Cycle-Consistent Adversarial Learning as Approximate Bayesian Inference

Louis C. Tiao [1] **Edwin V. Bonilla** [2] **Fabio Ramos** [1]

Abstract

We formalize the problem of learning inter-domain correspondences in the absence of paired data as Bayesian inference in a latent variable model (LVM), where one seeks the underlying hidden representations of entities from one domain as entities from the other domain. First, we introduce *implicit latent variable model*s, where the prior over hidden representations can be specified flexibly as an *implicit* distribution. Next, we develop a new variational inference (VI) algorithm for this model based on minimization of the *symmetric* Kullback-Leibler (KL) divergence between a variational joint and the exact joint distribution. Lastly, we demonstrate that the state-of-the-art cycle-consistent adversarial learning (CYCLEGAN) models can be derived as a special case within our proposed VI framework, thus establishing its connection to approximate Bayesian inference methods.

1. Introduction

Learning correspondences between entities from different domains is an important and challenging problem in machine learning, 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 realization 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 approaches.

[1]University of Sydney, Sydney, Australia [2]University of New South Wales, Sydney, Australia. Correspondence to: Louis Tiao <louis.tiao@sydney.edu.au>.

Presented at the ICML 2018 Workshop on Theoretical Foundations and Applications of Deep Generative Models. Stockholm, Sweden, 2018.

Within these unsupervised approaches, a number of recently proposed cycle-consistent adversarial learning (CYCLEGAN) methods have achieved remarkable success in addressing this problem (Kim et al., 2017; Zhu et al., 2017). As their name suggests, these approaches are based upon two heuristics: (i) adversarial learning and (ii) cycle consistency. The former, adversarial learning (Goodfellow et al., 2014), 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 minimization framework motivated intuitively by the two conceptually simple principles mentioned above, the original CYCLEGAN formulation lacks any further theoretical justification. This hinders (i) understanding of the distributional assumptions of all the variables of interest; (ii) how these are combined in terms of prior knowledge and observational assumptions; and (iii) whether more general methods can be used for estimating their parameters. In contrast, LVMs offer a principled framework for probabilistic reasoning about random variables, their statistical properties and dependency structures, with the goal of capturing the underlying data-generation process. Besides providing sound quantification of uncertainty, a LVM allows us to disentangle our modeling assumptions from the inference machinery we use to reason about the variables in the model. Interpreting standard methods from a Bayesian perspective has contributed significantly to the understanding of these methods and to the development of new approaches. Examples of this include the seminal work on probabilistic principal component analysis (PPCA) (Tipping & Bishop, 1999) and more recently, the advances in understanding Dropout (Gal & Ghahramani, 2015). Our contributions are threefold, and are summarized below.

First contribution (§ 2). We formulate the problem of learning correspondences between domains using unpaired data from a LVM perspective, where entities in one domain are latent representations of entities in the other domain. Crucial to our approach is to consider an *implicit* prior over these hidden representations, i.e. a prior that is not given by a prescribed distribution such as a Gaussian, but instead, is provided via samples from an unknown process.

Second contribution (§§ 3 and 4). We develop a new scalable variational inference algorithm for these types of models. Unlike standard VI approaches that approximate the posterior over the latent variables, we approximate the exact joint distribution over latent and observed variables with a variational joint. Furthermore, we carry out this approximation via the minimization of the symmetric KL divergence between these joints. This is in stark contrast with traditional VI approaches that minimize the forward KL divergence, as the symmetric KL divergence between the posteriors is intractable.

Third contribution (§ 5). Finally, we show that CYCLEGANs, as proposed by (Kim et al., 2017; Zhu et al., 2017) independently, can be recovered by applying our approximate inference algorithm in a LVM for domain correspondence with unpaired data. In this model, the prior is specified by an implicit *empirical* distribution and the observed variables are generated by a nonlinear function of its underlying latent variable. Intuitively, while using the symmetric KL divergence between the approximate and true joint distributions yields generative adversarial network (GAN)-like objectives between the domains, different specifications of the likelihood and the approximate posterior yield the cycle-consistency part of the loss in CYCLEGANs.

2. Implicit Latent Variable Models

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

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}) p(\mathbf{z}). \tag{1}$$

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 factorizes as,

$$p_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Z}) = \prod_{n=1}^{N} p_{\boldsymbol{\theta}}(\mathbf{x}_n, \mathbf{z}_n). \tag{2}$$

The graphical representation of this model is depicted in fig. A.2.

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 $\mathbf{z}$,

$$\begin{aligned} &\mathbf{x} \sim p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}) \\ \Leftrightarrow\ &\mathbf{x} = \mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{\xi}; \mathbf{z}), \quad \boldsymbol{\xi} \sim p(\boldsymbol{\xi}). \end{aligned} \tag{3}$$

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 w.r.t. $\boldsymbol{\xi}$ 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 a class of familiar Gaussian observation models. In appendix B, we give concrete examples of common latent variable models that can be instantiated under this definition.

2.2. Implicit Prior

In LVMs, the prior is almost invariably specified as a prescribed distribution, most typically a factorized 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,

$$\mathbf{z}_m^* \sim p^*(\mathbf{z}). \tag{4}$$

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 (Jaynes, 1968).

2.3. 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 $p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})$—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 $p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})$—a conditional density of $\mathbf{z}$ given $\mathbf{x}$—should produce renderings of landscape scenery in Van Gogh's iconic style.

Parameter learning and inference in implicit latent variable models (ILVMs) is paved with intractabilities. For moderately complicated likelihoods, the marginal likelihood $p_{\boldsymbol{\theta}}(\mathbf{x})$ is intractable, making it infeasible to perform maximum likelihood estimation of $\boldsymbol{\theta}$ and to compute the posterior exactly. Furthermore, the intractability of the prior renders most existing approximate inference methods futile. The remainder of this paper will be devoted to accurate and scalable approximate inference in ILVMs.

3. Variational Inference with Implicit Priors

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 $p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})$ (Blei et al., 2017; Jordan et al., 1999; Wainwright & Jordan, 2008).

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 (Gershman & Goodman, 2014). Namely, 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 optimize a single set of variational parameters $\boldsymbol{\phi}$ across all $\mathbf{x} \sim q^*(\mathbf{x})$.

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

$$\begin{aligned} &\mathbf{z} \sim q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \\ \Leftrightarrow\ &\mathbf{z} = \mathcal{G}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}, \mathbf{x}), \quad \boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}). \end{aligned} \tag{5}$$

Just as mapping $\mathcal{F}_{\boldsymbol{\theta}}$ underpins the generative model, mapping $\mathcal{G}_{\boldsymbol{\phi}}$ underpins the *recognition model* (Dayan et al., 1995). As with the likelihood, we restrict our attention to prescribed variational distributions.

As depicted in fig. A.2b, the dependency relationship between the variational parameters and the latent variables now mirrors that of the model parameters and observed variables. This symmetry is crucial to the derivation of CYCLEGAN later in § 5.3.2.

3.2. Reverse KL Variational Objective

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

$$\begin{aligned} \mathcal{L}_{\text{NELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \coloneqq\ &\mathbb{E}_{q^*(\mathbf{x}) q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}[-\log p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})] \\ &+ \mathbb{E}_{q^*(\mathbf{x})} \mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \parallel p^*(\mathbf{z})\right]. \end{aligned} \tag{6}$$

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

$$\mathcal{L}_{\text{NELL}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \coloneqq \mathbb{E}_{q^*(\mathbf{x}) q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}[-\log p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})]. \tag{7}$$

It is easy to perform stochastic gradient-based optimization of this term by applying the reparameterization trick (Kingma & Welling, 2014; Rezende et al., 2014),

$$\mathcal{L}_{\text{NELL}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{q^*(\mathbf{x}) p(\boldsymbol{\epsilon})}[-\log p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathcal{G}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}, \mathbf{x}))]. \tag{8}$$

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

$$\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \parallel p^*(\mathbf{z})\right] \coloneqq \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}[\log r^*(\mathbf{z}; \mathbf{x})], \tag{9}$$

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

$$r^*(\mathbf{z}; \mathbf{x}) \coloneqq q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) / p^*(\mathbf{z}). \tag{10}$$

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 *density ratio estimation (*DRE*)* (Mohamed & Lakshminarayanan, 2017; Sugiyama et al., 2012).

3.3. Approximate Divergence Minimization

Although we are primarily interested in estimating the KL divergence of eq. 9, we give a generalized treatment that is applicable to all f-divergences (Ali & Silvey, 1966; Ciszar, 1967). We denote a generic member of the family of f-divergences between distributions p and q as $\mathcal{D}_f\left[p \parallel q\right] \coloneqq \mathbb{E}_p[f(q/p)]$, for some convex lower-semicontinuous function $f : \mathbb{R}_+ \to \mathbb{R}$.

Leveraging results from convex analysis, Nguyen et al. (2010) devise a variational lower bound that estimates an f-divergence through samples when either or both of the densities are unavailable. Nowozin et al. (2016) extend this framework to derive GAN objectives that minimize arbitrary f-divergences. These results are underpin our methodology, and we restate a variant of it here for completeness.

Theorem 1 (Nguyen et al. 2010)**.** *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 $q(\mathbf{u})$,

$$\mathcal{D}_f\left[p(\mathbf{u}) \parallel q(\mathbf{u})\right] \geq \max_{\hat{r}\in\mathcal{R}}\{\mathbb{E}_{q(\mathbf{u})}[f'(\hat{r}(\mathbf{u}))] - \mathbb{E}_{p(\mathbf{u})}[f^\star(f'(\hat{r}(\mathbf{u})))]\}, \tag{11}$$

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 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,

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

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 optimize a separate density ratio estimator with parameters $\boldsymbol{\omega}_n$ for each observed $\mathbf{x}_n$. Instead, as with the posterior approximation, we also amortize 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 maximize an instance of the following generalized objective,

$$\boxed{\mathcal{L}_f^{\text{latent}}(\boldsymbol{\alpha}\,|\,\boldsymbol{\phi}) \coloneqq \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})}[f'(r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}))] - \mathbb{E}_{q^*(\mathbf{x})p^*(\mathbf{z})}[f^\star(f'(r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})))].} \tag{12}$$

Corollary 1.1. *We have the lower bound,*

$$\mathbb{E}_{q^*(\mathbf{x})}\mathcal{D}_f\left[p^*(\mathbf{z}) \parallel q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})\right] \geq \max_{\boldsymbol{\alpha}} \mathcal{L}_f^{\text{latent}}(\boldsymbol{\alpha}\,|\,\boldsymbol{\phi}), \tag{13}$$

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 *maximized*, thus tightening the bound of eq. 13 and the estimate of the density ratio $r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})$.

Divergence minimization loss. Conversely, the *divergence minimization (*DM*)* loss, denoted as $\mathcal{L}_f^{\text{latent}}(\boldsymbol{\phi}\,|\,\boldsymbol{\alpha})$, is *minimized* w.r.t. $\boldsymbol{\phi}$ while $\boldsymbol{\alpha}$ remains fixed, thus approximately minimizing the f-divergence. In theory, this should be symmetric to the DRE objective, $\mathcal{L}_f^{\text{latent}}(\boldsymbol{\phi}\,|\,\boldsymbol{\alpha}) \coloneqq \mathcal{L}_f^{\text{latent}}(\boldsymbol{\alpha}\,|\,\boldsymbol{\phi})$. However, alternative settings are often used in practice to alleviate the problem of vanishing gradients, as we shall see in § 5.

By applying corollary 1.1 for the setting $f_{\text{KL}}(u) \coloneqq u\log u$, we instantiate a lower bound on the KL divergence of eq. 9 in the following objective,

$$\mathcal{L}_{\text{KL}}^{\text{latent}}(\boldsymbol{\alpha}\,|\,\boldsymbol{\phi}) \coloneqq \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})}[\log r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})] - \mathbb{E}_{q^*(\mathbf{x})p^*(\mathbf{z})}[r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}) - 1]. \tag{14}$$

As we discuss in appendix G, maximization of the objective in eq. 14 is closely related to the *KL importance estimation procedure (*KLIEP*)* (Sugiyama et al., 2008).

Now, we define the DM loss symmetrically to the DRE objective in eq. 14—terms not involving $\boldsymbol{\phi}$ are omitted,

$$\mathcal{L}_{\text{KL}}^{\text{latent}}(\boldsymbol{\phi}\,|\,\boldsymbol{\alpha}) \coloneqq \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})}[\log r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})] \simeq \mathbb{E}_{q^*(\mathbf{x})}\text{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x}) \parallel p^*(\mathbf{z})\right]. \tag{15}$$

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

$$\max_{\boldsymbol{\alpha}} \quad \mathcal{L}_{\text{KL}}^{\text{latent}}(\boldsymbol{\alpha}\,|\,\boldsymbol{\phi}), \tag{16a}$$

$$\min_{\boldsymbol{\phi},\boldsymbol{\theta}} \quad \mathcal{L}_{\text{KL}}^{\text{latent}}(\boldsymbol{\phi}\,|\,\boldsymbol{\alpha}) + \mathcal{L}_{\text{NELL}}(\boldsymbol{\theta},\boldsymbol{\phi}), \tag{16b}$$

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

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* $q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z})$.

4.1. Variational Joint

Recall that $q^*(\mathbf{x})$ denotes the empirical data distribution. We define a variational approximation to the exact joint distribution of eq. 1 as

$$q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z}) \coloneqq q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})q^*(\mathbf{x}). \tag{17}$$

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

$$\text{KL}_{\text{SYMM}}\left[p \parallel q\right] \coloneqq \text{KL}\left[p \parallel q\right] + \text{KL}\left[q \parallel p\right]. \tag{18}$$

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,

$$\begin{aligned}&\text{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z}) \parallel p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\right] \\ &\quad\coloneqq \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z})}\left[\log q_{\boldsymbol{\phi}}(\mathbf{x},\mathbf{z}) - \log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\right] \\ &\quad= \mathcal{L}_{\text{NELBO}}(\boldsymbol{\theta},\boldsymbol{\phi}) - \mathbb{H}[q^*(\mathbf{x})],\end{aligned} \tag{19, 20}$$

where $\mathbb{H}[q^*(\mathbf{x})] \coloneqq \mathbb{E}_{q^*(\mathbf{x})}[-\log q^*(\mathbf{x})]$ is the entropy of $q^*(\mathbf{x})$, a constant w.r.t. parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. Hence,

minimizing the KL divergence of eq. 19 can be reduced to minimizing $\mathcal{L}_{\text{NELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ of eq. 6, without modification.

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,

$$\begin{aligned}&\text{KL}\left[p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) \parallel q_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{z})\right] \\ &\quad\coloneqq \mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) - \log q_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{z})\right] \qquad (21)\\ &\quad= \mathbb{E}_{p^*(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})}[\log p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}) - \log q_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{z})] \\ &\qquad- \mathbb{H}[p^*(\mathbf{z})]. \qquad (22)\end{aligned}$$

In analogy with the ELBO, we introduce a new variational objective that is minimized when the forward KL divergence of eq. 21 is minimized. First we define the recognition model analog to the marginal likelihood—the *marginal posterior*, or *aggregate posterior*, given by $q_\phi(\mathbf{z}) \coloneqq \int q_\phi(\mathbf{z} \mid \mathbf{x}) q^*(\mathbf{x}) d\mathbf{x}$. It can be approximated by the *aggregate posterior lower bound (*APLBO*)*. For consistency, we give its *negative*, written as

$$\begin{aligned}\mathcal{L}_{\text{NAPLBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \coloneqq \mathbb{E}_{p^*(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})}[-\log q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})] \\ + \mathbb{E}_{p^*(\mathbf{z})}\text{KL}\left[p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}) \parallel q^*(\mathbf{x})\right].\end{aligned} \qquad (23)$$

Furthermore, minimizing the KL divergence of eq. 21 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

$$\mathcal{L}_{\text{NELP}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \coloneqq \mathbb{E}_{p^*(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})}[-\log q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})]. \qquad (24)$$

We emphasize 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 eq. 22. Furthermore, as with the ELBO, we can perform stochastic gradient-based optimization of the ELP term by applying the same reparameterization trick as in eq. 8.

Now, the KL divergence term of the APLBO in eq. 23 can also be expressed as the expected logarithm of a density ratio $r^*(\mathbf{x}; \mathbf{z}) \coloneqq p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}) / q^*(\mathbf{x})$ that involves an intractable density $q^*(\mathbf{x})$—the empirical data distribution. To overcome this, we adopt the same approach as outlined in § 3.3. Namely, we apply theorem 1 to $q^*(\mathbf{x})$ and $p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}^*)$, and fit an amortized density ratio estimator $r_{\boldsymbol{\beta}}(\mathbf{x}; \mathbf{z})$ to $r^*(\mathbf{x}; \mathbf{z})$ by maximizing an instance of the generalized objective,

$$\begin{aligned}\boxed{\mathcal{L}_f^{\text{observed}}(\boldsymbol{\beta} \mid \boldsymbol{\theta}) \coloneqq \mathbb{E}_{p^*(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})}[f'(r_{\boldsymbol{\beta}}(\mathbf{x}; \mathbf{z}))]} \\ - \mathbb{E}_{p^*(\mathbf{z})q^*(\mathbf{x})}[f^\star(f'(r_{\boldsymbol{\beta}}(\mathbf{x}; \mathbf{z})))].\end{aligned} \qquad (25)$$

Corollary 1.2. *We have the lower bound,*

$$\mathbb{E}_{p^*(\mathbf{z})}\mathcal{D}_f\left[q^*(\mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})\right] \geq \max_{\boldsymbol{\beta}} \mathcal{L}_f^{\text{observed}}(\boldsymbol{\beta} \mid \boldsymbol{\theta}), \qquad (26)$$

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

By applying corollary 1.2 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 eq. 23, 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 eqs. 14 and 15, respectively. See table 4 for a summary of explicit definitions.

Hence, in addition to the bi-level optimization problems of eq. 16 we have,

$$\max_{\boldsymbol{\beta}} \quad \mathcal{L}_{\text{KL}}^{\text{observed}}(\boldsymbol{\beta} \mid \boldsymbol{\theta}), \qquad (27a)$$

$$\min_{\boldsymbol{\phi}, \boldsymbol{\theta}} \quad \mathcal{L}_{\text{KL}}^{\text{observed}}(\boldsymbol{\theta} \mid \boldsymbol{\beta}) + \mathcal{L}_{\text{NELP}}(\boldsymbol{\theta}, \boldsymbol{\phi}). \qquad (27b)$$

As shown, the minimizations in eqs. 16b and 27b corresponds to minimization of the symmetric KL over the joints $\text{KL}_{\text{SYMM}}\left[p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) \parallel q_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{z})\right]$, while the maximizations in eqs. 16a and 27a approximates the divergences, or more precisely, the density ratios involving implicit distributions.

5. CycleGAN as a Special Case

In this section, we demonstrate that cycle-consistent adversarial learning (CYCLEGAN) methods (Kim et al., 2017; Zhu et al., 2017) can be instantiated under our proposed VI framework.

5.1. Basic CycleGAN Framework

To address the problem of unpaired image-to-image translation as described in § 2.3, 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 types 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 maximizing an adversarial

objective w.r.t. parameters $\boldsymbol{\alpha}$,

$$\ell_{\text{GAN}}^{\text{reverse}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi}) \coloneqq \mathbb{E}_{p^*(\mathbf{z})}[\log \mathbf{D}_{\boldsymbol{\alpha}}(\mathbf{z})] + \mathbb{E}_{q^*(\mathbf{x})}[\log(1 - \mathbf{D}_{\boldsymbol{\alpha}}(\mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x})))], \tag{28}$$

while minimizing it w.r.t. 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 $\mathbf{z}^* \sim p^*(\mathbf{z})$. An adversarial objective $\ell_{\text{GAN}}^{\text{forward}}(\boldsymbol{\beta} \mid \boldsymbol{\theta})$ is defined for mapping $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ in like manner,

$$\ell_{\text{GAN}}^{\text{forward}}(\boldsymbol{\beta} \mid \boldsymbol{\theta}) \coloneqq \mathbb{E}_{p^*(\mathbf{x})}[\log \mathbf{D}_{\boldsymbol{\beta}}(\mathbf{x})] + \mathbb{E}_{p^*(\mathbf{z})}[\log(1 - \mathbf{D}_{\boldsymbol{\beta}}(\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z})))]. \tag{29}$$

Cycle-consistency. The second type 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,

$$\ell_{\text{CONST}}^{\text{reverse}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \coloneqq \mathbb{E}_{q^*(\mathbf{x})}[\|\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x}))\|_\rho^\rho], \tag{30}$$

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 $\mathbf{z}$,

$$\ell_{\text{CONST}}^{\text{forward}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \coloneqq \mathbb{E}_{p^*(\mathbf{z})}[\|\mathbf{z} - \mathbf{m}_{\boldsymbol{\phi}}(\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}))\|_\rho^\rho]. \tag{31}$$

These objectives are summarized in the following set of optimization problems,

$$\max_{\boldsymbol{\alpha}} \ell_{\text{GAN}}^{\text{reverse}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi}), \quad \max_{\boldsymbol{\beta}} \ell_{\text{GAN}}^{\text{forward}}(\boldsymbol{\beta} \mid \boldsymbol{\theta}), \tag{32a}$$

$$\min_{\boldsymbol{\phi}, \boldsymbol{\theta}} \ell_{\text{GAN}}^{\text{reverse}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) + \ell_{\text{CONST}}^{\text{reverse}}(\boldsymbol{\theta}, \boldsymbol{\phi}), \tag{32b}$$

$$\min_{\boldsymbol{\phi}, \boldsymbol{\theta}} \ell_{\text{GAN}}^{\text{forward}}(\boldsymbol{\theta} \mid \boldsymbol{\beta}) + \ell_{\text{CONST}}^{\text{forward}}(\boldsymbol{\theta}, \boldsymbol{\phi}). \tag{32c}$$

We now highlight the correspondences between these objectives and those of our proposed VI framework, as summarized in the optimization problems of eqs. 16 and 27.

5.2. Cycle-consistency as Conditional Probability Maximization

We now demonstrate that minimizing the cycle-consistency losses corresponds to maximizing the expected log likelihood and variational posterior of eqs. 7 and 24. This can be shown by instantiating specific classes of $p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})$ and $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$ that recover $\ell_{\text{CONST}}^{\text{reverse}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ and $\ell_{\text{CONST}}^{\text{forward}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ from $\mathcal{L}_{\text{NELL}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ and $\mathcal{L}_{\text{NELP}}(\boldsymbol{\theta}, \boldsymbol{\phi})$, respectively.

Proposition 1. *Consider a typical case where the likelihood and the posterior approximation are both Gaussians,*

$$p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}) \coloneqq \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}), \tau^2 \mathbf{I}),$$
$$q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \coloneqq \mathcal{N}(\mathbf{z} \mid \mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x}), t^2 \mathbf{I}).$$

Then, in the limit as the posterior variance tends to zero, $\mathcal{L}_{\text{NELL}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ approaches $\ell_{\text{CONST}}^{\text{reverse}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ for $\rho = 2$, up to constants[1]*. Formally stated,*

$$\mathcal{L}_{\text{NELL}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \to \gamma_1 \ell_{\text{CONST}}^{\text{reverse}}(\boldsymbol{\theta}, \boldsymbol{\phi}) + \delta_1 \quad \text{as } t \to 0$$
$$\propto \ell_{\text{CONST}}^{\text{reverse}}(\boldsymbol{\theta}, \boldsymbol{\phi}).$$

where $\gamma_1 = \frac{1}{2\tau^2}$ and $\delta_1 = \frac{D}{2} \log \frac{\pi}{\gamma_1}$. Likewise,

$$\mathcal{L}_{\text{NELP}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \to \gamma_2 \ell_{\text{CONST}}^{\text{forward}}(\boldsymbol{\theta}, \boldsymbol{\phi}) + \delta_2 \quad \text{as } \tau \to 0$$
$$\propto \ell_{\text{CONST}}^{\text{forward}}(\boldsymbol{\theta}, \boldsymbol{\phi}).$$

where $\gamma_2 = \frac{1}{2t^2}$ and $\delta_2 = \frac{K}{2} \log \frac{\pi}{\gamma_2}$.

The proof is given in appendix C. Hence, roughly speaking, the cycle-consistency losses can be seen as specific cases of the ELL and ELP with *degenerate* conditional distributions. Furthermore, this sheds new light on the roles of the cycle-consistency losses. In particular, the reverse consistency loss—like the ELL term—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.

5.3. Distribution Matching as Approximate Divergence Minimization

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 $\ell_{\text{GAN}}^{\text{forward}}(\boldsymbol{\beta} \mid \boldsymbol{\theta})$.

5.3.1. As Density Ratio Estimation by Probabilistic Classification

Firstly, the connections between GANs, divergence minimization and DRE are well-established (Mohamed & Lakshminarayanan, 2017; Nowozin et al., 2016; Uehara et al., 2016). Although $\ell_{\text{GAN}}^{\text{reverse}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$ is a scoring rule for probabilistic classification (Gneiting & Raftery, 2007), one can readily show that it can also be subsumed as an instance of the generalized 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 eq. 10.

Lemma 1. *By setting $f_{\text{GAN}}(u) = u \log u - (u+1) \log(u+1)$ in the generalized objective $\mathcal{L}_f^{\text{latent}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$ of eq. 12, we*

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

instantiate the objective

$$\mathcal{L}^{\text{reverse}}_{\text{GAN}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi}) \coloneqq \mathbb{E}_{q^*(\mathbf{x})p^*(\mathbf{z})}[\log \mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})] + \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z}\mid\mathbf{x})}[\log(1-\mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}))], \tag{33}$$

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

Lemma 2. *By specifying a discriminator* $\mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}) = \mathbf{D}_{\boldsymbol{\alpha}}(\mathbf{z})$ *which ignores auxiliary input* $\mathbf{x}$*, and mapping* $\mathcal{G}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon};\mathbf{x}) = \mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x})$ *which ignores noise input* $\boldsymbol{\epsilon}$*,* $\mathcal{L}^{\text{reverse}}_{\text{GAN}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$ *reduces to* $\ell^{\text{reverse}}_{\text{GAN}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$.

Proposition 2. *The reverse adversarial objective* $\ell^{\text{reverse}}_{\text{GAN}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$ *can be subsumed as an instance of the generalized variational lower bound* $\mathcal{L}^{\text{latent}}_{f}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$.

Proposition 2 follows directly from lemmas 1 and 2. Their proofs are given in appendices D and E, respectively. Now, by corollary 1.1, objective $\mathcal{L}^{\text{reverse}}_{\text{GAN}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$ is maximized 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 $\mathcal{L}^{\text{reverse}}_{\text{GAN}_{\text{A}}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) \coloneqq \mathcal{L}^{\text{reverse}}_{\text{GAN}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi})$. Omitting terms not involving $\boldsymbol{\phi}$, this is given by

$$\mathcal{L}^{\text{reverse}}_{\text{GAN}_{\text{A}}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) \coloneqq \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z}\mid\mathbf{x})}[\log(1-\mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}))]. \tag{34}$$

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 minimize the KL divergence of eq. 9. Hence, the minimization problem of eq. 32b does not correspond to that of eq. 16b, and so does not maximize the ELBO, or any known VI objective.

5.3.2. Recovering KL Through Alternative Divergence Minimization 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 eq. 9.

Firstly, of the CYCLEGAN methods, Kim et al. (2017) adopt the widely-used DM loss originally suggested by Goodfellow et al. (2014),

$$\mathcal{L}^{\text{reverse}}_{\text{GAN}_{\text{B}}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) \coloneqq \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z}\mid\mathbf{x})}[-\log \mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})], \tag{35}$$

while Zhu et al. (2017) optimize the Least-Squares GAN (LSGAN) objectives of Mao et al. (2016).

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

$$\begin{aligned}\mathcal{L}^{\text{reverse}}_{\text{GAN}_{\text{C}}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) &\coloneqq \mathcal{L}^{\text{reverse}}_{\text{GAN}_{\text{A}}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) + \mathcal{L}^{\text{reverse}}_{\text{GAN}_{\text{B}}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) \\ &= \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z}\mid\mathbf{x})}\left[-\log \frac{\mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})}{1-\mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})}\right].\end{aligned} \tag{36}$$

Proposition 3. *We have* $\mathcal{L}^{\text{reverse}}_{\text{GAN}_{\text{C}}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) = \mathcal{L}^{\text{latent}}_{\text{KL}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha})$.

Proposition 3 was originally observed by Sønderby et al. (2016) and is shown in appendix F. Thus, for the setting of the DM loss $\mathcal{L}^{\text{reverse}}_{\text{GAN}_{\text{C}}}(\boldsymbol{\phi} \mid \boldsymbol{\alpha})$, the minimization problem of eq. 32b corresponds to that of eq. 16b, and thus maximizes 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 minimize the KL divergence of eq. 9. Such an approach is prevalent among existing implicit VI methods (Huszár, 2017; Mescheder et al., 2017; Pu et al., 2017; Tran et al., 2017).

In summary, 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 maximization of which can be seen as density ratio estimation by probabilistic classification. By explicitly setting the corresponding divergence minimization loss such that it leads to minimization of the required KL divergence terms in the ELBO and APLBO, we subsume the CYCLEGAN model under our proposed VI framework. See appendix H for a succinct summary of the relationships.

6. Related Work

This paper is closely related to the recent works that seek to extend the scope of VI to implicit distributions, making it feasible in scenarios where one or more of the densities that constitute the ELBO are not explicitly available. A recurring theme throughout this line of work is approximation of the ELBO by exploiting the formal connection between density ratio estimation and GANs (Mohamed & Lakshminarayanan, 2017; Uehara et al., 2016). The major variation is in the choice of the target density ratio being estimated, which is dictated by the problem setting. Makhzani et al. (2015); Mescheder et al. (2017) estimate the density ratio $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})/p(\mathbf{z})$ so as to allow for arbitrarily expressive sample-based posterior approximations $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$. This corresponds to the reverse KL minimization component of our approach, wherein we also estimate the same density ratio, but instead to allow for implicit prior distributions $p(\mathbf{z})$.

Similar to BIGAN (Dumoulin et al., 2016) and ALI (Donahue et al., 2016), Tran et al. (2017, 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 minimize 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. (2017), which also minimizes 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.

Next, similar to InfoGAN (Chen et al., 2016) and VEEGAN (Srivastava et al., 2017), the forward KL minimization component of our method also optimizes a model of the latent variables, which is reminiscent of the wake-sleep algorithm for training Helmholtz machines (Dayan et al., 1995). This is discussed further by Hu et al. (2017), 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.

7. Experiments

To empirically assess the performance of our approach, 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 fig. 1a). This "banana-shaped distribution" is a commonly used testbed for adaptive MCMC methods (Haario et al., 1999; Titsias, 2017). Its samples can be generated by drawing from a bivariate Gaussian with unit variances and correlation $\rho = 0.95$, and transforming them through mapping $H(z_1, z_2) \coloneqq [z_1, z_2 - z_1^2 - 1]^T$. While the density of this distribution can be computed, it is withheld from our algorithm and used only in the variational autoencoder (VAE) baseline, which does not permit implicit distributions. Figure 1b shows, for every observation $\mathbf{x}$ from a held-out test set, the mean $\mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x})$ of the posterior $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$ over its underlying latent representation $\mathbf{z}$. 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 fig. 1c.

Table 1. Mean-squared errors of reconstructions.

METHOD	MSE $\mathbf{z}$	MSE $\mathbf{x}$
SJMVI (OURS)	**0.17**	**0.04**
VAE (KINGMA & WELLING, 2014)	0.88	**0.04**
AVB (MESCHEDER ET AL., 2017)	0.29	**0.04**

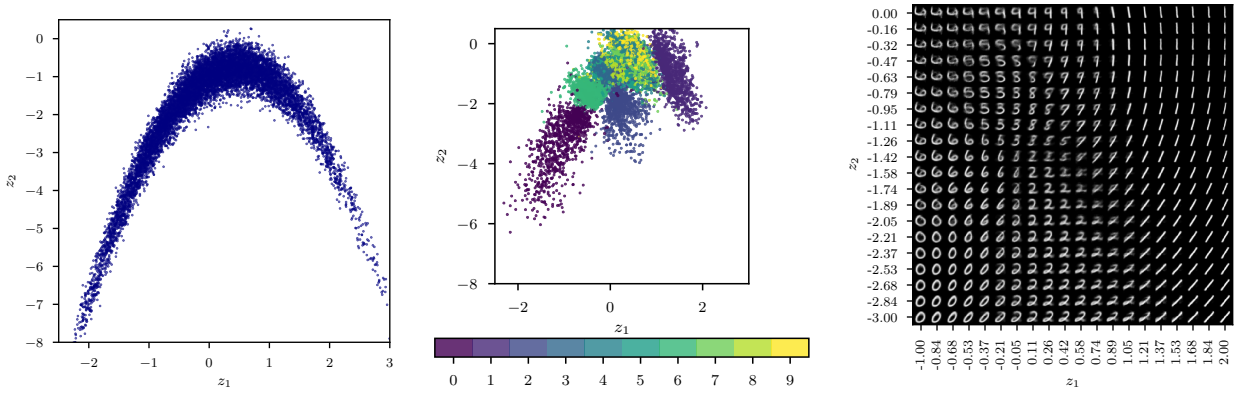
In table 1, we report the mean-squared error (MSE) on the reconstructions of observations from the held-out test set and benchmark against VAE / AVB. Also, for the joint approximation to properly match the support of the exact joint, the latent codes should also be representable by its corresponding 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.

8. Conclusion

We introduced implicit latent variable models, which offer the greatest extent of flexibility in treatment of prior information, and can be used in to model problems ranging from dimensionality reduction to unpaired image-to-image translation. For their parameter estimation and inference, we developed a variational inference framework that augments traditional VI approaches by minimizing the symmetric KL between the exact joint distribution and an approximate distribution.

Additionally, we 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 ILVM. Parameter learning in CYCLEGANs corresponds to approximate inference in this 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 requirement 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.



(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 1. Visualization of 2D latent space and the corresponding observed space manifold.

Acknowledgements

We are grateful to Taeksoo Kim, Alistair Reid, Kelvin Hsu, Hisham Husain and Harrison Nguyen and anonymous reviewers for insightful discussion and feedback. Louis Tiao is partially supported by the CSIRO Data61 Postgraduate Scholarship.

References

Ali, S. M. and Silvey, S. D. A General Class of Coefficients of Divergence of One Distribution from Another, 1966.

Bartholomew, David J, Knott, Martin, and Moustaki, Irini. *Latent variable models and factor analysis: A unified approach*, volume 904. John Wiley & Sons, 2011.

Blei, David M, Kucukelbir, Alp, and McAuliffe, Jon D. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518): 859–877, 2017. doi: 10.1080/01621459.2017.1285773.

Chen, Xi, Duan, Yan, Houthooft, Rein, Schulman, John, Sutskever, Ilya, and Abbeel, Pieter. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. jun 2016.

Ciszar, I. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

Dayan, Peter, Hinton, Geoffrey E, Neal, Radford M, and Zemel, Richard S. The Helmholtz machine. *Neural Computation*, 7(5):889–904, 1995. doi: 10.1162/neco.1995.7.5.889.

Donahue, Jeff, Krähenbühl, Philipp, and Darrell, Trevor. Adversarial Feature Learning. may 2016.

Dumoulin, Vincent, Belghazi, Ishmael, Poole, Ben, Mastropietro, Olivier, Lamb, Alex, Arjovsky, Martin, and Courville, Aaron. Adversarially Learned Inference. jun 2016.

Frey, Brendan J and Hinton, Geoffrey E. Variational learning in nonlinear Gaussian belief networks. *Neural Computation*, 11(1):193–213, 1999.

Gal, Yarin and Ghahramani, Zoubin. Dropout as a Bayesian Approximation: Appendix. jun 2015.

Gershman, Samuel and Goodman, Noah. Amortized inference in probabilistic reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36, 2014.

Gneiting, Tilmann and Raftery, Adrian E. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. ISSN 01621459. doi: 10.2307/27639845.

Goodfellow, Ian J., Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative Adversarial Networks. In Ghahramani, Z, Welling, M, Cortes, C, Lawrence, N D, and Weinberger, K Q (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., jun 2014.

Haario, Heikki, Saksman, Eero, and Tamminen, Johanna. Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics*, 14(3):375, 1999. ISSN 09434062. doi: 10.1007/s001800050022.

Hu, Zhiting, Yang, Zichao, Salakhutdinov, Ruslan, and

Xing, Eric P. On Unifying Deep Generative Models. jun 2017.

Huszár, Ferenc. Variational Inference using Implicit Distributions. feb 2017.

Jaynes, Edwin. Prior Probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4(3):227–241, 1968. ISSN 0536-1567. doi: 10.1109/TSSC.1968.300117.

Jordan, Michael I., Ghahramani, Zoubin, Jaakkola, Tommi S., and Saul, Lawrence K. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183–233, 1999. ISSN 08856125. doi: 10.1023/A:1007665907178.

Kim, Taeksoo, Cha, Moonsu, Kim, Hyunsoo, Lee, Jung Kwon, and Kim, Jiwon. Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pp. 1857–1865, 2017.

Kingma, Diederik P and Welling, Max. Auto-Encoding Variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR) 2014*, Dec 2014.

Lappalainen, Harri and Honkela, Antti. Bayesian nonlinear independent component analysis by multi-layer perceptrons. In *Advances in independent component analysis*, pp. 93–121. Springer, 2000.

Makhzani, Alireza, Shlens, Jonathon, Jaitly, Navdeep, Goodfellow, Ian, and Frey, Brendan. Adversarial Autoencoders. nov 2015.

Mao, Xudong, Li, Qing, Xie, Haoran, Lau, Raymond Y. K., Wang, Zhen, and Smolley, Stephen Paul. Least Squares Generative Adversarial Networks. nov 2016.

Mescheder, Lars, Nowozin, Sebastian, and Geiger, Andreas. Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pp. 2391–2400, Jan 2017.

Mohamed, Shakir and Lakshminarayanan, Balaji. Learning in Implicit Generative Models. In *The 5th International Conference on Learning Representations*, oct 2017.

Nguyen, XuanLong, Wainwright, Martin J, and Jordan, Michael I. Estimating Divergence Functionals and the Likelihood Ratio by Convex Risk Minimization. *IEEE Trans. Information Theory*, 56(11):5847–5861, 2010. doi: 10.1109/TIT.2010.2068870.

Nowozin, Sebastian, Cseke, Botond, and Tomioka, Ryota. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization. In Lee, D D, Sugiyama, M, Luxburg, U V, Guyon, I, and Garnett, R (eds.), *Advances in Neural Information Processing Systems 29*, pp. 271–279. Curran Associates, Inc., jun 2016.

Pu, Yuchen, Wang, Weiyao, Henao, Ricardo, Chen, Liqun, Gan, Zhe, Li, Chunyuan, and Carin, Lawrence. Adversarial Symmetric Variational Autoencoder. In Guyon, I, Luxburg, U V, Bengio, S, Wallach, H, Fergus, R, Vishwanathan, S, and Garnett, R (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4330–4339. Curran Associates, Inc., 2017.

Rezende, Danilo Jimenez, Mohamed, Shakir, and Wierstra, Daan. Stochastic backpropagation and approximate inference in deep generative models. In Xing, Eric P and Jebara, Tony (eds.), *Proceedings of The 31st ...*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1278–1286, Bejing, China, jan 2014. PMLR. ISBN 9781634393973. doi: 10.1051/0004-6361/201527329.

Sønderby, Casper Kaae, Caballero, Jose, Theis, Lucas, Shi, Wenzhe, and Huszár, Ferenc. Amortised map inference for image super-resolution. *arXiv preprint arXiv:1610.04490*, oct 2016.

Srivastava, Akash, Valkoz, Lazar, Russell, Chris, Gutmann, Michael U., Sutton, Charles, Valkov, Lazar, Russell, Chris, Gutmann, Michael U., and Sutton, Charles. Veegan: Reducing mode collapse in gans using implicit variational learning. In *Advances in Neural Information Processing Systems*, pp. 3310–3320, may 2017.

Sugiyama, Masashi, Suzuki, Taiji, Nakajima, Shinichi, Kashima, Hisashi, von Bünau, Paul, and Kawanabe, Motoaki. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, Dec 2008. ISSN 1572-9052. doi: 10.1007/s10463-008-0197-x.

Sugiyama, Masashi, Suzuki, Taiji, and Kanamori, Takafumi. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012. doi: 10.1017/CBO9781139035613.

Tipping, Michael E. and Bishop, Christopher M. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61:611–622, 1999. doi: 10.2307/2680726.

Titsias, Michalis K. Learning Model Reparametrizations: Implicit Variational Inference by Fitting MCMC distributions. aug 2017.

Tran, Dustin, Ranganath, Rajesh, and Blei, David. Hierarchical Implicit Models and Likelihood-Free Variational Inference. In Guyon, I, Luxburg, U V, Bengio, S, Wallach, H, Fergus, R, Vishwanathan, S, and Garnett, R

(eds.), *Advances in Neural Information Processing Systems 30*, pp. 5527–5537. Curran Associates, Inc., 2017.

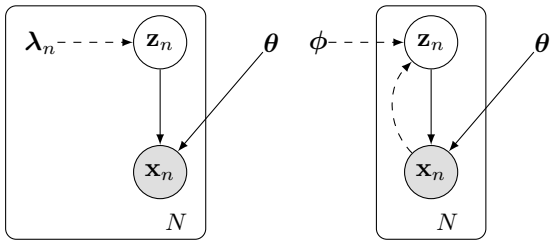
Uehara, Masatoshi, Sato, Issei, Suzuki, Masahiro, Nakayama, Kotaro, and Matsuo, Yutaka. Generative Adversarial Nets from a Density Ratio Estimation Perspective. oct 2016.

Wainwright, Martin J. and Jordan, Michael I. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, nov 2008. ISSN 1935-8237. doi: 10.1561/2200000001.

Zhu, Jun-Yan, Park, Taesung, Isola, Phillip, and Efros, Alexei A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. pp. 2223–2232, mar 2017.

A. Graphical Representation

The graphical representation of implicit latent variable models is depicted below in fig. A.2.



(a) Without amortized 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 A.2. Graphical representation of the *generative model* (**solid**) and the *recognition model* (**dashed**).

B. Recovering Common Latent Variable 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 $\boldsymbol{\xi}$,

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

for functions $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$ that take $\mathbf{z}$ as input. To simplify matters further, assume $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is constant w.r.t. 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 $\mathbf{z}$,

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

and the covariance matrix is diagonal $\boldsymbol{\Psi} = \operatorname{diag}(\psi_1^2, \ldots, \psi_D^2)$, we recover *factor analysis* (FA) (Bartholomew et al., 2011). Furthermore, when the covariance matrix is isotropic $\boldsymbol{\Psi} = \psi^2\mathbf{I}$, we recover PPCA (Tipping & Bishop, 1999). 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 (Lappalainen & Honkela, 2000), nonlinear Gaussian sigmoid belief networks (Frey & Hinton, 1999), and other more sophisticated variants of deep latent variable models. When the mapping is defined by a multilayer perceptron (MLP), we can recover simple instances of a VAE with a Gaussian probabilistic decoder (Kingma & Welling, 2014; Rezende et al., 2014).

C. Proof of Proposition 1

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

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

and,

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

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

$$\begin{aligned} \mathcal{L}_{\text{NELL}}(\boldsymbol{\theta},\boldsymbol{\phi}) &= \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})}[-\log p_{\boldsymbol{\theta}}(\mathbf{x}\,|\,\mathbf{z})] \\ &= \mathbb{E}_{q^*(\mathbf{x})p(\boldsymbol{\epsilon})}[-\log p_{\boldsymbol{\theta}}(\mathbf{x}\,|\,\mathcal{G}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon};\mathbf{x}))] \\ &= \frac{1}{2\tau^2}\mathbb{E}_{q^*(\mathbf{x})p(\boldsymbol{\epsilon})}[\|\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathcal{G}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon};\mathbf{x}))\|_2^2] \\ &\quad + \frac{D}{2}\log 2\pi\tau^2 \\ &= \frac{1}{2\tau^2}\mathbb{E}_{q^*(\mathbf{x})p(\boldsymbol{\epsilon})}[\|\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x}) + t\boldsymbol{\epsilon})\|_2^2] \\ &\quad + \frac{D}{2}\log 2\pi\tau^2 \\ &\to \frac{1}{2\tau^2}\mathbb{E}_{q^*(\mathbf{x})}[\|\mathbf{x} - \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x}))\|_2^2] \\ &\quad + \frac{D}{2}\log 2\pi\tau^2, \qquad \text{as } t \to 0 \\ &= \gamma_1 \ell^{\text{reverse}}_{\text{CONST}}(\boldsymbol{\theta},\boldsymbol{\phi}) + \delta_1 \\ &\propto \ell^{\text{reverse}}_{\text{CONST}}(\boldsymbol{\theta},\boldsymbol{\phi}) \end{aligned}$$

where $\gamma_1 = \frac{1}{2\tau^2}$ and $\delta_1 = \frac{D}{2}\log\frac{\pi}{\gamma_1}$.

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

D. Proof of Lemma 1

Proof. To instantiate $\mathcal{L}^{\text{reverse}}_{\text{GAN}}(\boldsymbol{\alpha}\,|\,\boldsymbol{\phi})$ of eq. 33, 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 $\mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}) \coloneqq 1 - \sigma(\log r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}))$. First we compute the first derivative f'_{GAN} and the convex dual f^*_{GAN} of f_{GAN}, which involve straightforward calculations,

$$\begin{aligned} f'_{\text{GAN}}(u) &= \log\sigma(\log u), \\ f^*_{\text{GAN}}(t) &= -\log(1 - \exp t). \end{aligned}$$

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

$$\begin{aligned} f^*_{\text{GAN}}(f'_{\text{GAN}}(u)) &= -\log(1 - \exp f'_{\text{GAN}}(u)) \\ &= -\log(1 - \sigma(\log u)). \end{aligned}$$

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

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

and

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

respectively, as required. □

E. Proof of Lemma 2

Proof. Through reparameterization 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 $\mathcal{D}_{\boldsymbol{\alpha}}(\mathbf{z}; \mathbf{x}) = \mathbf{D}_{\boldsymbol{\alpha}}(\mathbf{z})$ which ignores auxiliary input $\mathbf{x}$, and mapping $\mathcal{G}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}; \mathbf{x}) = \mathbf{m}_{\boldsymbol{\phi}}(\mathbf{x})$ which ignores noise input $\boldsymbol{\epsilon}$, this reduces to

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

as required. □

F. Proof of Proposition 3

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

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

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

G. Relation to KL Importance Estimation Procedure (KLIEP)

We now discuss the connections to KLIEP (Sugiyama et al., 2008). Consider the same problem setting as in § 3.3 where we wish to use a parameterized function $r_{\boldsymbol{\alpha}}$ to estimate the exact density ratio,

$$r_{\boldsymbol{\alpha}}(\mathbf{z}; \mathbf{x}) \simeq r^*(\mathbf{z}; \mathbf{x}) \coloneqq \frac{q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}{p^*(\mathbf{z})}.$$

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

$$q_{\boldsymbol{\alpha}}(\mathbf{z} \mid \mathbf{x}) \coloneqq r_{\boldsymbol{\alpha}}(\mathbf{z}; \mathbf{x}) p^*(\mathbf{z}) \simeq q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}).$$

Although in our specific problem setting, the density $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$ is tractable, we nonetheless fit an auxiliary model $q_{\boldsymbol{\alpha}}(\mathbf{z} \mid \mathbf{x})$ 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 $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$ and $q_{\boldsymbol{\alpha}}(\mathbf{z} \mid \mathbf{x})$ with respect to $\boldsymbol{\alpha}$,

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

Hence, this is equivalent to *maximizing*

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

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

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

Rewriting this integral, we have the constraint

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

Combined, we have the following constrained optimization 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 optimization problem with objective,

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

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

H. Summary of Definitions

In this section, we summarize the definitions of the losses defined in the proposed VI framework of §§ 3 and 4, and underscore the relationships to their respective counterparts in the CYCLEGAN framework of § 5.

Table 2 summarizes 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 Jensen-Shannon (JS) divergence (up to constants) that GANs are known to minimize.

Table 3 gives the calculations of the terms necessary to explicitly write down instances of the generalized 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 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 generalized *stochastic* formulations of the GAN objectives in the CYCLEGAN framework, while table 5 lists their *deterministic* counterpart.

Lastly, table 6 gives forward and reverse cycle-consistency constraints in the CYCLEGAN framework, and the specific class of Gaussian likelihoods and posteriors that instantiates these constraints (in the limit).

Table 2. 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	$\mathbb{E}_{q^*(\mathbf{x})} \mathcal{D}_f \left[p^*(\mathbf{z}) \parallel q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \right]$	$\mathbb{E}_{q^*(\mathbf{x})} \mathrm{KL} \left[q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \parallel p^*(\mathbf{z}) \right]$	$2 \cdot \mathbb{E}_{q^*(\mathbf{x})} \mathrm{JS} \left[p^*(\mathbf{z}) \parallel q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \right] - \log 4$
OBSERVED	$\mathbb{E}_{p^*(\mathbf{z})} \mathcal{D}_f \left[q^*(\mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}) \right]$	$\mathbb{E}_{p^*(\mathbf{z})} \mathrm{KL} \left[p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}) \parallel q^*(\mathbf{x}) \right]$	$2 \cdot \mathbb{E}_{p^*(\mathbf{z})} \mathrm{JS} \left[q^*(\mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z}) \right] - \log 4$

Table 3. 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. 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	$\mathcal{L}_f^{\text{latent}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi}) \coloneqq \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}[f'(r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}))] - \mathbb{E}_{q^*(\mathbf{x})p^*(\mathbf{z})}[f^\star(f'(r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})))]$	$\mathcal{L}_{\text{KL}}^{\text{latent}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi}) \coloneqq \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}[\log r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})] - \mathbb{E}_{q^*(\mathbf{x})p^*(\mathbf{z})}[r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}) - 1]$	$\mathcal{L}_{\text{GAN}}^{\text{reverse}}(\boldsymbol{\alpha} \mid \boldsymbol{\phi}) \coloneqq \mathbb{E}_{q^*(\mathbf{x})q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})}[\log \sigma(\log r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x}))] + \mathbb{E}_{q^*(\mathbf{x})p^*(\mathbf{z})}[\log(1 - \sigma(\log r_{\boldsymbol{\alpha}}(\mathbf{z};\mathbf{x})))]$
OBSERVED	$\mathcal{L}_f^{\text{observed}}(\boldsymbol{\beta} \mid \boldsymbol{\theta}) \coloneqq \mathbb{E}_{p^*(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})}[f'(r_{\boldsymbol{\beta}}(\mathbf{x};\mathbf{z}))] - \mathbb{E}_{p^*(\mathbf{z})q^*(\mathbf{x})}[f^\star(f'(r_{\boldsymbol{\beta}}(\mathbf{x};\mathbf{z})))]$	$\mathcal{L}_{\text{KL}}^{\text{observed}}(\boldsymbol{\beta} \mid \boldsymbol{\theta}) \coloneqq \mathbb{E}_{p^*(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})}[\log r_{\boldsymbol{\beta}}(\mathbf{x};\mathbf{z})] - \mathbb{E}_{p^*(\mathbf{z})q^*(\mathbf{x})}[r_{\boldsymbol{\beta}}(\mathbf{x};\mathbf{z}) - 1]$	$\mathcal{L}_{\text{GAN}}^{\text{forward}}(\boldsymbol{\beta} \mid \boldsymbol{\theta}) \coloneqq \mathbb{E}_{p^*(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})}[\log \sigma(\log r_{\boldsymbol{\beta}}(\mathbf{x};\mathbf{z}))] + \mathbb{E}_{p^*(\mathbf{z})q^*(\mathbf{x})}[\log(1 - \sigma(\log r_{\boldsymbol{\beta}}(\mathbf{x};\mathbf{z})))]$

Table 5. General stochastic GAN objectives and their deterministic counterparts.

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

Table 6. Negative expected log conditionals and the cycle-consistency constraints.

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