The Information Autoencoding Family: A Lagrangian Perspective on Latent Variable Generative Models

Shengjia Zhao
Computer Science Department
Stanford University
sjzhao@stanford.edu

Jiaming Song
Computer Science Department
Stanford University
tsong@stanford.edu

Stefano Ermon
Computer Science Department
Stanford University
ermon@stanford.edu

Abstract

A large number of objectives have been proposed to train latent variable generative models. We show that many of them are Lagrangian dual functions of the same primal optimization problem. The primal problem optimizes the mutual information between latent and visible variables, subject to the constraints of accurately modeling the data distribution and performing correct amortized inference. By choosing to maximize or minimize mutual information, and choosing different Lagrange multipliers, we obtain different objectives including InfoGAN, ALI/BiGAN, ALICE, CycleGAN, beta-VAE, adversarial autoencoders, AVB, AS-VAE and InfoVAE. Based on this observation, we provide an exhaustive characterization of the statistical and computational trade-offs made by all the training objectives in this class of Lagrangian duals. Next, we propose a dual optimization method where we optimize model parameters as well as the Lagrange multipliers. This method achieves Pareto optimal solutions in terms of optimizing information and satisfying the constraints.

1 INTRODUCTION

Latent variable generative models are designed to accomplish a wide variety of tasks in computer vision (Radford et al., 2015; Kuleshov & Ermon, 2017), natural language processing (Yang et al., 2017), reinforcement learning (Li et al., 2017b), compressed sensing Dhar et al. (2018),etc. Prominent examples include Variational Autoencoders (VAE, Kingma & Welling (2013); Rezende et al. (2014)), with extensions such as β-VAE (Higgins et al., 2016), Adversarial Autoencoders (Makhzani et al., 2015), and InfoVAE (Zhao et al., 2017); Generative Adversarial Networks (Goodfellow et al., 2014), with extensions such as ALI/BiGAN (Dumoulin et al., 2016a; Donahue et al., 2016), InfoGAN (Chen et al., 2016a) and ALICE (Li et al., 2017a); hybrid objectives such as CycleGAN (Zhu et al., 2017), DiscoGAN (Kim et al., 2017), AVB (Mescheder et al., 2017) and AS-VAE (Pu et al., 2017). All these models attempt to fit an empirical data distribution, but differ in multiple ways: how they measure the similarity between distributions; whether or not they allow for efficient (amortized) inference; whether the latent variables should retain or discard information about the data; and how the model is optimized, which can be likelihood-based or likelihood-free (Mohamed & Lakshminarayanan, 2016; Grover et al., 2018).

In this paper, we generalize existing training objectives for latent variable generative models. We show that all the above training objectives can be viewed as *Lagrangian dual functions* of a constrained optimization problem (primal problem). The primal problem optimizes over the parameters of a generative model and an (amortized) inference distribution. The optimization objective is to maximize or minimize mutual information between latent and observed variables; the constraints (which we term "*consistency constraints*") are to accurately model the data distribution and to perform correct amortized inference. By considering the Lagrangian dual function and different settings of the Lagrange multipliers, we can obtain all the aforementioned generative modeling training objectives. Surprisingly, under mild assumptions, the aforementioned objectives can be linearly combined to produce every possible primal objective/multipliers in this model family.

In Lagrangian dual optimization, the dual function is maximized with respect to the Lagrange multipliers, and minimized with respect to the primal parameters. Under strong duality, the optimal parameters found by this procedure also solve the original primal problem. However, the aforementioned objectives use fixed (rather than maximized) multipliers. As a consequence, strong duality does not generally hold.

To overcome this problem, we propose a new learning approach where the Lagrange multipliers are also optimized. We show that strong duality holds in distribution space,

so this optimization procedure is guaranteed to optimize the primal objective while satisfying the consistency constraints. As an application of this approach, we propose *Lagrangian VAE*, a Lagrangian optimization algorithm for the InfoVAE (Zhao et al., 2017) objective. Lagrangian VAE can explicitly trade-off optimization of the primal objective and consistency constraint satisfaction. In addition, both theoretical properties (of Lagrangian optimization) and empirical experiments show that solutions obtained by Lagrangian VAE *Pareto dominate* solutions obtained with InfoVAE: Lagrangian VAE either obtains better mutual information or better constraint satisfaction, regardless of the hyper-parameters used by either method.

2 BACKGROUND

We consider two groups of variables: observed variables $\boldsymbol{x} \in \mathcal{X}$ and latent variables $\boldsymbol{z} \in \mathcal{Z}$. Our algorithm receives input distributions $q(\boldsymbol{x}), p(\boldsymbol{z})$ over $\boldsymbol{x}$ and $\boldsymbol{z}$ respectively. Each distribution is either specified *explicitly* through a tractable analytical expression such as $\mathcal{N}(0, I)$, or *implicitly* through a set of samples. For example, in latent variable generative modeling of images (Kingma & Welling, 2013; Goodfellow et al., 2014), $\mathcal{X}$ is the space of images, and $\mathcal{Z}$ is the space of latent features. $q(\boldsymbol{x})$ is a dataset of sample images, and $p(\boldsymbol{z})$ is a simple "prior" distribution, e.g., a Gaussian; in unsupervised image translation (Zhu et al., 2017), $\mathcal{X}$ and $\mathcal{Z}$ are both image spaces and $q(\boldsymbol{x})$, $p(\boldsymbol{z})$ are sample images from two different domains (e.g., pictures of horses and zebras).

The underlying joint distribution on $(\boldsymbol{x}, \boldsymbol{z})$ is not known, and we are not given any sample from it. Our goal is to nonetheless learn some model of the joint distribution $r_{\text{model}}(\boldsymbol{x}, \boldsymbol{z})$ with the following desiderata:

Desideratum 1. Matching Marginal The marginals of $r_{\text{model}}(\boldsymbol{x}, \boldsymbol{z})$ over $\boldsymbol{x}, \boldsymbol{z}$ respectively match the provided distributions $q(\boldsymbol{x}), p(\boldsymbol{z})$.

Desideratum 2. Meaningful Relationship $r_{\text{model}}(\boldsymbol{x}, \boldsymbol{z})$ captures a meaningful relationship between $\boldsymbol{x}$ and $\boldsymbol{z}$. For example, in latent variable modeling of images, the latent variables $\boldsymbol{z}$ should correspond to semantically meaningful features describing the image $\boldsymbol{x}$. In unsupervised image translation, $r_{\text{model}}(\boldsymbol{x}, \boldsymbol{z})$ should capture the "correct" pairing between $\boldsymbol{x}$ and $\boldsymbol{z}$.

We address desideratum 1 in this section, and desideratum 2 in section 3. The joint distribution $r_{\text{model}}(\boldsymbol{x}, \boldsymbol{z})$ can be represented in factorized form by chain rule. To do so, we define conditional distribution families $\{p_{\theta^p}(\boldsymbol{x}|\boldsymbol{z}), \theta^p \in \Theta^p\}$ and $\{q_{\theta^q}(\boldsymbol{z}|\boldsymbol{x}), \theta^q \in \Theta^q\}$. We require that for any $\boldsymbol{z}$ we can both efficiently sample from $p_{\theta^p}(\boldsymbol{x}|\boldsymbol{z})$ and compute $\log p_{\theta^p}(\boldsymbol{x}|\boldsymbol{z})$, and similarly for $q_{\theta^q}(\boldsymbol{z}|\boldsymbol{x})$. For compactness we use $\theta = (\theta^p, \theta^q)$ to denote the parameters of both distributions p_θ and q_θ. We define the joint distribution $r_{\text{model}}(\boldsymbol{x}, \boldsymbol{z})$ in two ways:

$$r_{\text{model}}(\boldsymbol{x}, \boldsymbol{z}) \stackrel{\text{def}}{=} p_\theta(\boldsymbol{x}, \boldsymbol{z}) \stackrel{\text{def}}{=} p(\boldsymbol{z}) p_\theta(\boldsymbol{x}|\boldsymbol{z}) \tag{1}$$

and symmetrically

$$r_{\text{model}}(\boldsymbol{x}, \boldsymbol{z}) \stackrel{\text{def}}{=} q_\theta(\boldsymbol{x}, \boldsymbol{z}) \stackrel{\text{def}}{=} q(\boldsymbol{x}) q_\theta(\boldsymbol{z}|\boldsymbol{x}) \tag{2}$$

Defining the model in two (redundant) ways seem unusual but has significant computational advantages: given $\boldsymbol{x}$ we can tractably sample $\boldsymbol{z}$, and vice versa. For example, in latent variable models, given observed data $\boldsymbol{x}$ we can sample latent features from $\boldsymbol{z} \sim q_\theta(\boldsymbol{z}|\boldsymbol{x})$ (amortized inference), and given latent feature $\boldsymbol{z}$ we can generate novel samples from $\boldsymbol{x} \sim p_\theta(\boldsymbol{x}|\boldsymbol{z})$ (ancestral sampling).

If the two definitions (1), (2) are **consistent**, which we define as $p_\theta(\boldsymbol{x}, \boldsymbol{z}) = q_\theta(\boldsymbol{x}, \boldsymbol{z})$, we automatically satisfy desideratum 1:

$$r_{\text{model}}(\boldsymbol{x}) = \int_{\boldsymbol{z}} r_{\text{model}}(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z} = \int_{\boldsymbol{z}} q_\theta(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z} = q(\boldsymbol{x})$$
$$r_{\text{model}}(\boldsymbol{z}) = \int_{\boldsymbol{x}} r_{\text{model}}(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{x} = \int_{\boldsymbol{x}} p_\theta(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{x} = p(\boldsymbol{z})$$

Based on this observation, we can design objectives that encourage consistency. Many latent variable generative models fit into this framework. For example, variational autoencoders (VAE, Kingma & Welling (2013)) enforce consistency by minimizing the KL divergence:

$$\min_\theta D_{\text{KL}}(q_\theta(\boldsymbol{x}, \boldsymbol{z}) \| p_\theta(\boldsymbol{x}, \boldsymbol{z}))$$

This minimization is equivalent to maximizing the evidence lower bound ($\mathcal{L}_{\text{ELBO}}$) (Kingma & Welling, 2013):

$$\max_\theta -D_{\text{KL}}(q_\theta(\boldsymbol{x}, \boldsymbol{z}) \| p_\theta(\boldsymbol{x}, \boldsymbol{z})) \tag{3}$$
$$= -\mathbb{E}_{q_\theta(\boldsymbol{x}, \boldsymbol{z})} \left[\log(q_\theta(\boldsymbol{z}|\boldsymbol{x}) q(\boldsymbol{x})) - \log(p_\theta(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}))\right]$$
$$= \mathbb{E}_{q_\theta(\boldsymbol{x}, \boldsymbol{z})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] + H_q(\boldsymbol{x}) - \mathbb{E}_{q(\boldsymbol{x})} \left[D_{\text{KL}}(q_\theta(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z}))\right]$$
$$\equiv \left.\begin{array}{l} \mathbb{E}_{q_\theta(\boldsymbol{x}, \boldsymbol{z})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] \\ -\mathbb{E}_{q(\boldsymbol{x})} \left[D_{\text{KL}}(q_\theta(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z}))\right] \end{array}\right\} \mathcal{L}_{\text{ELBO}} \tag{4}$$

where $H_q(\boldsymbol{x})$ is the entropy of $q(\boldsymbol{x})$ and is a constant that can be ignored for the purposes of optimization over model parameters θ (denoted $\equiv$).

As another example, BiGAN/ALI (Donahue et al., 2016; Dumoulin et al., 2016b) use an adversarial discriminator to approximately minimize the Jensen-Shannon divergence

$$\min_\theta D_{\text{JS}}(q_\theta(\boldsymbol{x}, \boldsymbol{z}) \| p_\theta(\boldsymbol{x}, \boldsymbol{z}))$$

Many other ways of enforcing consistency are possible. Most generally, we can enforce consistency with a vector of divergences $\mathcal{D} = [D_1, \ldots, D_m]$, where each D_i takes two

probability measures as input, and outputs a non-negative value which is zero if and only if the two input measures are the same. Examples of possible divergences include Maximum Mean Discrepancy (MMD, Gretton et al. (2007)), denoted D_{MMD}; Wasserstein distance (Arjovsky et al., 2017), denoted D_{W}; f-divergences (Nowozin et al., 2016), denoted D_f; and Jensen-Shannon divergence (Goodfellow et al., 2014), denoted D_{JS}.

Each D_i can be any divergence applied to a pair of probability measures. The pair of probability measures can be defined over either both variables $(\boldsymbol{x}, \boldsymbol{z})$, a single variable $\boldsymbol{x}$, $\boldsymbol{z}$, or conditional $\boldsymbol{x}|\boldsymbol{z}$, $\boldsymbol{z}|\boldsymbol{x}$. If the probability measure is defined over a conditional $\boldsymbol{x}|\boldsymbol{z}$, $\boldsymbol{z}|\boldsymbol{x}$, we also take expectation over the conditioning variable with respect to p_θ or q_θ. Some examples of D_i are:

$$\begin{gathered}\mathbb{E}_{q_\theta(z)}[D_{\mathrm{KL}}(q_\theta(x|z)\|p_\theta(x|z))] \\ D_{\mathrm{MMD}}(q_\theta(\boldsymbol{z})\|p(\boldsymbol{z})) \\ D_{\mathrm{W}}(p_\theta(\boldsymbol{x},\boldsymbol{z})\|q_\theta(\boldsymbol{x},\boldsymbol{z})) \\ \mathbb{E}_{q(x)}[D_f(q_\theta(\boldsymbol{z}|\boldsymbol{x})\|p_\theta(\boldsymbol{z}|\boldsymbol{x}))] \\ D_{\mathrm{JS}}(q(\boldsymbol{x})\|p_\theta(\boldsymbol{x}))\end{gathered}$$

We only require that

$$D_i = 0, \forall i \in \{1,\dots,m\} \iff p_\theta(\boldsymbol{x},\boldsymbol{z}) = q_\theta(\boldsymbol{x},\boldsymbol{z})$$

so $\mathcal{D} = \mathbf{0}$ implies consistency. Note that each D_i implicitly depends on the parameters θ through p_θ and q_θ, but notationally we neglect this for simplicity.

Enforcing consistency $p_\theta(\boldsymbol{x},\boldsymbol{z}) = q_\theta(\boldsymbol{x},\boldsymbol{z})$ by $\mathcal{D} = \mathbf{0}$ satisfies desideratum 1 (matching marginal), but does not directly address desideratum 2 (meaningful relationship). A large number of joint distributions can have the same marginal distributions $p(\boldsymbol{z})$ and $q(\boldsymbol{x})$ (including ones where $\boldsymbol{z}$ and $\boldsymbol{x}$ are independent), and only a small fraction of them encode meaningful models.

3 GENERATIVE MODELING AS CONSTRAINT OPTIMIZATION

To address desideratum 2, we modify the training objective and specify additional preferences among consistent $p_\theta(\boldsymbol{x},\boldsymbol{z})$ and $q_\theta(\boldsymbol{x},\boldsymbol{z})$. Formally we solve the following primal optimization problem

$$\min_\theta f(\theta) \quad \text{subject to } \mathcal{D} = \mathbf{0} \tag{5}$$

where $f(\theta)$ encodes our preferences over consistent distributions, and depends on θ through $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ and $q_\theta(\boldsymbol{x}|\boldsymbol{z})$.

An important preference is the mutual information between $\boldsymbol{x}$ and $\boldsymbol{z}$. Depending on the downstream application, we may maximize mutual information (Chen et al., 2016b; Zhao et al., 2017; Li et al., 2017a; Chen et al., 2016a) so that the features (latent variables) $\boldsymbol{z}$ can capture as much information as possible about $\boldsymbol{x}$, or minimize mutual information (Zhao et al., 2017; Higgins et al., 2016; Tishby & Zaslavsky, 2015; Shamir et al., 2010) to achieve compression. To implement mutual information preference we consider the following objective

$$f_I(\theta;\alpha_1,\alpha_2) = \alpha_1 I_{q_\theta}(\boldsymbol{x};\boldsymbol{z}) + \alpha_2 I_{p_\theta}(\boldsymbol{x};\boldsymbol{z}) \tag{6}$$

where $I_{p_\theta}(\boldsymbol{x};\boldsymbol{z}) = \mathbb{E}_{p_\theta(\boldsymbol{x},\boldsymbol{z})}[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log p_\theta(\boldsymbol{x})p(\boldsymbol{z})]$ is the mutual information under $p_\theta(\boldsymbol{x},\boldsymbol{z})$, and $I_{q_\theta}(\boldsymbol{x};\boldsymbol{z})$ is their mutual information under $q_\theta(\boldsymbol{x},\boldsymbol{z})$.

The optimization problem in Eq.(5) with mutual information $f(\theta)$ in Eq.(6) has the following Lagrangian dual function:

$$\alpha_1 I_{q_\theta}(\boldsymbol{x};\boldsymbol{z}) + \alpha_2 I_{p_\theta}(\boldsymbol{x};\boldsymbol{z}) + \boldsymbol{\lambda}^\top \mathcal{D} \tag{7}$$

where $\boldsymbol{\lambda} = [\lambda_1,\dots,\lambda_m]$ is a vector of Lagrange multipliers, one for each of the m consistency constraints in $\mathcal{D} = [D_1,\dots,D_m]$.

In the next section, we will show that many existing training objectives for generative models minimize the Lagrangian dual in Equation 7 for some *fixed* α_1, α_2, $\mathcal{D}$ and $\boldsymbol{\lambda}$. However, dual optimization requires maximization over the dual parameters $\boldsymbol{\lambda}$, which should *not* be kept fixed. We discuss dual optimization in Section 5.

4 GENERALIZING OBJECTIVES WITH FIXED MULTIPLIERS

Several existing objectives for latent variable generative models can be rewritten in the dual form of Equation 7 with *fixed* Lagrange multipliers. We provide several examples here and provide more in Appendix A.

VAE (Kingma & Welling, 2013) Per our discussion in Section 2, the VAE training objective commonly written as ELBO maximization in Eq.(4) is actually equivalent to Equation 3. This is a dual form where we set $\mathcal{D} = [D_{\mathrm{KL}}(q_\theta(\boldsymbol{x},\boldsymbol{z})\|p_\theta(\boldsymbol{x},\boldsymbol{z})]$, $\alpha_1 = \alpha_2 = 0$ and $\boldsymbol{\lambda} = 1$. Because $\alpha_1 = \alpha_2 = 0$, this objective has no information preference, confirming previous observations that the learned distribution can have high, low or zero mutual information between x and z. Chen et al. (2016b); Zhao et al. (2017).

β-VAE (Higgins et al., 2016) The following objective $\mathcal{L}_{\beta-\mathrm{VAE}}$ is proposed to learn disentangled features $\boldsymbol{z}$:

$$-\mathbb{E}_{q_\theta(\boldsymbol{x},\boldsymbol{z})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] + \beta\mathbb{E}_{q(\boldsymbol{x})}\left[D_{\mathrm{KL}}(q_\theta(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z}))\right]$$

This is equivalent to the following dual form:

$$\begin{aligned}&\mathcal{L}_{\beta-\mathrm{VAE}} \\ &\equiv \mathbb{E}_{q_\theta(\boldsymbol{x},\boldsymbol{z})}\left[\log\frac{q_\theta(\boldsymbol{x}|\boldsymbol{z})q(\boldsymbol{x})}{p_\theta(\boldsymbol{x}|\boldsymbol{z})q_\theta(\boldsymbol{x}|\boldsymbol{z})} + \beta\log\frac{q_\theta(\boldsymbol{z}|\boldsymbol{x})q_\theta(\boldsymbol{z})}{q_\theta(\boldsymbol{z})p(\boldsymbol{z})}\right] \\ &\equiv (\beta-1)I_{q_\theta}(\boldsymbol{x};\boldsymbol{z}) && \text{(primal)} \\ &\quad +\beta D_{\mathrm{KL}}(q_\theta(\boldsymbol{z})\|p(\boldsymbol{z}))) && \text{(consistency)} \\ &\quad +\mathbb{E}_{q_\theta(\boldsymbol{z})}[D_{\mathrm{KL}}(q_\theta(\boldsymbol{x}|\boldsymbol{z})\|p_\theta(\boldsymbol{x}|\boldsymbol{z}))]\end{aligned}$$

$f(p, q)$	Likelihood Based	Unary Likelihood Free	Binary Likelihood Free
0	VAE (Kingma & Welling, 2013)	VAE-GAN (Makhzani et al., 2015)	ALI (Dumoulin et al., 2016b)
$\alpha_1 I_q$	β-VAE (Higgins et al., 2016)	InfoVAE (Zhao et al., 2017)	ALICE (Li et al., 2017a)
$\alpha_2 I_p$	VMI (Barber & Agakov, 2003)	InfoGAN (Chen et al., 2016a)	-
$\alpha_1 I_q + \alpha_2 I_p$	-	CycleGAN (Zhu et al., 2017)	AS-VAE (Pu et al., 2017)

Table 1: For each choice of α and computability class (Definition 2) we list the corresponding existing model. Several other objectives are also Lagrangian duals, but they are not listed because they are similar to models in the table. These objectives include DiscoGAN (Kim et al., 2017), BiGAN (Donahue et al., 2016), AAE (Makhzani et al., 2015), WAE (Tolstikhin et al., 2017).

where we use $\equiv$ to denote "equal up to a value that does not depend on θ". In this case,

$$\alpha_1 = \beta - 1, \qquad \alpha_2 = 0$$
$$\boldsymbol{\lambda} = [\beta, 1]$$
$$\mathcal{D} = [KL(q_\theta(\boldsymbol{z})\|p(\boldsymbol{z}))), \mathbb{E}_{q_\theta(\boldsymbol{z})}[D_{\mathrm{KL}}(q_\theta(\boldsymbol{x}|\boldsymbol{z})\|p_\theta(\boldsymbol{x}|\boldsymbol{z}))]$$

When $\alpha_1 > 0$ or equivalently $\beta > 1$, there is an incentive to minimize mutual information between $\boldsymbol{x}$ and $\boldsymbol{z}$.

InfoGAN (Chen et al., 2016a) As another example, the InfoGAN objective [1] :

$$\mathcal{L}_{\mathrm{InfoGAN}} = D_{\mathrm{JS}}(q(\boldsymbol{x})\|p_\theta(\boldsymbol{x})) - \mathbb{E}_{p_\theta(\boldsymbol{x},\boldsymbol{z})}[\log q_\theta(\boldsymbol{z}|\boldsymbol{x})]$$

is equivalent to the following dual form:

$$\begin{aligned}
&\mathcal{L}_{\mathrm{InfoGAN}} \equiv \mathbb{E}_{p_\theta(\boldsymbol{x},\boldsymbol{z})}[-\log p_\theta(\boldsymbol{z}|\boldsymbol{x}) + \log p(\boldsymbol{z}) \\
&+ \log p_\theta(\boldsymbol{z}|x) - \log q_\theta(\boldsymbol{z}|\boldsymbol{x})] + D_{\mathrm{JS}}(q(\boldsymbol{x})\|p_\theta(\boldsymbol{x})) \\
&\equiv -I_{p_\theta}(\boldsymbol{x}; \boldsymbol{z}) && \text{(primal)} \\
&\quad +\mathbb{E}_{p_\theta(\boldsymbol{x})}[D_{\mathrm{KL}}(p_\theta(\boldsymbol{z}|\boldsymbol{x})\|q_\theta(\boldsymbol{z}|\boldsymbol{x}))] && \text{(consistency)} \\
&\quad +D_{\mathrm{JS}}(q(\boldsymbol{x})\|p_\theta(\boldsymbol{x}))
\end{aligned}$$

In this case $\alpha_1 = 0$, and $\alpha_2 = -1 < 0$, the model maximizes mutual information between $\boldsymbol{x}$ and $\boldsymbol{z}$.

In fact, all objectives in Table 1 belong to this class[2]. Derivations for additional models can be found in Appendix A.

4.1 ENUMERATION OF ALL OBJECTIVES

The Lagrangian dual form of an objective reveals its mutual information preference (α_1, α_2), type of consistency constraints ($\mathcal{D}$), and weighting of the constraints ($\boldsymbol{\lambda}$). This suggests that the Lagrangian dual perspective may unify many existing training objectives. We wish to identify and categorize *all* objectives that have Lagrangian dual form as in Eq.7). However, this has two technical difficulties that we proceed to resolve.

[1]For conciseness we use $\boldsymbol{z}$ to denote structured latent variables, which is represented as $\boldsymbol{c}$ in (Chen et al., 2016a).

[2]Variational Mutual Information Maximization (VMI) is not truly a Lagrangian dual because it does not enforce consistency constraints ($\boldsymbol{\lambda} = 0$).

1. Equivalence: Many objectives appear different, but are actually identical for the purposes of optimization (as we have shown). To handle this we characterize "equivalent objectives" with a set of pre-specified transformations.

Definition 1. ***Equivalence (Informal):*** *An objective $\mathcal{L}$ is equivalent to $\mathcal{L}'$ when there exists a constant C, so that for all parameters θ, $\mathcal{L}(\theta) = \mathcal{L}'(\theta) + C$. We denote this as $\mathcal{L} \equiv \mathcal{L}'$.*

$\mathcal{L}$ and $\mathcal{L}'$ are elementary equivalent if $\mathcal{L}'$ can be obtained from $\mathcal{L}$ by applying chain rule or Bayes rule to probabilities in $\mathcal{L}$, and addition/subtraction of constants $\mathbb{E}_{q(\boldsymbol{x})}[\log q(\boldsymbol{x})]$ and $\mathbb{E}_{p(\boldsymbol{z})}[\log p(\boldsymbol{z})]$.

A more formal but verbose definition is in Appendix B, Definition 1.

Elementary equivalences define simple yet flexible transformations for deriving equivalent objectives. For example, all the transformations in Section 4 (VAE, β-VAE and InfoGAN) and Appendix A are elementary. This implies that all objectives in Table 1 are elementary equivalent to a Lagrangian dual function in Eq.(7) . However, these transformations are not exhaustive. For example, tranforming $\mathbb{E}_{p_\theta}[g(\boldsymbol{x})]$ into $\mathbb{E}_{q_\theta}[g(\boldsymbol{x})p_\theta(\boldsymbol{x})/q_\theta(\boldsymbol{x})]$ via importance sampling is not accounted for, hence the two objectives are not considered to be elementary equivalent.

2. Optimization Difficulty: Some objectives are easier to evaluate/optimize than others. For example, variational autoencoder training is robust and stable, adversarial training is less stable and requires careful hyper-parameter selection (Kodali et al., 2018), and direct optimization of the log-likelihood $\log p_\theta(\boldsymbol{x})$ is very difficult for latent variable models and almost never used Grover et al. (2018).

To assign a "hardness score" to each objective, we first group the "terms" (an objective is a sum of terms) from easy to hard to optimize. An objective belongs to a "hardness class" if it cannot be transformed into an objective with easier terms. This is formalized below:

Definition 2. ***Effective Optimization:*** *We define*

1. Likelihood-based terms as the following set

$$T_1 = \{\mathbb{E}_{p_\theta(\boldsymbol{x},\boldsymbol{z})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})], \mathbb{E}_{p_\theta(\boldsymbol{x},\boldsymbol{z})}[\log p_\theta(\boldsymbol{x},\boldsymbol{z})], \\ \mathbb{E}_{p_\theta(\boldsymbol{z})}[\log p(\boldsymbol{z})], \mathbb{E}_{p_\theta(\boldsymbol{x},\boldsymbol{z})}[\log q_\theta(\boldsymbol{z}|\boldsymbol{x})] \\ \mathbb{E}_{q_\theta(\boldsymbol{x},\boldsymbol{z})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})], \mathbb{E}_{q_\theta(\boldsymbol{x},\boldsymbol{z})}[\log p_\theta(\boldsymbol{x},\boldsymbol{z})], \\ \mathbb{E}_{q_\theta(\boldsymbol{z})}[\log p(\boldsymbol{z})], \mathbb{E}_{q_\theta(\boldsymbol{x},\boldsymbol{z})}[\log q_\theta(\boldsymbol{z}|\boldsymbol{x})]\}$$

2. Unary likelihood-free terms as the following set

$$T_2 = \{D(q(\boldsymbol{x})\|p_\theta(\boldsymbol{x})), D(q_\theta(\boldsymbol{z})\|p(\boldsymbol{z}))\}$$

3. Binary likelihood-free terms as the following set

$$T_3 = \{D(q_\theta(\boldsymbol{x},\boldsymbol{z})\|p_\theta(\boldsymbol{x},\boldsymbol{z}))\}$$

where each D can be f-divergence, Jensen Shannon divergence, Wasserstein distance, or Maximum Mean Discrepancy. An objective $\mathcal{L}$ is likelihood-based computable if $\mathcal{L}$ is elementary equivalent to some $\mathcal{L}'$ that is a linear combination of elements in T_1; unary likelihood-free computable if $\mathcal{L}'$ is a linear combination of elements in $T_1 \cup T_2$; binary likelihood-free computable if $\mathcal{L}'$ is a linear combination of elements in $T_1 \cup T_2 \cup T_3$.

The rationale of this categorization is that elements in T_1 can be estimated by Monte-Carlo estimators and optimized by stochastic gradient descent effectively in practice (with low bias and variance) (Kingma & Welling, 2013; Rezende et al., 2014). In contrast, elements in T_2 are optimized by likelihood-free approaches such as adversarial training (Goodfellow et al., 2014) or kernelized methods such as MMD (Gretton et al., 2007) or Stein variational gradient (Liu & Wang, 2016). These optimization procedures are known to suffer from stability problems (Arjovsky et al., 2017) or cannot handle complex distributions in high dimensions (Ramdas et al., 2015). Finally, elements in T_3 are over both $\boldsymbol{x}$ and $\boldsymbol{z}$, and they are empirically shown to be even more difficult to optimize (Li et al., 2017a). We do not include terms such as $\mathbb{E}_{q(x)}[\log p_\theta(x)]$ because they are seldom feasible to compute or optimize for latent variable generative models.

Now we are able to fully characterize all Lagrangian dual objectives in Eq.(7) that are likelihood-based / unary likelihood free / binary likelihood free computable in Table 1.

In addition, Table 1 contains essentially *all* possible models for each optimization difficulty class in Definition 2. This is shown in the following theorem (informal, formal version and proof in Appendix B, Theorem 3,4,5)

Theorem 1. ***Closure theorem (Informal)****: Denote a Lagrangian objectives in the form of Equation 7 where all divergences are D_{KL} a KL Lagrangian objective. Under elementary equivalence defined in Definition 1,*

1) *Any KL Lagrangian objective that is elementary equivalent to a likelihood based computable objective is equivalent to a linear combination of VMI and β-VAE.*

2) *Any KL Lagrangian objective that is elementary equivalent to a unary likelihood computable objective is equivalent to a linear combination of InfoVAE and InfoGAN.*

3) *Any KL Lagrangian objective that is elementary equivalent to a binary likelihood computable objective is equivalent to a linear combination of ALICE, InfoVAE and InfoGAN.*

We also argue in the Appendix (without formal proof) that this theorem holds for other divergences including D_{MMD}, D_{W}, D_f or D_{JS}.

Intuitively, this suggests a rather negative result: if a new latent variable model training objective contains mutual information preference and consistency constraints (defined through D_{KL}, D_{MMD}, D_{W}, D_f or D_{JS}), and this objective can be effectively optimized as in Definition 1 and Definition 2, then this objective is a linear combination of existing objectives. Our limitation is that we are restricted to elementary transformations and the set of terms defined in Definition 2. To derive new training objectives, we should consider new transformations, non-linear combinations and/or new terms.

5 DUAL OPTIMIZATION FOR LATENT VARIABLE GENRATIVE MODELS

While existing objectives for latent variable generative models have dual form in Equation 7, they are not solving the dual problem exactly because the Lagrange multipliers $\boldsymbol{\lambda}$ are predetermined instead of optimized. In particular, if we can show strong duality, the optimal solution to the dual is also an optimal solution to the primal (Boyd & Vandenberghe, 2004). However if the Lagrange multipliers are fixed, this property is lost, and the parameters θ obtained via dual optimization may be suboptimal for $\min_\theta f(\theta)$, or violate the consistency conditions $\mathcal{D} = \mathbf{0}$.

5.1 RELAXATION OF CONSISTENCY CONSTRAINTS

This observation motivates us to directly solve the dual optimization problem where we also *optimize* the Lagrange multipliers.

$$\max_{\boldsymbol{\lambda}\geq 0} \min_\theta f(\theta) + \boldsymbol{\lambda}^T \mathcal{D}$$

Unfortunately, this is usually impractical because the consistency constrains are difficult to satisfy when the model has finite capacity, so in practice the primal optimization problem is actually infeasible and $\boldsymbol{\lambda}$ will be optimized to $+\infty$.

One approach to this problem is to use relaxed consistency constraints, where compared to Eq.(5) we require consis-

tency up to some error $\epsilon > 0$:

$$\min_{\theta} f(\theta) \quad \text{subject to} \quad \mathcal{D} \leq \boldsymbol{\epsilon} \tag{8}$$

For a sufficiently large $\boldsymbol{\epsilon}$, the problem is feasible. This has the corresponding dual problem:

$$\max_{\boldsymbol{\lambda} \geq 0} \min_{\theta} f(\theta) + \boldsymbol{\lambda}^{\top}(\mathcal{D} - \boldsymbol{\epsilon}) \tag{9}$$

When $\boldsymbol{\lambda}$ is constant (instead of maximized), Equation 9 still reduces to existing latent variable generative modeling objectives since $\boldsymbol{\lambda}^{\top}\boldsymbol{\epsilon}$ is a constant, so the objective simply becomes

$$\min_{\theta} f(\theta) + \boldsymbol{\lambda}^{T}\mathcal{D} + \text{constant}$$

In contrast, we propose to find $\boldsymbol{\lambda}^*, \theta^*$ that optimize the Lagrangian dual in Eq.(9). If we additionally have strong duality, θ^* is also the optimal solution to the primal problem in Eq.(8).

5.2 STRONG DUALITY WITH MUTUAL INFORMATION OBJECTIVES

This section aims to show that strong duality for Eq.(8) holds in distribution space if we replace mutual informations in f with upper and lower bounds. We prove this via Slater's condition (Boyd & Vandenberghe, 2004), which has three requirements: 1. $\forall D \in \mathcal{D}$, D is convex in θ; 2. $f(\theta)$ is convex for $\theta \in \Theta$; 3. the problem is strictly feasible: $\exists \theta$ s.t. $\mathcal{D} < \boldsymbol{\epsilon}$. We propose weak conditions to satisfy all three in distribution space, so strong duality is guaranteed.

For simplicity we focus on discrete $\mathcal{X}$ and $\mathcal{Z}$. We parameterize $q_\theta(\boldsymbol{z}|\boldsymbol{x})$ with a parameter matrix $\theta^q \in \mathbb{R}^{|\mathcal{X}||\mathcal{Z}|}$ (we add the superscript q to distinguish parameters of q_θ from that of p_θ) where

$$q_\theta(\boldsymbol{z} = j|\boldsymbol{x} = i) = \theta^q_{ij}, \forall i \in \mathcal{X}, j \in \mathcal{Z} \tag{10}$$

The only restriction is that θ^q must correspond to valid conditional distributions. More formally, we require that $\theta^q \in \Theta^q$, where

$$\Theta^q = \left\{ \theta^q \in \mathbb{R}^{|\mathcal{X}||\mathcal{Z}|} \; s.t. \; 0 \leq \theta^q_{ij} \leq 1, \sum_j \theta^q_{ij} = 1 \right\} \tag{11}$$

Similarly we can define $\theta^p \in \Theta^p$ for p_θ. We still use

$$\theta = [\theta^q, \theta^p], \quad \Theta = \Theta^q \times \Theta^p \tag{12}$$

to denote both sets of parameters.

1) Constraints $D \in \mathcal{D}$ are convex: We show that some divergences used in existing models are convex in distribution space.

Lemma 1 (Convex Constraints (Informal)). *D_{KL}, D_{MMD}, or D_f over any marginal distributions on x or z or joint distributions on (x, z) are convex with respect to $\theta \in \Theta$ as defined in Eq.(12).*

Therefore if one only uses these convex divergences, the first requirement for Slater's condition is satisfied.

2) Convex Bounds for $f(\theta)$: $f(\theta) = \alpha_1 I_{q_\theta}(x; z) + \alpha_2 I_{p_\theta}(x; z)$ is not itself guaranteed to be convex in general. However we observe that mutual information has a convex upper bound, and a concave lower bound, which we denote as $\overline{I}_{q_\theta}$ and $\underline{I}_{q_\theta}$ respectively:

$$\begin{aligned} & I_{q_\theta}(\boldsymbol{x}; \boldsymbol{z}) \\ &= \mathbb{E}_{q(\boldsymbol{x})}[D_{\mathrm{KL}}(q_\theta(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z}))] && \text{convex upper bound } \overline{I}_{q_\theta} \\ &\quad - D_{\mathrm{KL}}(q_\theta(\boldsymbol{z})\|p(\boldsymbol{z})) && \text{bound gap } \overline{I}_{q_\theta} - I_{q_\theta} \\ &= \mathbb{E}_{q_\theta(\boldsymbol{x},\boldsymbol{z})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] + H_q(\boldsymbol{x}) && \text{concave lower bound } \underline{I}_{q_\theta} \\ &\quad + \mathbb{E}_{p(\boldsymbol{z})} D_{\mathrm{KL}}(q(\boldsymbol{x}|\boldsymbol{z})\|p_\theta(\boldsymbol{x}|\boldsymbol{z})) && \text{bound gap } I_{q_\theta} - \underline{I}_{q_\theta} \end{aligned} \tag{13}$$

The convexity/concavity of these bounds is shown by the following lemma, which we prove in the appendix

Lemma 2 (Convex/Concave Bounds). *$\overline{I}_{q_\theta}$ is convex with respect to $\theta \in \Theta$ as defined in Eq.(12), and $\underline{I}_{q_\theta}$ is concave with respect to $\theta \in \Theta$.*

A desirable property of these bounds is that if we look at the bound gaps (difference between bound and true value) in Eq.(13), they are 0 if the consistency constraint is satisfied (i.e., $p_\theta(\boldsymbol{x}, \boldsymbol{z}) = q_\theta(\boldsymbol{x}, \boldsymbol{z})$). They will be tight (bound gaps are small) when consistency constraints are approximately satisfied (i.e., $p_\theta(\boldsymbol{x}, \boldsymbol{z}) \approx q_\theta(\boldsymbol{x}, \boldsymbol{z})$). In addition we also denote identical bounds for I_{p_θ} as $\overline{I}_{p_\theta}$ and $\underline{I}_{p_\theta}$ Similar bounds for mutual information have been discussed in (Alemi et al., 2017).

3) Strict Feasibility: the optimization problem has non empty feasible set, which we show in the following lemma:

Lemma 3 (Strict Feasibility). *For discrete $\mathcal{X}$ and $\mathcal{Z}$, and $\boldsymbol{\epsilon} > 0$, $\exists \theta \in \Theta$ such that $\mathcal{D} < \boldsymbol{\epsilon}$.*

Therefore we have shown that for convex/concave upper and lower bounds on f, all three of Slater's conditions are satisfied, so strong duality holds. We summarize this in the following theorem.

Theorem 2 (Strong Duality). *If $\mathcal{D}$ contains only divergences in Lemma 1, then for all $\epsilon > 0$:*

If $\alpha_1, \alpha_2 \geq 0$ strong duality holds for the following problems:

$$\min_{\theta \in \Theta} \alpha_1 \overline{I}_{q_\theta} + \alpha_2 \overline{I}_{p_\theta} \quad \textit{subject to} \quad \mathcal{D} \leq \epsilon \tag{14}$$

If $\alpha_1, \alpha_2 \leq 0$, strong duality holds for the following problem

$$\min_{\theta \in \Theta} \alpha_1 \underline{I}_{q_\theta} + \alpha_2 \underline{I}_{p_\theta} \quad \textit{subject to} \quad \mathcal{D} \leq \epsilon \tag{15}$$

Algorithm 1 Dual Optimization for Latent Variable Generative Models

Input: Analytical form for $p(\boldsymbol{z})$ and samples from $q(\boldsymbol{x})$; constraints $\mathcal{D}$; α_1, α_2 that specify maximization / minimization of mutual information; $\boldsymbol{\epsilon} > 0$ which specifies the strength of constraints; step size ρ_θ, ρ_λ for θ and $\boldsymbol{\lambda}$.
Output: θ (parameters for $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ and $q_\theta(\boldsymbol{z}|\boldsymbol{x})$).

Initialize θ randomly
Initialize the Lagrange multipliers $\boldsymbol{\lambda} := 1$
if $\alpha_1, \alpha_2 > 0$ **then**
  $f(\theta) \leftarrow \alpha_1 \overline{I}_{q_\theta} + \alpha_2 \overline{I}_{p_\theta}$
else
  $f(\theta) \leftarrow \alpha_1 \underline{I}_{q_\theta} + \alpha_2 \underline{I}_{p_\theta}$
end if
for $t = 0, 1, 2, \ldots$ **do**
  $\theta \leftarrow \theta - \rho_\theta(\nabla_\theta f(\theta) + \boldsymbol{\lambda}^\top \nabla_\theta \mathcal{D})$
  $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} + \rho_\lambda(\mathcal{D} - \boldsymbol{\epsilon})$
end for

5.3 DUAL OPTIMIZATION

Because the problem is convex in distribution space and satisfies Slater's condition, the θ^* that achieves the saddle point

$$\boldsymbol{\lambda}^\star, \theta^\star = \arg\max_{\boldsymbol{\lambda} \geq 0} \arg\min_\theta f(\theta) + \boldsymbol{\lambda}^T(\mathcal{D} - \boldsymbol{\epsilon}) \quad (16)$$

is also a solution to the original optimization problem Eq.(8) (Boyd & Vandenberghe, 2004)(Chapter 5.4). In addition the max-min problem Eq.(16) is convex with respect to θ and concave (linear) with respect to $\boldsymbol{\lambda}$, so one can apply iterative gradient descent/ascent over θ (minimize) and $\boldsymbol{\lambda}$ (maximize) and achieve stable convergence to saddle point (Holding & Lestas, 2014). We describe the iterative algorithm in Algorithm 1.

In practice, we do not optimize over distribution space and $\{p_\theta(\boldsymbol{x}|\boldsymbol{z})\}, \{q_\theta(\boldsymbol{z}|\boldsymbol{x})\}$ are some highly complex and non-convex families of functions. We show in the experimental section that this scheme is stable and effective despite non-convexity.

6 LAGRANGIAN VAE

In this section we consider a particular instantiation of the general dual problem proposed in the previous section. Consider the following primal problem, with $\alpha_1 \in \mathbb{R}$:

$$\min_\theta \alpha_1 I_{q_\theta}(\boldsymbol{x}; \boldsymbol{z}) \quad (17)$$
$$\text{subject to } D_{\mathrm{KL}}(q_\theta(\boldsymbol{x}, \boldsymbol{z}) \| p_\theta(\boldsymbol{x}, \boldsymbol{z}))) \leq \epsilon_1$$
$$D_{\mathrm{MMD}}(q_\theta(\boldsymbol{z}) \| p(\boldsymbol{z})) \leq \epsilon_2$$

For mutual information minimization / maximization, we respectively replace the (possibly non-convex) mutual information by upper bound $\overline{I}_{q_\theta}$ if $\alpha_1 \geq 0$ and lower bound $\underline{I}_{q_\theta}$ if $\alpha_1 < 0$. The corresponding dual optimization problem can be written as:

$$\max_{\boldsymbol{\lambda} \geq 0} \min_\theta \begin{cases} \alpha_1 \overline{I}_{q_\theta} + \boldsymbol{\lambda}^\top(\mathcal{D}_{\mathrm{InfoVAE}} - \boldsymbol{\epsilon}), & \alpha_1 \geq 0 \\ \alpha_1 \underline{I}_{q_\theta} + \boldsymbol{\lambda}^\top(\mathcal{D}_{\mathrm{InfoVAE}} - \boldsymbol{\epsilon}), & \alpha_1 < 0 \end{cases} \quad (18)$$

where $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2]$, $\boldsymbol{\lambda} = [\lambda_1, \lambda_2]$ and

$$\mathcal{D}_{\mathrm{InfoVAE}} = [D_{\mathrm{KL}}(q_\theta(\boldsymbol{x}, \boldsymbol{z}) \| p_\theta(\boldsymbol{x}, \boldsymbol{z}))), D_{\mathrm{MMD}}(q_\theta(\boldsymbol{z}) \| p(\boldsymbol{z}))]$$

We call the objective in 18 Lagrangian (Info)VAE (LagVAE). Note that setting a constant $\boldsymbol{\lambda}$ for the dual function recovers the InfoVAE objective (Zhao et al., 2017). By Theorem 2 strong duality holds for this problem and finding the max-min saddle point of LagVAE in Eq.(18) is identical to finding the optimal solution to original problem of Eq.(17).

The final issue is choosing the $\boldsymbol{\epsilon}$ hyper-parameters so that the constraints are feasible. This is non-trivial since selecting $\boldsymbol{\epsilon}$ that describe feasible constraints depends on the task and model structure. We introduce a general strategy that is effective in all of our experiments. First we learn a parameter θ^* that satisfies the consistency constraints "as well as possible" without considering mutual information maximization/minimization. Formally this is achieved by the following optimization (for any choice of $\boldsymbol{\lambda} > 0$),

$$\theta^* = \arg\min_\theta \quad \boldsymbol{\lambda}^T \mathcal{D}_{\mathrm{InfoVAE}} \quad (19)$$

This is the original training objective for InfoVAE with $\alpha_1 = 0$ and can be optimized by

$$\begin{aligned} &\min_\theta \boldsymbol{\lambda}^T \mathcal{D}_{\mathrm{InfoVAE}} \\ &= \lambda_1 D_{\mathrm{KL}}(q_\theta(\boldsymbol{x}, \boldsymbol{z}) \| p_\theta(\boldsymbol{x}, \boldsymbol{z}))) + \lambda_2 D_{\mathrm{MMD}}(q_\theta(\boldsymbol{z}) \| p(\boldsymbol{z})) \\ &\equiv \lambda_1 \mathcal{L}_{\mathrm{ELBO}}(\theta) + \lambda_2 D_{\mathrm{MMD}}(q_\theta(\boldsymbol{z}) \| p(\boldsymbol{z})) \end{aligned} \quad (20)$$

where $\mathcal{L}_{\mathrm{ELBO}}(\theta)$ is the evidence lower bound defined in Eq.(4). Because we only need a rough estimate of how well consistency constraints can be satisfied, the selection of weighing λ_1 and λ_2 is unimportant. The recommendation in (Zhao et al., 2017) works well in all our experiments ($\lambda_1 = 1, \lambda_2 = 100$).

Now we introduce a "slack" to specify how much we are willing to tolerate consistency error to achieve higher/lower mutual information. Formally, we define $\hat{\boldsymbol{\epsilon}}$ as the divergences $\mathcal{D}_{\mathrm{InfoVAE}}$ evaluated at the above θ^*. Under this $\hat{\boldsymbol{\epsilon}}$ the following constraint must be feasible (because θ^* is a solution):

$$\mathcal{D}_{\mathrm{InfoVAE}} \leq \hat{\boldsymbol{\epsilon}}$$

Now we can safely set $\boldsymbol{\epsilon} = \boldsymbol{\gamma} + \hat{\boldsymbol{\epsilon}}$, where $\boldsymbol{\gamma} > 0$, and the constraint

$$\mathcal{D}_{\mathrm{InfoVAE}} \leq \boldsymbol{\epsilon}$$

must still be feasible (and strictly feasible). $\boldsymbol{\gamma}$ has a very nice intuitive interpretation: it is the "slack" that we are willing to accept. Compared to tuning α_1 and $\boldsymbol{\lambda}$ for InfoVAE, tuning $\boldsymbol{\gamma}$ is much more interpretable: we can anticipate the final consistency error before training.

Another practical consideration is that the one of the constraints $D_{\mathrm{KL}}(q_\theta(\boldsymbol{x},\boldsymbol{z})\|p_\theta(\boldsymbol{x},\boldsymbol{z}))$ is difficult to estimate. However, we have

$$D_{\mathrm{KL}}(q_\theta(\boldsymbol{x},\boldsymbol{z})\|p_\theta(\boldsymbol{x},\boldsymbol{z})) = -\mathcal{L}_{\mathrm{ELBO}} - H_q(\boldsymbol{x})$$

where $\mathcal{L}_{\mathrm{ELBO}}$ is again, the evidence lower bound in Eq.(4) of Section 2, and $H_q(\boldsymbol{x})$ is the entropy of the true distribution $q(x)$. $\mathcal{L}_{\mathrm{ELBO}}$ is empirically easy to estimate, and $H_q(\boldsymbol{x})$ is a constant irrelevant to the optimization problem. The optimization problem is identical if we replacing the more difficult constraint $D_{\mathrm{KL}}(q_\theta(\boldsymbol{x},\boldsymbol{z})\|p_\theta(\boldsymbol{x},\boldsymbol{z})) \le \epsilon_1$ with the easier-to-optimize/estimate constraint $-\mathcal{L}_{\mathrm{ELBO}} \le \epsilon_1'$ (where $\epsilon_1' = \epsilon_1 + H_q(\boldsymbol{x})$). In addition, ϵ_1' can be selected by the technique in the previous paragraph.

7 EXPERIMENTS

We compare the performance of **LagVAE**, where we learn $\boldsymbol{\lambda}$ automatically, and **InfoVAE**, where we set $\boldsymbol{\lambda}$ in advance (as hyperparameters). Our primal problem is to find solutions that maximize / minimize mutual information under the consistency constraints. Therefore, we consider two performance metrics:

- $I_q(\boldsymbol{x},\boldsymbol{z})$ the mutual information between $\boldsymbol{x}$ and $\boldsymbol{z}$. We can estimate the mutual information via the identity:

  $$I_q(\boldsymbol{x};\boldsymbol{z}) = \mathbb{E}_{q_\theta(\boldsymbol{x},\boldsymbol{z})}\left[\log q_\theta(\boldsymbol{z}|\boldsymbol{x}) - \log q_\theta(\boldsymbol{z})\right] \quad (21)$$

  where we approximate $q_\theta(\boldsymbol{z})$ with a kernel density estimator.

- the consistency divergences $D_{\mathrm{KL}}(q_\theta(\boldsymbol{x},\boldsymbol{z})\|p_\theta(\boldsymbol{x},\boldsymbol{z}))$ and $D_{\mathrm{MMD}}(q_\theta(\boldsymbol{z})\|p(\boldsymbol{z}))$. As stated in Section 6, we replace $D_{\mathrm{KL}}(q_\theta(\boldsymbol{x},\boldsymbol{z})\|p_\theta(\boldsymbol{x},\boldsymbol{z}))$ with the evidence lower bound $\mathcal{L}_{\mathrm{ELBO}}$.

In the remainder of this section we demonstrate the following empirical observations:

- LagVAE reliably maximizes/minimizes mutual information without violating the consistency constraints. InfoVAE, on the other hand, makes unpredictable and task-specific trade-offs between mutual information and consistency.
- LagVAE is Pareto optimal, as no InfoVAE hyperparameter choice is able to achieve both better mutual information and better consistency (measured by D_{MMD} and $\mathcal{L}_{\mathrm{ELBO}}$) than LagVAE.

7.1 VERIFICATION OF DUAL OPTIMIZATION

We first verify that LagVAE reliably maximizes/minimizes mutual information subject to consistency constraints. We train LagVAE on MNIST according to Algorithm 1. $\boldsymbol{\epsilon}$ is selected according to Section 6, where we first compute $\hat{\boldsymbol{\epsilon}} = (\hat{\epsilon}_1, \hat{\epsilon}_2)$ without information maximization/minimization by Eq.(20). Next we choose slack variables $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$, and set $\boldsymbol{\epsilon} = \hat{\boldsymbol{\epsilon}} + \boldsymbol{\gamma}$. For γ_1 we explore values from 0.1 to 4.0, and for γ_2 we use the fixed value $0.5\hat{\epsilon}_2$.

The results are shown in Figure 1, where mutual information is estimated according to Eq.(21). For any given slack $\boldsymbol{\gamma}$, setting α_1 to positive values and negative values respectively minimizes or maximizes the mutual information within the feasible set $\mathcal{D} \le \boldsymbol{\epsilon}$. In particular, the absolute value of α_1 does not affect the outcome, and only the sign matters. This is consistent with the expected behavior (Figure 1 Left) where the model finds the maximum/minimum mutual information solution within the feasible set.

7.2 VERIFICATION OF PARETO IMPROVEMENTS

In this section we verify Pareto optimality of LagVAE. We evaluate LagVAE and InfoVAE on the MNIST dataset with a wide variety of hyper-parameters. For LagVAE, we set ϵ_1 for $\mathcal{L}_{\mathrm{ELBO}}$ to be $\{83, 84, \ldots, 95\}$ and ϵ_2 for D_{MMD} to be 0.0005. For InfoVAE, we set $\alpha \in \{1, -1\}$, $\lambda_1 \in \{1, 2, 5, 10\}$ and $\lambda_2 \in \{1000, 2000, 5000, 10000\}$ [3].

Figure 2 plots the mutual information and $\mathcal{L}_{\mathrm{ELBO}}$ achieved by both methods. Each point is the outcome of one hyperparameter choice of LagVAE / InfoVAE. Regardless of the hyper-parameter choice of both models, no InfoVAE hyperparameter lead to better performance on both mutual information and $\mathcal{L}_{\mathrm{ELBO}}$ on the training set. This is expected because LagVAE always finds the maximum/minimum mutual information solution out of all solutions with given consistency value. The same trend is true even on the test set, indicating that it is not an outcome of over-fitting.

8 CONCLUSION

Many existing objectives for latent variable generative modeling are Lagrangian dual functions of the same type of constrained optimization problem with fixed Lagrangian multipliers. This allows us to explore their statistical and computational trade-offs, and characterize all models in this class. Moreover, we propose a practical dual optimization method that optimizes both the Lagrange multipliers and the model parameters, allowing us to specify interpretable constraints and achieve Pareto-optimality empirically.

[3] Code for this set of experiments is available at `https://github.com/ermongroup/lagvae`

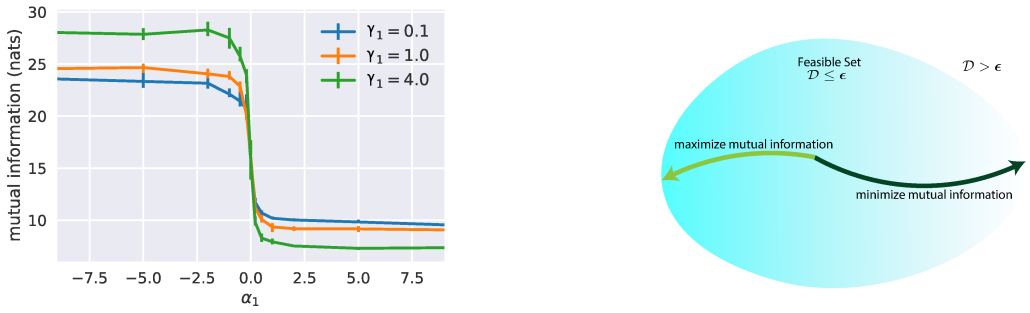

Figure 1: **Left:** Effect of α_1 and γ_1 on the primal objective (mutual information). When α_1 is positive we minimize mutual information within the feasible set, and when α_1 is negative we maximize mutual information. When α_1 is zero the preference is undetermined, and mutual information varies depending on initialization. Note that mutual information does not depend on the absolute value of α_1 but only on its sign. **Right:** An illustration of this effect. Lagrangian dual optimization finds the maximum/minimum mutual information solution in the feasible set $\mathcal{D} \leq \boldsymbol{\epsilon}$.

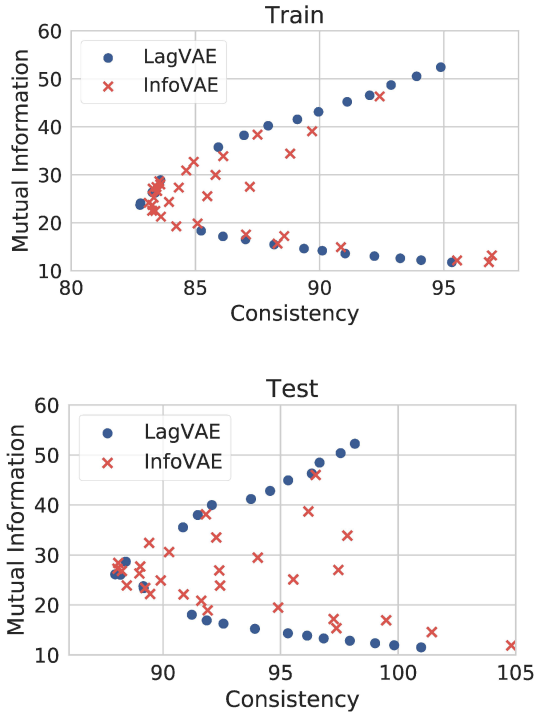

Figure 2: LagVAE Pareto dominates InfoVAE with respect to Mutual information and consistency ($\mathcal{L}_{\text{ELBO}}$ values) on train (top) and test (bottom) set. Each point is the outcome of one hyper-parameter choice for LagVAE / InfoVAE. When we maximize mutual information ($\alpha_1 < 0$), for any given $\mathcal{L}_{\text{ELBO}}$ value, LagVAE always achieve similar or larger mutual information; when we minimize mutual information ($\alpha_1 > 0$), for any given ELBO value, LagVAE always achieve similar or smaller mutual information.

In this work, we only considered Lagrangian (Info)VAE, but the method is generally applicable to other Lagrangian dual objectives. In addition we only considered mutual information preference. Exploring different preferences is a promising future directions.

Acknolwledgements

This research was supported by Intel Corporation, TRI, NSF (#1651565, #1522054, #1733686) and FLI (#2017-158687).

References

Alemi, Alexander A., Poole, Ben, Fischer, Ian, Dillon, Joshua V., Saurous, Rif A., and Murphy, Kevin. An information-theoretic analysis of deep latent-variable models. *CoRR*, abs/1711.00464, 2017. URL `http://arxiv.org/abs/1711.00464`.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. *ArXiv e-prints*, January 2017.

Barber, David and Agakov, Felix. The im algorithm: a variational approach to information maximization. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, pp. 201–208. MIT Press, 2003.

Boyd, Stephen and Vandenberghe, Lieven. *Convex optimization*. Cambridge university press, 2004.

Chen, Xi, Duan, Yan, Houthooft, Rein, Schulman, John, Sutskever, Ilya, and Abbeel, Pieter. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv preprint arXiv:1606.03657*, 2016a.

Chen, Xi, Kingma, Diederik P, Salimans, Tim, Duan, Yan, Dhariwal, Prafulla, Schulman, John, Sutskever, Ilya, and Abbeel, Pieter. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016b.

Dhar, Manik, Grover, Aditya, and Ermon, Stefano. Sparsegen: Modeling sparse deviations for compressed sensing using generative models. *International Conference on Machine Learning*, 2018.

Donahue, Jeff, Krähenbühl, Philipp, and Darrell, Trevor. Adversarial feature learning. *CoRR*, abs/1605.09782, 2016. URL http://arxiv.org/abs/1605.09782.

Dumoulin, Vincent, Belghazi, Ishmael, Poole, Ben, Lamb, Alex, Arjovsky, Martin, Mastropietro, Olivier, and Courville, Aaron. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016a.

Dumoulin, Vincent, Belghazi, Ishmael, Poole, Ben, Lamb, Alex, Arjovsky, Martin, Mastropietro, Olivier, and Courville, Aaron. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016b.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Gretton, Arthur, Borgwardt, Karsten M, Rasch, Malte, Schölkopf, Bernhard, and Smola, Alex J. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pp. 513–520, 2007.

Grover, Aditya, Dhar, Manik, and Ermon, Stefano. Flow-GAN: Combining maximum likelihood and adversarial learning in generative models. In *AAAI Conference on Artificial Intelligence*, 2018.

Higgins, Irina, Matthey, Loic, Pal, Arka, Burgess, Christopher, Glorot, Xavier, Botvinick, Matthew, Mohamed, Shakir, and Lerchner, Alexander. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

Holding, Thomas and Lestas, Ioannis. On the convergence to saddle points of concave-convex functions, the gradient method and emergence of oscillations. In *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, pp. 1143–1148. IEEE, 2014.

Kim, Taeksoo, Cha, Moonsu, Kim, Hyunsoo, Lee, Jung Kwon, and Kim, Jiwon. Learning to discover cross-domain relations with generative adversarial networks. *CoRR*, abs/1703.05192, 2017. URL http://arxiv.org/abs/1703.05192.

Kingma, D. P and Welling, M. Auto-Encoding Variational Bayes. *ArXiv e-prints*, December 2013.

Kodali, Naveen, Hays, James, Abernethy, Jacob, and Kira, Zsolt. On convergence and stability of gans. 2018.

Kuleshov, Volodymyr and Ermon, Stefano. Deep hybrid models: Bridging discriminative and generative approaches. In *Proceedings of the Conference on Uncertainty in AI (UAI)*, 2017.

Li, Chunyuan, Liu, Hao, Chen, Changyou, Pu, Yunchen, Chen, Liqun, Henao, Ricardo, and Carin, Lawrence. Towards understanding adversarial learning for joint distribution matching. 2017a.

Li, Yunzhu, Song, Jiaming, and Ermon, Stefano. Inferring the latent structure of human decision-making from raw visual inputs. 2017b.

Liu, Qiang and Wang, Dilin. Stein variational gradient descent: A general purpose bayesian inference algorithm. *arXiv preprint arXiv:1608.04471*, 2016.

Makhzani, Alireza, Shlens, Jonathon, Jaitly, Navdeep, and Goodfellow, Ian. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Mescheder, Lars, Nowozin, Sebastian, and Geiger, Andreas. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. *arXiv preprint arXiv:1701.04722*, 2017.

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

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

Pu, Yuchen, Wang, Weiyao, Henao, Ricardo, Chen, Liqun, Gan, Zhe, Li, Chunyuan, and Carin, Lawrence. Adversarial symmetric variational autoencoder. In *Advances in Neural Information Processing Systems*, pp. 4333–4342, 2017.

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

Ramdas, Aaditya, Reddi, Sashank Jakkam, Póczos, Barnabás, Singh, Aarti, and Wasserman, Larry A. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *AAAI*, pp. 3571–3577, 2015.

Rezende, D., Mohamed, S., and Wierstra, D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *ArXiv e-prints*, January 2014.

Shamir, Ohad, Sabato, Sivan, and Tishby, Naftali. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29-30):2696–2711, 2010.

Tishby, Naftali and Zaslavsky, Noga. Deep learning and the information bottleneck principle. *CoRR*, abs/1503.02406, 2015. URL http://arxiv.org/abs/1503.02406.

Tolstikhin, Ilya, Bousquet, Olivier, Gelly, Sylvain, and Schoelkopf, Bernhard. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

Yang, Zichao, Hu, Zhiting, Salakhutdinov, Ruslan, and Berg-Kirkpatrick, Taylor. Improved variational autoencoders for text modeling using dilated convolutions. *CoRR*, abs/1702.08139, 2017. URL http://arxiv.org/abs/1702.08139.

Zhao, Shengjia, Song, Jiaming, and Ermon, Stefano. Infovae: Information maximizing variational autoencoders. *CoRR*, abs/1706.02262, 2017. URL http://arxiv.org/abs/1706.02262.

Zhu, Jun-Yan, Park, Taesung, Isola, Phillip, and Efros, Alexei A. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017. URL http://arxiv.org/abs/1703.10593.