4 CYCLE-CONSISTENT GENERATIVE ADVERSARIAL NETWORKS AS A BAYESIAN APPROXIMATION

4.1 INTRODUCTION

Learning correspondences between entities from different domains is an important and challenging problem in ML, especially in the *absence of paired data*. Consider for example the task of image-to-image translation where we want to learn a mapping from an image in a source domain, such as a photograph of a natural scene, to a corresponding image in a target domain, such as the realisation of such a scene in an 1860s celebrated artist's signature impressionistic style. The shortage of ground-truth pairings from the source domain to the target domain renders standard supervised approaches infeasible, thus motivating the need for unsupervised learning.

Within unsupervised approaches, a number of recently proposed CYCLEGAN methods have achieved remarkable success in addressing this problem [122, 294]. As their name suggests, these approaches are based upon two heuristics: (i) adversarial learning and (ii) cycle consistency. The former, adversarial learning [84], allows images in the source domain to be translated to output images that, to an auxiliary discriminator, are indistinguishable from images in the target domain, thereby matching their distributions. However, while distribution matching is necessary, it is insufficient to guarantee one-to-one mappings between the images, as the problem is heavily under-constrained. Briefly stated, the cycle-consistency is the constraint that an image mapped to a target domain should be *representable* in the original domain. It is this constraint that significantly shrinks the space of possible solutions.

Beyond the empirical risk minimisation framework motivated intuitively by the two principles mentioned above, the original CYCLEGAN formulation lacks any further theoretical justification. Besides providing sound quantification of uncertainty, a LVM allows us to disentangle our modeling assumptions from the inference machinery used to reason about the model variables. Interpreting standard methods from a Bayesian perspective has contributed significantly to the understanding of these methods and to the development of new approaches [73, 255].

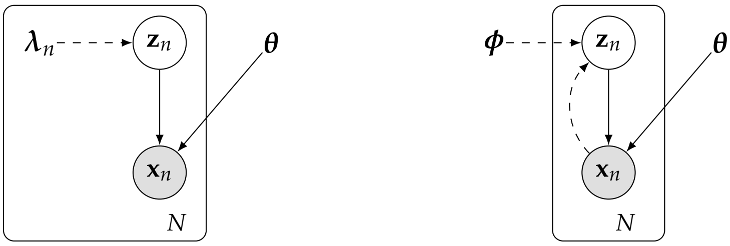
In this chapter, we introduce *implicit* LVMS, where the prior over hidden representations can be specified flexibly as an *implicit* distribution. We develop a VI algorithm for this model based on minimisation of the *symmetric* KL divergence between a *variational joint* and the exact joint distribution, in contrast to traditional reverse KL minimisation, which notoriously underestimate posterior's / exact distribution's support. Lastly, we demonstrate that the state-of-the-art CYCLEGANS as proposed contemporaneously by Kim et al. [122] and Zhu et al. [294] can be derived as a special case within our proposed VI framework, thus establishing its connection to approximate Bayesian inference methods.

4.2 IMPLICIT LATENT VARIABLE MODELS

LVMS are an indispensable tool for uncovering the hidden representations of observed data. In a LVM, an observation $\mathbf{x}$ is assumed governed by its underlying hidden variable $\mathbf{z}$, which is drawn from a prior $p(\mathbf{z})$ and related to $\mathbf{x}$ through the likelihood $p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})$. Accordingly, the joint density of $\mathbf{x}$ and $\mathbf{z}$ is given by

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}) p(\mathbf{z}). \tag{4.1}$$

Given data distribution $q^*(\mathbf{x})$ and a finite collection $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N$ of observations $\mathbf{x}_n \sim q^*(\mathbf{x})$, and the set of corresponding latent variables $\mathbf{Z} = \{\mathbf{z}_n\}_{n=1}^N$, the joint over all variables factorises as, $p_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Z}) = \prod_{n=1}^N p_{\boldsymbol{\theta}}(\mathbf{x}_n, \mathbf{z}_n)$.



(a) Without amortised inference, each local latent variable is governed by its own local variational parameters.

(b) With amortized inference, we condition on observed variables and employ a single set of global variational parameters.

Figure 4.1: Graphical representation of the *generative model* (solid) and the *recognition model* (dashed).

The graphical representation of implicit LVMS is depicted in Figure 4.1. Instead of approximating the exact posterior $p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x}_n)$ for each $\mathbf{x}_n$, using a separate variational distribution $q(\mathbf{z}; \boldsymbol{\lambda}_n)$ with local variational parameters $\boldsymbol{\lambda}_n$, we condition on $\mathbf{x}$ and optimise a single set of variational parameters $\boldsymbol{\phi}$ across all $\mathbf{x} \sim q^*(\mathbf{x})$. Accordingly, the variational distribution is denoted $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \triangleq q(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\phi})$.

4.2.1 *Prescribed Likelihood*

We specify the likelihood through a mapping $\mathcal{F}_{\boldsymbol{\theta}}$ that takes as input random noise $\boldsymbol{\xi}$ and latent variable $\mathbf{z}$,

$$\mathbf{x} \sim p_{\boldsymbol{\theta}}(\mathbf{x} \,|\, \mathbf{z}) \quad \Leftrightarrow \quad \mathbf{x} = \mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{\xi}; \mathbf{z}), \quad \boldsymbol{\xi} \sim p(\boldsymbol{\xi}). \tag{4.2}$$

We shall restrict our attention to *prescribed* likelihoods, where evaluation of their density is tractable. This requires that $\mathcal{F}_{\boldsymbol{\theta}}(\,\cdot\,; \mathbf{z})$ be a diffeomorphism wrt $\boldsymbol{\xi}$ and density $p(\boldsymbol{\xi})$ be tractable. For example, when $\mathcal{F}_{\boldsymbol{\theta}}(\,\cdot\,; \mathbf{z})$ is a location-scale transform of noise $\boldsymbol{\xi}$ and $p(\boldsymbol{\xi})$ is Gaussian, we recover Gaussian observation models.

Our model specification is sufficiently general for encapsulating a broad range of familiar latent variable models, even when we make simplifying assumptions on the mapping $\mathcal{F}_{\boldsymbol{\theta}}(\,\cdot\,; \mathbf{z})$. In particular, consider the special case where the mapping is an affine transformation of the noise vector $\boldsymbol{\xi}$,

$$\mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{\xi}; \mathbf{z}) \triangleq \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}) + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\mathbf{z})^{\frac{1}{2}} \boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

for functions $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ parameterised by $\boldsymbol{\theta}$ that take $\mathbf{z}$ as input. To simplify matters further, assume $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is constant wrt to its input, i. e. $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\mathbf{z}) = \boldsymbol{\Psi}$ for all $\mathbf{z}$. The likelihood can then be written explicitly as

$$p_{\boldsymbol{\theta}}(\mathbf{x} \,|\, \mathbf{z}) = \mathcal{N}(\mathbf{x} \,|\, \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}), \boldsymbol{\Psi}).$$

factor analysis & probabilistic pca. In the case where the mean function $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ is an affine transformation of $\mathbf{z}$,

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}) \triangleq \mathbf{W}\mathbf{z} + \mathbf{b},$$

and the covariance matrix is diagonal $\boldsymbol{\Psi} = \operatorname{diag}(\psi_1^2, \dots, \psi_D^2)$, we recover fa [8]. Furthermore, when the covariance matrix is isotropic $\boldsymbol{\Psi} = \psi^2 \mathbf{I}$, we recover *probabilistic principal component analysis* (ppca) [255]. In this example, the parameters $\boldsymbol{\theta}$ consist of the factor loading matrix $\mathbf{W}$, the bias vector $\mathbf{b}$ and the covariance matrix $\boldsymbol{\Psi}$.

deep and nonlinear latent variable models. By introducing nonlinearities to the mean function, we are able to recover nonlinear factor analysis [135], nonlinear Gaussian sigmoid belief networks [71], and other more sophisticated variants of deep latent variable models. When the mapping is defined by a mlp, we can recover simple instances of a variational autoencoder (vae) with a Gaussian probabilistic decoder [124, 206].

4.2.2 *Implicit Prior*

In lvms, the prior typically specified as a prescribed distribution, e. g. a factorised Gaussian centered at zero. Oftentimes, however, the practitioner possesses prior knowledge that simply cannot be embodied within a prescribed distribution. To address this limitation, we introduce *implicit* LVMS, wherein the prior over latent variables is specified as an implicit distribution $p^*(\mathbf{z})$, given only by a finite collection $\mathbf{Z}^* = \{\mathbf{z}^*_m\}_{m=1}^M$ of its samples,

$$\mathbf{z}^*_m \sim p^*(\mathbf{z}). \tag{4.3}$$

This formulation offers the utmost degree of flexibility in the treatment of prior information, the difficulties of which have hindered the application of Bayesian statistics since the time of Laplace [112].

EXAMPLE: UNPAIRED IMAGE-TO-IMAGE TRANSLATION . Suppose we have collections of images $\mathbf{X}$ and $\mathbf{Z}^*$, which are assumed to be draws from the data distribution $q^*(\mathbf{x})$ and implicit prior distribution $p^*(\mathbf{z})$, respectively. For example, these might be photographs of natural landscapes and the paintings of Van Gogh. The goal of unpaired image-to-image translation is to learn the correspondence between variables $\mathbf{x}$ and $\mathbf{z}$ by capturing the underlying generative process specified by mapping $\mathcal{F}_{\boldsymbol{\theta}}$. This defines the likelihood $p_{\boldsymbol{\theta}}(\mathbf{x} \,|\, \mathbf{z})$ –a conditional density of $\mathbf{x}$ given $\mathbf{z}$. Continuing with the above example, the problem amounts to learning parameters $\boldsymbol{\theta}$ of the mapping such that this conditional yields photorealistic renderings of scenes portrayed in Van Gogh's paintings. Furthermore, the resulting posterior on the latent representation $p_{\boldsymbol{\theta}}(\mathbf{z} \,|\, \mathbf{x})$ – a conditional density of $\mathbf{z}$ given $\mathbf{x}$ – should produce renderings of landscape scenery in Van Gogh's iconic style.

4.3 VARIATIONAL INFERENCE

In this section, we describe the first component of our bipartite VI framework. In traditional VI, one specifies a family $\mathcal{Q}$ of densities over the latent variables and seeks the member $q \in \mathcal{Q}$ closest in KL divergence to the exact posterior $p_{\boldsymbol{\theta}}(\mathbf{z} \,|\, \mathbf{x})$ [17, 116, 270].

4.3.1 *Prescribed Variational Posterior*

We begin by describing the variational family $q \in \mathcal{Q}$. We adopt the common practice of *amortizing* inference using an inference network [81]. Namely, instead of approximating the exact posterior $p_{\boldsymbol{\theta}}(\mathbf{z} \,|\, \mathbf{x}_n)$ for each $\mathbf{x}_n$, using a separate variational distribution $q(\mathbf{z}; \boldsymbol{\lambda}_n)$ with local variational parameters $\boldsymbol{\lambda}_n$, we condition on $\mathbf{x}$ and optimise a single set of variational parameters $\boldsymbol{\phi}$ across all $\mathbf{x} \sim q^*(\mathbf{x})$.

The variational distribution $q_{\boldsymbol{\phi}}(\mathbf{z} \,|\, \mathbf{x})$ is specified through an inverse mapping $\mathcal{G}_{\boldsymbol{\phi}}$ that takes as input random noise $\boldsymbol{\epsilon}$ and observed variable $\mathbf{x}$,

$$\mathbf{z} \sim q_{\boldsymbol{\phi}}(\mathbf{z} \,|\, \mathbf{x}) \quad \Leftrightarrow \quad \mathbf{z} = \mathcal{G}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}; \mathbf{x}), \quad \boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}). \tag{4.4}$$

Just as mapping $\mathcal{F}_{\boldsymbol{\theta}}$ underpins the generative model, mapping $\mathcal{G}_{\boldsymbol{\phi}}$ underpins the *recognition model* [51]. As with the likelihood, *we restrict our attention to prescribed variational distributions.*

As depicted in Figure 4.1b, the dependency relationship between the variational parameters and the latent variables mirrors that of the model parameters and observed variables. This symmetry is crucial to the derivation of CYCLEGAN later in Section 4.5.3.2.

4.3.2 *Reverse KL Variational Objective*

Minimizing the reverse KL between the exact and variational posterior is equivalent to maximizing the ELBO, or minimizing its *negative*, defined as

$$\begin{aligned}\mathcal{L}_{\text{NELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \triangleq \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}[-\log p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})] \\ + \mathbb{E}_{q^*(\mathbf{x})}\text{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \parallel p^*(\mathbf{z})\right].\end{aligned} \tag{4.5}$$

The first term of the ELBO is the (negative) ELL, defined as

$$\mathcal{L}_{\text{NELL}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \triangleq \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}[-\log p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})]. \tag{4.6}$$

It is easy to perform stochastic gradient-based optimisation of this term by applying the reparameterisation trick [124, 206],

$$\mathcal{L}_{\text{NELL}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{q^*(\mathbf{x})p(\boldsymbol{\epsilon})}[-\log p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathcal{G}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}; \mathbf{x}))]. \tag{4.7}$$

However, the second term – the KL divergence between $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$ and implicit prior $p^*(\mathbf{z})$ – is not so straightforward. In particular, the KL divergence can be expressed as

$$\text{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \parallel p^*(\mathbf{z})\right] \triangleq \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}[\log r^*(\mathbf{z}; \mathbf{x})], \tag{4.8}$$

where $r^*(\mathbf{z}; \mathbf{x})$ is defined as the ratio of densities,

$$r^*(\mathbf{z}; \mathbf{x}) \triangleq \frac{q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}{p^*(\mathbf{z})}. \tag{4.9}$$

The dependence on this density ratio is problematic since the prior $p^*(\mathbf{z})$ is implicit and cannot be evaluated directly. To overcome this, we resort to methods for approximating f-divergences between implicit distributions, which are inextricably tied to DRE [172, 245].

4.3.3 *Approximate Divergence Minimisation*

Although we are primarily interested in estimating the KL divergence of Equation (4.8), we give a generalised treatment that is applicable to all f-divergences [2, 46]. We denote a generic member of the family of f-divergences between distributions p and q as $\mathcal{D}_f\left[p \parallel q\right] \triangleq \mathbb{E}_p[f(q/p)]$, for some convex lower-semicontinuous function $f : \mathbb{R}_+ \to \mathbb{R}$.

Leveraging results from convex analysis, Nguyen, Wainwright, and Jordan [183] devise a variational lower bound that estimates an f-divergence through samples when either or both of the densities are unavailable. Nowozin, Cseke, and Tomioka [185] extend this framework to derive GAN objectives that minimise arbitrary f-divergences. These results underpin our methodology, and we restate a variant of it here for completeness.

Theorem 4.3.1 (Nguyen, Wainwright, and Jordan [183]). *Let $f^\star$ be the convex dual of f and $\mathcal{R}$ a class of functions with codomains equivalent to the domain of f'. We have the following lower bound on the f-divergence between distributions $p(\mathbf{u})$ and $q(\mathbf{u})$,*

$$\mathcal{D}_f\left[p(\mathbf{u}) \parallel q(\mathbf{u})\right] \geq \max_{\hat{r}\in\mathcal{R}}\{\mathbb{E}_{q(\mathbf{u})}[f'(\hat{r}(\mathbf{u}))] \\ - \mathbb{E}_{p(\mathbf{u})}[f^\star(f'(\hat{r}(\mathbf{u})))]\},$$

where equality is attained when $\hat{r}(\mathbf{u})$ is exactly the true density ratio $\hat{r}(\mathbf{u}) = {}^{q(\mathbf{u})}/_{p(\mathbf{u})}$.

Applying Theorem 4.3.1 to $p^*(\mathbf{z})$ and $q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x}_n)$ for a given $\mathbf{x}_n$, and optimizing over a class of functions indexed by parameters $\boldsymbol{\omega}_n$, we obtain the following lower bound on their divergence,

$$\mathcal{D}_f\left[p^*(\mathbf{z}) \parallel q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x}_n)\right] \geq \max_{\boldsymbol{\omega}_n}\left\{\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x}_n)}[f'(r_{\boldsymbol{\omega}_n}(\mathbf{z}))] - \mathbb{E}_{p^*(\mathbf{z})}[f^\star(f'(r_{\boldsymbol{\omega}_n}(\mathbf{z})))]\right\}.$$

While this provides a way to estimate any f-divergence between implicit prior $p^*(\mathbf{z})$ and variational distribution $q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x}_n)$ with only samples, it also requires us to optimise a separate density ratio estimator with parameters $\boldsymbol{\omega}_n$ for each observed $\mathbf{x}_n$. Instead, as with the posterior approximation, we also amortise the density ratio estimator by conditioning on $\mathbf{x}$ and optimizing a single set of parameters $\boldsymbol{\alpha}$ across all $\mathbf{x} \sim q^*(\mathbf{x})$. Accordingly, the estimator becomes $r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})$, taking also $\mathbf{x}$ as input. We now maximise an instance of the following generalised objective,

$$\mathcal{L}_f^{\text{latent}}(\boldsymbol{\alpha}\,|\,\boldsymbol{\phi}) \triangleq \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})}[f'(r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}))] \\ - \mathbb{E}_{q^*(\mathbf{x})p^*(\mathbf{z})}[f^\star(f'(r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})))]. \tag{4.10}$$

Corollary 4.3.2. *We have the lower bound,*

$$\mathbb{E}_{q^*(\mathbf{x})}\mathcal{D}_f\left[p^*(\mathbf{z}) \parallel q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})\right] \geq \max_{\boldsymbol{\alpha}} \mathcal{L}_f^{\text{latent}}(\boldsymbol{\alpha}\,|\,\boldsymbol{\phi}), \tag{4.11}$$

with equality at $r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}) = r^*(\mathbf{z};\mathbf{x})$.

DENSITY RATIO ESTIMATION OBJECTIVE. We write $\mathcal{L}_f^{\text{latent}}(\boldsymbol{\alpha}\,|\,\boldsymbol{\phi})$ to denote the DRE objective, wherein $\boldsymbol{\phi}$ is fixed, while $\boldsymbol{\alpha}$ is a free parameter that varies as this objective is *maximised*, thus tightening the bound of Equation (4.11) and the estimate of the density ratio $r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})$.

DIVERGENCE MINIMISATION LOSS. Conversely, the *divergence minimisation* (DM) loss, denoted as $\mathcal{L}_f^{\text{latent}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha})$, is *minimised* wrt $\boldsymbol{\phi}$ while $\boldsymbol{\alpha}$ remains fixed, thus approximately minimizing the f-divergence. In theory, this should be symmetric to the DRE objective, $\mathcal{L}_f^{\text{latent}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) \triangleq \mathcal{L}_f^{\text{latent}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$. However, alternative settings are often used in practice to alleviate the problem of vanishing gradients, as we shall see in Section 4.5.

By applying Theorem 4.3.2 for the setting $f_{\text{KL}}(u) \triangleq u \log u$, we instantiate a lower bound on the KL divergence of Equation (4.8) in the following objective,

$$\mathcal{L}_{\text{KL}}^{\text{latent}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi}) \triangleq \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}[\log r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})] - \mathbb{E}_{q^*(\mathbf{x})p^*(\mathbf{z})}[r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}) - 1]. \tag{4.12}$$

As we discuss in Section 4.A, maximisation of the objective in Equation (4.12) is closely related to the KLIEP [244].

Now, we define the DM loss symmetrically to the DRE objective in Equation (4.12) – terms not involving $\boldsymbol{\phi}$ are omitted,

$$\begin{aligned} \mathcal{L}_{\text{KL}}^{\text{latent}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) &\triangleq \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}[\log r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})] \qquad (4.13)\\ &\approx \mathbb{E}_{q^*(\mathbf{x})} \text{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \parallel p^*(\mathbf{z})\right]. \end{aligned}$$

Combined with the ELL, this estimate of the KL divergence yields an approximation to the ELBO where all terms are tractable. These objectives are summarised in the bi-level optimisation problem below,

$$\max_{\boldsymbol{\alpha}} \quad \mathcal{L}_{\text{KL}}^{\text{latent}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi}), \tag{4.14a}$$

$$\min_{\boldsymbol{\phi},\boldsymbol{\theta}} \quad \mathcal{L}_{\text{KL}}^{\text{latent}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) + \mathcal{L}_{\text{NELL}}(\boldsymbol{\theta}, \boldsymbol{\phi}), \tag{4.14b}$$

thus concluding the reverse KL minimisation component of our VI framework.

4.4 SYMMETRIC JOINT-MATCHING VARIATIONAL INFERENCE

We now complete the remaining component of our VI framework. In the previous section, we gave an extension to classical VI, which is fundamentally concerned with approximating the exact posterior. Now, let us instead consider directly approximating the *exact joint* $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})$ through a *variational joint* $q_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{z})$.

4.4.1 *Variational Joint*

Recall that $q^*(\mathbf{x})$ denotes the empirical data distribution. We define a variational approximation to the exact joint distribution of Equation (4.1) as

$$q_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{z}) \triangleq q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})q^*(\mathbf{x}). \tag{4.15}$$

We approximate the exact joint by seeking a variational joint closest in *symmetric* KL divergence, $\text{KL}_{\text{SYM}}\left[p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}) \parallel q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z})\right]$, where

$$\text{KL}_{\text{SYM}}\left[p \parallel q\right] \triangleq \text{KL}\left[p \parallel q\right] + \text{KL}\left[q \parallel p\right]. \tag{4.16}$$

We first look at the reverse KL divergence ($\text{KL}\left[q \parallel p\right]$) term. When expanded, we see that it is equivalent to the negative ELBO up to additive constants,

$$\begin{aligned}\text{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z}) \parallel p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\right] &\triangleq \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z})}\left[\log q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z}) - \log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\right] &(4.17)\\ &= \mathcal{L}_{\text{NELBO}}(\boldsymbol{\theta},\boldsymbol{\phi}) - \mathbb{H}[q^*(\mathbf{x})], &(4.18)\end{aligned}$$

where $\mathbb{H}[q^*(\mathbf{x})] \triangleq \mathbb{E}_{q^*(\mathbf{x})}[-\log q^*(\mathbf{x})]$ is the entropy of $q^*(\mathbf{x})$, a constant wrt parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. Hence, minimizing the KL divergence of Equation (4.17) can be reduced to minimizing $\mathcal{L}_{\text{NELBO}}(\boldsymbol{\theta},\boldsymbol{\phi})$ of Equation (4.5), without modification.

4.4.2 *Forward KL Variational Objective*

As for the forward KL divergence ($\text{KL}\left[p \parallel q\right]$) term, we have a similar expansion,

$$\begin{aligned}&\text{KL}\left[p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}) \parallel q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z})\right] &(4.19)\\ &\triangleq \mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}) - \log q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z})\right] &(4.20)\\ &= \mathbb{E}_{p^*(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x}\,|\,\mathbf{z})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}\,|\,\mathbf{z}) - \log q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z})] - \mathbb{H}[p^*(\mathbf{z})]. &(4.21)\end{aligned}$$

In analogy with the ELBO, we introduce a new variational objective that is minimised when the forward KL divergence of Equation (4.19) is minimised. First we define the recognition model analog to the marginal likelihood – the *marginal posterior*, or *aggregated posterior*, given by $q_\phi(\mathbf{z}) \triangleq \int q_\phi(\mathbf{z}\,|\,\mathbf{x})q^*(\mathbf{x})d\mathbf{x}$. It can be approximated by the *aggregate posterior lower bound (APLBO)*. For consistency, we give its *negative*, written as

$$\begin{aligned}\mathcal{L}_{\text{NAPLBO}}(\boldsymbol{\theta},\boldsymbol{\phi}) \triangleq\; &\mathbb{E}_{p^*(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x}\,|\,\mathbf{z})}[-\log q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})]\\ &+ \mathbb{E}_{p^*(\mathbf{z})}\text{KL}\left[p_{\boldsymbol{\theta}}(\mathbf{x}\,|\,\mathbf{z}) \parallel q^*(\mathbf{x})\right].\end{aligned} \tag{4.22}$$

Furthermore, minimizing the KL divergence of Equation (4.19) can be reduced to minimizing $\mathcal{L}_{\text{NAPLBO}}(\boldsymbol{\theta},\boldsymbol{\phi})$,

$$\text{KL}\left[p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}) \parallel q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z})\right] = \mathcal{L}_{\text{NAPLBO}}(\boldsymbol{\theta},\boldsymbol{\phi}) - \mathbb{H}[p^*(\mathbf{z})].$$

The first term of the negative APLBO is the (negative) *expected log posterior (ELP)*, defined as

$$\mathcal{L}_{\text{NELP}}(\boldsymbol{\theta},\boldsymbol{\phi}) \triangleq \mathbb{E}_{p^*(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x}\,|\,\mathbf{z})}[-\log q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})]. \tag{4.23}$$

We emphasise a key advantage of having considered the KL between the joint distributions instead of between the *posteriors*. Computing the *forward* KL divergence between the exact and approximate *posterior* distribution is problematic, since it requires evaluating expectations over the exact posterior $p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})$, the intractability of which is the reason we appealed to approximate inference in the first place.

In contrast, the forward KL divergence between the exact and approximate *joint* poses no such difficulties – we are able to sidestep the dependency on the exact posterior by expanding it into the form of Equation (4.21). Furthermore, as with the ELBO, we can perform stochastic gradient-based optimisation of the ELP term by applying the same reparameterisation trick as in Equation (4.7).

Now, the KL divergence term of the APLBO in Equation (4.22) can also be expressed as the expected logarithm of a density ratio $r^*(\mathbf{x}; \mathbf{z}) \triangleq p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}) / q^*(\mathbf{x})$ that involves an intractable density $q^*(\mathbf{x})$ – the empirical data distribution. To overcome this, we adopt the same approach as outlined in Section 4.3.3. Namely, we apply Theorem 4.3.1 to $q^*(\mathbf{x})$ and $p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}^*)$, and fit an amortised density ratio estimator $r_{\boldsymbol{\beta}}(\mathbf{x}; \mathbf{z})$ to $r^*(\mathbf{x}; \mathbf{z})$ by maximizing an instance of the generalised objective,

$$\begin{aligned} \mathcal{L}_f^{\text{observed}}(\boldsymbol{\beta} \mid \boldsymbol{\theta}) \triangleq \; & \mathbb{E}_{p^*(\mathbf{z}) p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})}[f'(r_{\boldsymbol{\beta}}(\mathbf{x}; \mathbf{z}))] \\ & - \mathbb{E}_{p^*(\mathbf{z}) q^*(\mathbf{x})}[f^{\star}(f'(r_{\boldsymbol{\beta}}(\mathbf{x}; \mathbf{z})))]. \end{aligned} \tag{4.24}$$

Corollary 4.4.1. *We have the lower bound,*

$$\mathbb{E}_{p^*(\mathbf{z})} \mathcal{D}_f \left[q^*(\mathbf{x}) \,\|\, p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}) \right] \geq \max_{\boldsymbol{\beta}} \mathcal{L}_f^{\text{observed}}(\boldsymbol{\beta} \mid \boldsymbol{\theta}), \tag{4.25}$$

with equality at $r_{\boldsymbol{\beta}}(\mathbf{x}; \mathbf{z}) = r^*(\mathbf{x}; \mathbf{z})$.

By applying Theorem 4.4.1 with the previously defined $f_{\text{KL}}(u)$, we obtain lower bound objective $\mathcal{L}_{\text{KL}}^{\text{observed}}(\boldsymbol{\beta} \mid \boldsymbol{\theta})$ on the KL divergence term in Equation (4.22), and a corresponding DM loss $\mathcal{L}_{\text{KL}}^{\text{observed}}(\boldsymbol{\theta} \mid \boldsymbol{\beta})$, analogous to the definitions of $\mathcal{L}_{\text{KL}}^{\text{latent}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$ and $\mathcal{L}_{\text{KL}}^{\text{latent}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha})$ in Equations (4.12) and (4.13), respectively. See Table 4.B.3 for a summary of explicit definitions.

Hence, in addition to the bi-level optimisation problems of Equation (4.14) we have,

$$\max_{\boldsymbol{\beta}} \quad \mathcal{L}_{\text{KL}}^{\text{observed}}(\boldsymbol{\beta} \mid \boldsymbol{\theta}), \tag{4.26a}$$

$$\min_{\boldsymbol{\phi}, \boldsymbol{\theta}} \quad \mathcal{L}_{\text{KL}}^{\text{observed}}(\boldsymbol{\theta} \mid \boldsymbol{\beta}) + \mathcal{L}_{\text{NELP}}(\boldsymbol{\theta}, \boldsymbol{\phi}). \tag{4.26b}$$

As shown, the minimisations in Equations (4.14b) and (4.26b) corresponds to minimisation of the symmetric KL over the joints $\text{KL}_{\text{SYM}} \left[p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) \,\|\, q_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{z}) \right]$, while the maximisations in Equations (4.14a) and (4.26a) approximates the divergences, or more precisely, the density ratios involving implicit distributions.

4.5 CYCLEGAN AS A SPECIAL CASE

In this section, we demonstrate that CYCLEGAN methods [122, 294] can be instantiated under our proposed VI framework.

4.5.1 *Basic CycleGAN Framework*

To address the problem of unpaired image-to-image translation as described in Section 4.2.2, the CYCLEGAN model learns two mappings $\boldsymbol{\mu}_{\boldsymbol{\theta}} : \mathbf{z} \mapsto \mathbf{x}$ and $\mathbf{m}_{\boldsymbol{\phi}} : \mathbf{x} \mapsto \mathbf{z}$ by optimizing two complementary classes of objectives.

DISTRIBUTION MATCHING. The first are the adversarial objectives, which help match the output of mapping $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ to the empirical distribution $q^*(\mathbf{x})$, and the output of $\mathbf{m}_{\boldsymbol{\phi}}$ to $p^*(\mathbf{z})$. In particular, for mapping $\mathbf{m}_{\boldsymbol{\phi}}$, this involves introducing a discriminator $\mathbf{D}_{\boldsymbol{\alpha}}$ and the saddle-point adversarial objective,

$$\begin{aligned}\ell_{\text{GAN}}^{\text{reverse}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi}) \triangleq \mathbb{E}_{p^*(\mathbf{z})}[\log \mathbf{D}_{\boldsymbol{\alpha}}(\mathbf{z})] \\ + \mathbb{E}_{q^*(\mathbf{x})}[\log(1 - \mathbf{D}_{\boldsymbol{\alpha}}(\mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x})))],\end{aligned} \tag{4.27}$$

while minimizing it wrt parameters $\boldsymbol{\phi}$. This encourages $\mathbf{m}_{\boldsymbol{\phi}}$ to produce realistic outputs $\mathbf{z} = \mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x}), \mathbf{x} \sim q^*(\mathbf{x})$ which, to the discriminator $\mathbf{D}_{\boldsymbol{\alpha}}$, are "indistinguishable" from $\mathbf{z}^* \sim p^*(\mathbf{z})$. A similar adversarial objective is defined for mapping $\boldsymbol{\mu}_{\boldsymbol{\theta}}$,

$$\ell_{\text{GAN}}^{\text{forward}}(\boldsymbol{\beta} \mid \boldsymbol{\theta}) \triangleq \mathbb{E}_{p^*(\mathbf{x})}[\log \mathbf{D}_{\boldsymbol{\beta}}(\mathbf{x})] + \mathbb{E}_{p^*(\mathbf{z})}[\log(1 - \mathbf{D}_{\boldsymbol{\beta}}(\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z})))]. \tag{4.28}$$

CYCLE-CONSISTENCY. Next are the cycle-consistency losses, which enforce tight correspondence between domains by ensuring that reconstruction $\mathbf{x}' = \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x}))$ is close to the input $\mathbf{x}$, and likewise for $\mathbf{m}_{\boldsymbol{\phi}}(\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}))$. This is achieved by minimizing a reconstruction loss,

$$\ell_{\text{CONST}}^{\text{reverse}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \triangleq \mathbb{E}_{q^*(\mathbf{x})}[\|\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x}))\|_{\rho}^{\rho}], \tag{4.29}$$

where $\|\cdot\|_{\rho}$ denotes the ℓ_{ρ}-norm. A similar loss $\ell_{\text{CONST}}^{\text{forward}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ is defined for the reconstruction of $\mathbf{z}$,

$$\ell_{\text{CONST}}^{\text{forward}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \triangleq \mathbb{E}_{p^*(\mathbf{z})}[\|\mathbf{z} - \mathbf{m}_{\boldsymbol{\phi}}(\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}))\|_{\rho}^{\rho}]. \tag{4.30}$$

These objectives are summarised in the following set of optimisation problems,

$$\max_{\boldsymbol{\alpha}} \ell_{\text{GAN}}^{\text{reverse}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi}), \quad \max_{\boldsymbol{\beta}} \ell_{\text{GAN}}^{\text{forward}}(\boldsymbol{\beta} \mid \boldsymbol{\theta}), \tag{4.31a}$$

$$\min_{\boldsymbol{\phi},\boldsymbol{\theta}} \ell_{\text{GAN}}^{\text{reverse}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) + \ell_{\text{CONST}}^{\text{reverse}}(\boldsymbol{\theta}, \boldsymbol{\phi}), \tag{4.31b}$$

$$\min_{\boldsymbol{\phi},\boldsymbol{\theta}} \ell_{\text{GAN}}^{\text{forward}}(\boldsymbol{\theta} \mid \boldsymbol{\beta}) + \ell_{\text{CONST}}^{\text{forward}}(\boldsymbol{\theta}, \boldsymbol{\phi}). \tag{4.31c}$$

We now highlight the correspondences between these objectives and those of our proposed VI framework, as summarised in the optimisation problems of Equations (4.14) and (4.26).

4.5.2 *Cycle-consistency as Conditional Entropy Maximisation*

We now demonstrate that minimizing the cycle-consistency losses corresponds to maximizing the expected log likelihood and variational posterior of Equations (4.6) and (4.23). This can be shown by instantiating specific classes of $p_{\boldsymbol{\theta}}(\mathbf{x} \,|\, \mathbf{z})$ and $q_{\boldsymbol{\phi}}(\mathbf{z} \,|\, \mathbf{x})$ that recover $\ell^{\text{reverse}}_{\text{CONST}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ and $\ell^{\text{forward}}_{\text{CONST}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ from $\mathcal{L}_{\text{NELL}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ and $\mathcal{L}_{\text{NELP}}(\boldsymbol{\theta}, \boldsymbol{\phi})$, respectively.

Proposition 4.5.1. *Consider a typical case where the likelihood and the posterior approximation are both Gaussians,*

$$p_{\boldsymbol{\theta}}(\mathbf{x} \,|\, \mathbf{z}) \triangleq \mathcal{N}(\mathbf{x} \,|\, \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}), \tau^{-1}\mathbf{I}), \qquad q_{\boldsymbol{\phi}}(\mathbf{z} \,|\, \mathbf{x}) \triangleq \mathcal{N}(\mathbf{z} \,|\, \mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x}), t^{-1}\mathbf{I}).$$

In the limit as the posterior precision t tends to ∞, $\mathcal{L}_{\text{NELL}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ approaches $\ell^{\text{reverse}}_{\text{CONST}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ for $\rho = 2$, up to constants[1]. More precisely,

$$\mathcal{L}_{\text{NELL}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \to \frac{\tau}{2}\mathcal{L}^{\text{reverse}}_{\text{CONST}}(\boldsymbol{\theta}, \boldsymbol{\phi}) + \text{const}, \quad \text{as } t \to \infty$$

Similarly, we have,

$$\mathcal{L}_{\text{NELP}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \to \frac{t}{2}\mathcal{L}^{\text{forward}}_{\text{CONST}}(\boldsymbol{\theta}, \boldsymbol{\phi}) + \text{const}, \quad \text{as } \tau \to \infty$$

Proof. First, note the generative mappings underlying the given Gaussian likelihood and approximate posterior are

$$\begin{aligned} & \mathbf{z} \sim p_{\boldsymbol{\theta}}(\mathbf{x} \,|\, \mathbf{z}) \triangleq \mathcal{N}(\mathbf{x} \,|\, \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}), \tau^{-1}\mathbf{I}), \\ \Leftrightarrow \quad & \mathbf{z} = \mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{\xi}; \mathbf{z}) \triangleq \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}) + \tau^{-\frac{1}{2}}\boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \end{aligned}$$

and,

$$\begin{aligned} & \mathbf{x} \sim q_{\boldsymbol{\phi}}(\mathbf{z} \,|\, \mathbf{x}) \triangleq \mathcal{N}(\mathbf{z} \,|\, \mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x}), t^{-1}\mathbf{I}), \\ \Leftrightarrow \quad & \mathbf{x} = \mathcal{G}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}; \mathbf{x}) \triangleq \mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x}) + t^{-\frac{1}{2}}\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \end{aligned}$$

respectively. Thus, expanding out $\mathcal{L}_{\text{NELL}}(\boldsymbol{\theta}, \boldsymbol{\phi})$, we have

$$\begin{aligned} & \mathcal{L}_{\text{NELL}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \\ & \quad = \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z} \,|\, \mathbf{x})}[-\log p_{\boldsymbol{\theta}}(\mathbf{x} \,|\, \mathbf{z})] \\ & \quad = \mathbb{E}_{q^*(\mathbf{x})p(\boldsymbol{\epsilon})}[-\log p_{\boldsymbol{\theta}}(\mathbf{x} \,|\, \mathcal{G}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}; \mathbf{x}))] \\ & \quad = \frac{\tau}{2}\mathbb{E}_{q^*(\mathbf{x})p(\boldsymbol{\epsilon})}[\|\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathcal{G}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}; \mathbf{x}))\|_2^2] + \frac{D}{2}\log\frac{2\pi}{\tau} \\ & \quad = \frac{\tau}{2}\mathbb{E}_{q^*(\mathbf{x})p(\boldsymbol{\epsilon})}[\|\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x}) + t^{-\frac{1}{2}}\boldsymbol{\epsilon})\|_2^2] + \text{const} \\ & \quad \to \frac{\tau}{2}\mathbb{E}_{q^*(\mathbf{x})}[\|\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x}))\|_2^2] + \text{const}, \quad \text{as } t \to \infty \\ & \quad = \frac{\tau}{2}\mathcal{L}^{\text{reverse}}_{\text{CONST}}(\boldsymbol{\theta}, \boldsymbol{\phi}) + \text{const}. \end{aligned}$$

A similar analysis can be carried out for $\mathcal{L}_{\text{NELP}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ and its deterministic counterpart $\ell^{\text{forward}}_{\text{CONST}}(\boldsymbol{\theta}, \boldsymbol{\phi})$. □

[1] we obtain the same result for the case $\rho = 1$ by instead setting both the likelihood and approximate posterior to be Laplace distributions.

Hence, the cycle-consistency losses can be seen as special cases of the ELL and ELP with *degenerate* conditional distributions. Furthermore, this sheds new light on the roles of the cycle-consistency losses. For example, similar to the ELL, the reverse consistency loss encourages the conditional $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$ to place its mass on configurations of latent variables that can explain, or in this case, *represent* the data well.

4.5.3 *Distribution Matching as Approximate Divergence Minimisation*

We now discuss how the adversarial objectives $\ell_{\text{GAN}}^{\text{reverse}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$ and $\ell_{\text{GAN}}^{\text{forward}}(\boldsymbol{\beta} \mid \boldsymbol{\theta})$ relate to the KL variational lower bounds of our framework, $\mathcal{L}_{\text{KL}}^{\text{latent}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$ and $\mathcal{L}_{\text{KL}}^{\text{observed}}(\boldsymbol{\beta} \mid \boldsymbol{\theta})$, respectively. To reduce clutter, we restrict our discussion to the reverse objective $\ell_{\text{GAN}}^{\text{reverse}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$, as the same reasoning readily applies to the forward $\ell_{\text{GAN}}^{\text{forward}}(\boldsymbol{\beta} \mid \boldsymbol{\theta})$.

4.5.3.1 *As Density Ratio Estimation by Probabilistic Classification*

Firstly, the connections between GANs, divergence minimisation and DRE are well-established [172, 185, 245]. Although $\ell_{\text{GAN}}^{\text{reverse}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$ is a scoring rule for probabilistic classification [83], one can readily show that it can also be subsumed as an instance of the generalised variational lower bound $\mathcal{L}_{f}^{\text{latent}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$. Furthermore, similar to $\mathcal{L}_{\text{KL}}^{\text{latent}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$, maximizing $\ell_{\text{GAN}}^{\text{reverse}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$ corresponds estimating the intractable density ratio $r^*(\mathbf{z}; \mathbf{x})$ of Equation (4.9).

Lemma 4.5.2. *By setting $f_{\text{GAN}}(u) = u \log u - (u+1)\log(u+1)$ in the generalised objective $\mathcal{L}_{f}^{\text{latent}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$ of Equation (4.10), we instantiate the objective*

$$\begin{aligned}\mathcal{L}_{\text{GAN}}^{\text{reverse}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi}) \triangleq\ & \mathbb{E}_{q^*(\mathbf{x})p^*(\mathbf{z})}[\log \mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z}; \mathbf{x})] \\ & + \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}[\log(1 - \mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z}; \mathbf{x}))],\end{aligned} \tag{4.32}$$

where $\mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z}; \mathbf{x}) \triangleq 1 - \sigma(\log r_{\boldsymbol{\alpha}}(\mathbf{z}; \mathbf{x}))$, and σ is the logistic sigmoid function.

Proof. To instantiate $\mathcal{L}_{\text{GAN}}^{\text{reverse}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$ of Equation (4.32), it suffices to show that $-f_{\text{GAN}}^*(f_{\text{GAN}}'(r_{\boldsymbol{\alpha}}(\mathbf{z}; \mathbf{x}))) = \log \mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z}; \mathbf{x})$ and $f_{\text{GAN}}'(r_{\boldsymbol{\alpha}}(\mathbf{z}; \mathbf{x})) = \log(1 - \mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z}; \mathbf{x}))$, where $\mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z}; \mathbf{x}) \triangleq 1 - \sigma(\log r_{\boldsymbol{\alpha}}(\mathbf{z}; \mathbf{x}))$. First we compute the first derivative f_{GAN}' and the convex dual f_{GAN}^* of f_{GAN}, which involve straightforward calculations,

$$f_{\text{GAN}}'(u) = \log \sigma(\log u), \quad f_{\text{GAN}}^*(t) = -\log(1 - \exp t).$$

Thus, the composition $(f_{\text{GAN}}^* \circ f_{\text{GAN}}') : u \mapsto f_{\text{GAN}}^*(f_{\text{GAN}}'(u))$ can be simplified as

$$f_{\text{GAN}}^*(f_{\text{GAN}}'(u)) = -\log(1 - \exp f_{\text{GAN}}'(u)) = -\log(1 - \sigma(\log u)).$$

Applying f_{GAN}' and $f_{\text{GAN}}^* \circ f_{\text{GAN}}'$ to $r_{\boldsymbol{\alpha}}(\mathbf{z}; \mathbf{x})$, we have

$$f_{\text{GAN}}'(r_{\boldsymbol{\alpha}}(\mathbf{z}; \mathbf{x})) = \log \sigma(\log r_{\boldsymbol{\alpha}}(\mathbf{z}; \mathbf{x})) = \log(1 - \mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z}; \mathbf{x})),$$

and

$$f^*_{\text{GAN}}(f'_{\text{GAN}}(r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}))) = -\log(1-\sigma(\log r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}))) = -\log \mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}),$$

respectively, as required. □

Lemma 4.5.3. *By specifying a discriminator $\mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}) = \mathbf{D}_{\boldsymbol{\alpha}}(\mathbf{z})$ that ignores auxiliary input $\mathbf{x}$, and mapping $\mathcal{G}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon};\mathbf{x}) = \mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x})$ that ignores noise input $\boldsymbol{\epsilon}$, $\mathcal{L}^{\text{reverse}}_{\text{GAN}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$ reduces to $\ell^{\text{reverse}}_{\text{GAN}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$.*

Proof. Through reparameterisation of $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$, we have

$$\begin{aligned}\mathcal{L}^{\text{reverse}}_{\text{GAN}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi}) = \mathbb{E}_{q^*(\mathbf{x})p^*(\mathbf{z})}&[\log \mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})] \\ &+ \mathbb{E}_{q^*(\mathbf{x})p(\boldsymbol{\epsilon})}[\log(1-\mathcal{D}_{\boldsymbol{\alpha}}(\mathcal{G}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon};\mathbf{x});\mathbf{x}))].\end{aligned}$$

By specifying a discriminator $\mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}) = \mathbf{D}_{\boldsymbol{\alpha}}(\mathbf{z})$ that ignores auxiliary input $\mathbf{x}$, and mapping $\mathcal{G}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon};\mathbf{x}) = \mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x})$ that ignores noise input $\boldsymbol{\epsilon}$, this reduces to

$$\begin{aligned}\mathcal{L}^{\text{reverse}}_{\text{GAN}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi}) &= \mathbb{E}_{p^*(\mathbf{z})}[\log \mathbf{D}_{\boldsymbol{\alpha}}(\mathbf{z})] + \mathbb{E}_{q^*(\mathbf{x})}[\log(1-\mathbf{D}_{\boldsymbol{\alpha}}(\mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x})))] \\ &= \ell^{\text{reverse}}_{\text{GAN}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi}),\end{aligned}$$

as required. □

Proposition 4.5.4. *The reverse adversarial objective $\ell^{\text{reverse}}_{\text{GAN}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$ can be subsumed as an instance of the generalised variational lower bound $\mathcal{L}^{\text{latent}}_{f}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$.*

Theorem 4.5.4 follows directly from Theorems 4.5.2 and 4.5.3.

Now, by Theorem 4.3.2, objective $\mathcal{L}^{\text{reverse}}_{\text{GAN}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$ is maximised exactly when $r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}) = r^*(\mathbf{z};\mathbf{x})$. Hence, we can interpret $\mathcal{L}^{\text{reverse}}_{\text{GAN}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$ as an objective for density-ratio estimation based on probabilistic classification, while $\mathcal{L}^{\text{latent}}_{\text{KL}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$ is an objective based on KLIEP.

Now, the default choice of DM loss is $\mathcal{L}^{\text{reverse}}_{\text{GAN}_\text{A}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) \triangleq \mathcal{L}^{\text{reverse}}_{\text{GAN}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$. Omitting terms not involving $\boldsymbol{\phi}$, this is given by

$$\mathcal{L}^{\text{reverse}}_{\text{GAN}_\text{A}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) \triangleq \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}[\log(1-\mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}))]. \tag{4.33}$$

Unlike $\mathcal{L}^{\text{latent}}_{\text{KL}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha})$, minimizing $\mathcal{L}^{\text{reverse}}_{\text{GAN}_\text{A}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha})$ does not minimise the KL divergence of Equation (4.8). Hence, the minimisation problem of Equation (4.31b) does not correspond to that of Equation (4.14b), and so does not maximise the ELBO, or any known VI objective.

4.5.3.2 *Recovering KL Through Alternative Divergence Minimisation Losses*

Although the default choice of DM loss does not yield a tight correspondence to VI, the existing CYCLEGAN frameworks – and indeed most GAN-based approaches – arbitrarily select an alternative DM loss that avoids vanishing gradients, and work well in practice. Hence, one need only choose an alternative that *does* correspond to minimizing the KL divergence of Equation (4.8).

Firstly, of the CYCLEGAN methods, Kim et al. [122] adopt the widely-used DM loss originally suggested by Goodfellow et al. [84],

$$\mathcal{L}^{\text{reverse}}_{\text{GAN}_\text{B}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) \triangleq \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}[-\log \mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})], \tag{4.34}$$

while [294] optimise the Least-Squares GAN (LSGAN) objectives of [158].

Consider the *combination* of losses $\mathcal{L}^{\text{reverse}}_{\text{GAN}_\text{A}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha})$ and $\mathcal{L}^{\text{reverse}}_{\text{GAN}_\text{B}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha})$,

$$\begin{aligned}\mathcal{L}^{\text{reverse}}_{\text{GAN}_\text{C}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) &\triangleq \mathcal{L}^{\text{reverse}}_{\text{GAN}_\text{A}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) + \mathcal{L}^{\text{reverse}}_{\text{GAN}_\text{B}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) \\ &= \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}\left[-\log \frac{\mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})}{1-\mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})}\right].\end{aligned} \tag{4.35}$$

Proposition 4.5.5. *We have* $\mathcal{L}^{\text{reverse}}_{\text{GAN}_\text{C}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) = \mathcal{L}^{\text{latent}}_{\text{KL}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha})$.

Theorem 4.5.5 was originally noted by Sønderby et al. [233] and is shown below.

Proof. Expanding out $\mathcal{L}^{\text{reverse}}_{\text{GANC}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha})$, we have

$$\begin{aligned}\mathcal{L}^{\text{reverse}}_{\text{GANC}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) &= \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}\left[-\log \frac{\mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})}{1-\mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})}\right] \\ &= \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}\left[\log \frac{\sigma(\log r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}))}{1-\sigma(\log r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}))}\right] \\ &= \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}[\log r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})] \triangleq \mathcal{L}^{\text{latent}}_{\text{KL}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha}).\end{aligned}$$

Hence, $\mathcal{L}^{\text{reverse}}_{\text{GANC}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) = \mathcal{L}^{\text{latent}}_{\text{KL}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha})$ as required. □

Thus, for the setting of the DM loss $\mathcal{L}^{\text{reverse}}_{\text{GAN}_\text{C}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha})$, the minimisation problem of Equation (4.31b) corresponds to that of Equation (4.14b), and thus maximises the ELBO. This is equivalent to fitting the density ratio estimator $r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})$ by maximizing the objective $\mathcal{L}^{\text{reverse}}_{\text{GAN}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$ instead of $\mathcal{L}^{\text{latent}}_{\text{KL}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$, and plugging it back into $\mathcal{L}^{\text{latent}}_{\text{KL}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha})$ to approximately minimise the KL divergence of Equation (4.8). Such an approach is prevalent among existing implicit VI methods [107, 167, 196, 261]

SUMMARY OF THEORETICAL CONNECTIONS We have that the cycle-consistency losses are a specific instance of the ELL and ELP, while the adversarial objectives are a specific instance of the variational lower bound for divergence estimation, the maximisation of which can be seen as density ratio estimation by probabilistic classification. By explicitly setting the corresponding divergence minimisation loss such that it leads to minimisation of the required KL divergence terms in the ELBO and APLBO, we subsume the CYCLEGAN model under our proposed VI framework. See Section 4.B for a succinct summary of the relationships.

4.6 RELATED WORK

This paper is closely related to the recent works that seek to extend the scope of VI to implicit distributions. A recurring theme throughout this line of work is approximation of the ELBO by exploiting the formal connection between density ratio estimation and GANS [172, 265]. The major variation is in the choice of the target density ratio being estimated, which is dictated by the problem setting. Makhzani et al. [156] and Mescheder, Nowozin, and Geiger [167] estimate the density ratio $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) / p(\mathbf{z})$ so as to allow for expressive sample-based posterior approximations $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$. This corresponds to the reverse KL minimisation component of our approach, wherein we allow for implicit priors $p(\mathbf{z})$.

Similar to BIGAN [61] and ALI [58], Tran, Ranganath, and Blei [261, LFVI] match a variational joint to an exact joint distribution by estimating the density ratio $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) / q_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{z})$ and using it to approximately minimise the KL divergence. Although this formulation relaxes the requirement of having *any* tractable densities, their focus is on inference for models with intractable likelihoods $p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})$, and also incorporate the implicit posteriors of AVB. In our setting, the joint's intractability is due instead to the implicit prior $p^*(\mathbf{z})$. While we also approximate the exact joint, we do so by minimizing a *symmetric* KL divergence. Furthermore, since both $p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})$ and $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$ are prescribed, we incorporate them explicitly within our loss functions, and estimate a different set of density ratios. This closely resembles the approach of Pu et al. [196], which also minimises the symmetric KL divergence between the joints. However, the focus of their method is not on implicit distributions, and thus specify a different set of losses than ours – one that requires solving more complicated density ratio estimation problems. More importantly, their method does not yield a tight correspondence to CYCLEGAN models.

A consequence of solely minimizing the forward KL (as in frameworks such as adversarially learned inference (ALI)) in contrast to minimizing the symmetric KL (as in our framework), is that of non-identifiability. This issue has been addressed by Li et al. [140] who proposed a conditional-entropy regulariser to ALI's objective. Although Li et al. [140] relate their method to CYCLEGANS, such a relationship is not made explicit mathematically as we do in this work. Furthermore, unlike Li et al. [140], we derive our full objective within an approximate Bayesian inference perspective. More recently, Chen et al. [31] also highlight inference issues associated with ALI regarding the quality of data generated from the inferred latent variables. They propose a symmetric variational autoencoder that inherits the realistic image generation feature of adversarial approaches while overcoming the asymmetry limitations of the forward KL divergence, characteristic of standard VAEs. Unlike our approach, they do not provide an explicit relationship with CYCLEGANS.

Finally, similar to InfoGAN [35] and VEEGAN [239], the forward KL minimisation component of our method also optimises a model of the latent variables, which is reminiscent of the wake-sleep algorithm for training Helmholtz machines [51]. This is discussed further by Hu et al. [106], who provide a comprehensive treatment of the links between the work mentioned in this section, and importantly, the symmetric perspective of generation on recognition that is fundamental to our approach.

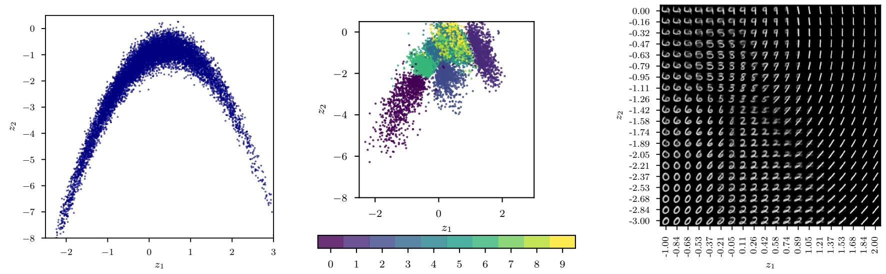
4.7 EXPERIMENTS

SYNTHETIC DATA. First we consider the problem of reducing the dimensionality of the MNIST dataset to a 2D latent space, wherein the prior distribution on the latent representations is specified by its samples (shown in Figure 4.2a). This "banana-shaped distribution" is a commonly used testbed for adaptive MCMC methods [91, 258]. Its samples can be generated by drawing from a bivariate Gaussian with unit variances and correlation $\rho = 0.95$, and transforming them through mapping $H(z_1, z_2) \triangleq [z_1, z_2 - z_1^2 - 1]^T$. While the density of this distribution can be computed, it is withheld from our algorithm and used only in the VAE baseline, which does not permit implicit distributions.

Qualitative and quantitative results are given Figure 4.2 and Table 4.1, which demonstrate superior performance to competing approaches. Observe that instances of the various digit classes are disentangled in this latent space, while still closely matching the shape of the prior distribution, despite having only access to its samples. The resulting manifold of reconstructions is depicted in Figure 4.2c.

In Table 4.1, we report the mean-squared error (MSE) on the reconstructions of observations from the held-out test set and benchmark against VAE / AVB. Also, for the joint approximation to properly match the support of the exact joint, the latent codes should also be representable by its corresponding observation. Hence, we also report the MSE between samples from the prior and their reconstructions. While we find no improvements on reconstruction quality of observations, our method significantly outperforms others in reconstructing latent codes, suggesting our method has greater capacity to faithfully approximate the exact joint.

IMAGE-TO-IMAGE TRANSLATION. We apply our method to the task of transferring features between images of faces on the CelebA dataset [148]. We consider the case where one feature differs between domains. In particular, distributions $q^*(\mathbf{x})$ and $q^*(\mathbf{z})$ are specified by images of women with blond and black hair, respectively. We spec-



(a) 10k samples from implicit prior $p^*(\mathbf{z})$.

(b) Mean of $q_\phi(\mathbf{z} \mid \mathbf{x})$ for every $\mathbf{x}$ from held-out test set of size 10k, colored by digit class.

(c) Mean of $p_\theta(\mathbf{x} \mid \mathbf{z})$ for 20×20 values of $\mathbf{z}$ along a uniform grid.

Figure 4.2: Visualisation of 2D latent space and the corresponding observed space manifold.Instances of the various digit classes are disentangled in this latent space, while still closely matching the shape of the prior distribution, despite having only access to its samples .

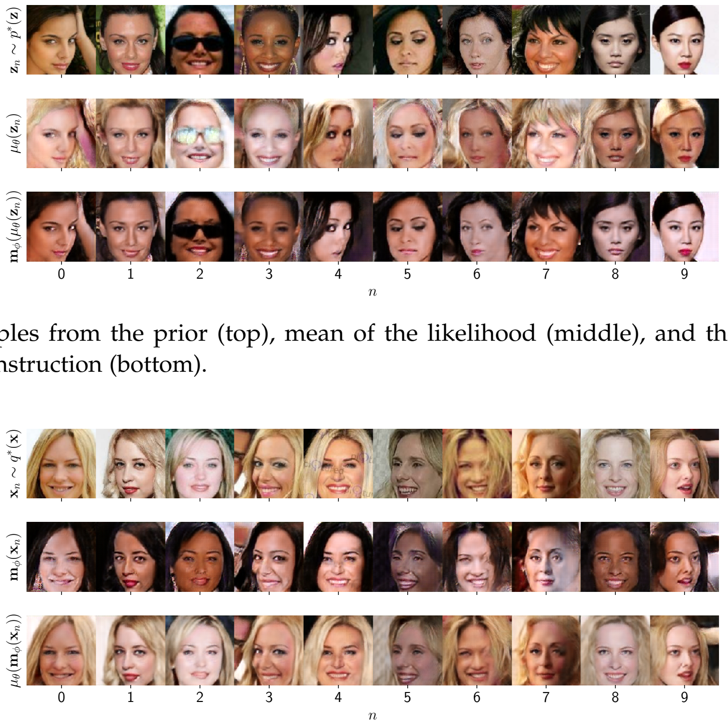
Table 4.1: Mean-squared errors of reconstructions.

METHOD	MSE z	MSE x
SJMVI (OURS)	**0.17**	**0.04**
VAE [124]	0.88	**0.04**
AVB [167]	0.29	**0.04**

ify both $p_\theta(\mathbf{x} \mid \mathbf{z})$ and $q_\phi(\mathbf{z} \mid \mathbf{x})$ as a Laplace distribution, with fixed variance, and mean functions $\boldsymbol{\mu}_\theta(\mathbf{z})$ and $\mathbf{m}_\phi(\mathbf{x})$ defined by neural networks. Their architectures, as well as those of discriminators $\mathcal{D}_\alpha(\mathbf{z}; \mathbf{x})$ and $\mathcal{D}_\beta(\mathbf{x}; \mathbf{z})$ are defined in the same way as in [294]. In Figure 4.3, we show the outputs of the mean functions on samples from a hold-out test set, after training for 10 epochs. From Figure 4.3b, we see that given datapoints $\mathbf{x}$ (top) we're able to learn a posterior over latent representations $\mathbf{z}$ in the other domain (mean is shown in middle). Furthermore, these latent representations are configured so as to maximise the likelihood of observing the original data, as evident from the reconstructions (bottom). Refer to appendix Figure 4.4 for qualitative results produced by the CYCLEGAN baseline approach [122, 294].

4.8 SUMMARY

we have provided a theoretical treatment of the link between CYCLEGANS and approximate Bayesian inference. In short, samples from the two domains correspond respectively to those drawn from the data and implicit prior distribution in a implicit LVM (ILVM). Parameter learning in CYCLEGANS corresponds to approximate inference in this

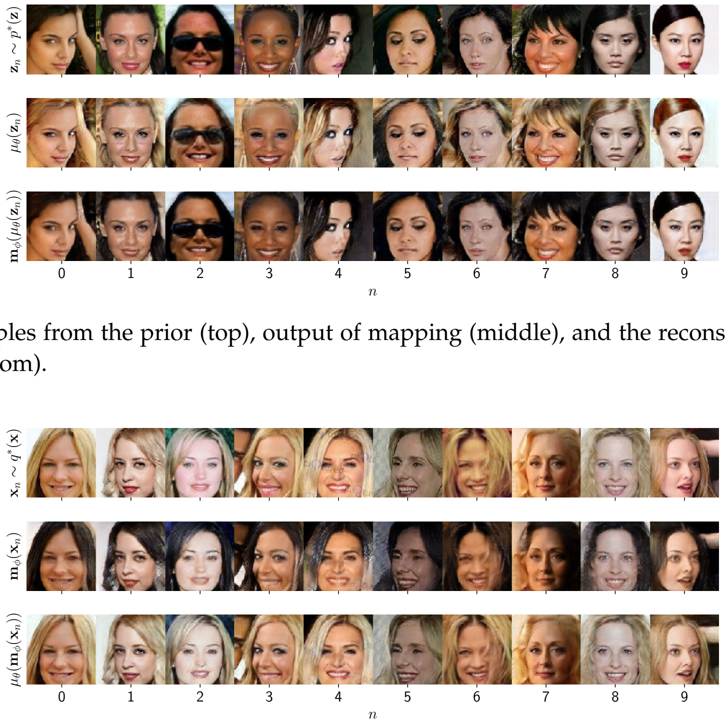

$\mathbf{z}_n \sim p^*(\mathbf{z})$
$\mu_\theta(\mathbf{z}_n)$
$\mathbf{m}_\phi(\mu_\theta(\mathbf{z}_n))$
0
1
2
3
4
5
6
7
8
9
n
$\mathbf{x}_n \sim q^*(\mathbf{x})$
$\mathbf{m}_\phi(\mathbf{x}_n)$
$\mu_\theta(\mathbf{m}_\phi(\mathbf{x}_n))$
0
1
2
3
4
5
6
7
8
9
n

(a) Samples from the prior (top), mean of the likelihood (middle), and the mean reconstruction (bottom).

(b) Samples from the data (top), mean of the posterior (middle), and the mean reconstruction (bottom).

Figure 4.3: Image-to-image translation (blond to black hair) on CelebA dataset [148] performed by our proposed approach.

ILVM under our proposed VI framework. The forward and reverse mappings in CYCLEGANS arise naturally in the generative and recognition models, while the cycle-consistency constraints correspond to their log probabilities, and the adversarial losses are approximations to an f-divergence. By lifting the limitations of prescribed prior distributions in favor of arbitrarily flexible implicit distributions, we can discover different perspectives on existing learning methods and provide more flexible approaches to probabilistic modeling.


$\mathbf{z}_n \sim p^*(\mathbf{z})$
$\mu_\theta(\mathbf{z}_n)$
$\mathbf{m}_\phi(\mu_\theta(\mathbf{z}_n))$
0
1
2
3
4
5
6
7
8
9
n
$\mathbf{x}_n \sim q^*(\mathbf{x})$
$\mathbf{m}_\phi(\mathbf{x}_n)$
$\mu_\theta(\mathbf{m}_\phi(\mathbf{x}_n))$
0
1
2
3
4
5
6
7
8
9
n

(a) Samples from the prior (top), output of mapping (middle), and the reconstruction (bottom).

(b) Samples from the data (top), output of mapping (middle), and the reconstruction (bottom).

Figure 4.4: Image-to-image translation (blond to black hair) on CelebA dataset [148] performed by the baseline CYCLEGAN method.

ADDENDUM

4.A RELATION TO KL IMPORTANCE ESTIMATION PROCEDURE (KLIEP)

We now discuss the connections to KLIEP [244]. Consider the same problem setting as in Section 4.3.3 where we wish to use a parameterised function $r_{\boldsymbol{\alpha}}$ to estimate the exact density ratio,

$$r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}) \approx r^*(\mathbf{z};\mathbf{x}) \triangleq \frac{q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})}{p^*(\mathbf{z})}.$$

We can view $r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})$ as the correction factor required for $p^*(\mathbf{z})$ to match $q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})$. This gives rise to an estimator of $q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})$,

$$q_{\boldsymbol{\alpha}}(\mathbf{z}\,|\,\mathbf{x}) \triangleq r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})p^*(\mathbf{z}) \approx q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x}).$$

Although in our specific problem setting, the density $q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})$ is tractable, we nonetheless fit an auxiliary model $q_{\boldsymbol{\alpha}}(\mathbf{z}\,|\,\mathbf{x})$ to it as a means of fitting the underlying density ratio estimator $r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})$.

In particular, consider *minimizing* the KL divergence between $q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})$ and $q_{\boldsymbol{\alpha}}(\mathbf{z}\,|\,\mathbf{x})$ with respect to $\boldsymbol{\alpha}$,

$$\begin{aligned}
&\mathbb{E}_{q^*(\mathbf{x})}\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x}) \parallel q_{\boldsymbol{\alpha}}(\mathbf{z}\,|\,\mathbf{x})\right] \\
&\quad\triangleq \mathbb{E}_{q^*(\mathbf{x})}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})}\left[\log\frac{q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})}{q_{\boldsymbol{\alpha}}(\mathbf{z}\,|\,\mathbf{x})}\right], \\
&\quad= \mathbb{E}_{q^*(\mathbf{x})}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})}\left[\log\frac{q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})}{p^*(\mathbf{z})r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})}\right], \\
&\quad= -\mathbb{E}_{q^*(\mathbf{x})}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})}[\log r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})] + \text{const.}
\end{aligned}$$

Hence, this is equivalent to *maximizing*

$$\mathbb{E}_{q^*(\mathbf{x})}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})}[\log r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})].$$

Now, for the conditional $q_{\boldsymbol{\alpha}}(\mathbf{z}\,|\,\mathbf{x})$ to be a probability density function, its integral must sum to one,

$$\int q_{\boldsymbol{\alpha}}(\mathbf{z}\,|\,\mathbf{x})q^*(\mathbf{x})d\mathbf{x}d\mathbf{z} = 1.$$

Rewriting this integral, we have the constraint

$$\begin{aligned}
\int q_{\boldsymbol{\alpha}}(\mathbf{z}\,|\,\mathbf{x})q^*(\mathbf{x})d\mathbf{x}d\mathbf{z} &= \int r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})p^*(\mathbf{z})q^*(\mathbf{x})d\mathbf{x}d\mathbf{z} \\
&= \mathbb{E}_{q^*(\mathbf{x})p^*(\mathbf{z})}[r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})] = 1.
\end{aligned}$$

Table 4.B.1: Relevant latent and observed space f-divergences instantiated for particular settings of f.

		Reverse KL	GAN
	$f(u)$	$u \log u$	$u \log u - (u+1)\log(u+1)$
Latent	$\mathbb{E}_{q^*(\mathbf{x})}\mathcal{D}_f\left[p^*(\mathbf{z}) \parallel q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})\right]$	$\mathbb{E}_{q^*(\mathbf{x})}\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \parallel p^*(\mathbf{z})\right]$	$2 \cdot \mathbb{E}_{q^*(\mathbf{x})}\mathcal{D}_{\mathrm{JS}}\left[p^*(\mathbf{z}) \parallel q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})\right] - \log 4$
Observed	$\mathbb{E}_{p^*(\mathbf{z})}\mathcal{D}_f\left[q^*(\mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})\right]$	$\mathbb{E}_{p^*(\mathbf{z})}\mathrm{KL}\left[p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}) \parallel q^*(\mathbf{x})\right]$	$2 \cdot \mathbb{E}_{p^*(\mathbf{z})}\mathcal{D}_{\mathrm{JS}}\left[q^*(\mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})\right] - \log 4$

Combined, we have the following constrained optimisation problem,

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \mathbb{E}_{q^*(\mathbf{x})}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}[\log r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})] \\ \text{subject to} \quad & \mathbb{E}_{q^*(\mathbf{x})p^*(\mathbf{z})}[r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}) - 1] = 0. \end{aligned}$$

Through the method of Lagrange multipliers, this can be cast as an unconstrained optimisation problem with objective,

$$\begin{aligned} \mathcal{L}_{\mathrm{KLIEP}}^{\mathrm{latent}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi}) \triangleq {} & \mathbb{E}_{q^*(\mathbf{x})}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}[\log r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})] \\ & - \lambda\mathbb{E}_{q^*(\mathbf{x})p^*(\mathbf{z})}[r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}) - 1], \end{aligned}$$

where λ is the Lagrange multiplier. For $\lambda = 1$, $\mathcal{L}_{\mathrm{KLIEP}}^{\mathrm{latent}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$ trivially reduces to $\mathcal{L}_{\mathrm{KL}}^{\mathrm{latent}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$.

4.B SUMMARY OF DEFINITIONS

In this section, we summarise the definitions of the losses defined in the proposed VI framework of Sections 4.3 and 4.4, and underscore the relationships to their respective counterparts in the CYCLEGAN framework of Section 4.5.

Table 4.B.1 summarises the settings of convex function $f : \mathbb{R}_+ \to \mathbb{R}$ that recover the reverse KL divergence terms within the ELBO and APLBO, and the JS divergence (up to constants) that GANs are known to minimise.

Table 4.B.2 gives the calculations of the terms necessary to explicitly write down instances of the generalised variational lower bound for particular convex functions f – namely the convex dual $f^\star$, the first derivative f' and the composition $f^\star \circ f'$.

Table 4.B.3 gives instances of the variational lower bound that approximate the latent and observed space KL divergences within the ELBO and APLBO, respectively. Additionally, it gives generalised *stochastic* formulations of the GAN objectives in the CYCLEGAN framework, while Table 4.B.4 lists their *deterministic* counterpart.

Lastly, Table 4.B.5 gives forward and reverse cycle-consistency constraints in the CYCLEGAN framework, and the specific class of Gaussian

Table 4.B.2: Calculations for convex functions.

	Reverse KL	GAN
$f(u)$	$u \log u$	$u \log u - (u+1)\log(u+1)$
$f^\star(t)$	$\exp(t-1)$	$-\log(1-\exp t)$
$f'(u)$	$1+\log u$	$\log \sigma(\log u)$
$f^\star(f'(u))$	u	$-\log(1-\sigma(\log u))$

Table 4.B.3: Instances of variational lower bounds on the relevant latent and observed space f-divergences.

		Reverse KL	GAN
	$f(u)$	$u \log u$	$u \log u - (u+1)\log(u+1)$
Latent	$\mathcal{L}_f^{\text{latent}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi}) \triangleq \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}[f'(r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}))]$ $- \mathbb{E}_{q^*(\mathbf{x})p^*(\mathbf{z})}[f^\star(f'(r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})))]$	$\mathcal{L}_{\text{KL}}^{\text{latent}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi}) \triangleq \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}[\log r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})]$ $- \mathbb{E}_{q^*(\mathbf{x})p^*(\mathbf{z})}[r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}) - 1]$	$\mathcal{L}_{\text{GAN}}^{\text{reverse}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi}) \triangleq \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}[\log \sigma(\log r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}))]$ $+ \mathbb{E}_{q^*(\mathbf{x})p^*(\mathbf{z})}[\log(1-\sigma(\log r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})))]$
Observed	$\mathcal{L}_f^{\text{observed}}(\boldsymbol{\beta} \mid \boldsymbol{\theta}) \triangleq \mathbb{E}_{p^*(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})}[f'(r_{\boldsymbol{\beta}}(\mathbf{x};\mathbf{z}))]$ $- \mathbb{E}_{p^*(\mathbf{z})q^*(\mathbf{x})}[f^\star(f'(r_{\boldsymbol{\beta}}(\mathbf{x};\mathbf{z})))]$	$\mathcal{L}_{\text{KL}}^{\text{observed}}(\boldsymbol{\beta} \mid \boldsymbol{\theta}) \triangleq \mathbb{E}_{p^*(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})}[\log r_{\boldsymbol{\beta}}(\mathbf{x};\mathbf{z})]$ $- \mathbb{E}_{p^*(\mathbf{z})q^*(\mathbf{x})}[r_{\boldsymbol{\beta}}(\mathbf{x};\mathbf{z}) - 1]$	$\mathcal{L}_{\text{GAN}}^{\text{forward}}(\boldsymbol{\beta} \mid \boldsymbol{\theta}) \triangleq \mathbb{E}_{p^*(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})}[\log \sigma(\log r_{\boldsymbol{\beta}}(\mathbf{x};\mathbf{z}))]$ $+ \mathbb{E}_{p^*(\mathbf{z})q^*(\mathbf{x})}[\log(1-\sigma(\log r_{\boldsymbol{\beta}}(\mathbf{x};\mathbf{z})))]$

likelihoods and posteriors that instantiates these constraints (in the limit).

Table 4.B.4: General stochastic GAN objectives and their deterministic counterparts.

	STOCHASTIC	DETERMINISTIC
REVERSE	$\mathcal{L}_{\text{GAN}}^{\text{reverse}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi}) \triangleq \mathbb{E}_{q^*(\mathbf{x})p^*(\mathbf{z})}[\log \mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z}; \mathbf{x})]$ $+ \mathbb{E}_{q^*(\mathbf{x})p(\boldsymbol{\epsilon})}[\log(1 - \mathcal{D}_{\boldsymbol{\alpha}}(\mathcal{G}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}; \mathbf{x}); \mathbf{x}))]$	$\ell_{\text{GAN}}^{\text{reverse}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi}) \triangleq \mathbb{E}_{p^*(\mathbf{z})}[\log \mathbf{D}_{\boldsymbol{\alpha}}(\mathbf{z})]$ $+ \mathbb{E}_{q^*(\mathbf{x})}[\log(1 - \mathbf{D}_{\boldsymbol{\alpha}}(\mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x})))]$
FORWARD	$\mathcal{L}_{\text{GAN}}^{\text{forward}}(\boldsymbol{\beta} \mid \boldsymbol{\theta}) \triangleq \mathbb{E}_{p^*(\mathbf{z})q^*(\mathbf{x})}[\log \mathcal{D}_{\boldsymbol{\beta}}(\mathbf{x}; \mathbf{z})]$ $+ \mathbb{E}_{p^*(\mathbf{z})p(\boldsymbol{\xi})}[\log(1 - \mathcal{D}_{\boldsymbol{\beta}}(\mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{\xi}; \mathbf{z}); \mathbf{z}))]$	$\ell_{\text{GAN}}^{\text{forward}}(\boldsymbol{\beta} \mid \boldsymbol{\theta}) \triangleq \mathbb{E}_{q^*(\mathbf{x})}[\log \mathbf{D}_{\boldsymbol{\beta}}(\mathbf{x})]$ $+ \mathbb{E}_{p^*(\mathbf{z})}[\log(1 - \mathbf{D}_{\boldsymbol{\beta}}(\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z})))]$

Table 4.B.5: Negative expected log conditionals and the cycle-consistency constraints.

GAUSSIAN		DEGENERATE	
$p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})$	$q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$	$p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})$	$q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$
$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}), \tau^{-1}\mathbf{I})$	$\mathcal{N}(\mathbf{z} \mid \mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x}), t^{-1}\mathbf{I})$	$\delta(\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}))$	$\delta(\mathbf{z} - \mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x}))$
$\mathcal{L}_{\text{NELL}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \triangleq \frac{\tau}{2}\mathbb{E}_{q^*(\mathbf{x})p(\boldsymbol{\epsilon})}[\lVert\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x}) + t^{-\frac{1}{2}}\boldsymbol{\epsilon})\rVert_2^2] + \frac{D}{2}\log\frac{2\pi}{\tau}$		$\ell_{\text{CONST}}^{\text{reverse}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \triangleq \mathbb{E}_{q^*(\mathbf{x})}[\lVert\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x}))\rVert_2^2]$	
$\mathcal{L}_{\text{NELP}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \triangleq \frac{t}{2}\mathbb{E}_{p^*(\mathbf{z})p(\boldsymbol{\xi})}[\lVert\mathbf{z} - \mathbf{m}_{\boldsymbol{\phi}}(\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}) + \tau^{-\frac{1}{2}}\boldsymbol{\xi})\rVert_2^2] + \frac{K}{2}\log\frac{2\pi}{t}$		$\ell_{\text{CONST}}^{\text{forward}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \triangleq \mathbb{E}_{p^*(\mathbf{z})}[\lVert\mathbf{z} - \mathbf{m}_{\boldsymbol{\phi}}(\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}))\rVert_2^2]$	