyper-parameters. The work in this thesis differs from the literature in a way that the challenges of learning a data representation are examined from both generative and information theoretic point of views. In fact, we demonstrate that both perspectives are simultaneously required to achieve a better understanding of the properties of different models.

Chapter 2

Learning Objectives for Generative Models

A number of the successful approaches for unsupervised representation learning currently build upon a generative perspective. This is because the ability of synthesizing new observations requires some form of understanding of the underlying characteristics and patterns of the data, which are hopefully captured in its representation. For this reason, we are going to approach the problem of unsupervised representation learning from a generative point of view by first presenting the techniques of information and moment projection, and how they can be used to model a data generating process. The second part of this chapters focuses on showing how the generative approach can be extended to model joint occurrences of data and codes to create a latent data representation.

2.1 Information and moment projection

The information and moment projections represent two different approaches to model probability distributions first introduced in the context of information theory. Considering a domain $\mathcal{X}$ and the set $S(\mathcal{X})$ of all the probability distributions defined on $\mathcal{X}$, the information projection (I-projection) $p^I_{\mathcal{C}}(X)$ of a probability distribution $p(X) \in S(\mathcal{X})$ onto a set $\mathcal{C} \subseteq S(\mathcal{X})$ is a probability distribution in $\mathcal{C}$ defined as:

$$p^I_{\mathcal{C}}(X) := \underset{q(X)\in\mathcal{C}}{\operatorname{argmin}} KL(q(X)||p(X)) \tag{2.1}$$

Where $KL(q(X)||p(X))$ represents the Kullback-Leibler divergence (KL-divergence) between $q(X) \in \mathcal{C}$ and the given distribution $p(X)$.

Analogously, the moment projection (M-projection) $p^M_{\mathcal{C}}(X)$ of a probability distribution $p(X) \in S(\mathcal{X})$ onto the set $\mathcal{C} \subseteq S(\mathcal{X})$ is defined as the minimizer of the other direction of the KL-divergence:

$$p^M_{\mathcal{C}}(X) := \underset{q(X)\in\mathcal{C}}{\operatorname{argmin}} KL(p(X)||q(X)) \tag{2.2}$$

As the KL-divergence is not symmetric, the information and moment projection generally do not coincide and capture different properties of the projected distribution.

The I-projection $p^I_{\mathcal{C}}(X)$ usually tends to underestimate the support of $p(X)$ by modeling only some of the modes of the projected distribution. In fact, the minimization described in equation 2.1 induces a **zero-forcing** behavior (Murphy, 2013), meaning that all the zeros in $p(X)$ must be zero in its I-projection. Contrarily, the M-projection $p^M_{\mathcal{C}}(X)$ induces a **zero-avoiding** behavior by inducing a distribution that assigns non-zero mass to every point x that has positive probability according to $p(X)$, which results in an overestimation the support of $p(X)$.

The aforementioned characteristics of the two projections are illustrated by Figure 2.1 that reports the I and M-projections of a bi-modal distribution $p(X)$ on the set $\mathcal{C}_{\mathcal{N}}$ of Normal distributions. The visualization of the parameters configuration (on

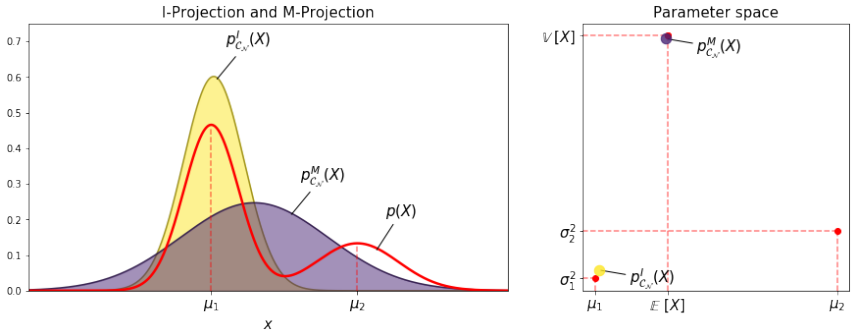

FIGURE 2.1: Visualization of the characteristics of the information (in yellow) and moment projection (in blue) of a mixture model (in red), onto the set of Normal distributions. The characteristics of the parameters' configurations are visualized on the right side of the picture.

the right side of the plot) underlines how the I-projection $p^I_{\mathcal{C}_{\mathcal{N}}}(X)$ models the characteristics of one of the two modes, while the M-projection matches both the mean and the variance of $p(X)$ ($\mathbb{E}[X]$ and $\mathbb{V}[X]$ respectively in Figure 2.1). This result underlines that the I-projection tends to model local characteristics of the projected distribution, while the M-projection focuses on a more global aspect (the first and second moment in the reported example). The same minimization procedure described in equation 2.1 and 2.2 can be applied for other divergences between probability distributions, which may lead to intermediate solutions that associate both the zero-avoiding and zero-forcing characteristics. A brief discussion regarding how the properties of the information and moment projections can be combined is reported in Appendix B.

2.1.1 Parametric models

The problem of finding a probability distribution in a given set $\mathcal{C}$ that achieves the minimal divergence is generally difficult. Nevertheless, one may look at the problem from a parametric perspective. By fixing a parameter space Θ and a functional form $q_\theta(X)$, one may cast the search of a probability distribution in $\mathcal{C}$ into the problem of identifying a parameters configuration in Θ. Defining $\mathcal{C}_\Theta$ as the set of all the probability distributions that are representable by some $\theta \in \Theta$, we have:

$$\hat{\theta}_I := \operatorname*{argmin}_{\theta\in\Theta} KL(q_\theta(X)||p(X)) \iff q_{\hat{\theta}_I}(X) = \operatorname*{argmin}_{q(X)\in\mathcal{C}_\Theta} KL(q(X)||p(X)) \tag{2.3}$$

$$\hat{\theta}_M := \operatorname*{argmin}_{\theta\in\Theta} KL(p(X)||q_\theta(X)) \iff q_{\hat{\theta}_M}(X) = \operatorname*{argmin}_{q(X)\in\mathcal{C}_\Theta} KL(p(X)||q(X)) \tag{2.4}$$

Thanks to the equivalences reported in equation 2.3 and 2.4, the I and M-projections can be computed directly using an optimization procedure on the parameter space. In particular, the Stochastic Gradient Descent procedure consent to approximately find the values for the optimal parameters $\hat{\theta}_I$ and $\hat{\theta}_M$ whenever the expression of the KL-divergence is differentiable with respect to the parameters. This generative technique is closely related to other approaches in the literature as shown in the next two sections.

M-projection and maximum log-likelihood

Given a dataset $\mathcal{D}$ of independent identically distributed samples from an unknown distribution $r(X)$, the empirical distribution $\tilde{r}_{\mathcal{D}}(X)$ is defined by equally splitting the probability mass among the observations $x \in \mathcal{D}$. One may easily show that the computation of the M-projection of $\tilde{r}_{\mathcal{D}}(X)$ on a set $\mathcal{C}_\Theta$ represented through a parameter space Θ corresponds to the maximum likelihood estimation for the parameter $\theta \in \Theta$:

$$\begin{aligned}\hat{\theta}_M &:= \underset{\theta\in\Theta}{\operatorname{argmin}}\, KL(\tilde{r}_{\mathcal{D}}(X)||q_\theta(X)) \\ &= \underset{\theta\in\Theta}{\operatorname{argmin}}\, CE(\tilde{r}_{\mathcal{D}}(X)||q_\theta(X)) \\ &= \underset{\theta\in\Theta}{\operatorname{argmax}} \sum_{x\in\mathcal{D}} \log q_\theta(X = x)\end{aligned} \tag{2.5}$$

Where $CE(\tilde{r}_{\mathcal{D}}(X)||q_\theta(X))$ in the second line of equation 2.5 represents the cross-entropy between $\tilde{r}_{\mathcal{D}}(X)$ and $q_\theta(X)$. If the parametrization is flexible enough, the optimization procedure will result in a distribution which almost perfectly matches the empirical distribution $\tilde{r}_{\mathcal{D}}(X)$. This is generally an unwanted feature since the model tends to represent the dataset rather then the underlying data-generating distribution $r(X)$.

One can avoid perfectly matching the empirical distribution by constraining the set $\mathcal{C}_\Theta$ to some $\mathcal{C}'_\Theta$ that does not contain $\tilde{r}_{\mathcal{D}}(X)$. In this case, the restriction on the parameter space might be interpreted as a regularization term. This interpretation is graphically visualized in Figure 2.2 where $q_{\hat{\theta'}_M}(X)$ represents the M-projection of the empirical distribution onto $\mathcal{C}'_\Theta$, while the projection of $\tilde{r}_{\mathcal{D}}(X)$ onto $\mathcal{C}_\Theta$ coincides with itself since $\tilde{r}_{\mathcal{D}}(X) \in \mathcal{C}_\Theta$.

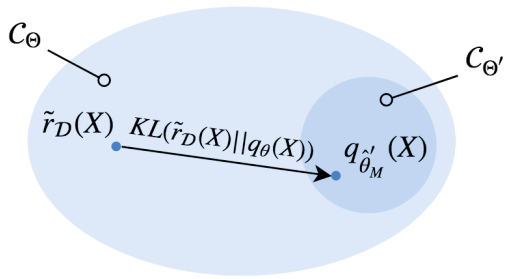

FIGURE 2.2: Visualization of the maximum likelihood interpretation of the Moment projection for two distinct parameterizations defined through the parameters spaces $\Theta \supseteq \Theta'$. In the reported example, $\hat{\theta}'_M$ represents parameter configuration in Θ' that achieves the maximum likelihood for the empirical distribution $\tilde{r}_{\mathcal{D}}(X)$.

I-projection and density ratio estimation

Fixing a parameter space Θ and defining $\mathcal{C}_\Theta$ as the set of probability distribution which can be modeled by some $\theta \in \Theta$, the parameter configuration corresponding to the I-projection of the empirical distribution is defined as:

$$\hat{\theta}_I = \underset{\theta\in\Theta}{\operatorname{argmin}}\, KL(q_\theta(X)||\tilde{r}_{\mathcal{D}}(X)) \tag{2.6}$$

One may notice a problem in equation 2.6 since the value of the reported KL-divergence is finite only when the support of $q_{\hat{\theta}_I}(X)$ is completely contained in the support of the empirical distribution. This characteristic is generally unwanted since the dataset $\mathcal{D}$ is usually under-representing the support of the real data-generating distribution $r(X)$. For this reason, instead of projecting $\tilde{r}_{\mathcal{D}}(X)$ directly, one may consider a smoother approximation with a more extensive support. This is obtained by approximately modeling the ratio between $\tilde{r}_{\mathcal{D}}(X)$ and $q_\theta(X)$ instead.

Among the several techniques to approximate the ratio between two probability distributions proposed in the literature (Sugiyama, Suzuki, and Kanamori, 2010; Sugiyama, Suzuki, and Kanamori, 2012; Mohamed and Lakshminarayanan, 2016) we will focus on the approach that makes use of an approximate binary classifier $\tilde{s}(Y|X = x)$ to estimate the probability that the samples x are drawn according to $q_\theta(X)$ $(Y = 0)$ rather then $\tilde{r}_{\mathcal{D}}(X)$ $(Y = 1)$:

$$\begin{aligned}\hat{\theta}_I &= \underset{\theta\in\Theta}{\operatorname{argmin}}\, \mathbb{E}_{x\sim q_\theta(X)}\left[\log \frac{q_\theta(X)}{\tilde{r}_{\mathcal{D}}(X)}\right] \\ &\approx \underset{\theta\in\Theta}{\operatorname{argmin}}\, \mathbb{E}_{x\sim q_\theta(X)}\left[\log \frac{1-\tilde{s}(Y=0|X=x)}{\tilde{s}(Y=0|X=x)}\right]\end{aligned} \tag{2.7}$$

This approximation, known the **density-ratio trick** (Rosca et al., 2017), is closely related to the adversarial training procedure that is used for Generative Adversarial Networks (Goodfellow et al., 2014a) and has widely been used in literature as a "likelihood free" alternative for Variational Inference (Karaletsos, 2016; Huszár, 2017; Mescheder, Nowozin, and Geiger, 2017; Tran, Ranganath, and Blei, 2017). Further details concerning the computation of this approximation can be found in Appendix C.

2.2 Modeling joint distributions

In the previous section we described two approaches to model a distribution $r(X)$ starting from a set of empirical observations by using the methodologies of information and moment projection. Even if the proposed approaches are effectively defining some strategies to design a generative model, they lack of an explicit latent representation. This section aims to extend the method to model distributions jointly defined on the domain $\mathcal{X} \times \mathcal{Z}$ determined by data and latent representations.

Given a distribution $r(Z)$ defined over a code domain $\mathcal{Z}$ and an empirical approximation $\tilde{r}_{\mathcal{D}}(X)$ for a data-generating distribution $r(X)$, we define the set $\mathcal{C}(r(X), r(Z))$ of distributions jointly defined over $\mathcal{X} \times \mathcal{Z}$ that have $r(X)$ and $r(Z)$ as the two marginals[1]:

$$\mathcal{C}(r(X), r(Z)) := \{p(X, Z) \in S(\mathcal{X} \times \mathcal{Z}) \mid p(X) = r(X) \wedge p(Z) = r(Z)\} \tag{2.8}$$

Representing the set of probability distributions that have both marginals $r(X)$ and $r(Z)$ fixed is hard since it requires to impose a constraint on quantities obtained

[1]The view proposed in this section is consistent to the one presented in Zhao, Song, and Ermon, 2017b even if the two works have been originally developed independently.

through the marginalization of $r(X, Z)$:

$$r(X,Z) \in \mathcal{C}(r(X), r(Z)) \iff \begin{cases} \sum_{x \in \mathcal{X}} r(X = x, Z) = r(Z) \\ \sum_{z \in \mathcal{Z}} r(X, Z = z) = r(X) \end{cases} \tag{2.9}$$

For this reason, in order to identify the distribution with the given marginals, we may express that $\mathcal{C}(r(X), r(Z))$ as the intersection between the sets $\mathcal{C}(r(X))$ and $\mathcal{C}(r(Z))$ of the distributions with marginal $r(X)$ and $r(Z)$ respectively:

$$\mathcal{C}(r(X)) := \{q(X,Z) \in S(\mathcal{X} \times \mathcal{Z}) \mid q(X) = r(X)\} \tag{2.10}$$
$$\mathcal{C}(r(Z)) := \{p(X,Z) \in S(\mathcal{X} \times \mathcal{Z}) \mid p(Z) = r(Z)\} \tag{2.11}$$

Other than allowing for an alternative description of $\mathcal{C}(r(X), r(Z))$, the definition of the two sets facilitates an easier representation, since the elements of $\mathcal{C}(r(X))$ and $\mathcal{C}(r(Z))$ can be expressed as a product of the fixed marginal and a conditional distribution:

$$p(X,Z) = p(X|Z)r(Z) \implies \sum_{x \in \mathcal{X}} p(X,Z) = r(Z) \implies p(X,Z) \in \mathcal{C}(r(Z)) \tag{2.12}$$
$$q(X,Z) = r(X)q(Z|X) \implies \sum_{z \in \mathcal{Z}} q(X,Z) = r(X) \implies q(X,Z) \in \mathcal{C}(r(X)) \tag{2.13}$$

Note that since $r(X)$ and $r(Z)$ are fixed, the problem of modeling joint distributions over $\mathcal{X} \times \mathcal{Z}$ is cast into the problem of modeling the two conditionals $p(X|Z)$ and $q(Z|X)$.

In order to identify the intersection between the two sets $\mathcal{C}(r(X))$ and $\mathcal{C}(r(Z))$, we may exploit the properties that the value of any divergence is zero if and only if the two distributions are identical. Fixing a divergence $D(\cdot||\cdot)$, the set $\mathcal{C}(r(X), r(Z))$ can be represented as the subset of $\mathcal{C}(r(Z))$ for which there exists a corresponding distribution in $\mathcal{C}(r(X))$ that achieves zero divergence:

$$\mathcal{C}(r(X), r(Z)) = \{p(X,Z) \in \mathcal{C}(r(Z)) \mid \exists q(X,Z) \in \mathcal{C}(r(X)), D(p(X,Z)||q(X,Z)) = 0\} \tag{2.14}$$

The expression for the set $\mathcal{C}(r(X), r(Z))$ in equation 2.14 is equivalent to the original definition (equation 2.8). However, the new formulation eases the identification of the elements of $\mathcal{C}(r(X), r(Z))$. This is because the problem of finding the elements of $\mathcal{C}(r(X), r(Z))$ can be expressed as a joint minimization over $\mathcal{C}(r(X))$ and $\mathcal{C}(r(Z))$:

$$(\hat{p}(X,Z), \hat{q}(Z,X)) := \underset{p(X,Z) \in \mathcal{C}(r(Z)),\, q(X,Z) \in \mathcal{C}(r(X))}{\operatorname{argmin}} D(p(X,Z)||q(X,Z))$$
$$\hat{p}(X,Z) = \hat{q}(Z,X) \in \mathcal{C}(r(X), r(Z)) \tag{2.15}$$

2.2.1 A parametric view

From a practical perspective, searching in the space of all the conditional distributions is not practical, therefore, we will restrict the sets $\mathcal{C}(r(X))$ and $\mathcal{C}(r(Z))$ to the set of probability distributions that can be represented parametrically. Furthermore, since we are not given access to the data-generating distribution $r(X)$ we will consider its empirical approximation $\tilde{r}_{\mathcal{D}}(X)$ instead.

Fixing the parameter domains Θ and Φ and a functional form for the two conditional distributions, we can define a parametric representation for the two joint distributions:

$$p_\phi(X,Z) = p_\phi(X|Z)r(Z) \qquad \phi \in \Phi \tag{2.16}$$
$$q_\theta(X,Z) = \tilde{r}_\mathcal{D}(X)q_\theta(Z|X) \qquad \theta \in \Theta \tag{2.17}$$

The parametrization of two conditional distributions implicitly restricts the two original sets $\mathcal{C}(\tilde{r}_\mathcal{D}(X))$ and $\mathcal{C}(r(Z))$. This results in the definition of two sets $\mathcal{C}_\Theta(\tilde{r}_\mathcal{D}(X))$ and $\mathcal{C}_\Phi(r(Z))$ that depend on the specific parametric representation for $p_\phi(X|Z)$ and $q_\theta(Z|X)$:

$$p_\phi(X,Z) \in \mathcal{C}_\Phi(r(Z)) \subseteq \mathcal{C}(r(Z)) \tag{2.18}$$
$$q_\theta(X,Z) \in \mathcal{C}_\Theta(\tilde{r}_\mathcal{D}(X)) \subseteq \mathcal{C}(\tilde{r}_\mathcal{D}(X)) \tag{2.19}$$

As the size of the two sets of parametric distributions is determined by the flexibility of the two conditionals, there are no guarantees that the intersection between $\mathcal{C}_\Phi(r(Z))$ and $\mathcal{C}_\Theta(\tilde{r}_\mathcal{D}(X))$ is non-empty. In such in which $\mathcal{C}_\Phi(r(Z))$ and $\mathcal{C}_\Theta(\tilde{r}_\mathcal{D}(X))$ are disjoint, the minimization of any divergence $D(q_\theta(X,Z)||p_\phi(X,Z))$ determines two distinct joint distributions, which depend on the choice of the divergence and the definition two sets as shown in Figure 2.3.

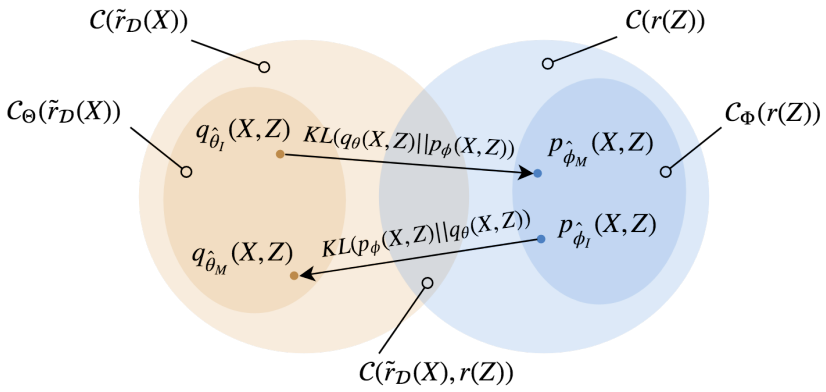

FIGURE 2.3: Graphical visualization of the sets $\mathcal{C}(\tilde{r}_\mathcal{D}(X))$ and $\mathcal{C}(r(Z))$ of distributions with marginals $\tilde{r}_\mathcal{D}(X)$ and $r(Z)$ respectively. Since their parametric restriction $\mathcal{C}_\Theta(\tilde{r}_\mathcal{D}(X))$ and $\mathcal{C}_\Phi(r(Z))$ do not intersect, the probability distributions $q_\theta(X,Z)$ and $p_\phi(X,Z)$, identified through the minimization of a joint divergence, do not coincide. Note that in this scenario the choice of different divergences influences the parameter configuration of the minimum.

Among all the possible choices for $D(q_\theta(X,Z)||p_\phi(X,Z))$, the Kullback-Leibler divergence presents interesting properties. In fact, by selecting $KL(q_\theta(X,Z)||p_\phi(X,Z))$ as the measure of divergence, one may observe that the minimizers represent the information and moment projections of each-other. Defining $\hat{\theta}_I \in \Theta$ and $\hat{\phi}_M \in \Phi$ as

the parameter configuration that achieve the minimal KL-divergence we have:

$$\hat{\theta}_I = \underset{\theta \in \Theta}{\operatorname{argmin}}\, KL(q_\theta(X,Z)||p_{\hat{\phi}_M}(X,Z)) \tag{2.20}$$

$$\hat{\phi}_M = \underset{\phi \in \Phi}{\operatorname{argmin}}\, KL(q_{\hat{\theta}_I}(X,Z)||p_\phi(X,Z)) \tag{2.21}$$

Therefore, we can expect $p_{\hat{\theta}_I}(X,Z)$ to over-estimate the support of $q_{\hat{\phi}_M}(X,Z)$ due to the induced zero-avoiding behavior. Contrarily, the minimizers $\hat{\theta}_M$ and $\hat{\phi}_I$ of the opposite direction of the Kullback-Leibler divergence $KL(p_\phi(X,Z)||q_\theta(X,Z))$ lead to a configuration in which $p_{\hat{\phi}_I}(X,Z)$ represents the I-projection of $q_{\hat{\theta}_M}(X,Z)$ instead, promoting the zero-forcing characteristic:

$$\hat{\theta}_M = \underset{\theta \in \Theta}{\operatorname{argmin}}\, KL(p_{\hat{\phi}_I}(X,Z)||q_\theta(X,Z)) \tag{2.22}$$

$$\hat{\phi}_I = \underset{\phi \in \Phi}{\operatorname{argmin}}\, KL(p_\phi(X,Z)||q_{\hat{\theta}_M}(X,Z)) \tag{2.23}$$

In the presented scenarios, the divergence minimization procedure allows to identify two joint distributions, $q_\theta(X,Z)$ and $p_\phi(X,Z)$, which model the data distribution together with a latent representation Z. Unfortunately, as shown in the next chapter, the generative objective is not sufficient for representation learning. For this reason, Chapter 3 will focus on the information theoretical interpretation of the two distributions $p_\phi(X,Z)$ and $q_\theta(X,Z)$, targeting their role in terms of representation learning. Further discussion regarding the modeling choices for the two conditional distributions $q_\theta(Z|X)$ and $p_\phi(X|Z)$ is reported in Appendix D.

Chapter 3

Information Theoretical Quantities for Representation Learning

Chapter 2 discussed how to model joint distributions over $\mathcal{X} \times \mathcal{Z}$, but the inclusion of a latent variable Z is not a sufficient condition to obtain a meaningful data representation. In fact, considering the empirical distribution $\tilde{r}_{\mathcal{D}}(X)$ and the code distribution $r(Z)$, one can define a joint distribution $r_{ind}(X, Z)$ by considering their product:

$$r_{ind}(X, Z) := \tilde{r}_{\mathcal{D}}(X)r(Z) \tag{3.1}$$

According to $r_{ind}(X, Z)$, the latent representation Z and the data X are completely independent, nevertheless it satisfies the requirements of having $\tilde{r}_{\mathcal{D}}(X)$ and $r(Z)$ as the two marginals. In order to better understand what are the characteristic of a meaningful latent representation, we will take into account the effect of the mutual information between the data X and the code Z, which intuitively expresses the amount of information regarding the observation that is preserved in the latent representation. For this reason, we will study the characteristics of both encoding $q_\theta(X, Z)$ and decoding $p_\phi(X, Z)$ joint distributions from an information theoretical perspective with a main focus on the role of the mutual information induced by the two distributions, namely $I(q_\theta(X, Z))$ and $I(p_\phi(X, Z))$. The analysis proposed in this chapter aims to determine a strategy to compute the two mutual information by exploiting their relation with other measurable quantities. This is done by furnishing an intuitive explanation of the interaction between $q_\theta(X, Z)$ and $p_\phi(X, Z)$ inspired from a coding perspective. The two main sections address the computation of **encoding mutual information**, $I(q_\theta(X, Z))$, and **decoding mutual information**, $I(p_\phi(X, Z))$, respectively.

3.1 Encoding mutual information

The encoding mutual information represents a quantity of interest when designing a latent representation since it denotes the amount of information that is preserved from the data to the latent codes by the encoder $q_\theta(Z|X)$. The mutual information is notoriously hard to compute, for this reason, some information theoretical properties will be considered to define an upper and a lower bound for $I(q_\theta(X, Z))$, and identify a possible rough approximation. This is done by first considering the measures of data distortion and code rate that arise by interpreting $q_\theta(Z|X)$ and $p_\phi(X|Z)$ from a data-encoding perspective (Alemi et al., 2018).

The first quantity that one can take into account is the entropy of the empirical distribution $H(q(X))$, which expresses the amount of information that is needed to describe the observations. Since the marginal $q(X)$ is fixed and defined by the

empirical distribution, the entropy $H(q(X))$ is constant and upper bounded by the log-size of the dataset (property A.2):

$$H(q(X)) = H(\tilde{r}_{\mathcal{D}}(X)) \leq \log |\mathcal{D}| \tag{3.2}$$

As a consequence of property A.6, the mutual information $I(q_\theta(X, Z))$ is also upper bounded by the same quantity. Since this value is fixed, the different parameter configurations $\theta \in \Theta$ of the encoding distribution can trade-off between mutual information $I(q_\theta(X, Z))$ and conditional entropy $H(q_\theta(X|Z))$:

$$H(q(X)) = I(q_\theta(X, Z)) + H(q_\theta(X|Z)) \tag{3.3}$$

Unfortunately, both terms in equation 3.3 cannot be evaluated directly. However, the expression for $H(q_\theta(X|Z))$ can be re-written as the difference between the cross-entropy and the KL-divergence from $q_\theta(X|Z)$ to $p_\phi(X|Z)$:

$$H(q_\theta(X|Z)) = CE(q_\theta(X, Z)||p_\phi(X|Z)) - KL(q_\theta(X|Z)||p_\phi(X|Z)) \tag{3.4}$$

The divergence reported in equation 3.4 is intractable since it requires the evaluation of the density of $q_\theta(X|Z)$. Nevertheless, the cross-entropy term can be easily computed, providing the foundations for the estimation of the encoding mutual information.

3.1.1 Data distortion

The cross-entropy introduced in equation 3.4 can be interpreted as the expected negative log-likelihood of the data reconstructions. Its value expresses the discrepancies that are measured when the data is encoded according to $q_\theta(Z|X)$ and decoded using $p_\phi(X|Z)$. For this reason, it will be referred to as the **data distortion** (Alemi et al., 2018) $D_X(\theta, \phi)$:

$$D_X(\theta, \phi) = \mathbb{E}_{x\sim q(X)}\mathbb{E}_{z\sim q_\theta(Z|X)}\left[-\log p_\phi(X = x|Z = z)\right] \tag{3.5}$$

By re-arranging the terms of equation 3.4, we observe that the distortion induced by the use of the encoding distribution $q_\theta(Z|X)$ and the decoding distribution $p_\phi(X|Z)$ is composed by two terms:

$$D_X(\theta, \phi) = H(q_\theta(X|Z)) + KL(q_\theta(X|Z)||p_\phi(X|Z)) \tag{3.6}$$

The first term, $H(q_\theta(X|Z))$, represents the amount of information regarding the observations that is lost through the encoding procedure, while the second component, $KL(q_\theta(X|Z)||p_\phi(X|Z))$, reports the amount of information that is wasted due to the discrepancies between $p_\phi(X|Z)$ and $q_\theta(X|Z)$. Expressing the conditional entropy as the difference between the entropy of the data and the mutual information (property A.8), one can write $D_X(\theta, \phi)$ as a function of $I(q_\theta(X, Z))$:

$$D_X(\theta, \phi) = H(\tilde{r}_{\mathcal{D}}(X)) - I(q_\theta(X, Z)) + KL(q_\theta(X|Z)||p_\phi(X|Z)) \tag{3.7}$$

The three terms in equation 3.7 indicate that the data distortion can be decreased either by improving the matching between the conditional distributions $q_\theta(X|Z)$ and $p_\phi(X|Z)$ or by increasing the amount of information that is captured in the latent representation through $q_\theta(Z|X)$. Rearranging the terms in equation 3.7 and considering that the KL-divergence is always positive, one may identify a lower bound for

the encoding mutual information (Barber and Agakov, 2003):

$$\begin{aligned} I(q_\theta(X,Z)) &= H(\tilde{r}_{\mathcal{D}}(X)) - D_X(\theta,\phi) + KL(q_\theta(X|Z)||p_\phi(X|Z)) \\ &\geq H(\tilde{r}_{\mathcal{D}}(X)) - D_X(\theta,\phi) \end{aligned} \tag{3.8}$$

3.1.2 Code rate

The mutual information induced by the encoding distribution, $I(q_\theta(X,Z))$, can be also expressed as a difference between two KL-divergences (property A.23):

$$I(q_\theta(X,Z)) = \underbrace{KL(q_\theta(Z|X)||p(Z))}_{R_Z(\theta)} - KL(q_\theta(Z)||p(Z)) \tag{3.9}$$

The first term in equation 3.9 will be referred to as the **code rate** $R_Z(\theta)$ (Alemi et al., 2018). The code rate represents the average amount of extra information that is needed to represents samples from $q_\theta(Z|X)$ when a coding scheme optimized for the prior, $p(Z)$, is used instead of the optimal one. The second term in equation 3.9, on the other hand, represents the discrepancy between $q_\theta(Z)$ and $p(Z)$ expressed as the amount of extra information that is required to encode samples from $q_\theta(Z)$ using a coding scheme designed for $p(Z)$.

Contrarily to the encoding mutual information, the code rate $R_Z(\theta)$ can be computed directly and, since the $KL(q_\theta(Z)||p(Z))$ is always positive, it represents an upper bound for the encoding mutual information:

$$I(q_\theta(X,Z)) \leq R_Z(\theta) \tag{3.10}$$

3.1.3 Computation

The additive relations between the terms described in this section can be summarized by the graphical representation in Figure 3.1. One may notice that when only $D_X(\theta,\phi)$, $R_Z(\theta)$ and $H(q(X))$ are given, it is not possible to compute $I(q_\theta(X,Z))$. However, by incorporating an estimation of $KL(q_\theta(Z)||p(Z))$ obtained using the density-ratio trick (Appendix C.1), all the terms visualized in Figure 3.1 can be roughly determined. As a result, it is possible approximate not only the mutual information, but also the joint divergence $KL(q_\theta(X,Z)||p_\phi(X,Z))$, and its two components, $KL(q_\theta(Z)||p(Z))$ and $KL(q_\theta(X|Z)||p_\phi(X|Z))$. The measures of $KL(q_\theta(Z)||p(Z))$ and $KL(q_\theta(X|Z)||p_\phi(X|Z))$ are particularly insightful since they respectively express how much the two code distributions, $q_\theta(Z)$ and $p(Z)$, and the conditional distributions, $q_\theta(X|Z)$ and $p_\phi(X|Z)$, differ from each other. A more detailed description of the computation and refinement of the quantities enumerated in Figure 3.1 is reported in appendix E.

3.2 Decoding mutual information

Similarly to the analysis presented for the encoding mutual information, it is possible to interpret and describe the quantities induced by the parametric distribution $p_\phi(X,Z)$. In particular, the decoding mutual information, $I(p_\phi(X,Z))$, represents a quantity of interest since it denotes the expected amount of information regarding the codes that is preserved through the decoding procedure determined by $p_\phi(X|Z)$. Analogously to the analysis proposed for the encoding mutual information, we determine the bounds for its decoding counterpart by considering a coding scheme

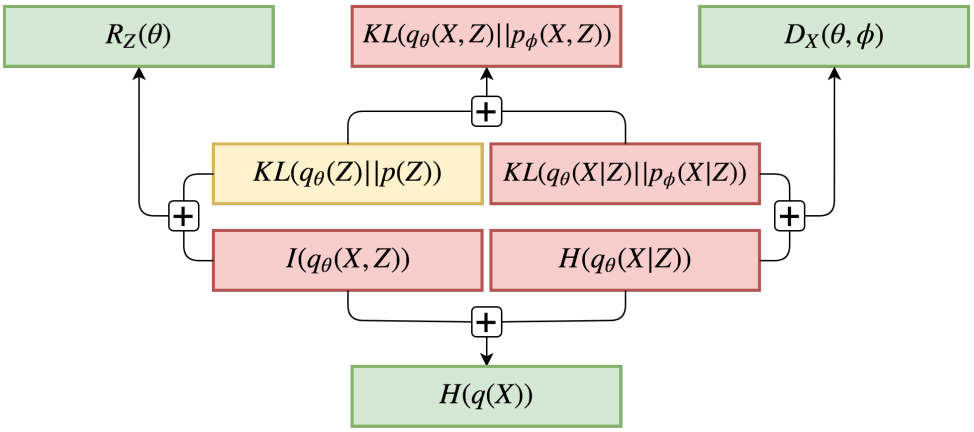

FIGURE 3.1: Graphical representation of the relationships between the information theoretical quantities defined by the encoding distribution, $q_\theta(X,Z)$, and its interaction with the decoding distribution, $p_\phi(X,Z)$. The coloring used for the different terms represents their tractability. A red color surrounds the terms that cannot be approximated directly, while the green box denotes the components that can be easily computed. The KL-divergence $KL(q_\theta(Z)||p(Z))$ is colored in yellow since we have access to a rough estimation obtained by using the density-ratio trick.

that is defined symmetrically with respect to the one reported in the previous section. In this "inverse" coding scheme, the role of the codes Z and the data X is swapped. In fact, the reconstruction error is measured by encoding the codes using $p_\phi(X|Z)$ and decoding them according to $q_\theta(Z|X)$. This novel formulation allows to interpret different information theoretical quantities that define a lower bound for $I(p_\phi(X,Z))$.

Starting from the entropy of the codes $H(p(Z))$, one may notice that its value is fixed and determined by the given distribution $r(Z)$ (equation 2.16). This implies that the sum of mutual information $I(p_\phi(X,Z))$ and conditional entropy $H(p_\phi(Z|X))$ is constant for every parameter configuration $\phi \in \Phi$. The value of $H(p_\phi(Z|X))$ cannot be computed directly, however, it is possible to write $H(p_\phi(Z|X))$ as the difference of the cross-entropy $CE(p_\phi(X,Z)||q_\theta(Z|X))$ and the KL-divergence between $p_\phi(Z|X)$ and $q_\theta(Z|X)$:

$$H(p_\phi(Z|X)) = CE(p_\phi(X,Z)||q_\theta(Z|X)) - KL(p_\phi(Z|X)||q_\theta(Z|X)) \tag{3.11}$$

In accordance with the analysis presented in the previous section, the only term of equation 3.11 that can be computed directly is the cross-entropy $CE(p_\phi(X,Z)||q_\theta(Z|X))$. Its value is fundamental to determine a lower bound for $I(p_\phi(X,Z))$ that is presented in the next subsection.

3.2.1 Code distortion

Because of the symmetry with the definition of data distortion (equation 3.5), we will refer to the cross-entropy $CE(p_\phi(X,Z)||q_\theta(Z|X))$ as the **code distortion** $D_Z(\theta,\phi)$, which represents the expected log-likelihood of the codes once they are decoded using $p_\phi(X|Z)$ and encoded according to $q_\theta(Z|X)$.

By re-arranging the terms in equation 3.11 and writing the $H(p_\phi(Z|X))$ as a function

of the mutual information, we can observe that the distortion of the codes $D_Z(\theta, \phi)$ can be decreased either by matching the two conditional distributions ,$p_\phi(Z|X)$ and $q_\theta(Z|X)$, or by increasing the amount of information regarding the codes that is preserved through the decoding procedure:

$$D_Z(\theta, \phi) = H(p(Z)) - I(p_\phi(X, Z)) + KL(p_\phi(Z|X)||q_\theta(Z|X)) \tag{3.12}$$

The terms of equation 3.12 directly determine a lower bound for the decoding mutual information:

$$I(p_\phi(X, Z)) \geq H(p(Z)) - D_Z(\theta, \phi) \tag{3.13}$$

Note that the bound reported in equation 3.13 has also been considered in recent literature (Chen et al., 2016; Phuong et al., 2018), however, the derivation proposed in this work provides additional interpretability for the terms involved by the inequality.

3.2.2 Data rate

The mutual information determined by the decoding distribution, $I(p_\phi(X, Z))$, can be expressed, analogously to the counterpart defined through the encoding distribution, as the difference between KL-divergences:

$$I(p_\phi(X, Z)) = \underbrace{KL(p_\phi(X|Z)||q(X))}_{R_X(\phi)} - KL(p_\phi(X)||q(X)) \tag{3.14}$$

The KL-divergence from $p_\phi(X|Z)$ to $q(X)$ can be interpreted as the **data rate** $R_X(\phi)$, symmetrically to the code rate counterpart. In this scenario, $R_X(\phi)$ denotes an upper bound for the mutual information determined by $p_\phi(X, Z)$.
The computation of $R_X(\phi)$ represents the main difference with the interpretation proposed in the previous section. In fact, since the empirical distribution assigns zero probability to every un-observed $x \in \mathcal{X}$, $KL(p_\phi(X|Z)||q(X))$ is generally not finite. As a consequence, the tightest upper bound for $I(p_\phi(X, Z))$ that one may consider is given by the entropy of the prior distribution:

$$I(p_\phi(X, Z)) \leq H(p(Z)) \tag{3.15}$$

Since $H(p(Z))$ is fixed, the inequality reported in equation 3.15 is generally not practically useful for the computation of the decoding mutual information.

3.2.3 Computation

Figure 3.2 reports the relation between the quantities described in this section. Since $R_X(\phi)$ cannot be computed, the approximation of $KL(p_\phi(X)||q(X))$ obtained by applying the density-ratio trick does not suffice to determine an estimation for the decoding mutual information. Nevertheless, one may exploit the value of the code distortion to obtain a lower bound for $I(p_\phi(X, Z))$, as reported in equation 3.13. The lack of an estimation for the data distortion also prevents the computation of the joint KL-divergence $KL(p_\phi(X, Z)||q_\theta(X, Z))$ and the conditional divergence $KL(p_\phi(Z|X)||q_\theta(Z|X))$, limiting the understanding of the interaction of the quantities mentioned in Figure 3.2. Nevertheless, the additive relations visualized in both Figure 3.1 and 3.2 represent a useful tool to interpret and rephrase several learning objectives proposed in the literature of unsupervised representation learning, as suggested in the next chapter.

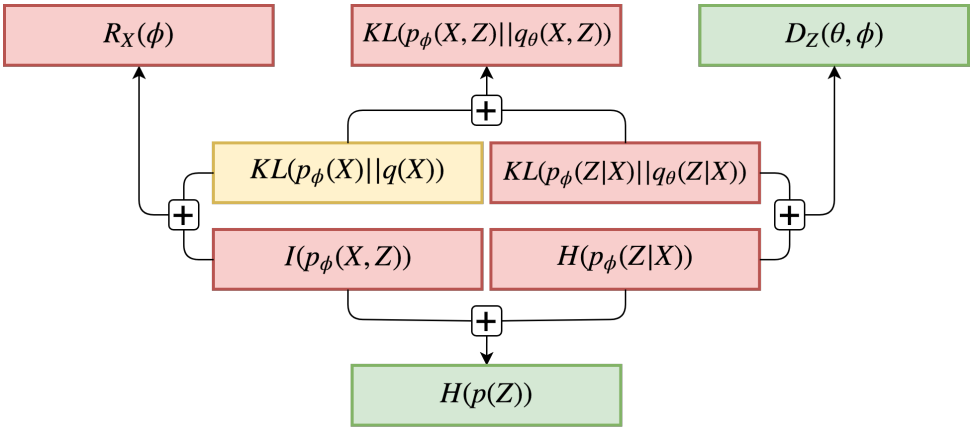

FIGURE 3.2: Graphical representation of the relationships between the information theoretical quantities defined by the decoding distribution, $p_\phi(X,Z)$, and its interaction with the encoding distribution, $q_\theta(X,Z)$. Analogously to the representation reported in Figure 3.1, the color scheme is used to represents the tractability of different quantities.

Chapter 4

Learning Objectives for Unsupervised Representation Learning

In Chapter 2 a methodology for modeling joint distributions over the data and the latent codes was introduced. In addition, Chapter 3 has shown that in order to obtain a meaningful data representation the mutual information between the data and the codes has to be taken into account. Considering the developed constraints, this chapter first introduces a novel formulation of the learning objective for unsupervised representation learning.
In addition, the second part of this chapter shows how the loss function of different popular models for unsupervised representation learning can be considered as a specific instance of the proposed generalized objective. The different models that are considered in this analysis are:

- Variational Auto-Encoder (Section 4.2.1)
- β Variational Auto-Encoder (Section 4.2.2)
- Information Variational Auto-Encoder (Section 4.2.3)
- Information Generative Adversarial Network (Section 4.2.4)
- Cycle-Consistent Adversarial Networks (Section 4.2.5)

To conclude this chapter, a summary of the proposed re-formulations applied to the discussed loss functions is provided.

4.1 A generalized learning objective

The minimization of a joint divergence is a necessary condition to obtain a meaningful generative model. We stressed that this minimization i) allows to express a joint distribution given the marginals over both the data and the codes and ii) forces the encoding and decoding distributions to be consistent. However, conditions i) and ii) have shown not to be sufficient for obtaining a meaningful data representation. For this reason, the learning objective proposed in this work combines i) and ii) with a constraint on encoding and decoding mutual information to control the amount of information that is preserved in the latent representation.

The general loss function $\mathcal{L}$ for unsupervised representation learning proposed in

this work is defined as the sum of a generative (in blue) and a representation component (in red):

$$\mathcal{L}(\theta, \phi; D, \alpha_p, \alpha_q) := \underbrace{D(q_\theta(X,Z)||p_\phi(X,Z))}_{\text{Generative Component}} + \underbrace{\alpha_p I(p_\phi(X,Z)) + \alpha_q I(q_\theta(X,Z))}_{\text{Representation Component}} \tag{4.1}$$

The first part, $D(q_\theta(X,Z)||p_\phi(X,Z))$, denotes a joint divergence between $q_\theta(X,Z)$ and $p_\phi(X,Z)$, which can be interpreted as the generative component of the loss; while the last two terms, $\alpha_p I(p_\phi(X,Z))$ and $\alpha_q I(q_\theta(X,Z))$, enforce the informativeness of the data representation. The coefficients α_p and α_q in equation 4.1 represent the weights for the mutual information induced by respectively the decoding and encoding joint distributions. Positive values of α_p and α_q penalize the mutual information while negative values will result in its maximization.
Since both $q_\theta(Z|X)$ and $p_\phi(X|Z)$ are modeled parametrically, we have no guarantee that there exists a parameter configuration that allows $p_\phi(X,Z)$ and $q_\theta(X,Z)$ to perfectly match. As a consequence, the minimum of equation 4.1 depends on the choice of the joint divergence and the two coefficients, α_p and α_q.

Note that the proposed objective, $\mathcal{L}(\theta, \phi; D, \alpha_p, \alpha_q)$, is generally intractable due to the fact that the three individual terms can not be computed directly. Nevertheless, by appropriately choosing the divergence $D(q_\theta(X,Z)||p_\phi(X,Z))$, the loss may be expressed by a weighted sum of tractable terms, i.e. the data distortion, code distortion, code-rate and other divergences between the marginal distributions which can be more easily approximated.

4.2 Analysis of the different training objectives

This section reports an analysis of the training objective of some of the most successful models in the context of unsupervised representation learning. Each subsection presents a short introduction of an architecture, its original loss definition accompanied by its alternative expression in the form proposed by equation 4.1. This reformulation provides insights regarding the effect of each model's hyper-parameters on its generative and representation performances.

4.2.1 Variational Autoencoder

The Variational Autoencoder (Kingma and Welling, 2013) was originally designed as an efficient method to perform approximate posterior inference by re-parameterizing the lower bound of the data log-likelihood. This architecture has successfully been applied as a generative model for text (Bowman et al., 2015), images and captions (Pu et al., 2016; Gao et al., 2017). It has also demonstrated to work well when applied semi-supervised learning tasks (Kingma et al., 2014).

Learning objective

The learning objective of the Variational Autoencoder, $\mathcal{L}_{VAE}(\theta, \phi)$, is inferred from the maximization of the log-likelihood of the dataset.

$$\sum_{x \in \mathcal{D}} \log p_\theta(X = x) \geq -\mathcal{L}_{VAE}(\theta, \phi) \tag{4.2}$$

The Variational Lower Bound, $\mathcal{L}_{VAE}(\theta, \phi)$, can be expressed as the sum of a reconstruction term and a regularization term. These two components may respectively be interpreted as the data distortion $D_X(\theta, \phi)$ and the code rate $R_Z(\theta)$ (Alemi et al.,

2018):

$$\mathcal{L}_{VAE}(\theta,\phi) = \underbrace{\mathbb{E}_{x,z\sim q_\theta(X,Z)}\left[-\log p_\phi(X=x|Z=z)\right]}_{D_X(\theta,\phi)} + \underbrace{KL(q_\theta(Z|X)||p(Z))}_{R_Z(\theta)} \tag{4.3}$$

By expressing the loss using equations 3.7 (data distortion) and 3.9 (code rate), one may notice that the Variational Autoencoder loss is equivalent the joint KL-divergence $KL(q_\theta(X,Z)||p_\phi(X,Z))$:

$$\begin{aligned}\mathcal{L}_{VAE}(\theta,\phi) &= \underbrace{KL(q_\theta(X|Z)||p_\phi(X|Z)) - I(q_\theta(X,Z)) + H(q(X))}_{D_X(\theta,\phi)} \\ &\quad + \underbrace{KL(q_\theta(Z)||p(Z)) + I(q_\theta(X,Z))}_{R_Z(\theta)} \\ &\equiv KL(q_\theta(X,Z)||p_\phi(X,Z))\end{aligned} \tag{4.4}$$

Where the equivalence follows from the fact that the entropy of the empirical distribution $q(X)$ is fixed and does not depend on the parameters.

Analysis

Considering the minimizers $\hat{q}_\theta(X,Z)$ and $\hat{p}_\phi(X,Z)$ of the Variational Autoencoder objective $\mathcal{L}_{VAE}(\theta,\phi)$, we may notice that the encoding distribution $\hat{q}_\theta(X,Z)$ represents the I-projection of the decoding distribution $\hat{p}_\phi(X,Z)$. At the same time, $\hat{p}_\phi(X,Z)$ also represents the M-projection of $\hat{q}_\theta(X,Z)$ (Section 2.2.1). As a consequence, $\hat{p}_\phi(X,Z)$ is encouraged to spread across every region of $\mathcal{X}\times\mathcal{Z}$ in which $\hat{q}_\theta(X,Z)$ has non-zero probability mass (zero-avoiding). The same behavior is enforced on the data marginal distribution as there is a latent representation, for every observation, that needs to be "covered" by the decoding distribution. More formally, for every data-point $x\in\mathcal{D}$ there exists at least one $z_x\in\mathcal{Z}$ that has non-zero probability according to $\hat{q}_\theta(Z|X=x)$. Due to the fact that $\hat{q}_\theta(X=x,Z=z_x)$ is positive, the optimal encoding distribution $\hat{p}_\phi(X=x,Z=z_x)$ is also encouraged have non-zero probability, which implies that its marginal $\hat{p}_\phi(X=x)$ is strictly greater then zero.

Similarly, we can show that the aggregated posterior $\hat{q}_\theta(Z)$ tends to fit in regions in which the prior $p(Z)$ is strictly positive. However, it does not necessarily cover its full support. This is because all the codes which have a positive prior probability and zero probability according to $q_\theta(Z)$ may be decoded in regions of the data space $\mathcal{X}$ that are distant from the observations. Consequently, a poor quality of the generated samples is expected whenever the mismatch between the two supports is large. This phenomenon has been widely observed when the number of latent dimensions is increased (Larsen et al., 2015; Theis and Bethge, 2016).

Note that in the Variational Autoencoder loss $\mathcal{L}_{VAE}(\theta,\phi)$ consists only of a generative component since both the coefficients α_p and α_q are set to zero. This implies that the learning objective is neither encouraging nor penalizing the flow of information from the data to the codes. As a result, the informativeness of the latent representation is not determined by the training objective itself but by the flexibility of the parametric representations. This issue of the Variational Autoencoder is know in literature as the **information preference** problem (Chen et al., 2016), which is manifested in a complete de-correlation between data and codes that is observed

whenever $p_\phi(X|Z)$ is more flexible. Therefore, the information preference problem shows that the original Variational Autoencoder objective is not sufficient for representation learning, underlying the importance of representation component of the proposed loss.

4.2.2 β-Variational Autoencoder

A simple extension of the Variational Autoencoder framework has recently been proposed by Higgins et al., 2017, by introducing a scaling factor β for the regularization term. The representations generated β-Variational Autoencoder (β-VAE) benefit from the stronger regularization induced by β which also disentangles the latent components (Burgess et al., 2017).

The learning objective

The learning objective of the β-VAE $\mathcal{L}_{\beta-VAE}(\theta, \phi; \beta)$ is formulated as an optimization of the data distortion $D_X(\theta, \phi)$ constrained by a maximum code-rate $R_Z(\theta)$:

$$\min_{\theta,\phi} \underbrace{\mathbb{E}_{x,z\sim q_\theta(X,Z)}\left[-\log p_\phi(X=x|Z=z)\right]}_{D_X(\theta,\phi)} \quad \text{subject to} \quad \underbrace{KL(q_\theta(X|Z)||p(Z))}_{R_Z(\theta)} \leq \epsilon \tag{4.5}$$

By expressing the constrained optimization as a Lagrangian, the learning objective can be expressed as:

$$\mathcal{L}_{\beta-VAE}(\theta, \phi; \beta) = D_X(\theta, \phi) + \beta\, R_Z(\theta) \tag{4.6}$$

Where the parameter β regulates the strength of the constraint in equation 4.5. We can formulate the β-VAE loss as a weighted sum of KL-divergences and mutual-information by normalizing the coefficients and re-arranging the terms according to equations 3.7 and 3.9:

$$\begin{aligned}\mathcal{L}_{\beta-VAE}(\theta, \phi; \beta) = & \underbrace{\frac{\beta}{\beta+1} KL(q_\theta(Z)||p(Z)) + \frac{1}{\beta+1} KL(q_\theta(X|Z)||p_\phi(X|Z))}_{D(q_\theta(X,Z)||p_\phi(X,Z))} \\ & + \underbrace{\frac{\beta-1}{\beta+1}}_{\alpha_q} I(q_\theta(X,Z))\end{aligned} \tag{4.7}$$

The first two terms in equation 4.7 represent a valid divergence between the two joint distributions $q_\theta(X, Z)$ and $p_\phi(X, Z)$ for any $\beta > 0$. For $\beta = 1$ the β-VAE loss function coincides with the Variational Autoencoder.

Analysis

From expression 4.7 we can observe that for $0 \leq \beta < 1$, the loss increases the flow of information from the data to the codes, whereas $\beta > 1$ penalizes $I(q_\theta(X, Z))$. Not only does the parameter β influence the mutual information, it also affects the weighting of the two components of the KL-divergence. In fact, when $\beta \to 0$ the divergence between the conditional distributions $q_\theta(X|Z)$ and $p_\phi(X|Z)$ gains more importance, while for $\beta \to \infty$ the majority of the weight shifts towards the divergence between the aggregated posterior, $q_\theta(Z)$, and the prior, $p(Z)$. A less strict-matching between $q_\theta(X|Z)$ and $p_\phi(X|Z)$, caused by high values of β, decreases the quality of the data reconstructions as observed in Higgins et al., 2017.

Analogously to the VAE objective, the β-VAE learning goal is minimizing the a re-scaled version of $KL(q_\theta(X,Z)||p_\phi(X,Z)$, therefore the same zero-avoiding behavior described in section 4.2.1 has to be expected.

4.2.3 Information Variational Autoencoder

The Information Variational Autoencoder (Info-VAE) has been presented by Zhao, Song, and Ermon, 2017a as a refinement of the VAE model. The Info-VAE learning objective addresses the information preference problem by modifying the regularization term in the original VAE formulation to ensure that the mutual information is maximized. Tolstikhin et al., 2018 independently suggests a similar objective (Wasserstein Auto-Encoder) based on a optimal transport formulation (Villani, 2003) of the problem. Both the Info-VAE and the Wassertein Auto-Encoder formulations can be considered as an extension of the Adversarial Autoencoder (AAE) originally proposed in Makhzani et al., 2014 that estimates the discrepancy between the aggregated posterior $q_\theta(Z)$ and the prior $p(Z)$ using an adversarial approximation of the Jensen-Shannon divergence. For the clarity of notation and symmetry with the Info-GAN training objective, the learning goal will be referred to as Info-VAE.

Learning objective

Observing that the code rate $R_Z(\theta)$ can be expressed as the sum of the encoding mutual information and the KL-divergence between the marginals $q_\theta(Z)$ and $p(Z)$ (equation 3.9), the Info-VAE objective $\mathcal{L}_{Info-VAE}(\theta,\phi;\lambda)$ is defined by adapting the regularization term to address only the divergence between the two distributions in the code space:

$$\mathcal{L}_{Info-VAE}(\theta,\phi;\lambda) = \underbrace{\mathbb{E}_{x,z\sim q_\theta(X,Z)}\left[-\log p(X=x|Z=z)\right]}_{D_X(\theta,\phi)} + \lambda D(q_\theta(Z)||p(Z)) \quad (4.8)$$

Zhao, Song, and Ermon, 2017a suggests three different choices for $D(q_\theta(Z)|p(Z))$:

- Adversarial approximation for the Jensen-Shannon divergence (Makhzani et al., 2014):
$$D(q_\theta(Z)||p(Z)) = JS(q_\theta(Z)||p(Z))$$
- Stein Variational Gradient approximation for the KL-divergence (Liu and Wang, 2016):
$$D(q_\theta(Z)||p(Z)) = KL(q_\theta(Z)||p(Z))$$
- Maximum Mean Discrepancy based on a positive definite kernel $k(\cdot||\cdot)$ (Li, Swersky, and Zemel, 2015; Dziugaite, Roy, and Ghahramani, 2015):
$$D(q_\theta(Z)||p(Z)) = MMD_k(q_\theta(Z)||p(Z))$$

In this work, we will consider both adversarial approximations of the Jensen-Shannon and Kullback-Leibler divergences estimated by using the density ratio trick described in Appendix C.

Normalizing the coefficients and re-writing the data distortion using equation 3.7

we can express the loss as:

$$\mathcal{L}_{Info-VAE}(\theta,\phi;\lambda) \equiv \underbrace{\frac{1}{\lambda+1}KL(q_\theta(X|Z)||p_\phi(X|Z)) + \frac{\lambda}{\lambda+1}D(q_\theta(Z)|p(Z))}_{D(q_\theta(X,Z)||p_\phi(X,Z))} + \underbrace{\frac{-1}{\lambda+1}}_{\alpha_q} I(q_\theta(X,Z)) \tag{4.9}$$

Note that the first two term of equation 4.9 represent a divergence between the joint distributions $q_\theta(X,Z)$ and $p_\phi(X,Z)$ for any $\lambda > 0$.

Analysis

Equation 4.9 suggests that by increasing the value of λ, we can expect two different effects. Similarly to the β-VAE loss, whenever the value of λ is increased, the weight shifts from the divergence between conditional distributions $q_\theta(X|Z)$ and $p_\phi(X|Z)$ to the one between the latent code marginals $q_\theta(Z)$ and $p(Z)$. Furthermore, the hyper-parameter λ influences the value of the coefficient α_q. In fact, its value decreases to zero as the value of λ is incremented. Contrarily to the β-VAE objective, the coefficient α_q is negative for any value of the hyper-parameter λ. This means that possible reductions of the mutual information are not caused by the training objective itself but by lack of flexibility of the parametric representation of the encoding distribution $q_\theta(X|Z)$.

The divergence between the joint distributions $q_\theta(X,Z)$ and $p_\phi(X,Z)$ in the Info-VAE objective depends on the specific choice for $D(q_\theta(Z)||p(Z))$. Selecting the Kullback-Leibler divergence we may observe the same zero-avoiding behavior discussed for the VAE and β-VAE models.

4.2.4 Information Generative Adversarial Network

The Information Generative Adversarial Network (Info-GAN) model introduced by Chen et al., 2016 represents an information theoretic extension of the GAN architecture designed for representation learning. While the original GAN objective does not provide a direct strategy to produce a latent data embedding, Info-GAN associates the generator with an encoder that is trained to reconstruct the original latent codes sampled used for the generative process. The framework has been shown to produce state-of-the-art results in terms of quality of the learned latent representation and sharpness of the generative samples.

Learning Objective

The traditional GAN training focuses only on the minimization of the Jensen-Shannon divergence between the generated and the real data. On the other hand, the Info-GAN model combines this original objective with an additional term that ensures the maximization of the mutual information induced by the decoding distribution, $I(p_\phi(X,Z))$. This is done by exploiting a lower bound on the mutual information known as the Info-Max bound (Barber and Agakov, 2003) which allows to write the loss $\mathcal{L}_{Info-GAN}(\theta,\phi;\lambda)$ as weighted sum of the code distortion $D_Z(\theta,\phi)$ and a

divergence, $D(p_\phi(X)||q(X))$, between the two marginal distributions:

$$\mathcal{L}_{Info-GAN}(\theta,\phi;\lambda) = D(p_\phi(X)||q(X)) + \lambda \underbrace{\mathbb{E}_{x,z\sim p_\phi(X,Z)}\left[-\log q_\theta(Z=z|X=x)\right]}_{D_Z(\theta,\phi)} \tag{4.10}$$

In the original work, $D(p_\phi(X)||q(X))$ is selected to be the adversarial approximation of the Jensen-Shannon divergence, consistently with the original GAN formulation. In the scope of this work, both the Jensen-Shannon and Kullback-Leibler adversarial approximations will taken into account.

Re-arranging the terms in the Info-GAN formulation as prescribed by equation 3.12 one can express $\mathcal{L}_{Info-GAN}(\theta,\phi;\lambda)$ as a weighed sum of decoding mutual information and divergence between the two joint distributions:

$$\begin{aligned}\mathcal{L}_{Info-GAN}(\theta,\phi;\lambda) = &\underbrace{\frac{1}{\lambda+1}D(p_\phi(X)||q(X)) + \frac{\lambda}{\lambda+1}KL(p_\phi(Z|X)||q_\theta(Z|X))}_{D(q_\theta(X,Z)||p_\phi(X,Z))} \\ &+ \underbrace{\frac{-\lambda}{\lambda+1}}_{\alpha_p} I(p_\phi(X,Z))\end{aligned} \tag{4.11}$$

The expression in equation 4.11 is a valid divergence between the joint distributions for any $\lambda > 0$.

Analysis

Symmetrically to the Info-VAE loss, whenever the value of λ is small, the learning objective prioritizes the minimization of the divergence between the data marginal $p_\theta(X)$ and the empirical distribution $q(X)$, yielding strong generative performances. Contrarily, when $\lambda \to \infty$, the KL-divergence $KL(p_\phi(Z|X)||q_\theta(Z|X))$ is considered instead, increasing the consistency of $q_\theta(Z|X)$ and $p_\phi(Z|X)$ but gradually ignoring the data marginals. As expected, for any $\lambda > 0$ the decoding mutual information is maximized since the coefficient α_p is strictly negative.

Among all the possible options for $D(p_\theta(X)||q(X))$, the choice Kullback-Leibler divergence, $KL(p_\theta(X)||q(X))$, is noteworthy. This is because by composing the conditional and the marginal divergence, the training objective addresses the minimization of the joint KL-divergence $KL(p_\phi(X,Z)||q_\theta(X,Z))$ in the opposite direction when compared to the VAE-based models. As a result, the minimizer $\hat{p}_\phi(X,Z)$ assumes the role of I-projection of the optimal encoding distribution $\hat{q}_\theta(X,Z)$. Thus, we can expect $p_\phi(X,Z)$ to avoid all the regions of $\mathcal{X}\times\mathcal{Z}$ in which $q_\theta(X,Z)$ is strictly zero. The zero-avoiding behavior of the decoding distribution also affects the data-marginal $p_\phi(X)$ which models regions of $\mathcal{X}$ that are densely populated by the empirical distribution $q(X)$, while potentially discarding some isolated observations. The characteristic of ignoring some of the data-points $x \in \mathcal{D}$ is known in literature as the problem of **mode-collapse** (Salimans et al., 2016; Goodfellow, 2016) and it has been are reported also when the Jensen-Shannon adversarial approximation is used instead of the Kullback-Leibler one for both GAN and Info-GAN models.

4.2.5 Cycle-Consistent Generative Adversarial Network

The Cycle-Consistent Ganerative Adversarial Network architecture (Zhu et al., 2017; Kim et al., 2017) has been originally proposed as an approach to discover cross-domain relation in an unsupervised fashion. By learning two conditional distributions, the Cycle-GAN model imposes a **cycle-consistency** term (Zhou et al., 2016) to ensure coherence between the translations determined by the conditional distributions.
Even if the Cycle-GAN model has been originally used for image-to-image translation, recent work (Zhao and Ermon, 2017; Tiao, Bonilla, and Ramos, 2018) has been considering its applications in the context of representation learning, by training a consistent mapping from the data observations to the latent codes.

Learning Objective

The Cycle-GAN loss $\mathcal{L}_{Cycle-GAN}(\theta, \phi; \lambda)$ is composed of two terms. The first part enforces the cycle-consistency while the second one ensure that both the data and code marginals are consistent with each other. More specifically, the cycle-consistency component consists of the sum of both the code and data distortion, while the coherence between the marginal distributions is achieved by considering the divergences $D(q_\theta(Z)||p(Z))$ and $D(p_\phi(X)||q(X))$. A scaling coefficient λ balances the interaction of the components:

$$\begin{aligned}\mathcal{L}_{Cycle-GAN}(\theta, \phi; \lambda) = D(q_\theta(Z)||p(Z)) + D(p_\phi(X)||q(X)) \\ + \lambda \underbrace{(D_X(\theta, \phi) + D_Z(\theta, \phi))}_{cycle-consistency}\end{aligned} \tag{4.12}$$

In the original objective proposed by Zhu et al., 2017, only the effect of the Jensen-Shannon adversarial approximation is studied. However, the more general formulation reported in equation 4.12 consents to consider other divergences. In particular, the use of the Kullback-Leibler adversarial approximation has recently shown promising results (Tiao, Bonilla, and Ramos, 2018). For this reason, consistently with the Info-VAE and Info-GAN losses, both options will be examined.

Considering that data and code distortion can be re-written employing equations 3.7 and 3.12, the loss expression in equation 4.12 may be formulated in the more general form:

$$\begin{aligned}\mathcal{L}_{Cycle-GAN}(\theta, \phi; \lambda) = D_{cycle}(q_\theta(X, Z)||p_\phi(X, Z)) \\ + \underbrace{\frac{-\lambda}{1+\lambda}}_{\alpha_p} I(p_\phi(X, Z)) + \underbrace{\frac{-\lambda}{1+\lambda}}_{\alpha_q} I(q_\theta(X, Z))\end{aligned} \tag{4.13}$$

Where the joint divergence $D_{cycle}(q_\theta(X, Z)||p_\phi(X, Z))$ is represented by:

$$\begin{aligned}D_{cycle}(q_\theta(X, Z)||p_\phi(X, Z)) = \frac{\lambda}{1+\lambda} KL(q_\theta(X|Z)||p_\phi(X|Z)) + \frac{1}{1+\lambda} D(q_\theta(Z)||p(Z)) \\ + \frac{\lambda}{1+\lambda} KL(p_\phi(Z|X)||q_\theta(Z|X)) + \frac{1}{1+\lambda} D(p_\phi(X)||q(X))\end{aligned} \tag{4.14}$$

Note that for every $\lambda > 0$ the sum of the four components reported in equation 4.14 is a valid divergence between $q_\theta(X, Z)$ and $p_\phi(X, Z)$. In particular, the joint divergence reported in equation 4.14 coincides with the symmetrized KL-divergence whenever the Kullback-Leibler divergence is selected to approximate the two marginals and $\lambda = 1$:

$$D_{cycle}(q_\theta(X, Z)||p_\phi(X, Z)) = \frac{1}{2}KL(p_\phi(X, Z)||q_\theta(X, Z)) + \frac{1}{2}KL(q_\theta(X, Z)||p_\phi(X, Z)) \tag{4.15}$$

Analysis

The Cycle-GAN loss can be seen as the sum of the Info-VAE and Info-GAN losses. For this reason, we can expect the hyper-parameter λ to regulate the consistency between the marginals (for both the data and the latent codes) and the conditional distributions. For small values of λ, the Cycle-GAN model ignores the divergence between the conditional distributions to focus on the consistency of the two marginals. In this situation, the divergence between the aggregated posterior $q_\theta(Z)$ and the prior $p(Z)$ is minimized together with the discrepancy between data-marginal $p_\phi(X)$ and the empirical distribution $q(X)$. This results the achievement of solid generative performances but an inconsistent latent representation.

When $\lambda \to \infty$, two effects may be observed. Firstly, most of the weight in the joint divergence $D_{cycle}(q_\theta(X, Z)||p_\phi(X, Z))$ is shifted to the divergences between the conditional distributions, enhancing their agreement. Secondly, as the value of the coefficients α_p and α_q goes to -1, the mutual information gains more importance in determining the optimal parameter configuration. Despite the achievement of a consistent coding scheme which captures the information from the observations, the lack of coherence between the marginals results in weak generative performances. As a result, poor generalization for new observations may be expected since the model is not able to explain the data generating process.

4.3 Summary and comparison

The β-VAE, Info-VAE, Info-GAN and Cycle-GAN objectives introduce an hyper-parameter that affects the weight of the encoding or decoding mutual information in the training loss. At the same time, the value of the hyper-parameter is shown to influence the generative performances by affecting the divergence between the $q_\theta(X, Z)$ and $p_\phi(X, Z)$. Even if this trade-off is implicit in the original formulations, one can exploit some information theoretical properties to re-write the model's learning goals in a more general form.

Table 4.1 reports a comparison of the loss functions described in this chapter in their original formulation as a weighted sum of data distortion $D_X(\theta, \phi)$, code distortion $D_Z(\theta, \phi)$, code rate $D_Z(\theta, \phi)$, divergence between the data marginals $D(p_\phi(X)||q(X))$ and code marginals $D(q_\theta(Z)||p(Z))$. The vanilla Auto-Encoder (AE) and Generative Adversarial Networks (GAN) objectives are also reported as a degenerate case of the previously listed models.

The different losses are also reported in Table 4.2 in the form specified by equation 4.1. In order to facilitate a comparison, each joint divergence $D(q_\theta(X, Z)||p_\phi(X, Z))$

Model	$D_X(\theta,\phi)$	$D_Z(\theta,\phi)$	$R_Z(\theta)$	$D(p_\phi(X)\|\|q(X))$	$D(q_\theta(Z)\|\|p(Z))$
VAE	1	0	1	0	0
β-VAE	1	0	β	0	0
Info-VAE	1	0	0	0	λ
Info-GAN	0	λ	0	1	0
Cycle-GAN	λ	λ	0	1	1
AE	1	0	0	0	0
GAN	0	0	0	1	0

TABLE 4.1: Comparison between the coefficient introduced by the different models in the original formulation. The different columns report the coefficients for the code-rate $R_Z(\theta)$, data distortion $D_X(\theta,\phi)$, code distortion $D_Z(\theta,\phi)$, divergence between the data-marginal and the empirical distribution $D(p_\phi(X)||q(X))$ and divergence between the prior and the aggregated posterior $D(q_\theta(Z)||p(Z))$.

is represented as a weighed sum of two components $M(\theta,\phi)$, the divergence between the marginals, and $C(\theta,\phi)$, the divergence between the conditional distributions:

$$D(q_\theta(X,Z)||p_\phi(X,Z)) = \pi_M\, M(\theta,\phi) + \pi_C\, C(\theta,\phi) \tag{4.16}$$

The coefficients π_M and π_C denote the two weighting factors for $M(\theta,\phi)$ and $C(\theta,\phi)$ respectively.

Model	π_M	π_C	α_q	α_p	$M(\theta,\phi)$	$C(\theta,\phi)$
VAE	1	1	0	0	$KL(q_\theta(Z)\|\|p(Z))$	$KL(q_\theta(X\|Z)\|\|p_\phi(X\|Z))$
β-VAE	$\frac{\beta}{1+\beta}$	$\frac{1}{1+\beta}$	$\frac{\beta-1}{1+\beta}$	0	$KL(q_\theta(Z)\|\|p(Z))$	$KL(q_\theta(X\|Z)\|\|p_\phi(X\|Z))$
Info-VAE	$\frac{\lambda}{1+\lambda}$	$\frac{1}{1+\lambda}$	$\frac{-1}{1+\lambda}$	0	$D(q_\theta(Z)\|\|p(Z))$	$KL(q_\theta(X\|Z)\|\|p_\phi(X\|Z))$
Info-GAN	$\frac{\lambda}{1+\lambda}$	$\frac{1}{1+\lambda}$	0	$\frac{-\lambda}{1+\lambda}$	$D(p_\phi(X)\|\|q(X))$	$KL(p_\phi(Z\|X)\|\|q_\theta(Z\|X))$
Cycle-GAN	$\frac{\lambda}{1+\lambda}$	$\frac{1}{1+\lambda}$	$\frac{-\lambda}{1+\lambda}$	$\frac{-\lambda}{1+\lambda}$	$D(p_\phi(X)\|\|q(X))$ $+D(q_\theta(Z)\|\|p(Z))$	$KL(p_\phi(Z\|X)\|\|q_\theta(Z\|X))$ $+KL(q_\theta(X\|Z)\|\|p_\phi(X\|Z))$
AE	0	1	-1	0	-	$KL(q_\theta(X\|Z)\|\|p_\phi(X\|Z))$
GAN	1	0	0	0	$D(p_\phi(X)\|\|q(X))$	-

TABLE 4.2: Comparison of the weighting of the components of the joint divergence (π_M and π_C) and the coefficients of the encoding and decoding mutual information (α_q and α_p) as a function of the model hyper-parameters. The joint divergence $D(q_\theta(X,Z)||p_\phi(X,Z))$ is expressed in the form $\pi_M\, M(\theta,\phi) + \pi_C\, C(\theta,\phi)$.

Chapter 5

Experiments

Previous chapters have presented several loss functions for the task of unsupervised representation learning. In their original definition, the learning objectives differ for the guiding principles and components. Therefore, this research proposed a general view which enables a direct comparison. This chapter addresses the employment of the proposed method, aiming for the following four aspects:

- Empirically validate the effect of the specific choice of the divergence between the encoding and decoding joint distributions by displaying some visualizations.
- Verify the effect of the hyper-parameters hypothesized in Chapter 4 by showing that the estimation for components of the joint divergence are consistent with the coefficient that have been determined theoretically.
- Confirm that the estimations for the values of encoding and decoding mutual information agree with the coefficient determined by the hyper-parameters, as described in Chapter 4, and investigate their interaction.
- Compare the presented models by showing that the information theoretic measurements can be insightful to understand the characteristic of different learning objectives.

Considering the aforementioned goals, this chapter is divided into three sections that aim to describe the experimental setup, present the empirical results and draw the final conclusions.

5.1 Experimental setup

This section describes the datasets and the modeling assumptions that have been selected to compare the different learning objectives; two main design principles have been considered. First of all, each experiment has been designed to minimize the influence of the optimization procedure, approximations and the modeling assumptions to focus on the investigation of loss function. Secondly, to qualitatively visualize the generative performances of different models, the experimentations have been performed on low dimensional spaces. This allows to directly plot marginals $p_\phi(X)$ and $q_\theta(Z)$. Furthermore, a reduced dimensionality of the data space $\mathcal{X}$ decreases the impact of the modeling choices for the decoder $p_\phi(X|Z)$.

In order to prevent the introduction of any external bias, the architectures used to model the conditional distributions $q_\theta(Z|X)$ and $p_\phi(X|Z)$ have been kept unaltered across multiple experiments, the initial parameter configurations θ_0 and ϕ_0 are identical and the same optimizer has been used. Further details on the neural networks used to model the two conditional distributions together with additional specifications for the training procedure can be found in appendix F.2.

5.1.1 1D-Mixture

The 1D-Mixture experiments have been designed to visualize and evaluate the generative performances. Both the data and code domains are selected to be unidimensional. On the one hand, this facilitates the visualization of two joint and marginal distributions. Furthermore, the one-dimensional domains grant an additional advantage since there is no need to make any modeling assumption concerning the conditional independence of the components. As a consequence, the final parameter configuration is mostly determined by the specific loss function in analysis.

Dataset

The dataset proposed for the 1D-mixture experiments consists of 20 samples drawn from a distribution $r(X)$ defined as the mixture of two Normal distributions with different variances, means and mixing proportions.

$$r(X) = \pi_1 \mathcal{N}(X|\mu_1, \sigma_1^2) + \pi_2 \mathcal{N}(X|\mu_2, \sigma_2^2)$$

With

$$\pi_1 = 0.7 \;\; \pi_2 = 0.3 \;\; \mu_1 = 0 \;\; \mu_2 = 3.2 \;\; \sigma_1 = 0.6 \;\; \sigma_2 = 0.9$$

The number of samples is extremely reduced to simulate a scenario in which the observations are sparse such that one may detect how the different models assign probability mass around each single sample. Figure 5.1 shows a representation of the probability density of the data-generating and empirical distributions.

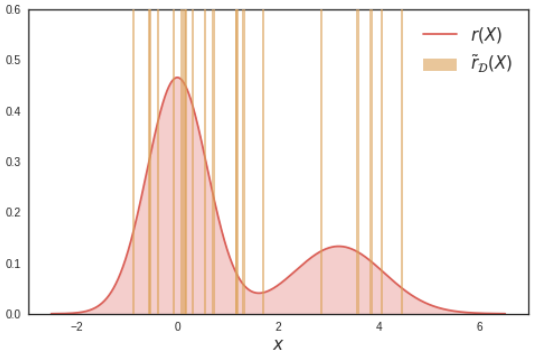

FIGURE 5.1: Visualization of the probability density of the data-generating distribution $r(X)$ (in red) and the empirical distribution $\tilde{r}_{\mathcal{D}}(X)$ (vertical orange lines) for the 1D-Mixture dataset.

Models

Both the encoding and the decoding conditional distribution are chosen to be Normal distributions. Their mean and variance are determined by parametric functions modeled as multi-layer perceptrons. More precisely:

$$q_\theta(Z|X = x) = \mathcal{N}(Z|\mu_\theta(x), \sigma_\theta^2(x))$$
$$p_\phi(X|Z = z) = \mathcal{N}(X|\mu_\phi(z), \sigma_\phi^2(z))$$

Where $\mu_\theta(x)$ and $\sigma_\theta^2(x)$ are functions that respectively map every data-point $x \in \mathcal{X}$ to the mean and the variance of the encoding distribution, while $\mu_\phi(z)$ and $\sigma_\phi^2(z)$

represents the counterparts for the decoding distribution $p_\phi(X|Z=z)$. Consistently with the literature, the prior $p(Z)$ is characterized with a Normal distribution with zero mean and unit variance:

$$p(Z) = \mathcal{N}(Z|0,1)$$

Note that while the prior and the empirical distribution are fixed, both the encoding and the decoding distributions can adapt their mean and variance by changing the values of the parameters θ and ϕ.

5.1.2 2D-Mixture

The 2D-Mixture experiments aim to show the quality of the reconstructions and the consistency between the marginals and conditional distributions. For this reason, both the data and the latent space are selected to be two-dimensional to be effectively visualized. The peculiarity of the 2D-Mixture experiments lies in the specific choice of the uniform prior distribution $p(Z)$ that allows to detect how the probability mass is distributed in the data space according to $p_\phi(X|Z)$ (see Figure 5.4).

Dataset

The dataset for the 2D-Mixture experiments consists of 500 samples from a mixture of multivariate Normal distributions. The three components present different characteristic to better underline the mode-covering capacities of the different models:

$$r(X) = \sum_{i=1}^{3} \pi_i \, \mathcal{N}(X|\boldsymbol{\mu}_i, \sigma_i^2 \mathbb{1})$$

With

$$\boldsymbol{\pi} = [0.6, 0.3, 0.1], \;\; \boldsymbol{\mu}_1 = [1,1], \;\; \boldsymbol{\mu}_2 = [5,4], \;\; \boldsymbol{\mu}_3 = [3,9], \;\; \boldsymbol{\sigma}^2 = [1, 0.5, 0.25]$$

The contour plot of of the log-density of the data-generating distribution $r(X)$ is visualized in Figure 5.2 together with the samples used for the experiments.

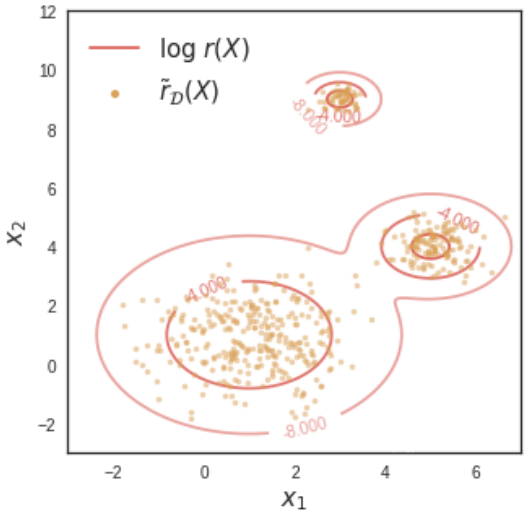

FIGURE 5.2: Visualization of the data generating distribution $r(X)$ (in red) and the empirical distribution $\tilde{r}_\mathcal{D}(X)$ (in orange) for the 2D-Mixture experiments.

Models

Similarly to the 1D-Mixture experiments, both the encoding and the decoding distributions are modeled as Multivariate Normal distributions whose position and scale depend on two parametric multi-layer perceptrons:

$$q_\theta(Z|X=x) = \mathcal{N}(Z|\boldsymbol{\mu}_\theta(x), \boldsymbol{\sigma}^2_\theta(x)^T\mathbb{1})$$
$$p_\phi(X|Z=z) = \mathcal{N}(X|\boldsymbol{\mu}_\phi(z), \boldsymbol{\sigma}^2_\phi(z)^T\mathbb{1})$$

Where $\boldsymbol{\mu}_\theta(x)$ and $\boldsymbol{\mu}_\phi(z)$ represent respectively the mean vector corresponding to the encoding and decoding distributions, while $\boldsymbol{\sigma}^2_\theta(x)$ and $\boldsymbol{\sigma}^2_\phi(z)$ refer to the diagonals of the respective covariance matrices. A relaxation of the uniform distribution over the interval $[-1, 1]$ is selected to model the prior $p(Z)$:

$$p(Z) = \prod_{i=1}^{2} \tilde{\mathcal{U}}(Z_i|m, s, \sigma^2)$$

The parameter m indicates the center of the distribution, s refer to the size of the uniform step while σ^2 represents the variance of the Normal distribution used to extend the support of the prior. In the experiments reported in this work we used:

$$m = 0, \quad s = 2, \quad \sigma^2 = 0.01$$

Further details regarding the functional form of the smoothed uniform distribution $\tilde{\mathcal{U}}(Z|m, s, \sigma^2)$ and its design can be found in Appendix F.1.

5.2 Experiments and Results

Empirical results obtained by the different models on the 1D and 2D-Mixture datasets are subdivided into four subsections corresponding to the goals defined at the beginning of this chapter. Each part will report the results together with a short conclusion that refer to the theory described in the previous chapters. A detailed description of the qualitative visualizations proposed in this section is reported in Appendix F.3, while Appendix F.4 displays the visualizations obtained for different settings of the hyper-parameters on the 2D-Mixture dataset.

5.2.1 Effect of the joint divergence

Figure 5.3 presents the result obtained by training the four models on the 1D-Mixture dataset. The visualization of the joint distributions $q_\theta(X, Z)$ and $p_\phi(X, Z)$ are consistent with the information and moment projection view hypothesized in Chapter 4. In fact, the decoding distributions $p_\phi(X, Z)$ trained using the β-VAE and Info-VAE objective clearly show the zero-avoiding behavior by completely covering the support of $q_\theta(X, Z)$. The Info-GAN architecture, on the other hand, exhibits a zero-forcing behavior that is induced by the minimization of the opposite direction of the KL-divergence. The architecture trained with the Cycle-GAN objective presents a more balanced matching induced by the sum of both the directions of the KL-Divergence.

The use the Jensen-Shannon adversarial approximation, instead of the Kullback-Leibler divergence, mitigates the zero-avoiding and zero forcing traits observed for the Info-VAE and Info-GAN learning objectives. A possible cause behind this phenomenon, which can be observed in the second row Figure 5.3 , may be traced back to symmetry of the Jensen-Shannon divergence. This hypothesis is also supported

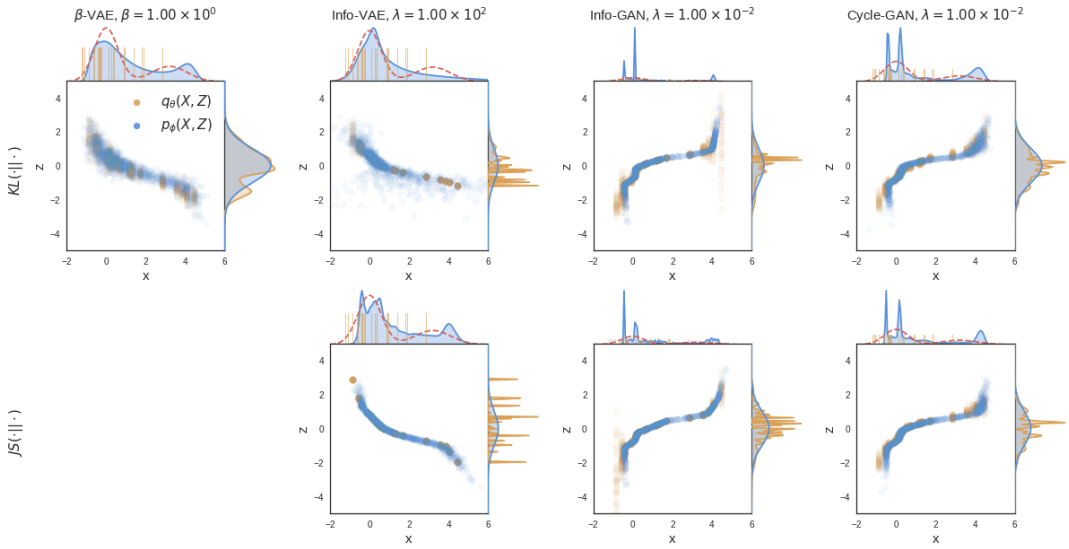

FIGURE 5.3: Visualization of joint encoding and decoding distributions $p_\phi(X, Z)$ and $q_\theta(X, Z)$ trained on the 1D-Mixture dataset for β-VAE, Info-VAE, Info-GAN and Cycle-GAN (one for each column). The two rows correspond to the choice of divergence used to match the marginal distributions (Kullback-Leibler divergence for the first row, Jensen-Shannon for the second one). Each plot presents an approximation of the joint distribution (in the center) together with the data marginals (on the top) and the latent code marginals (on the right). The data-generating distribution is represented trough a dashed red line. The plot shows that the for the β-VAE and Info-VAE models the decoding distribution completely covers the support of the encoding distribution. The Info-GAN architecture leads to the opposite scenario, while the Cycle-GAN loss results in a balance between the zero-avoiding and zero-forcing behavior.

by the reduced influence of divergence's choice for the Cycle-GAN model. This is because the Cycle-GAN loss is symmetric for both the Jensen-Shannon and Kullback-Leibler adversarial approximations of the two marginals.

The zero-forcing and zero-avoiding behavior induced on the joint distributions can be also observed in the data marginals. This characteristic is clearly noticeable in Figure 5.4 which compares the data marginal distributions $p_\phi(X)$ obtained by training the same model using the VAE and an Info-GAN learning objectives.

Note that the qualitative visualization reported in Figures 5.3 and 5.4 represent a selection of the full spectrum covered by the experiments. Nevertheless, the results presented in this section is consistent with the observed general trend.

5.2.2 Effect of the hyper-parameters on the generative performances

Figure 5.5 reports the estimated values of the two components of the joint KL-divergence $KL(q_\theta(X, Z)||p_\phi(X, Z))$ for the β-VAE, Info-VAE and Cycle-GAN architectures for different hyper-parameter configurations. In section 4.2 we have hypothesized that the importance of $KL(q_\theta(Z)||p(Z))$ and $KL(q_\theta(X|Z)||p_\phi(X|Z))$ is influenced by the values of the hyper-parameters. The trade-off between the values of $KL(q_\theta(Z)||p(Z))$ and $KL(q_\theta(X|Z)||p_\phi(X|Z))$ observed empirically is consistent with their scaling. In

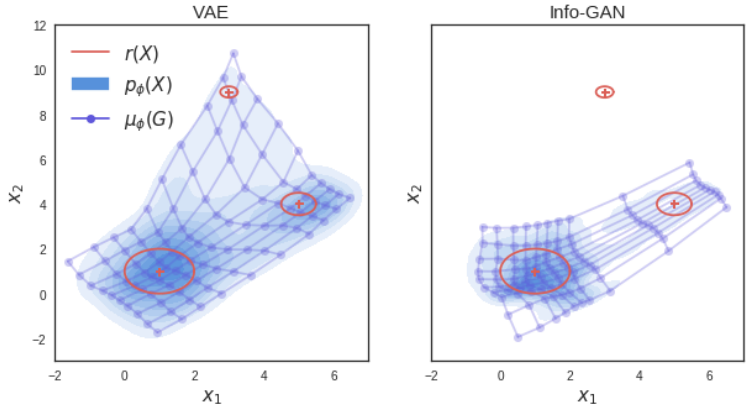

FIGURE 5.4: Visualization of the data marginals obtained by training the same model using the VAE (on the left) and Info-GAN (on the right) learning objectives on the 2D-Mixture dataset. The plot compares the modes of the data-generating distribution $r(X)$ (in red) with an approximation of the density computed by sampling the data marginal $p_\phi(X)$. By observing the area covered by the samples, one may notice that the VAE model over-estimates the support of $r(X)$, while the Info-GAN model does not cover one of the three modes of the data-generating distribution.

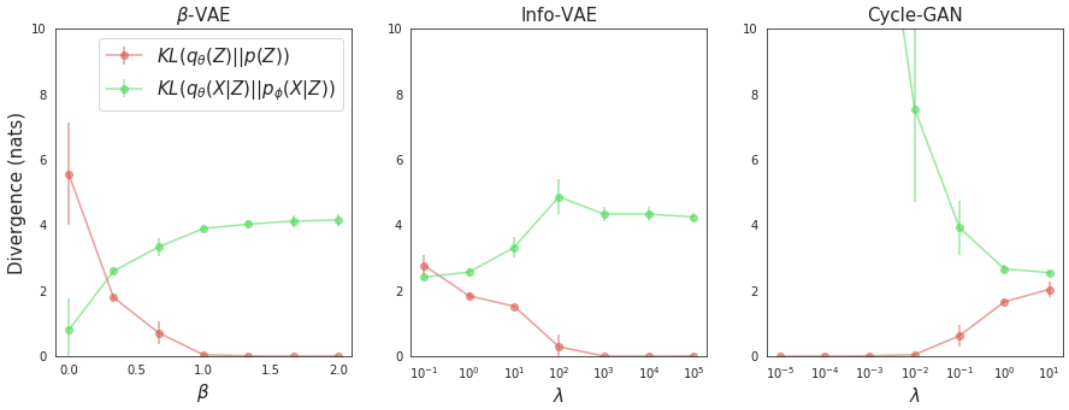

FIGURE 5.5: Plot of the trade-off between $KL(q_\theta(Z)||p(Z))$ (in red) and $KL(q_\theta(X|Z)||p_\phi(X|Z))$ (in green) measured for β-VAE, Info-VAE and Cycle-GAN by varying the respective hyper-parameter. Each model has been trained 5 times with different instances of the 1D-Mixture distribution. The average values of divergence achieved by the different architectures after convergence are reported together with the estimation of the standard deviations. The picture shows that the trade-off of $KL(q_\theta(Z)||p(Z))$ and $KL(q_\theta(X|Z)||p_\phi(X|Z))$ measured empirically is consistent with the scaling coefficient determined theoretically as a function of their respective hyper-parameter (Table 4.2). Furthermore the reduced variance validates the effectiveness of the proposed estimations for $KL(q_\theta(Z)||p(Z))$ and $KL(q_\theta(X|Z)||p_\phi(X|Z))$. Note that the Info-VAE and Info-GAN have been trained using the Kullback-Leibler adversarial approximation.

fact, when the value of the hyper-parameters increase, both the β-VAE and the Info-VAE prioritize matching the aggregated posterior $q_\theta(Z)$ with the prior $p(Z)$ and

reduce the importance of the divergence between $q_\theta(X|Z)$ and $p_\phi(X|Z)$. Contrarily, the graph for the Cycle-GAN model shows the opposite trend as theoretically predicted. Moreover, the agreement between the theoretical expectations and the empirical measurements across multiple experiments validates the estimations of the two components for the KL-divergence as proposed in Section 3.1.3.

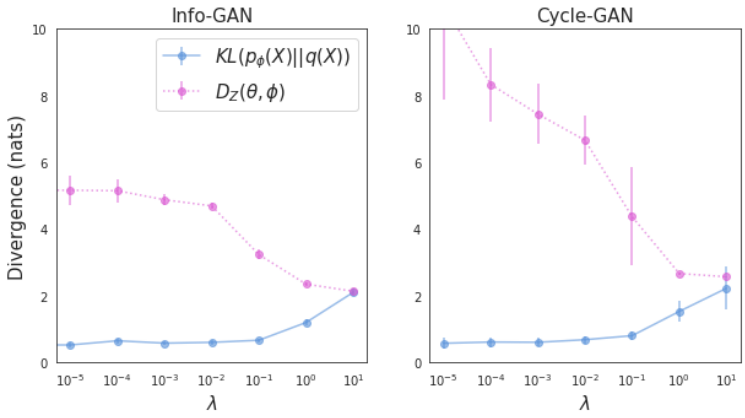

FIGURE 5.6: Plot of the trade-off between two components of the joint KL divergence $KL(p_\phi(X)||q(X))$ (in blue) and $KL(p_\phi(Z|X)||q_\theta(Z|X))$ obtained by the Info-GAN and Cycle-GAN by varying the respective hyper-parameters. The mean and the variance of the measurements have been estimated by following the same procedure described in Figure 5.5. Note that since $KL(p_\phi(Z|X)||q_\theta(Z|X))$ can not be estimated directly, the plot reports its upper bound that is represented by the code distortion $D_Z(\theta,\phi)$ instead (dotted line in pink). For both the Cycle-GAN and Info-GAN architectures, the empirical measurements are consistent with the values of the coefficients reported in Table 4.2.

Figure 5.6 reports the trade-off between the two components $KL(p_\phi(X)||q(X))$ and $KL(p_\phi(Z|X)||q_\theta(Z|X))$ of the KL-divergence for the Info-GAN and Cycle-GAN training objectives. Note that since we can not compute the data-rate $R_X(\phi)$, it is not possible to estimate the value of $KL(p_\phi(Z|X)||q_\theta(Z|X))$. For this reason, the plot employs the code distortion $D_Z(\theta,\phi)$ to represent its upper-bound instead.

The graphs visualized in Figures 5.5 and 5.6 consider 5 different instances of datasets, sampled from the distribution $r(X)$, for the 1D-mixture experiments. Other then representing a valid sanity-check for the theoretical analysis, the results show that the estimations for the components of the KL-divergence suggested in Appendix E are consistent across multiple runs.

5.2.3 Effect of the hyper-parameters on the mutual information

Figure 5.7 visualizes the estimated encoding and decoding mutual information for the β-VAE, Info-VAE, Info-GAN and Cycle-GAN training objectives with different hyper-parameter configurations. The scaling of the mutual information, induced by the coefficients α_q and α_p, is consistent with the empirical measurements. Particularly, the β-VAE directly penalizes the mutual information, since α_q is positive for every $\beta > 1$. Contrarily, the decreased mutual information observed for the Info-VAE objective is more likely to be determined by the specific choice of the encoding distribution $q_\theta(X|Z)$. As λ increases, $\mathcal{L}_{Info-VAE}(\theta,\phi,\lambda)$ assigns more importance to the

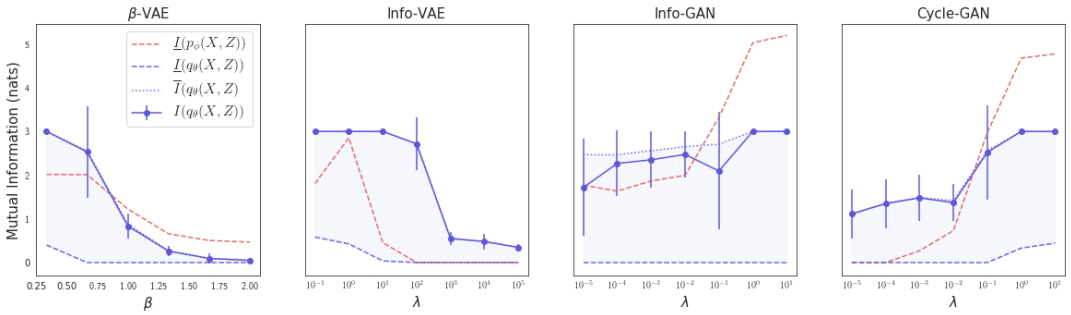

FIGURE 5.7: Comparison of encoding (in blue) and decoding (in red) mutual informations for the β-VAE, Info-VAE, Info-GAN and Cycle-GAN architectures obtained by varying the respective parameters on the 1D-Mixture dataset. Each plot reports the lower bound (dashed blue line), the upper bound (dotted blue line) and the estimation for the encoding mutual information (solid line) together with the lower bound of the decoding mutual information (dashed red line). The empirical observations reported in the four graphs are coherent with the coefficients α_q and α_p determined in Chapter 4. The procedure used to determine the mean and the variance of the measurement is identical to the one reported in Figures 5.6 and 5.5.

divergence between the aggregated posterior $q_\theta(Z)$ and the prior $p(Z)$. In this scenario, the model may sacrifice some of the encoding mutual information $I(q_\theta(X,Z))$ to achieve a better match between $q_\theta(Z)$ and $p(Z)$.

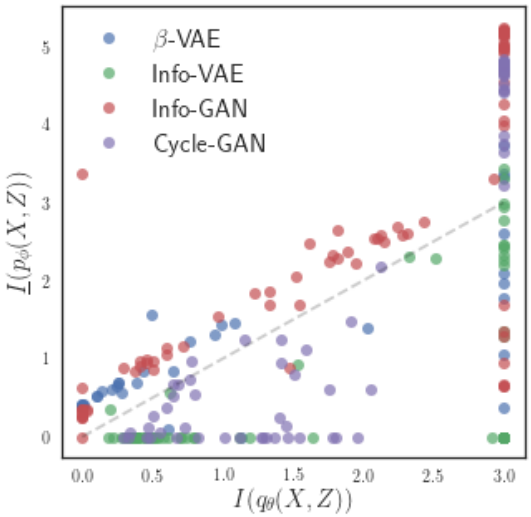

FIGURE 5.8: Visualization of the correlation between the encoding mutual information estimation $I(q_\theta(X,Z))$ (x-axis) and the lower bound of the decoding mutual information $\underline{I}(p_\phi(X,Z))$ (y-axis) measured for the different models, hyper-parameters and choices of marginal divergences on 5 different instances of the 1D-Mixture dataset. The two mutual information exhibit a strong correlation even when the training objective is addressing only one of the two components.

The Info-GAN objective, on the other hand, directly promotes the decoding mutual information $I(p_\phi(X,Z))$ whenever the parameter λ is raised. The same behavior can be observed for the Cycle-GAN learning objective where α_q and α_p influence $I(q_\theta(X,Z))$ and $I(p_\phi(X,Z))$ simultaneously. The Cycle-GAN loss exclusively addresses both $I(q_\theta(X,Z))$ and $I(p_\phi(X,Z))$, nevertheless all the four plots in Figure 5.7 report a strong correlation between the two quantities. This characteristic can clearly be observed in Figure 5.8. The estimation of the encoding mutual information $I(q_\theta(X,Z))$ is shown together with the lower-bound of the decoding mutual information $\underline{I}(p_\phi(X,Z))$ (equation 3.13) and has been evaluated across different runs with multiple hyper-parameter settings. The strong linear correlation between the two quantities reported in Figure 5.8 is likely to be caused by the minimization of the joint divergence $D(q_\theta(X,Z)||p_\phi(X,Z))$. In fact, when the divergence is minimal, $q_\theta(X,Z)$ and $p_\phi(X,Z)$ are "close" to each other. Consequently, the same constraint has to apply to $I(q_\theta(X,Z))$ and $I(p_\phi(X,Z))$.

If the joint divergence reaches the value of zero, the two mutual information have to match perfectly since $q_\theta(X,Z)$ and $p_\phi(X,Z)$ need to represent the same joint distribution.

5.2.4 Comparing the different models

Whereas the previous sections focused on validating the estimations of the information theoretical quantities, this section now aims to compare the different approaches. Figure 5.9 shows the trade-off between the data distortion $D_X(\theta,\phi)$ and code rate $R_Z(\theta)$ for different values of the model's hyper-parameters. The models

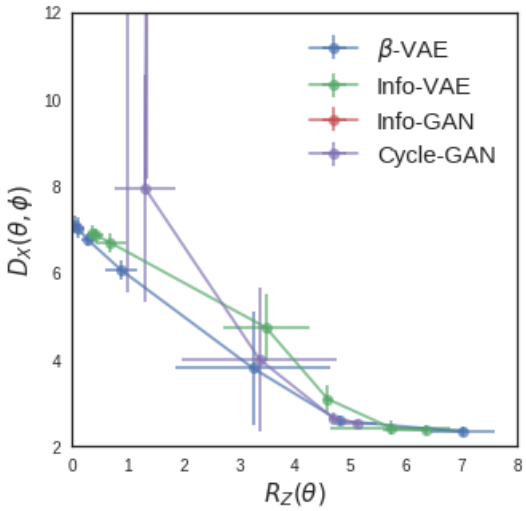

FIGURE 5.9: Plot of the trade-off between code rate $R_Z(\theta)$ and data distortion $D_X(\theta,\phi)$ achieved by the β-VAE, Info-VAE and Cycle-GAN objectives by varying respective parameters. The curve defined by the Info-GAN objective is not visualized as the data distortion is orders of magnitudes bigger then the ones achieved by the other architectures. The difference in scale underlines that the Rate-Distortion interpretation (Alemi et al., 2018) is inadequate to evaluate the different aspects of the 4 learning objectives.

trained with the β-VAE and the Info-VAE objectives smoothly interpolate between a regime of small data distortion (bottom right corner) and a configuration with small code-rate (top left corner). The Info-GAN and Cycle-GAN objectives denote different trajectories. In particular, the Info-GAN loss leads to configurations where the

data distortion $D_X(\theta, \phi)$ is more then 10 times bigger then the other models for all hyper-parameter configurations, causing it to be off-scale.

Info-GAN's elevated data distortion is caused by its training objective, since the rate-distortion visualization proposed in Alemi et al., 2018 only considers the values of $KL(q_\theta(X,Z)||p_\phi(X,Z))$. Thus, while all the other models focus on minimizing the visualized trade-off, Info-GAN optimizes for a different divergence (Section 4.2.4). More precisely, the minimization of the joint KL-divergence $KL(p_\phi(X,Z)||q_\theta(X,Z))$ suggested in the Info-GAN loss induces the zero-forcing behavior that can lead $p_\phi(X)$ to ignore some of the empirical observations. As reported in Figure 5.10, the reconstruction of the data-points ignored by $p_\phi(X)$ is not necessarily consistent, resulting in a drastic increase in the average data reconstruction error.

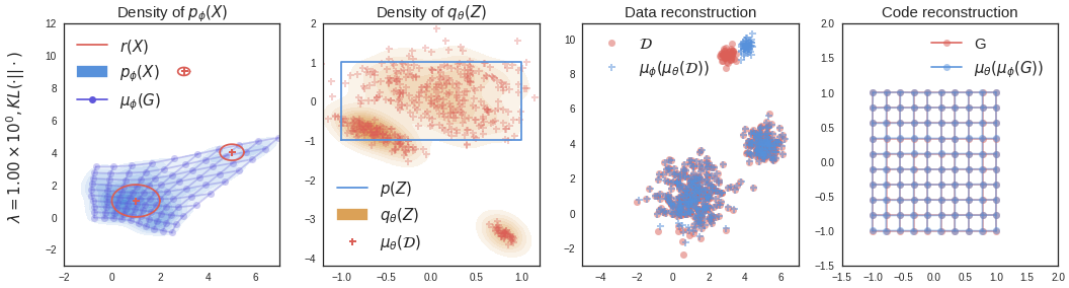

FIGURE 5.10: Visualization of the effect of the mode-collapse operated by the Info-GAN architecture. The plot visualizes both the estimations for the marginal $p_\phi(X)$ and $q_\theta(Z)$ (first and second columns from the left respectively) together with the data and the code reconstructions (last two columns on the right) obtained by encoding and decoding the dataset $\mathcal{D}$ and a uniform code grid G using the mean of the encoding $\mu_\theta(x)$ and decoding $\mu_\phi(z)$ distributions respectively. By observing that the decoding distribution does not model one of the modes of the real data distribution we can expect high values of data distortion $D_X(\theta, \phi)$. This is because the behavior of the encoder $q_\theta(Z|X)$ is un-constraint in the regions that are not modeled by $p_\phi(X)$. In fact, $q_\theta(Z|X)$ may map the observations that are not included in its support arbitrarily far from the prior (second figure from the left). At the same time, the decoder, $p_\phi(X|Z)$, is not necessarily consistent with the encoder in the areas that are unlikely to be sampled from $p(Z)$. As a consequence, the reconstruction for the data belonging to the missing modes lies arbitrarily far from the original observations (top-right mode in the third column) resulting in an increased data distortion $D_X(\theta, \phi)$.

Figure 5.11 illustrates the comparisons between the values of variational lower bound (ELBO), data distortion, code rate and code distortion that have been obtained by the four models, i.e. β-VAE, Info-VAE, Info-GAN and Cycle-GAN. All the architectures (except for β-VAE) have been trained by using both the Kulback-Leibler as the Jensen-Shannon adversarial approximation, with their respective hyper-parameter set to 1. By considering the values of data distortion $D_X(\theta, \phi)$, we can observe that Info-VAE on average achieves the best reconstructions with both the Jensen-Shannon and Kullback-Leibler approximations, closely followed by the models trained using the Cycle-GAN objective.

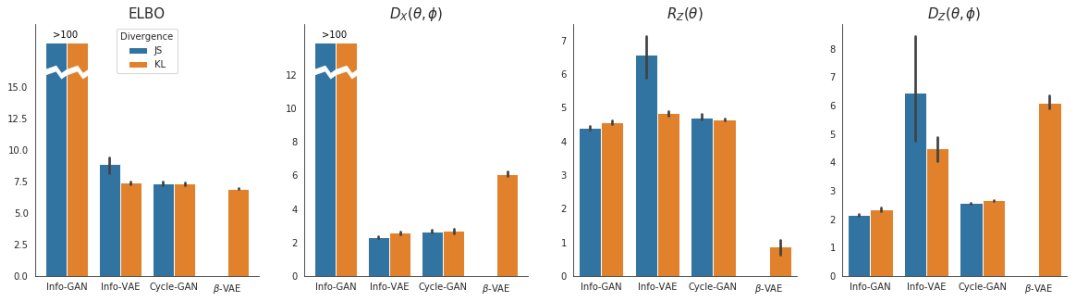

FIGURE 5.11: Comparison between the estimation of the values of the Variational Lower Bound (ELBO), data distortion ($D_X(\theta,\phi)$), code rate ($R_Z(\theta)$) and code distortion ($D_Z(\theta,\phi)$). Each model has been train on 5 different instances of the 1D-Mixture dataset with their respective hyper-parameter set to 1. The measures reported on the histograms refer the mean and standard deviation of the different quantities evaluated after convergence.

Looking at the values of the code distortion, $D_Z(\theta,\phi)$, one may notice that the Info-GAN and Cycle-GAN objectives consistently obtain much better results when compared the other models. This is, however, not surprising since both Info-GAN and Cycle-GAN losses directly minimize $D_Z(\theta,\phi)$; underlining a weak point of the Info-VAE and β-VAE models which are on average less consistent in reconstructing the codes sampled from the prior $p(Z)$.
The data rate, $R_Z(\theta)$, obtained by Info-GAN, contrarily to the measure of data distortion, is comparable with the values achieved by the other models. When looking at the third column of Figure 5.11 one may notice that the Info-VAE model trained using the Jensen-Shannon approximation results in significantly worse code rate when compared to the counterpart trained adopting the Kullback-Leibler approximation. This is because the value of $R_Z(\theta)$ can be decomposed as the sum of $KL(q_\theta(Z)|p(Z))$ and $I(q_\theta(X,Z))$ (equation 3.9). As a consequence, the Info-VAE variant that does not optimize the Kullback-Leibler divergence directly results in higher values of code rate, which are determined by $KL(q_\theta(Z)|p(Z))$ (Figure 5.12 second column). The reason behind the extremely reduced code rate achieved by the β-VAE model, on the other hand, can be traced back to the reduced encoding mutual information (Figure 5.13, first column).

Figure 5.12 reports the estimation of the components of the marginal and conditional KL-divergence obtained with the procedure presented in this work. The first two plots from the left compare the Kullback-Leibler divergence between the data marginal distribution. Unsurprisingly, the Info-GAN model that is trained Kullback-Leibler adversarial approximation achieves its minimal value. One may notice that the Info-GAN model trained with the Jensen-Shannon adversarial approximation performs much worse in terms of $KL(p_\phi(X)||q(X))$. The explanation behind this phenomenon is analogous to the $KL(q_\theta(Z)||p(Z))$ for the Info-VAE architecture: the Kullback-Leibler divergence is not an effective measure to estimate the distance of probability distributions that are minimizing the Jensen-Shannon divergence instead. The difference between the use of the two approximations is barely noticeable for Cycle-GAN , which presents consistent results for both choices of the adversarial approximations.
Considering the third column of Figure 5.12, one may notice that $q_\theta(X|Z)$ and $p_\phi(X|Z)$ are consistent for all the architectures but the Info-GAN. This measure supports the

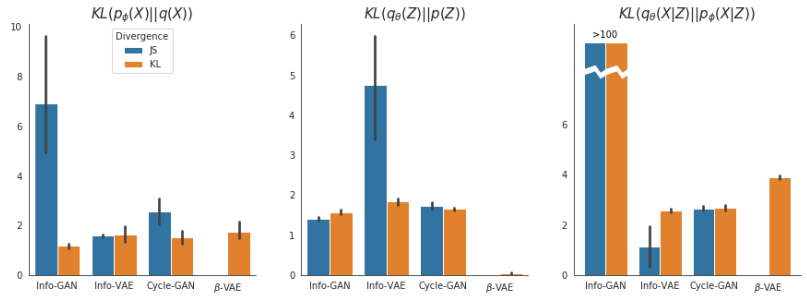

FIGURE 5.12: Comparison between the estimation of $KL(p_\phi(X)||q(X))$, $KL(q_\theta(Z)||p(Z))$ and $KL(q_\theta(X|Z)||p_\phi(X|Z))$ for the different models tested with $\lambda = 1$ ($\beta = 1$ for the β-VAE). The experimental setup is equivalent to the one used to produce Figure 5.11.

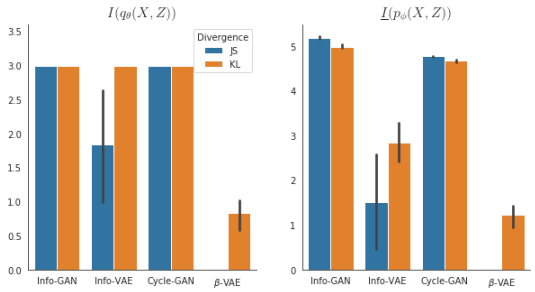

FIGURE 5.13: Visualization of the estimation for the values of encoding mutual information ($I(q_\theta(X,Z))$) and lower bound on the decoding mutual information ($\underline{I}(p_\phi(X,Z))$) obtained with the same experimental setup described in Figure 5.11.

hypothesis that the discrepancy between the two conditional distributions is caused by the problem of mode-collapse (Figure 5.10).
Even if the values of the KL-divergence between $p_\phi(Z|X)$ and the encoding distribution $q_\theta(Z|X)$ can not be computed directly, we may suppose, by symmetry, that most of the code distortion $D_Z(\theta,\phi)$ measured for the Info-VAE and β-VAE objectives can be traced back to the discrepancy between the two conditional distributions rather then a defect of the decoding mutual information $I(p_\phi(X,Z))$.

Figure 5.13 shows the comparison between encoding and decoding mutual information obtained by the different models. The β-VAE model consistently results in low values for both encoding and decoding mutual information. According to the analysis reported in Chapter 4 this is due to the effect of the hyper-parameter β, which does not positively affect $I(q_\theta(X,Z))$. One may notice that Info-GAN and Cycle-GAN obtain the maximum value for both quantities reported in Figure 5.13, while the Info-VAE training objective results in a reduced information flow when trained using the Jensen-Shannon adversarial approximation. This result underlines that the mutual information between data and codes depends not only on the scale of the coefficient α_p and α_q, but also on the specific joint divergence. In fact, as the Info-VAE models compared in Figure 5.13 differ only for the choice of $D(q_\theta(Z)||p(Z))$, we may conclude that the Jensen-Shannon divergence forces

$q_\theta(X, Z)$ and $p_\phi(X, Z)$ in regions of low mutual information. This behavior is difficult to identify or predict since it depends on the interaction between the joint divergence, $D(q_\theta(X, Z)||p_\phi(X, Z))$, and the specific parametric form of $q_\theta(Z|X)$ and $p_\phi(X|Z)$.

5.3 Conclusions

This section summarizes the empirical findings which have been reported. The main goal of the experiments was to show that the different models can be interpreted from both a generative and an information theoretical perspective. These two views give complementary information regarding unsupervised representation models. From the quantitative and qualitative results reported in this section we may therefore conclude the following:

1. The characteristics of the joint encoding and decoding distributions can be better understood by considering the generative component of the loss separately.
    - For the particular choice of the Kullback-Leibler divergence we can empirically observe the zero-avoiding and zero-forcing characteristic of the joint distributions on the data and code marginals.
    - By summing $KL(q_\theta(X, Z)||p_\phi(X, Z))$ and $KL(p_\phi(X, Z)||q_\theta(X, Z))$, the zero-forcing and zero-avoiding characteristics are counter-balanced in terms of the generative performances of the joint distributions.
    - The loss Info-VAE, Info-GAN and Cycle-GAN depends on the choice of marginal distribution. For both Info-VAE and Info-GAN the selection of the Jensen-Shannon adversarial approximation has shown to alleviate the zero-avoiding and zero-forcing behavior. Since the joint divergence for the Cycle-GAN is symmetric, the choice of the marginal divergence approximation is less relevant.
2. The hyper-parameters introduced by the different objectives influence both mutual information and the components of their respective joint divergence.
    - Each hyper-parameter trades-off the consistency between marginal and conditional distributions, defined by $q_\theta(X, Z)$ and $p_\phi(X, Z)$.
    - The empirical measurements of the components of the KL-divergence are consistent with the coefficients determined theoretically, supporting the validity of the proposed estimations.
3. Encoding and decoding mutual information are strongly correlated. Even when the training loss specifically targets only one of the two, the other follows. This might be due to the fact that the two joint distributions are matched by $D(p_\phi(X, Z)||q_\theta(X, Z))$.
4. The value of the Variational Lower bound is inadequate to evaluate the performances of models which do not directly minimize $KL(q_\theta(X, Z)||p_\phi(X, Z))$. As a consequence, the analysis of the code rate and data distortion expresses only limited aspects of the models making a direct comparison difficult.
5. The analysis proposed in this work identified some weak points of the compared objective functions.
    - The Info-GAN objective may cause mode collapse. This phenomenon induces inconsistent representations for the observations that are not modeled by $p_\phi(X)$. By symmetry, Info-VAE and β-VAE may be inconsistent

for codes z which are not modeled by the aggregated posterior $q_\theta(Z)$. Thanks to the cyclic consistency, the Cycle-GAN loss results in a good match for both conditional distributions, resulting in a more stable behavior.

- β-VAE and Info-VAE differ in the way they address mutual information. However, they present similar generative performances when the Info-VAE loss includes the adversarial approximation of the Kullback-Leibler divergence. The β-VAE explicitly penalizes the mutual information for increasing values of the hyper-parameter while the Info-VAE does not. For this reason, the reduction of the encoding mutual information observed empirically is likely to be related the parametrization of $q_\theta(Z|X)$.

The general learning goal provided in this works explicitly differentiates the generative aspect from the representation component of the loss function. This distinction can be useful to understand which architecture is more effective when considering different typologies of datasets. In contexts in which the observations are extremely sparse, the VAE-based models can result in a better representation, since their generative objective forces the model to consider all the observations. In particular, the β-VAE can represent an optimal solution whenever a general trend of the data needs to be captured. The Info-VAE alternative, on the other hand, allows to represent finer details at the price of a less stable adversarial training procedure. The Info-GAN model might represent a better alternative for dense or noisy datasets. This is because its loss encourages modeling local characteristics of the data, investing most of the representative capacity of the codes in regions that are densely populated with observations and ignoring possible outliers. The Cycle-GAN objective represents a good compromise between the two options, resulting in a latent representation that attempts to address all the observation without focusing on regions of the data space that are scarcely populated.

Once the learning objective has been fixed according to the characteristic of the data, the specific loss hyper-parameter can be tuned by considering its effect on the mutual information and the impact on conditional and marginal divergences. The quantitative estimation of the components of the Kullback-Leibler divergence proposed in this work can effectively help to diagnose potential inconsistencies between the two joint distributions that may be corrected by adjusting the scaling of the different loss components. At the same time, the presented measurement for $I(q_\theta(X, Z))$ may determine the amount of information that is preserved through the encoding procedure, which represents a fundamental indicator of the quality of the learned representation.

Chapter 6

Discussion and Future Work

In this work, we presented a generalized training objective for unsupervised representation learning which combines the minimization of a divergence measure between two parametric distributions with a constraint on the mutual information. An information theoretical analysis investigates the role played by different components induced by the two joint distributions, suggesting bounds and equivalences that allow to rephrase the learning objective of several well-known models in the literature to the proposed form. The unified loss representation consents to interpret and compare different design choices. In particular, the selection of the joint divergence is shown to have a crucial impact on both the consistency of the conditional distributions as the generative performances.

Other than unifying the different learning objectives, the proposed representation of the loss function provides valuable insights regarding the role of the hyper-parameters introduced by different models, which have not been investigated in the literature. An experimental analysis confirms that the hyper-parameters i) influence a trade-off between the matching of the marginals and the conditional distributions and ii) affect either one or both encoding and decoding mutual information. Furthermore, the empirical measurements reveal a strong correlation between the mutual information induced by encoding distribution and the one defined by the decoding distribution. This observation suggests that increasing the flow of information from the observations to the codes, or vice-versa, is sufficient to produce an informative data representation.

By comparing the β-VAE, Info-VAE, Info-GAN and Cycle-GAN learning objectives, both qualitatively and quantitatively, we were able to determine some of the strengths and weaknesses of the different approaches. Notably, we realized that the Cycle-GAN loss, which was not originally proposed as an objective for representation learning, combines the characteristic observed for Info-VAE and Info-GAN, leading to a more balanced behavior. The experimental results also suggest that the rate-distortion interpretation proposed by Alemi et al., 2018 presents only a limited view on the characteristic of the analyzed models. Nevertheless, one may improve the understanding of different learning objectives' aspects by considering other measurements. Specifically, the value of code distortion and the approximations for the components of KL-divergence between the joint distributions might expose potential inconsistencies between marginals and conditionals distributions. The proposed measurements not only represent a useful diagnostic tool, but also suggest how each model's strengths could be best expressed for different tasks and datasets.

6.1 Future work

This thesis focused on understanding the learning objectives for unsupervised representation learning. To this extent, the main goal addressed the analysis of different models in literature rather than defining of new learning goals. However, a number of variations based on linear combinations of the presented models arise naturally. The interpolation of the Info-VAE and Info-GAN learning objectives, for instance, would allow expressing a more general form of the Cycle-GAN loss, which induces different trade-offs between the zero-avoiding and zero-forcing generative characteristics. Furthermore, other popular models presented in the literature of unsupervised representation learning such as VAE/GAN (Larsen et al., 2015) consistently fit in the proposed framework as a weighted sum of terms considered in this analysis.

Other possible extensions address the inclusion of different constraints on the mutual information. The proposed generalized objective consents to maximize or minimize the amount of information represented in the latent codes. However, recent work (Alemi et al., 2018; Phuong et al., 2018) demonstrates that the achievement of a target level of mutual information can be beneficial for the quality of the latent representations. These approaches may consistently be integrated with the framework suggested in this work by simply updating the constraint imposed on the mutual information. Additionally, the refined mutual information estimation presented in this thesis suggests a possible alternative strategy to directly control the flow of information by itself.

Another direction of investigation involves the use of different divergences to match the joint distributions. Thanks to recent advancement in the field of likelihood-free approaches, an increasing number of algorithms propose new strategies to approximately minimize wider families of divergences (Nowozin, Cseke, and Tomioka, 2016; Regli and Silva, 2018). The use of alternatives to the Jensen-Shannon and Kullback-Leibler approximations considered in this work can be beneficial from multiple perspectives. In fact, the additional control on the loss function may result in increased training stability and more effective gradient estimations on multi-dimensional domains.

Lastly, future research could address extensions of the proposed generalized loss to create consistent embeddings in multi-modal and semi-supervised settings. The improved understanding of different objectives, together with the diagnostic estimations proposed in this work, result in better comprehension of the interaction between the joint parametric distributions. This additional knowledge allows to study the problem of matching multiple probability distributions, which encode either full or partial observations, at the same time. Since the goal of creating a joint representation for different modalities requires an in-depth understanding of the interaction between the data and the latent codes, this proposed extension represents a possible interesting path for future explorations.

Bibliography

Akata, Zeynep et al. (2013). "Label-Embedding for Attribute-Based Classification." In: *CVPR*. IEEE Computer Society, pp. 819–826. ISBN: 978-0-7695-4989-7. URL: http://dblp.uni-trier.de/db/conf/cvpr/cvpr2013.html#AkataPHS13.

Alemi, Alex et al. (2018). *An information-theoretic analysis of deep latent-variable models*. URL: https://openreview.net/forum?id=H1rRWl-Cb.

Arjovsky, Martin, Soumith Chintala, and Léon Bottou (2017). "Wasserstein Generative Adversarial Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia: PMLR, pp. 214–223. URL: http://proceedings.mlr.press/v70/arjovsky17a.html.

Barber, David and Felix Agakov (2003). "The IM Algorithm: A Variational Approach to Information Maximization". In: *Proceedings of the 16th International Conference on Neural Information Processing Systems*. NIPS'03. Whistler, British Columbia, Canada: MIT Press, pp. 201–208. URL: http://dl.acm.org/citation.cfm?id=2981345.2981371.

Belghazi, Mohamed Ishmael et al. (2018). "Adversarially Learned Inference". In: pp. 1–18. ISSN: 1097-6256. DOI: 10.1371/journal.pcbi.1005045. arXiv: 1802.01071. URL: http://arxiv.org/abs/1802.01071.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent (2013). "Representation Learning: A Review and New Perspectives". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 35.8, pp. 1798–1828. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2013.50. URL: http://dx.doi.org/10.1109/TPAMI.2013.50.

Bowman, Samuel R. et al. (2015). "Generating Sentences from a Continuous Space". In: DOI: 10.18653/v1/K16-1002. arXiv: 1511.06349. URL: http://arxiv.org/abs/1511.06349.

Burgess, Christopher P et al. (2017). "Understanding disentangling in β -VAE". In: Nips. arXiv: arXiv:1804.03599v1.

Chen, Xi et al. (2016). "Variational Lossy Autoencoder". In: pp. 1–17. arXiv: 1611.02731. URL: http://arxiv.org/abs/1611.02731.

— (2017). "Variational Lossy Autoencoder". In: pp. 1–17. arXiv: 1611.02731. URL: http://arxiv.org/abs/1611.02731.

Cover, Thomas M. and Joy A. Thomas (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. New York, NY, USA: Wiley-Interscience. ISBN: 0471241954.

Dziugaite, Gintare Karolina, Daniel M. Roy, and Zoubin Ghahramani (2015). "Training generative neural networks via Maximum Mean Discrepancy optimization". In: arXiv: 1505.03906. URL: http://arxiv.org/abs/1505.03906.

Gao, Hongyang et al. (2017). "Pixel Deconvolutional Networks". In: pp. 1–11. arXiv: 1705.06820. URL: http://arxiv.org/abs/1705.06820.

Goodfellow, Ian (2016). "NIPS 2016 Tutorial: Generative Adversarial Networks". In: ISSN: 0253-0465. DOI: 10.1001/jamainternmed.2016.8245. arXiv: 1701.00160. URL: http://arxiv.org/abs/1701.00160.

Goodfellow, Ian J. et al. (2014a). "Generative Adversarial Networks". In: pp. 1–9. ISSN: 10495258. DOI: 10.1001/jamainternmed.2016.8245. arXiv: 1406.2661. URL: http://arxiv.org/abs/1406.2661.

— (2014b). "Generative Adversarial Networks". In: pp. 1–9. ISSN: 10495258. DOI: 10.1001/jamainternmed.2016.8245. arXiv: 1406.2661. URL: http://arxiv.org/abs/1406.2661.

Higgins, Irina et al. (2017). "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: *Iclr* July, pp. 1–13. URL: https://openreview.net/forum?id=Sy2fzU9gl.

Hinton, Geoffrey et al. (2012). "Deep Neural Networks for Acoustic Modeling in Speech Recognition". In: *Signal Processing Magazine*.

Hinton, Geoffrey E. et al. (1995). "The wake-sleep algorithm for unsupervised neural networks". In: *Science* 268, pp. 1158–1161.

Hu, Zhiting et al. (2017). "On Unifying Deep Generative Models". In: pp. 1–20. arXiv: 1706.00550. URL: http://arxiv.org/abs/1706.00550.

Huszár, Ferenc (2017). "Variational Inference using Implicit Distributions". In: ISSN: 1702.08235. arXiv: 1702.08235. URL: http://arxiv.org/abs/1702.08235.

— (2018). *Goals and Principles of Representation Learning*. https://www.inference.vc/maximum-likelihood-for-representation-learning-2/. Blog.

Karaletsos, Theofanis (2016). "Adversarial Message Passing For Graphical Models". In: Nips. ISSN: 1612.05048. arXiv: 1612.05048. URL: http://arxiv.org/abs/1612.05048.

Kim, Taeksoo et al. (2017). "Learning to Discover Cross-Domain Relations with Generative Adversarial Networks". In: ISSN: 1938-7228. arXiv: 1703.05192. URL: http://arxiv.org/abs/1703.05192.

Kingma, Diederik P. and Jimmy Ba (2014). "Adam: A Method for Stochastic Optimization." In: *CoRR* abs/1412.6980. URL: http://dblp.uni-trier.de/db/journals/corr/corr1412.html#KingmaB14.

Kingma, Diederik P. and Max Welling (2013). "Auto-Encoding Variational Bayes." In: *CoRR* abs/1312.6114. URL: http://dblp.uni-trier.de/db/journals/corr/corr1312.html#KingmaW13.

Kingma, Diederik P. et al. (2014). "Semi-Supervised Learning with Deep Generative Models". In: pp. 1–9. ISSN: 10495258. arXiv: 1406.5298. URL: http://arxiv.org/abs/1406.5298.

Kingma, Diederik P et al. (2017). "Improved Variational Inference with Inverse Autoregressive Flow". In: Nips. arXiv: arXiv:1606.04934v2.

Krizhevsky, Alex (2010). *Convolutional deep belief networks on cifar-10*.

Larsen, Anders Boesen Lindbo et al. (2015). "Autoencoding beyond pixels using a learned similarity metric". In: ISSN: 1938-7228. arXiv: 1512.09300. URL: http://arxiv.org/abs/1512.09300.

LeCun, Yann et al. (1998). "Efficient BackProp". In: *Neural Networks: Tricks of the Trade, This Book is an Outgrowth of a 1996 NIPS Workshop*. London, UK, UK: Springer-Verlag, pp. 9–50. ISBN: 3-540-65311-2. URL: http://dl.acm.org/citation.cfm?id=645754.668382.

Ledig, Christian et al. (2017). "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 105–114. DOI: 10.1109/CVPR.2017.19. URL: https://doi.org/10.1109/CVPR.2017.19.

Li, Chunyuan et al. (2017). "ALICE: Towards Understanding Adversarial Learning for Joint Distribution Matching". In: Nips. ISSN: 10495258. arXiv: 1709.01215. URL: http://arxiv.org/abs/1709.01215.

Li, Yujia, Kevin Swersky, and Richard Zemel (2015). "Generative Moment Matching Networks". In: ISSN: 9781510810587. arXiv: 1502.02761. URL: http://arxiv.org/abs/1502.02761.

Liu, Ming-Yu, Thomas Breuel, and Jan Kautz (2017). "Unsupervised Image-to-Image Translation Networks". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., pp. 700–708. URL: http://papers.nips.cc/paper/6672-unsupervised-image-to-image-translation-networks.pdf.

Liu, Qiang and Dilin Wang (2016). "Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm". In: pp. 4–7. ISSN: 10495258. arXiv: 1608.04471. URL: http://arxiv.org/abs/1608.04471.

Makhzani, Alireza and Brendan Frey (2017). "PixelGAN Autoencoders". In: arXiv: 1706.00531. URL: http://arxiv.org/abs/1706.00531.

Makhzani, Alireza et al. (2014). "Adversarial Autoencoders". In: arXiv: arXiv:1511.05644v2.

Mescheder, Lars, Sebastian Nowozin, and Andreas Geiger (2017). "Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks". In: ISSN: 1938-7228. arXiv: 1701.04722. URL: http://arxiv.org/abs/1701.04722.

Miao, Yishu, Lei Yu, and Phil Blunsom (2016). "Neural Variational Inference for Text Processing". In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. ICML'16. New York, NY, USA: JMLR.org, pp. 1727–1736. URL: http://dl.acm.org/citation.cfm?id=3045390.3045573.

Mikolov, Tomas et al. (2013). "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges et al. Curran Associates, Inc., pp. 3111–3119. URL: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Mohamed, Shakir and Balaji Lakshminarayanan (2016). "Learning in Implicit Generative Models". In: arXiv: 1610.03483. URL: http://arxiv.org/abs/1610.03483.

Murphy, Kevin P. (2013). *Machine learning : a probabilistic perspective*. 1st ed. MIT Press. ISBN: 0262018020. URL: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/0262018020.

Nowozin, Sebastian, Botond Cseke, and Ryota Tomioka (2016). "f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization". In: ISSN: 10495258. arXiv: 1606.00709. URL: http://arxiv.org/abs/1606.00709.

Oord, Aaron van den et al. (2016). "Conditional Image Generation with PixelCNN Decoders". In: ISSN: 10495258. arXiv: 1606.05328. URL: http://arxiv.org/abs/1606.05328.

Passos, Alexandre et al. (2012). "Flexible Modeling of Latent Task Structures in Multitask Learning". In: *Proceedings of the 29th International Coference on International Conference on Machine Learning*. ICML'12. Edinburgh, Scotland: Omnipress, pp. 1283–1290. ISBN: 978-1-4503-1285-1. URL: http://dl.acm.org/citation.cfm?id=3042573.3042738.

Paszke, Adam et al. (2017). "Automatic differentiation in PyTorch". In:

Phuong, Mary et al. (2018). *The Mutual Autoencoder: Controlling Information in Latent Code Representations*. URL: https://openreview.net/forum?id=HkbmWqxCZ.

Pu, Yunchen et al. (2016). "Variational Autoencoder for Deep Learning of Images, Labels and Captions". In: Nips. ISSN: 10495258. DOI: 10.1109/ICCV.2017.245. arXiv: 1609.08976. URL: http://arxiv.org/abs/1609.08976.

Radford, A., L. Metz, and S. Chintala (2015). "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks". In: *ArXiv e-prints*. arXiv: 1511.06434.

Regli, J.-B. and R. Silva (2018). "Alpha-Beta Divergence For Variational Inference". In: *ArXiv e-prints*. arXiv: 1805.01045 [stat.ML].

Rezende, Danilo Jimenez and Shakir Mohamed (2016). "Variational Inference with Normalizing Flows". In: 37. ISSN: 1938-7228. arXiv: 1505.05770. URL: http://arxiv.org/abs/1505.05770.

Rosca, Mihaela et al. (2017). "Variational Approaches for Auto-Encoding Generative Adversarial Networks". In: arXiv: 1706.04987. URL: http://arxiv.org/abs/1706.04987.

Salimans, Tim et al. (2016). "Improved Techniques for Training GANs". In: pp. 1–10. ISSN: 09246495. DOI: arXiv:1504.01391. arXiv: 1606.03498. URL: http://arxiv.org/abs/1606.03498.

Salimans, Tim et al. (2017). "PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications". In: pp. 1–10. arXiv: arXiv:1701.05517v1.

Sugiyama, Masashi, Taiji Suzuki, and Takafumi Kanamori (2010). "Density Ratio Estimation: A Comprehensive Review". In: *Workshop on Statistical Experiment and its Related Topics* x, pp. 10–31.

— (2012). *Density-ratio matching under the Bregman divergence: A unified framework of density-ratio estimation*. Vol. 64. 5, pp. 1009–1044. ISBN: 1046301103438. DOI: 10.1007/s10463-011-0343-8.

Theis, Lucas and Matthias Bethge (2016). "Note on the evaluation of generative models". In: pp. 1–10. arXiv: arXiv:1511.01844v3.

Tiao, Louis C., Edwin V. Bonilla, and Fabio Ramos (2018). "Cycle-Consistent Adversarial Learning as Approximate Bayesian Inference". In: arXiv: 1806.01771. URL: http://arxiv.org/abs/1806.01771.

Tolstikhin, Ilya et al. (2018). "Wasserstein Auto-Encoders". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=HkL7n1-0b.

Tomczak, Jakub M. and Max Welling (2017). "VAE with a VampPrior". In: arXiv: 1705.07120. URL: http://arxiv.org/abs/1705.07120.

Tran, Dustin, Rajesh Ranganath, and David M. Blei (2017). "Hierarchical Implicit Models and Likelihood-Free Variational Inference". In: Nips. ISSN: 1702.08896. arXiv: 1702.08896. URL: http://arxiv.org/abs/1702.08896.

Villani, C. (2003). *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society. ISBN: 9780821833124. URL: https://books.google.it/books?id=GqRXYFxe0l0C.

Yu, Y. et al. (2017). "Zero-Shot Learning via Latent Space Encoding". In: *ArXiv e-prints*. arXiv: 1712.09300 [cs.CV].

Zhao, Shengjia and Stefano Ermon (2017). "The Information-Autoencoding Family : A Lagrangian Perspective on Latent Variable Generative Modeling". In: *NIPS 2017 Bayesian Deep Learning Workshop* Nips. arXiv: 1806.06514. URL: http://bayesiandeeplearning.org/2017/papers/60.pdf.

Zhao, Shengjia, Jiaming Song, and Stefano Ermon (2017a). "InfoVAE: Information Maximizing Variational Autoencoders". In: arXiv: 1706.02262. URL: http://arxiv.org/abs/1706.02262.

— (2017b). "Towards Deeper Understanding of Variational Autoencoding Models". In: arXiv: 1702.08658. URL: http://arxiv.org/abs/1702.08658.

Zhou, Tinghui et al. (2016). "Learning Dense Correspondence via 3D-guided Cycle Consistency". In: *Computer Vision and Pattern Recognition (CVPR)*.

Zhu, Jun Yan et al. (2017). "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks". In: *Proceedings of the IEEE International Conference on Computer Vision* 2017-Octob, pp. 2242–2251. ISSN: 15505499. DOI: 10.1109/ICCV.2017.244. arXiv: 1703.10593.

Appendix A

Information Theoretical Quantities

This section aims to introduce different information theoretical quantities, their interpretation and properties.

A.1 Entropy and mutual information

Given a distribution $p(X, Z)$ defined over $\mathcal{X} \times \mathcal{Z}$ the fundamental quantity that can be considered is represented by the **entropy** $H(p(X))$:

$$H(p(X)) = -\sum_{x \in \mathcal{X}} p(X = x) \log p(X = x) \tag{A.1}$$

The entropy can be interpreted as the number of nats (or bits if the logarithm with base 2 is used) that are needed to identify a specific $x \in \mathcal{X}$. For any discrete random variable, the entropy is a positive quantity. Its maximum value is achieved by the uniform distribution over the elements of $\mathcal{X}$, while the minimum is given by a delta distribution that is assigning all the probability to one element. Therefore, defining $u(X)$ as the uniform distribution over $\mathcal{X}$ and $\delta_{\hat{x}}(X)$ as the delta distribution that collapses all the probability mass into one element $x \in \mathcal{X}$ we have that for every $p(X)$:

$$0 = H(\delta_{\hat{x}}(X)) \leq H(p(X)) \leq H(u(X)) = \log |\mathcal{X}| \tag{A.2}$$

Analogously, for every $p(X, Z)$ we can define the **joint entropy** $H(p(X, Z))$ as the entropy of the joint distribution of X and Z:

$$H(p(X, Z)) = -\sum_{x \in \mathcal{X}} \sum_{z \in \mathcal{Z}} p(X = x, Z = z) \log p(X = x, Z = z) \tag{A.3}$$

$H(p(X, Z))$ represents the average number of nats required to describe a joint occurrence of some $x \in \mathcal{X}$ and $z \in \mathcal{Z}$. For any distribution $p(X, Z) \in S(\mathcal{X} \times \mathcal{Z})$, the entropy of the joint distributions at least as big as the maximum between the entropy of the marginals and upper bounded by their sum:

$$\max\{H(p(X)), H(p(Z))\} \leq H(p(X, Z)) \leq H(p(X)) + H(p(Z)) \tag{A.4}$$

When X carries information about Z, the amount of nats required to describe the joint occurrence is less then the joint length of their descriptions. The amount of information that we save by using a joint description instead of describing the events separately is defined as the **mutual information** $I(p(X, Z))$ between X and Z according to p:

$$I(p(X, Z)) = H(p(X)) + H(p(Z)) - H(p(X, Z)) \tag{A.5}$$

By considering the bounds reported in equation A.4, one may notice that the mutual information is always positive and upper bounded the minimum between the two marginal entropies:

$$0 \leq I(p(X,Z)) \leq \min\{H(p(X)), H(p(Z))\} \tag{A.6}$$

Note that $I(p(X,Z)) = 0$ is achieved when knowing the value of X gives us no information about Z (and vice-versa). This occurs when the two events are independent or, equivalently, when the joint distribution factorizes as the product of the two marginals:

$$I(p(X,Z)) = 0 \iff p(X,Z) = p(X)p(Z) \tag{A.7}$$

Another quantity of interest is the amount of uncertainty associated to X once the value of Z is given. This quantity, called **conditional entropy** $H(p(X|Z))$, can be expressed as the difference between the entropy of that variable and the mutual information $I(p(X,Z))$:

$$H(p(X|Z)) = H(p(X)) - I(p(X,Z)) \tag{A.8}$$

By considering the bounds for the mutual information reported in equation A.6, we know that the conditional information is always positive and upper-bounded by the entropy:

$$0 \leq H(p(X|Z)) \leq H(X) \tag{A.9}$$

In particular, if the conditional information is zero, Z expressed full information about X. This means that when value of Z is observed, it is possible to uniquely identify the corresponding X. Therefore, there exists a deterministic function $f : \mathcal{Z} \to \mathcal{X}$ that maps from each $z \in \mathcal{Z}$ to its corresponding $x \in \mathcal{X}$:

$$H(p(X|Z)) = 0 \iff \exists f : \mathcal{Z} \to \mathcal{X}.p(X|Z=z) = \delta_{f(z)}(X) \tag{A.10}$$

The relations described in this section can be intuitively represented by the entropy diagram reported in figure A.1.

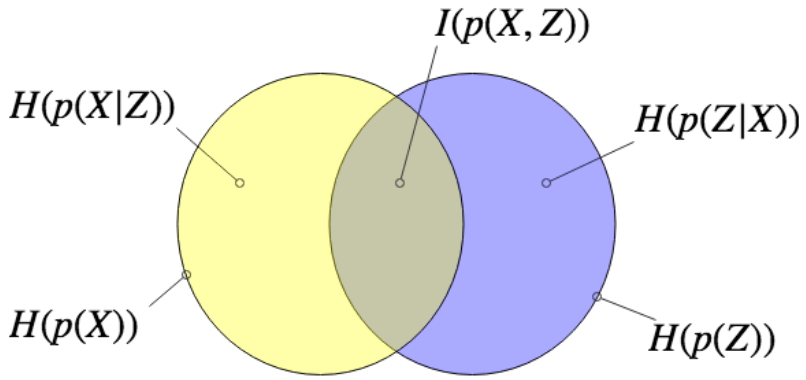

FIGURE A.1: Visualization of the entropy diagram for the random variables X and Z. The different areas correspond to the additive relationships that characterize the entropies, conditional entropies and mutual information.

A.2 Cross-entropy

All the quantities described up to this point, refer to one probability distribution, but other quantities of interest consider the interaction between different probability densities defined on the same support. In particular the **cross-entropy** $CE(p(X)||q(X))$ between the distributions $p(X)$ and $q(X)$ is defined as the average amount of nats that are required to describe events drawn according to $p(X)$ with a coding scheme that has been optimized according to $q(X)$:

$$CE(p(X)||q(X)) = -\sum_{x\in\mathcal{X}} p(X=x)\log q(X=x) \tag{A.11}$$

Clearly, since the coding scheme is sub-optimal, the cross-entropy is always bigger then the corresponding entropy:

$$0 \leq H(p(X)) \leq CE(p(X)||q(X)) \tag{A.12}$$

Analogously, we can define the **joint cross-entropy** $CE(p(X,Z)||q(X,Z))$ as the cross entropy between the joint distribution $p(X,Z)$ and $q(X,Z)$ as:

$$CE(p(X,Z)||q(X,Z)) = -\sum_{x\in\mathcal{X}}\sum_{z\in\mathcal{Z}} p(X=x,Z=z)\log q(X=x,Z=z) \tag{A.13}$$

And the **conditional cross entropy** $CE(p(X,Z)||q(X|Z))$ as the difference between the joint cross-entropy and the cross-entropy of one of the marginals:

$$CE(p(X,Z)||q(X|Z)) = CE(p(X,Z)||q(X,Z)) - CE(p(Z)||q(Z)) \tag{A.14}$$

Note that every cross-entropy can be interpreted as the corresponding entropy evaluated using a sub-optimal coding scheme defined by another distribution.

A.3 Kullback-Leibler divergence

The last quantity we will consider is the **Kullback Leibler-Divergence** $KL(p(X)||q(X))$ (KL-divergence in short) defined on two distributions $p(X)$ and $q(X)$ with the same support. It represents the overhead introduced by the use of a sub-optimal coding scheme defined by $q(X)$ to encode events distributed according to $p(X)$ and it can be obtained as the difference between the cost of using the coding scheme defined for $q(X)$ (which is represented by the cross-entropy) and the cost corresponding to the optimal coding scheme (which is represented by the entropy):

$$KL(p(X)||q(X)) = CE(p(X)||q(X)) - H(p(X)) \tag{A.15}$$

From equation A.12 we know that the cross-entropy is always bigger than the corresponding entropy ,therefore, we infer that the KL-divergence is always positive:

$$0 \leq KL(p(X)||q(X)) \tag{A.16}$$

In particular, the equality is achieved if only $p(X)$ perfectly matches $q(X)$:

$$KL(p(X)||q(X)) = 0 \iff p(X) = q(X) \tag{A.17}$$

Note that properties A.16 and A.17 suffice to demonstrate that the Kullback-Leibler divergence is effectively a divergence measure between probability distributions.

it is possible ton extend the definition of the Kullback-Leibler divergence to cover the joint and conditional scenarios as follows:

$$KL(p(X,Z)||q(X,Z)) = CE(p(X,Z)||q(X,Z)) - H(p(X,Z)) \tag{A.18}$$
$$KL(p(X|Z)||q(X|Z)) = CE(p(X,Z)||q(X|Z)) - H(p(X|Z)) \tag{A.19}$$
$$KL(p(X|Z)||q(X)) = CE(p(X,Z)||q(X)) - H(p(X|Z)) \tag{A.20}$$

Note that using the definition reported in equation A.14, the following chain decomposition rule holds:

$$KL(p(X,Z)||q(X,Z)) = KL(p(X)||q(X)) + KL(p(Z|X)||q(Z|X)) \tag{A.21}$$

Furthermore, because of the bound reported in equation A.9, the following bound holds:

$$KL(p(X)||q(X)) \leq KL(p(X|Z)||q(X)) \tag{A.22}$$

The difference between the two divergences reported in equation A.22 exactly represent the mutual information between X and Z according to $p(X,Z)$:

$$KL(p(X|Z)||q(X)) - KL(p(X)||q(X)) = I(p(X,Z)) \tag{A.23}$$

From equation A.22 and A.23, we infer not only that the mutual information is always positive, but that it can be expressed as a KL-divergence itself:

$$I(p(X,Z)) = KL(p(X,Z)||p(X)p(Z)) \tag{A.24}$$

Therefore, the mutual information can be considered as the overhead payed by encoding joint occurrences of X and Z by concatenating two optimal coding schemes (one optimized for $p(X)$ and one for $p(Z)$) instead of using a single scheme optimized for their joint occurrences.

A.4 Notes on differential entropy and cross-entropy

The definition reported in section A refer to the case in which both X and Z are discrete random variables. When X and Z are continuous, the values of entropy, cross-entropy and Kullback-Leibler divergence can be computed by replacing the summations with integrals over continuous domain. However, the values of the entropy and the cross entropy are not bounded to be positive and are referred to as **differential entropy**, $h(p(X))$ and **differential cross entropy**, $ce(p(X))$. Nevertheless, it is possible to extend the definition and properties reported in the previous section to continuous random variables by considering their discretization into bins of size Δ.

Given a value of differential entropy, $h(p(X))$, the corresponding entropy can be obtained by adjusting $h(p(X))$ with a constant that represents the number of bits (or nats) that are necessary to specify the bins used for the discretization (Cover and

Thomas, 2006, Theorem 8.3.1). More formally, defining X^Δ as the discretization of X, we have:

$$H(p(X^\Delta)) \approx \underbrace{-\int_{\mathcal{X}} p(X=x)\log p(X=x)dx}_{h(X)} + n \tag{A.25}$$

Where n represents the logarithm of the number of bins used for the discretization. The same discretization approach can be applied to discrete cross entropy:

$$CE(p(X^\Delta)||q(X^\Delta)) \approx \underbrace{-\int_{\mathcal{X}} p(X=x)\log q(X=x)dx}_{ce(p(X)||q(X))} + n \tag{A.26}$$

Note that the computation of the Kullback-Leibler divergence does not require the discretization reported in equation A.25 and A.26:

$$\begin{aligned} KL(p(X^\Delta)||q(X^\Delta)) &= CE(p(X^\Delta)||q(X^\Delta)) - H(p(X^\Delta)) \\ &= ce(p(X)||q(X^\Delta)) + n - h(p(X)) - n \\ &= KL(p(X)||q(X)) \end{aligned} \tag{A.27}$$

In order to simplify the notation used in this work, we omit the explicit discretization of the random variables by referring to $H(p(X^\Delta))$ directly with $H(p(X))$ even when X is continuous. Since the value of n is constant and depends on the numerical precision used to represent the floating point numbers, in the scope of this thesis, it will be considered as a simple translation constant that ensures the positivity of the considered quantities.

Appendix B

Interpolating the I-projection and M-projection

As discussed in section 2.1, the minimization of the two directions of the Kullback-Leibler divergence leads to projections with different characteristics. Since in general, we are interested in both the zero-avoiding and zero-forcing properties, we could consider balancing the them by considering the minimization of the convex combination of the two directions of the KL-divergence. Defining a parameter $\gamma \in [0, 1]$, we can build an interpolation $KL^\gamma(q(X)||p(X))$:

$$KL^\gamma(q(X)||p(X)) := \gamma KL(q(X)||p(X)) + (1-\gamma)KL(p(X)||q(X)) \tag{B.1}$$

Since the weighted sum of divergences is still a valid divergence, given a set $\mathcal{C} \subseteq S(\mathcal{X})$, we can define a γ-projection $p^\gamma_\mathcal{C}(X)$ of a distribution $p(X)$ onto $\mathcal{C}$ as:

$$p^\gamma_\mathcal{C}(X) := \operatorname*{argmin}_{q(X)\in\mathcal{C}} KL^\gamma(q(X)||p(X)) \qquad \gamma \in [0,1] \tag{B.2}$$

The parameter γ induces a trajectory in $\mathcal{C}$ that interpolates between the I-projection ($p^{\gamma=1}_{\mathcal{C}_\mathcal{N}}(X)$) and M-projection ($p^{\gamma=0}_{\mathcal{C}_\mathcal{N}}(X)$) by mediating between the zero-avoiding and zero-forcing behaviors.

Figure B.1 reports the projections of the same mixture distribution considered in Figure 2.1 for different values of γ, visualizing the trajectory in the parameter space.

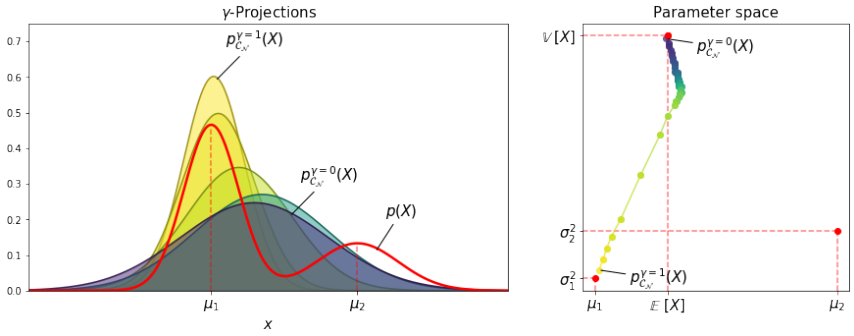

FIGURE B.1: γ-projections of the mixture distribution $p(X)$ obtained for different values of γ onto the set of Normal distributions.

Note that for $\gamma = 1/2$ the divergence $KL^\gamma(q(X)||p(X))$ is symmetric and is usually referred to as "symmetrized Kullback-Leibler divergence".

The convex combination proposed in equation B.2 is not the only way to interpolate between the two directions of the KL-divergence. In fact, by considering $m_\lambda(X)$ as the mixture of $q(X)$ and $p(X)$ according to a coefficient $\lambda \in [0, 1]$ we can define:

$$KL^\lambda(q(X)||p(X)) := \lambda KL(q(X)||m_\lambda(X)) + (1-\lambda)KL(p(X)||m_\lambda(X)) \tag{B.3}$$

With

$$m_\lambda(X) = (1-\lambda)q(X) + \lambda p(X) \tag{B.4}$$

Once again, for $\lambda = 1$ we have that $KL^\lambda(q(X)||p(X))$ coincides with $KL(q(X)||p(X))$, and when $\lambda = 0$ the interpolation gives $KL(p(X)||q(X))$. In this scenario, when $\lambda = 1/2$ the divergence is symmetric and represents the Jensen-Shannon divergence between $q(X)$ and $p(X)$.

Even if both the symmetrized Kullback-Leibler divergence and the Jensen-Shannon divergence are both symmetric, they balance the zero-avoiding and zero-forcing characteristics in different ways. In fact, since the Jensen-Shannon divergence considers the sum of the divergence from $p(X)$ and $q(X)$ to their mixture, the zero-forcing and zero-avoiding properties are manifested with less strength.

Appendix C

The Density-ratio Trick

We consider the ratio between a two distributions $p(X)$ and $q(X)$ evaluated for $x \in \mathcal{X}$:

$$ratio(x) := \frac{p(X = x)}{q(X = x)} \tag{C.1}$$

One can define a binary random variable Y and a distribution $s(X, Y)$ such that if $Y = 1$ then $s(X|Y = 1)$ represents the distribution $p(X)$, otherwise it models $q(X = x)$:

$$s(X|Y = 0) := p(X) \tag{C.2}$$
$$s(X|Y = 1) := q(X) \tag{C.3}$$

Using the Bayes rule and fixing $s(Y = 1) = s(Y = 0)$, we can re-write equation C.1 as:

$$\begin{aligned} ratio(x) &= \frac{s(X = x|Y = 0)}{s(X = x|Y = 1)} \\ &= \frac{s(Y = 0|X = x)s(X = x)}{s(Y = 0)} \Big/ \frac{s(Y = 1|X = x)s(X = x)}{s(Y = 1)} \\ &= \frac{s(Y = 0|X = x)}{1 - s(Y = 0|X = x)} \end{aligned} \tag{C.4}$$

By replacing $s(Y = 0|X = x)$ with an approximate binary classifier $\tilde{s}(Y = 0|X = x)$, we can compute the ratio between the two distributions:

$$ratio(x) \approx \frac{\tilde{s}(Y = 0|X = x)}{1 - \tilde{s}(Y = 0|X = x)} \tag{C.5}$$

C.1 Estimating the Kullback-Leibler divergence

The KL-divergence between two brobability distribution $p(X)$ and $q(X)$:

$$KL(p(X)||q(X)) = \mathbb{E}_{x \sim p(X)} \left[\log ratio(x)\right] \tag{C.6}$$

Replacing the expression of the ratio in the KL-divergence with the one reported in equation C.5 we obtain:

$$KL(p(X)||q(X)) \approx \mathbb{E}_{x \sim p(X)} \left[\log \frac{\tilde{s}(Y = 0|X = x)}{1 - \tilde{s}(Y = 0|X = x)}\right] \tag{C.7}$$

Note that since we are not considering the original classifier, the distribution that we are considering is not exactly $q(X)$ but rather an approximation determined by $\tilde{s}(Y=0|X)$ and $p(X)$:

$$q(X=x) = \frac{p(X=x)}{ratio(x)} \approx p(X=x)\frac{1-\tilde{s}(Y=0|X=x)}{\tilde{s}(Y=0|X=x)} \tag{C.8}$$

The closer $\tilde{s}(Y=0|X)$ is to $s(Y=0|X=x)$, the closer the expression on the right hand side of equation C.8 resembles $q(X)$.

Note that the expression of the ratio also easily allows for the computation of the opposite direction of the Kullback-Leibler divergence:

$$\begin{aligned} KL(q(X)||p(X)) &= \mathbb{E}_{x\sim q(X)}\left[\log\frac{1}{ratio(x)}\right] \\ &\approx \mathbb{E}_{x\sim q(X)}\left[\log\frac{1-\tilde{s}(Y=0|X=x)}{\tilde{s}(Y=0|X=x)}\right] \end{aligned} \tag{C.9}$$

C.2 Estimating the Jensen-Shannon divergence

Other then being useful for the estimation of the Kullback-Leibler divergence, the approximation between the ratio of $p(X)$ and $q(X)$ can be used to approximate the Jensen-Shannon divergence.

Considering $x \in \mathcal{X}$ the probability that x has been drawn from $p(X)$, represented by $s(Y=0|X=x)$, is given by the probability of observing x according to $p(X)$ divided by the the total probability of x:

$$s(Y=0|X=x) = \frac{p(X=x)}{p(X=x)+q(X=x)} \tag{C.10}$$

The Jensen-Shannon divergence, can then be expressed as a function of the binary classifier $s(Y|X=x)$:

$$\begin{aligned} JS(p(X)||q(X)) &= \frac{1}{2}\mathbb{E}_{x\sim p(X)}\left[\log\frac{p(X=x)}{\frac{1}{2}p(X=x)+\frac{1}{2}q(X=x)}\right] \\ &\quad + \frac{1}{2}\mathbb{E}_{x\sim q(X)}\left[\log\frac{q(X=x)}{\frac{1}{2}p(X=x)+\frac{1}{2}q(X=x)}\right] \\ &= \frac{1}{2}\mathbb{E}_{x\sim p(X)}\left[\log s(Y=0|X=x)\right] \\ &\quad + \frac{1}{2}\mathbb{E}_{x\sim q(X)}\left[\log 1 - s(Y=0|X=x)\right] + \log 2 \end{aligned} \tag{C.11}$$

Replacing the optimal classifier with its approximation, we obtain:

$$\begin{aligned} JS(p(X)||q(X)) &\approx \frac{1}{2}\mathbb{E}_{x\sim p(X)}\left[\log \tilde{s}(Y=0|X=x)\right] \\ &\quad + \frac{1}{2}\mathbb{E}_{x\sim q(X)}\left[\log 1 - \tilde{s}(Y=0|X=x)\right] + \log 2 \end{aligned} \tag{C.12}$$

C.3 Modeling the approximate classifier

The binary classifier $\tilde{s}(Y = 0|X)$ is generally modeled as a parametric distribution and it is usually referred to as "**discriminator**". By fixing a set $\mathcal{C}_\Psi$ of binary classifiers that can be represented by some parameters $\psi \in \Psi$, the discriminator is selected to be the minimizer of the cross-entropy with the conditional distribution $s(Y|X = x)$:

$$\hat{\psi} = \underset{\psi \in \Psi}{\operatorname{argmin}} \, CE(s(Y|X)||\tilde{s}_\psi(Y|X)) \tag{C.13}$$

Note that since $s(Y|X)$ is fixed, the minimization of the cross-entropy is equivalent to the minimization of the KL-divergence $KL(s(Y|X)||\tilde{s}_\psi(Y|X))$. For this reason, the approximate discriminator $\tilde{s}_\psi(Y|X)$ can be seen as the M-projection of $s(Y|X)$ onto $\mathcal{C}_\psi$.

Appendix D

Modeling the Parametric Distributions

D.1 The encoding distribution

The parametric conditional distribution $q_\theta(Z|X)$ can be defined as the **encoding** distribution since, in the context of latent variable models, it represents a mapping from the data domain $\mathcal{X}$ to the latent representations $\mathcal{Z}$. In order to chose a model for the encoding distribution, we are interested in the following characteristics:

- Flexible: since we want to match $p_\phi(X, Z)$ and $q_\theta(X, Z)$, increasing the number of representable distributions also allows to find a better minimum for $D(p_\theta(X, Z)||q_\phi(X, Z))$
- Easy to evaluate: most of the choices for $D(\cdot||\cdot)$ require to evaluate the value $q_\theta(Z = z|X = x)$. Furthermore, the differentiability with respect to θ allows for training procedures that involve the use of Stochastic Gradient Descent.
- Simple sampling procedure: the possibility of quickly generating samples is fundamental to compute Monte-Carlo approximations that are required for the computation of $D(\cdot||\cdot)$. A sampling procedure that allows for a re-parametrization trick that results in gradient for θ is preferable.

For this reason, one of most common choices for the modeling of $q_\phi(Z|X)$ consists on the use of a neural network that maps from an element of the input domain $\mathcal{X}$ and a weight configuration $\theta \in \Theta$ to the natural parameters $\tau \in T$ of some specified probability distribution:

$$f : \mathcal{X} \times \Theta \to T \tag{D.1}$$

$$q_\theta(Z|X = x) = q(Z|f(x, \theta)) \tag{D.2}$$

Whenever the domain $\mathcal{Z}$ is multi-dimensional ($\mathcal{Z} = \mathcal{Z}_1 \times \ldots \times \mathcal{Z}_d$), the encoding distribution is often modeled by assuming the conditional independence of the components to avoid the necessity of explicitly modeling the correlation between the different variables:

$$q_\theta(Z|X = x) = \prod_{i=1}^{d} q_\theta(Z_i|X = x) \tag{D.3}$$

D.2 The decoding distribution

The distribution $p_\phi(X|Z)$ plays a complementary role when compared to $q_\theta(Z|X)$, by mapping the latent representations to their corresponding interpretations in the

original space data domain $\mathcal{X}$. For this reason, the **decoding** distribution has similar requirements to the ones specified by the encoding conditional distribution.

A neural network g maps from values $z \in \mathcal{Z}$ and a weight configuration $\phi \in \Phi$ to the parameters $\kappa \in K$ of some chosen distribution defined on the $\mathcal{X}$ space.

$$g : \mathcal{Z} \times \Phi \to K \tag{D.4}$$

$$p_\phi(X|Z = z) = p(X|g(z, \phi)) \tag{D.5}$$

The literature presents two different options:

- The components of X are conditionally independent:
  for multi-dimensional $\mathcal{X}$ spaces, the probability $p_\phi(X = x|Z = z)$ factorizes as the product of the components. This option restricts the flexibility of the decoding distribution, potentially resulting in worse approximations when the data exhibits strong correlation between the dimensions of $\mathcal{X}$.

- The components of $\mathcal{X}$ are dependent:
  in this scenario, instead of modeling g as a simple feed-forward network, convolutional or recursive neural network are used to sequentially capture and represent the correlation between the components of X (Oord et al., 2016; Salimans et al., 2017).

Appendix E

Estimation of the Information Theoretical Quantities

E.1 Entropy

Since the distribution $q(X)$ and $p(Z)$ are fixed, the entropy $H(q(X))$ and $H(p(Z))$ are constant. In particular, the entropy of the empirical is determined by the size of the dataset.

$$H(q(X)) = \log |\mathcal{D}| - \frac{\sum_{x \in \mathcal{D}^*} n_x \log n_x}{|\mathcal{D}|} \tag{E.1}$$

$$\text{width } n_x := |\left\{x' \in \mathcal{D} | x = x'\right\}| \tag{E.2}$$

On the other hand, the entropy of the prior distribution can be either computed analytically or estimated through a Monte Carlo approximation:

$$H(p(Z)) \approx -\frac{1}{M} \sum_{i=1}^{M} \log p(Z = z_i) \qquad \text{with } z_i \sim p(Z) \tag{E.3}$$

Where M represents the number of samples utilized for the estimation.

E.2 Data distortion and code distortion

Both the data and the code distortion can be effectively estimated using a Monte Carlo approximation. Note that the two quantities requires to compute a double expectation. Fixing the amount of samples M_1 and M_2, the distortions can be estimated as:

$$D_X(\theta, \phi) \approx -\sum_{i=1}^{M_1} \sum_{j=1}^{M_2} \frac{\log p_\phi(X = x_i | Z = z_{ij})}{M_1 M_2} \quad \text{with } x_i \sim q(X), z_{ij} \sim q_\theta(Z | X = x_i) \tag{E.4}$$

$$D_Z(\theta, \phi) \approx -\sum_{i=1}^{M_1} \sum_{j=1}^{M_2} \frac{\log q_\theta(Z = z_i | X = x_{ij})}{M_1 M_2} \quad \text{with } z_i \sim p(Z), x_{ij} \sim p_\phi(X | Z = z_i) \tag{E.5}$$

Since the computational complexity of the procedure increases with the product of the number of samples $O(M_1 M_2)$, both approximation are computed by fixing $M_2 = 1$. This choice is commonly accepted in literature (Kingma and Welling, 2013) and it can be justified by the fact that $q_\theta(Z|X)$ and $p_\phi(X|Z)$ are usually unimodal distributions with small variance.

Note that when the encoding or decoding distributions are continuous, one may

add a fixed constant to approximately obtain their discretized versions (see Appendix A.4)

E.3 Data rate and the code rate

The code-rate $R_Z(\theta)$ can be effectively estimated using a Monte Carlo approximation:

$$R_Z(\theta) \approx \frac{1}{M}\sum_{i=1}^{M} KL(q_\theta(Z|X=x_i)||p(Z)) \qquad \text{with } x_i \sim q(X) \tag{E.6}$$

Since both the code distribution $p(Z)$ and the conditional posterior $q_\theta(Z|X)$ have a closed form expression, it is usually possible to analytically compute the value of the KL-divergence $KL(q_\theta(Z|X=x_i)||p(Z))$. In practice, the estimation is consistent even if the divergence in equation E.6 is estimated using Monte Carlo sampling:

$$R_Z(\theta) \approx \sum_{i=1}^{M_1}\sum_{j=1}^{M_2} \frac{\log q_\theta(Z=z_{ij}|X=x_i) - \log p(Z=z_{ij})}{M_1 M_2} \tag{E.7}$$
$$\text{with } x_i \sim q(X), z_{ij} \sim q_\theta(Z|X=x_i)$$

Analogously to the computation of the distortion, M_2 is usually fixed to one. If the computation of the code rate is straightforward, there is no effective way of computing the data rate $R_X(\phi)$. This is because it requires the computation of the density of the empirical distribution in regions that are not observed.

E.4 Kullback-Leibler divergence

The value of the Kullback-Leibler divergence between encoding and decoding joint distributions can be computed directly by considering the estimations for the data distortion, code rate and empirical entropy:

$$KL(q_\theta(X,Z)||p_\phi(X,Z)) = D_X(\theta,\phi) + R_Z(\theta) - H(q(X)) \tag{E.8}$$

The value of the opposite direction of the KL-divergence is intractable due to the lack of an approximation for the data rate.

Even if the value of the Kullback-Leibler divergence between the marginal distributions can not be computed directly, the density-ratio trick presents an indirect way to obtain a rough estimation (equation C.7).

$$KL(q_\theta(Z)||p(Z)) \approx \frac{1}{M}\sum_{i=1}^{M} \log \frac{\tilde{s}(Y=0|Z=z_i)}{1-\tilde{s}(Y=0|Z=z_i)} \qquad \text{with } z_i \sim q_\theta(Z) \tag{E.9}$$

Where M represents the number of samples used for the Monte Carlo estimation and $\tilde{s}(Y|Z)$ refers to the binary classifier trained to distinguish the source of code samples. Considering that the KL-divergence between the aggregated posterior and the prior can be expressed as a difference between the code rate and the mutual information (equation 3.9), we can define an upper-and a lower bound for $KL(q_\theta(Z)||p(Z))$:

$$R_Z(\theta) - \overline{I}(q_\theta(X,Z)) \leq KL(q_\theta(Z)||p(Z)) \leq R_Z(\theta) - \underline{I}(q_\theta(X,Z)) \tag{E.10}$$

Where $\overline{I}(q_\theta(X,Z))$ and $\underline{I}(q_\theta(X,Z))$ represent the upper bound and a lower bound of the encoding mutual information respectively. The two bounds reported in equation E.10 allow to correct the estimation for the divergence between the aggregated posterior and the prior described by equation E.9. Since we have access to both an estimation of $KL(q_\theta(X,Z)||p_\phi(X,Z))$ and $KL(q_\theta(Z)||p(Z))$ one can approximate the value of the divergence between the two conditional distributions $q_\theta(X|Z)$ and $p_\phi(X|Z)$ by considering their difference (equation A.21)

$$KL(q_\theta(X|Z)||p_\phi(X|Z)) = KL(q_\theta(X,Z)||p_\phi(X,Z)) - KL(q_\theta(Z)||p(Z)) \tag{E.11}$$

The strategy used for the estimation of the Kullback-Leibler divergence between the data marginal $p_\phi(X)$ and the empirical $q(X)$ is analogous. Considering a binary classifier $\tilde{s}(Y|X)$ and fixing a number of samples M, we have:

$$KL(p_\phi(X)||q(X)) \approx \frac{1}{M}\sum_{i=1}^{M} \log \frac{\tilde{s}(Y=0|X=x_i)}{1-\tilde{s}(Y=0|X=x_i)} \qquad \text{with } x_i \sim p_\phi(X) \tag{E.12}$$

Note that since we don not have access to any estimation of the data rate, the only bound that can be enforced to correct the estimation of $KL(p_\phi(X)||q(X))$ is represented by the positivity of the Kullback-Leibler divergence.

E.5 Mutual information

Before defining an estimation for the values of encoding and decoding mutual information we consider their respective upper and lower bounds. By considering equations 3.8 and 3.13, we can define the computation for the two lower-bounds:

$$\underline{I}(q_\theta(X,Z)) = \max(0, H(q(X)) - D_X(\theta,\phi)) \tag{E.13}$$
$$\underline{I}(p_\phi(X,Z)) = \max(0, H(p(Z)) - D_Z(\theta,\phi)) \tag{E.14}$$

The upper bounds of the two mutual information can be defined by considering that their value is constrained by both the entropy and the value of the code and data rates:

$$\overline{I}(q_\theta(X,Z)) = \min(H(q(X)), R_Z(\theta)) \tag{E.15}$$
$$\overline{I}(p_\phi(X,Z)) = H(p(Z)) \tag{E.16}$$

Where the difference in the formulation of the two upper bounds is given by the lack of an estimation for the data rate $R_X(\phi)$. Since we do have access to an estimation of both the code rate $R_Z(\theta,\phi)$ and the KL-divergence between the aggregated posterior and the prior, we can directly compute an estimation of the value of the encoding mutual information:

$$I(q_\theta(X,Z)) = R_Z(\theta,\phi) - KL(q_\theta(Z)||p(Z)) \tag{E.17}$$

E.6 Summary

The procedure for the computation of the information theoretical quantities described in this section is summarized by Algorithm 1. Note that the estimation for the measures of distortion assumes that both $q_\theta(Z|X)$ and $p_\phi(X|Z)$ are defined on a discrete domain. In the case in which the domain $\mathcal{X}$ and $\mathcal{Z}$ are continuous, the estimations can be adjusted by considering their approximate discretization as described in Appendix A.4.

Algorithm 1 Estimation of information theoretical quantities of interest

▷ Sample a batch of M elements from empirical and prior distributions

$x_1, .., x_M \sim q(X)$

$z_1, .., z_M \sim p(Z)$

for $i = 1, \ldots, M$ **do**

$z_{x_i} \sim q_\theta(Z|X = x_i)$ ▷ Sample the codes corresponding to the data batch

$x_{z_i} \sim p_\phi(X|Z = z_i)$ ▷ Sample the data corresponding to the code batch

▷ Data and code distortion

$D_X(\theta, \phi) \leftarrow -\frac{1}{M} \sum_{i=1}^{M} \log p_\phi(X = x_i|Z = z_{x_i})$

$D_Z(\theta, \phi) \leftarrow -\frac{1}{M} \sum_{i=1}^{M} \log q_\theta(Z = z_i|Z = x_{z_i})$

▷ Code rate

$R_Z(\theta) \leftarrow \frac{1}{M} \sum_{i=1}^{M} \log \frac{q_\theta(Z=z_{x_i}|X=x_i)}{p(Z=z_{x_i})}$

▷ KL-divergence between the joint distributions

$KL(q_\theta(X, Z)||p_\phi(X, Z)) \leftarrow D_X(\theta, \phi) + R_Z(\theta) - H(q(X))$

▷ Lower bound of the mutual information

$\underline{I}(q_\theta(X, Z)) \leftarrow \max(0, H(q(X)) - D_X(\theta, \phi))$

$\underline{I}(p_\phi(X, Z)) \leftarrow \max(0, H(p(Z)) - D_Z(\theta, \phi))$

▷ Upper bound of the mutual information

$\overline{I}(q_\theta(X, Z)) \leftarrow \min(H(q(X)), R_Z(\theta))$

$\overline{I}(p_\phi(X, Z)) \leftarrow H(p(Z))$

▷ Estimation of the KL-divergence between the marginals

$KL(q_\theta(Z)||p(Z)) \leftarrow \frac{1}{M} \sum_{i=1}^{M} \log \frac{s_\psi(Y=0|Z=z_{x_i})}{1-s_\psi(Y=0|Z=z_{x_i})}$

$KL(p_\psi(X)||q(X)) \leftarrow \frac{1}{M} \sum_{i=1}^{M} \log \frac{s_\psi(Y=0|X=x_{z_i})}{1-s_\psi(Y=0|X=x_{z_i})}$

▷ Correction of the estimations

$KL(q_\theta(Z)||p(Z)) \leftarrow \max(KL(q_\theta(Z)||p(Z)), R_Z(\theta) - \overline{I}(q_\theta(X, Z)))$

$KL(q_\theta(Z)||p(Z)) \leftarrow \min(KL(q_\theta(Z)||p(Z)), R_Z(\theta) - \underline{I}(q_\theta(X, Z)))$

$KL(p_\phi(X)||q(X)) \leftarrow \max(KL(p_\phi(X)||p(X)), 0)$

$KL(p_\phi(X)||q(X)) \leftarrow \min(KL(p_\phi(X)||q(X)), H(p(Z)))$

▷ Encoding mutual information

$I(q_\theta(X, Z)) \leftarrow R_Z(\theta) - KL(q_\theta(Z)||p(Z))$

▷ Conditional KL-divergence

$KL(q_\theta(X|Z)||p_\phi(X|Z)) \leftarrow KL(q_\theta(X, Z)||p_\phi(X, Z)) - KL(q_\theta(Z)||p(Z))$

Appendix F

Experimental Details

F.1 The Smoothed Uniform Distribution

The smoothed uniform distribution $\tilde{U}(m, z, \sigma^2)$ has been designed as a relaxation of the uniform distribution. While the uniform distribution assigns zero probability to every observation that is not included within the range specified by its parameters, the proposed relaxation uses two Normal tails to enforce a non-zero probability over the entire domain. This property ensures that quantities such as the cross-entropy and the Kullback-Leibler divergence that require the computation of the log-probability of the density are finite. Nevertheless, analogously to the uniform distribution, the **smoothed uniform distribution** $\tilde{U}(m, s, \sigma^2)$ denotes a region of constant probability.

The parameter m denotes the center of the distribution, s is used to specify the wideness of the uniform interval, while σ^2 described the variance of the Normal distributions used for the two tails. A visualization of the parameters is reported in Figure F.1.

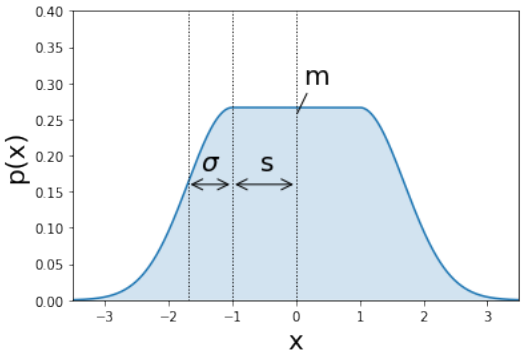

FIGURE F.1: Plot of the probability density function for the smoothed uniform distribution. The values of the parameters m, s and σ are graphically visualized on the picture.

Its functional form is expressed as:

$$\tilde{U}(Z = z|m, s, \sigma^2) = \begin{cases} \frac{\sqrt{2\pi\sigma^2}}{\sqrt{2\pi\sigma^2}+2s}\mathcal{N}(Z = z|m - s, \sigma^2) & \text{if } z < m - s \\ \frac{\sqrt{2\pi\sigma^2}}{\sqrt{2\pi\sigma^2}+2s}\mathcal{N}(Z = z|m + s, \sigma^2) & \text{if } z > m + s \\ \frac{1}{\sqrt{2\pi\sigma^2}+2s} & \text{o.w.} \end{cases} \tag{F.1}$$

Where $\mathcal{N}(Z = z|m - s, \sigma^2)$ represents the density of a Normal distribution with mean $m - s$ and variance σ^2. Note that the density defined by $\tilde{U}(m, s, \sigma^2)$ is continuous and differentiable on the whole domain. The smoothed uniform distribution has been used as a prior distribution since it allows to represent a region in which the codes are equally likely to occur.

F.2 Details on the experiments

All the parametric distributions considered in the experiments consists of multilayer perceptrons composed by 1 hidden layer of 100 units, with the only exception of the code and data discriminators that consist of 2 hidden layers of 100 units each. The size of the input and the output of the different architectures varies for the two datasets. Since both the encoder and the decoder model mean and variance of Normal distributions, their output size is doubled to represents the two vectors of parameters, namely mean and variance, separately. In order to constrain the values of the variance to be positive, an exponential function is applied to the corresponding activations of the last layer and a fixed constant of 10^3 is added to their value to ensure numerical stability.

Each architecture has been trained by using Adam optimizer (Kingma and Ba, 2014) and all the parameter have been initialized according to the strategy defined in LeCun et al., 1998. All the training procedures started with a warm-up period during which only the data and code discriminators are trained (1000 iterations for 1D-Mixture and 500 iterations for the 2D-Mixture datasets). Each training iteration consisted of one update step for the parameters of encoder and decoder networks followed by two steps of updates for both the parameters of code and data discriminators. The batches used to compute the gradients consisted of samples from empirical and prior distributions of size 10 for the 1D-Mixture dataset and 50 for the 2D-mixture one. In order to better represent the empirical distribution, the batches have been created by randomly sampling the observations without excluding possible repetitions. The training procedure involving the 1D-Mixture dataset consisted of 20000 iterations with batches of 10 observations each and a learning rate of 5×10^{-4}. The 2D-Mixture experiments, on the other hand, perform 5000 iterations with batches composed by 50 elements and a learning rate of 1×10^{-3}. The specific hyper-parameter configurations reported in this work are the result of a simple gird search that aimed to determine a configuration that produced consistent results for all the training losses across multiple runs. All the architectures have been implemented using the PyTorch framework (Paszke et al., 2017) for automatic gradient differentiation.

F.3 Details on the visualizations

In this section we describe the qualitative plots that have been produced for the different the experiments. The graphical visualizations have been developed to directly show different characteristic of the learning objectives such as the quality of the matching between the marginal and conditional distributions, and the quality of both data and code reconstructions.

The 1D-mixture plots (Figure 5.1) shows 1000 samples from the joint encoding and decoding distributions (with orange and blue coloring respectively). On the upper side of the plot, the marginal data distribution $p_\phi(X)$ is compared with the training samples represented by $q(X)$ and the real data distribution (dashed line in red). On the right part of the picture, the aggregated posterior $q_\theta(Z)$ is compared with

the prior distribution $p(Z)$. Both the marginals have been computed using a Monte Carlo approximation (10000 samples) by considering average probability observed in different regions of data and code spaces. This is done by fixing an interval ($[-4, 8]$ for the data and $[-5, 5]$ for the code space), a resolution (200 intervals) and evaluating the average probability associated to each subdivision.

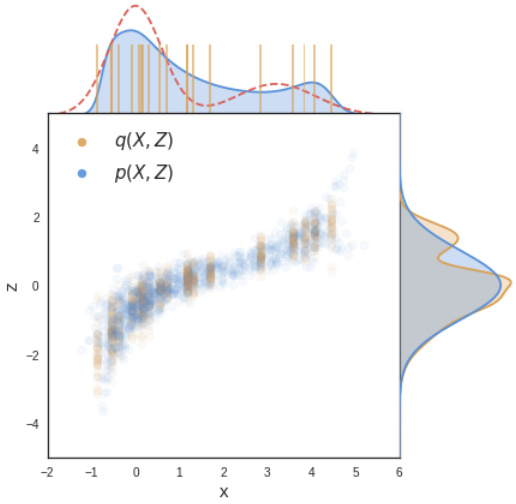

FIGURE F.2: Visualization of the encoding and decoding joint distributions induced by the Variational Autoencoder training objective $\mathcal{L}_{VAE}(\theta, \phi)$ on the 1D-Mixture dataset.

The 2D-mixture plots (Figure 5.2) show different characteristics of the encoding and decoding distributions through 4 visualizations. The first two visualizations from the left show the marginal distributions on the $\mathcal{X}$ and $\mathcal{Z}$ spaces. The mean and standard deviation of the mixture components used to produce the dataset (red crosses and lines) are compared with an approximation of the data-marginal $p_\phi(X)$ (represented with different shades of light-blue) obtained by smoothing 1000 samples from the data marginal $p_\phi(X)$ with a Gaussian kernel. This representation is enriched with the projection of a 10×10 grid G of points that are selected uniformly in $\mathcal{Z}$ and projected considering the means $\mu_\phi(G)$ defined by the decoder $p_\phi(X|Z = G)$. The lines between the decoded codes are drawn to connect points of the grid G that are neighbors in the latent space. Since the prior distribution has been selected to be uniform in the interval covered by the grid, this visualization has the advantage of directly showing how different region of the latent space $\mathcal{Z}$ are warped by the decoder to match the empirical distribution $q(X)$.

The second plot from the left in Figure 5.2 aims to evaluate the quality of the matching between the aggregated posterior $q_\theta(Z)$ (represented with different shades of orange) and the prior $p(Z)$ (depicted with a blue square). The red crosses in this visualization represent the means $\mu_\theta(\mathcal{D})$ of the conditional distributions $q_\theta(Z|X)$ for the different observations in the training set.

The last two plots on the right region of Figure 5.2 aim to show the quality of the data and the code reconstructions respectively. In particular, the first plot presents a direct comparison between the data-points $\mathcal{D}$ (in red) and the their most likely reconstruction, obtained by considering the mean of the decoding distribution given

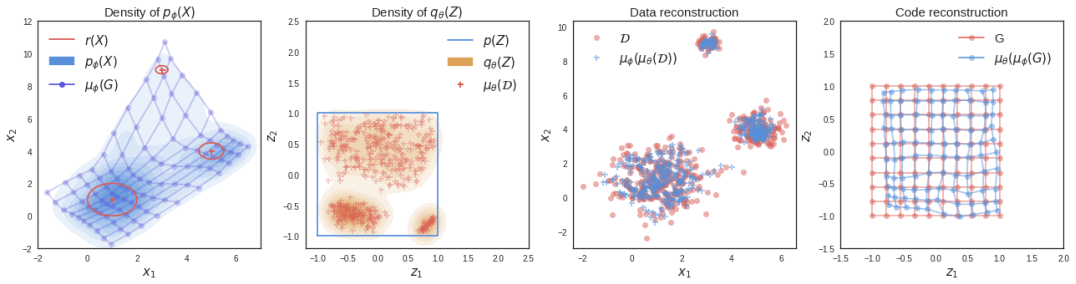

FIGURE F.3: Visualization of the encoding and decoding joint distributions induced by the Variational Autoencoder training objective $\mathcal{L}_{VAE}(\theta, \phi)$ on the 2D-Mixture dataset.

the average code for each data-point $\mu_\theta(\mu_\phi(\mathcal{D}))$. The last visualization on the right represents the quality of the reconstructions of different codes when they are decoded and re-encoded again. By considering the same 10×10 grid G (represented with red coloring) used for the first plot on the left, the last plot shows their most likely reconstruction $\mu_\phi(\mu_\theta(G))$ (in blue). Lines connect points that are neighbors in the original grid G to underline the warping effect induced by the encoding and decoding procedure.

F.4 Additional visualizations

In this section we report the results obtained for different parameter configurations on the 2D-Mixture dataset. Each plot show the configuration of $q_\theta(X, Z)$ and $p_\phi(X, Z)$ for three of the explored hyper-parameter configurations selected to underline their effect. The trend observed on the data and code marginals are consistent with the theoretical analysis reported in Chapter 4 and the quantitative measurement for the 1D-Mixture experiments reported in Chapter 5. Note that the legend has been omitted to increase the size of the representations. The reader can refer to the description reported in the previous section for further details regarding the visualizations.

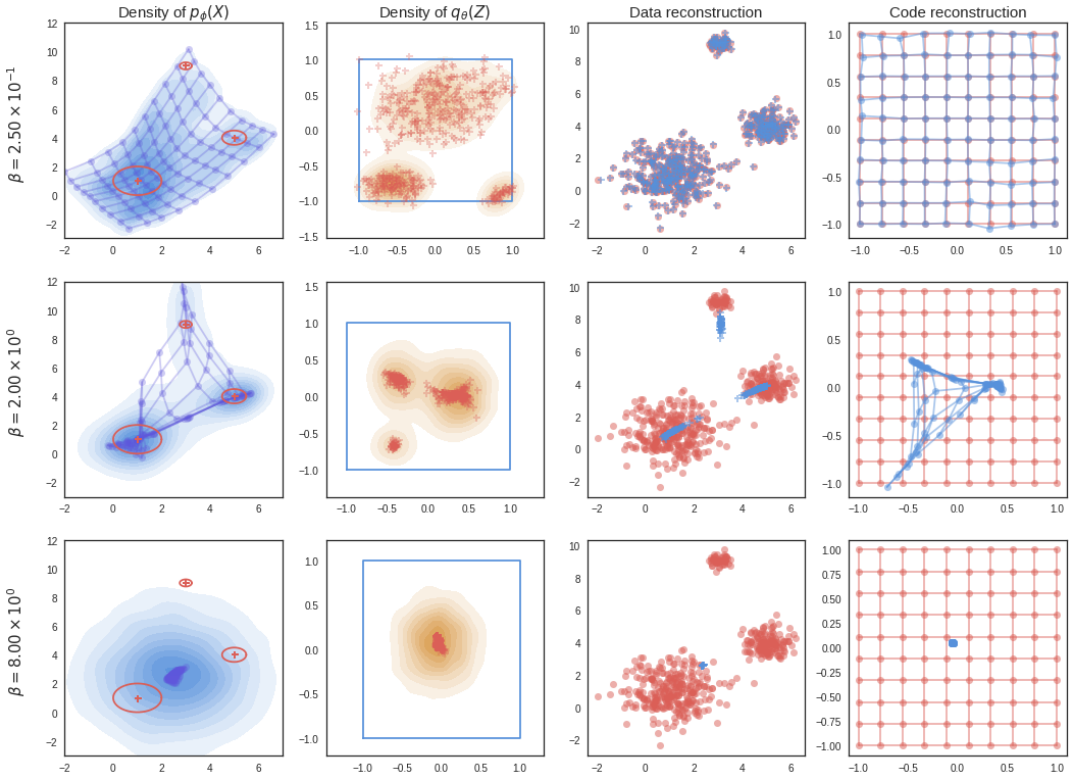

FIGURE F.4: Visualization of the generative and reconstruction performances for the Beta-VAE training objective.

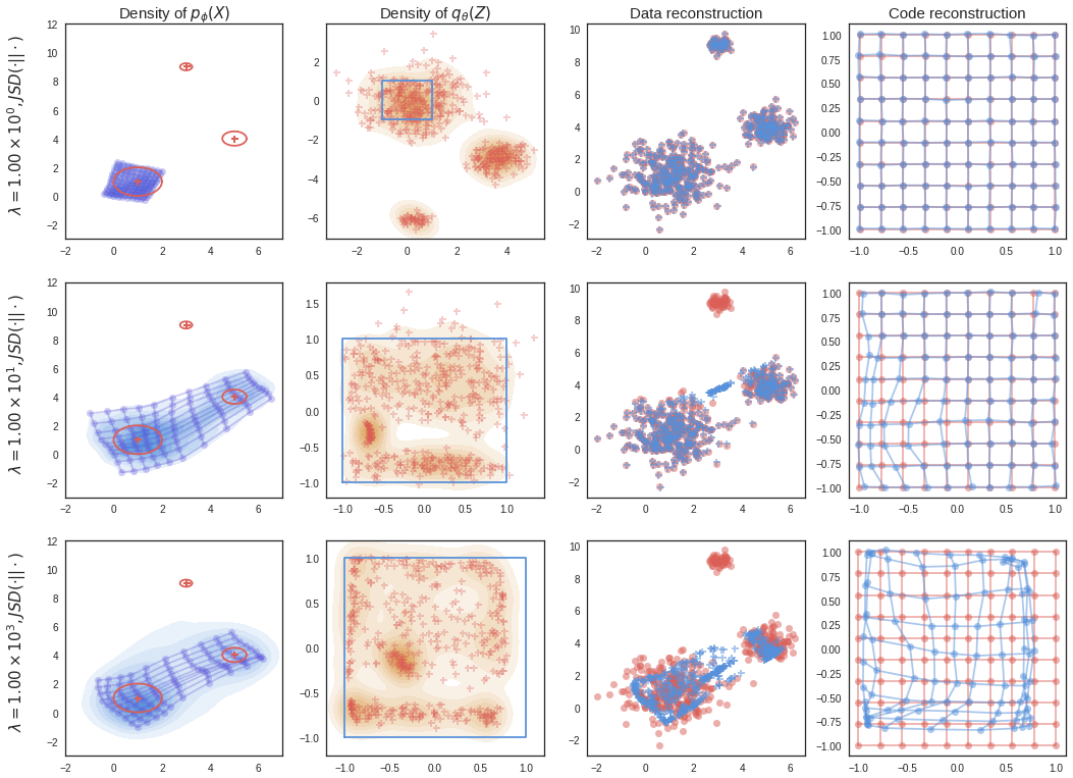

FIGURE F.5: Visualization of the generative and reconstruction performances for the Info-VAE training objective with the adversarial Jensen-Shannon divergence approximation.

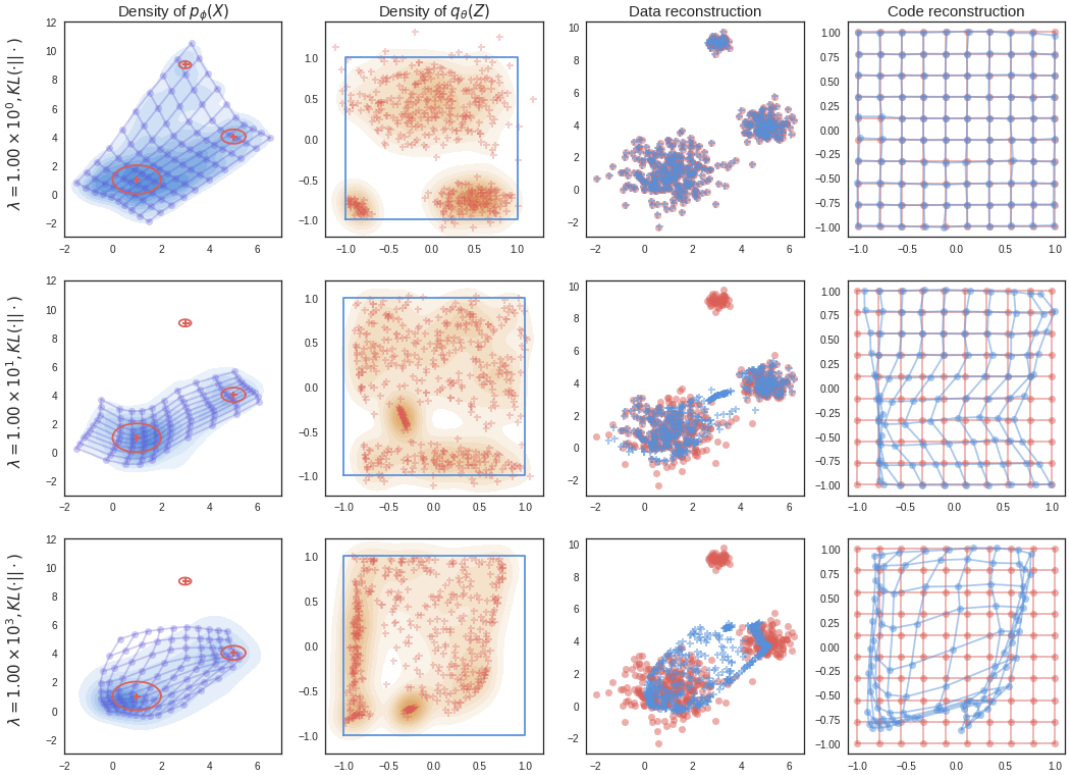

FIGURE F.6: Visualization of the generative and reconstruction performances for the Info-VAE training objective with the adversarial Kullback-Leibler divergence approximation.

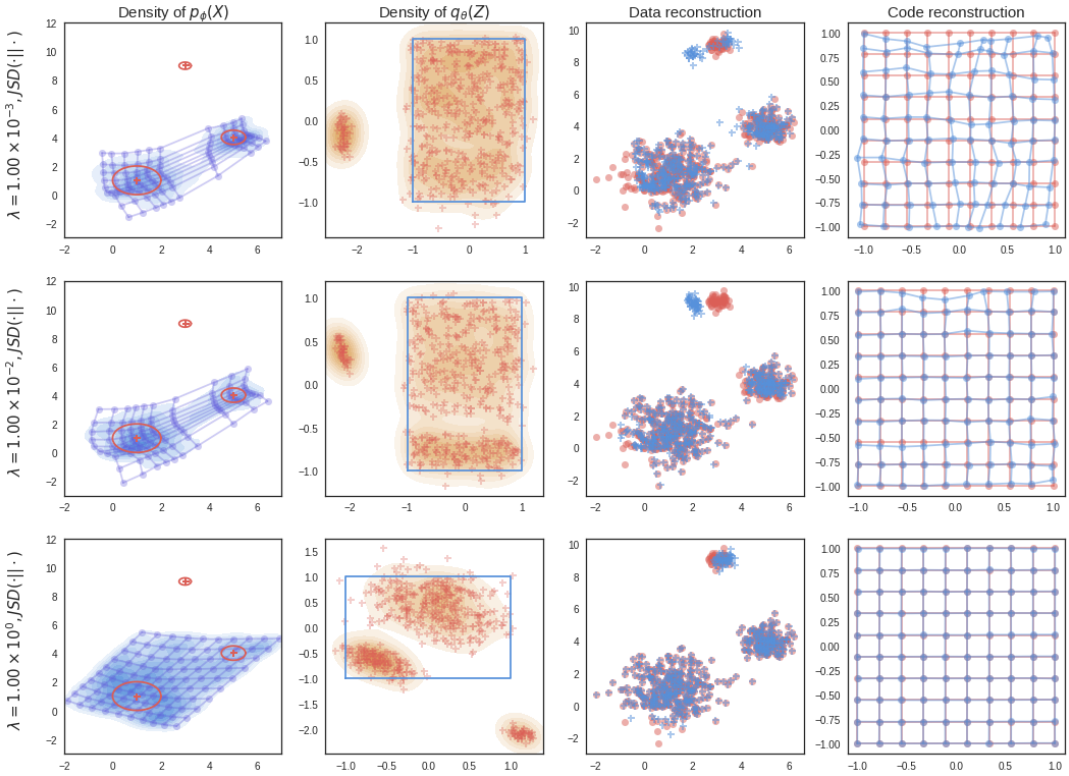

FIGURE F.7: Visualization of the generative and reconstruction performances for the Info-GAN training objective with the adversarial Jensen-Shannon divergence approximation.

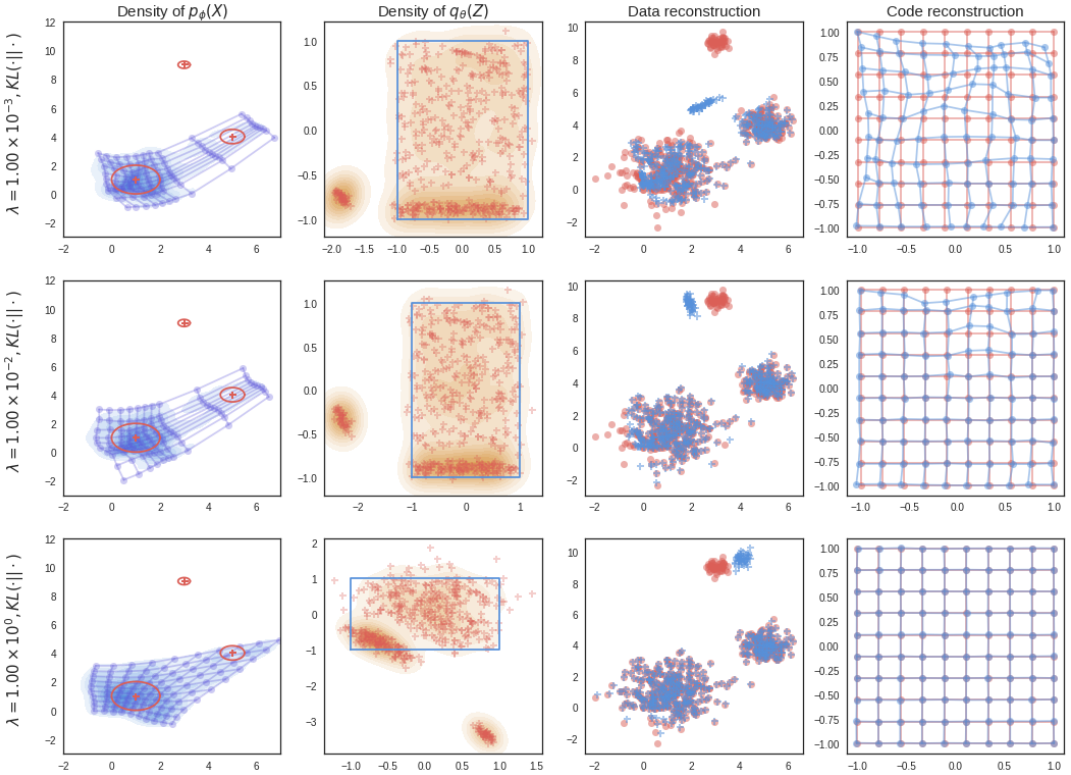

FIGURE F.8: Visualization of the generative and reconstruction performances for the Info-GAN training objective with the adversarial Kullback-Leibler divergence approximation.

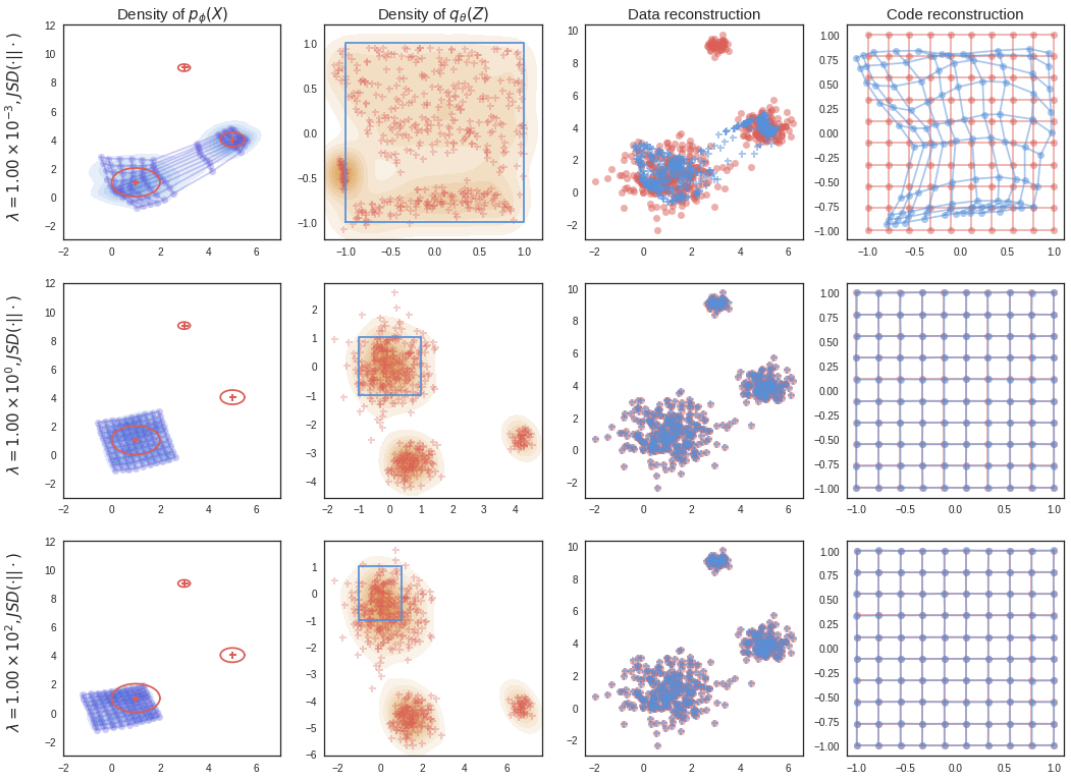

FIGURE F.9: Visualization of the generative and reconstruction performances for the Cycle-GAN training objective with the adversarial Jensen-Shannon divergence approximation.

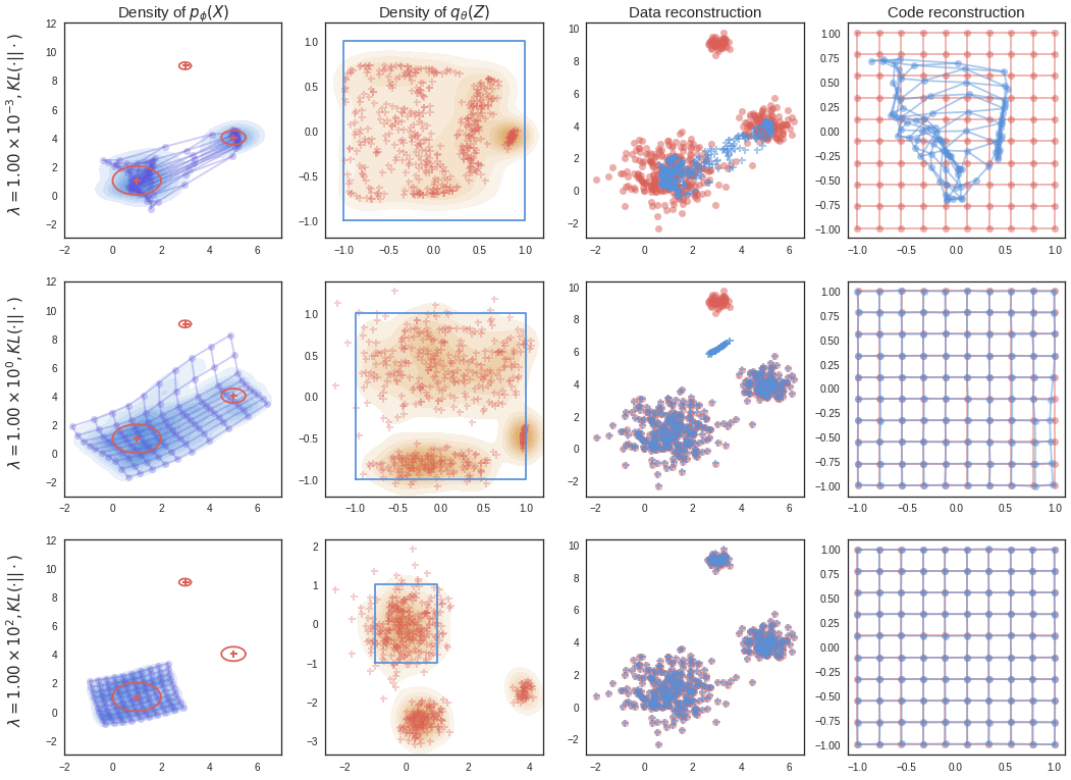

FIGURE F.10: Visualization of the generative and reconstruction performances for the Cycle-GAN training objective with the adversarial Kullback-Leibler divergence approximation.