The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)

InfoVAE: Balancing Learning and Inference in Variational Autoencoders

Shengjia Zhao, Jiaming Song, Stefano Ermon
Stanford University
sjzhao,tsong,ermon@stanford.edu

Abstract

A key advance in learning generative models is the use of amortized inference distributions that are jointly trained with the models. We find that existing training objectives for variational autoencoders can lead to inaccurate amortized inference distributions and, in some cases, improving the objective provably degrades the inference quality. In addition, it has been observed that variational autoencoders tend to ignore the latent variables when combined with a decoding distribution that is too flexible. We again identify the cause in existing training criteria and propose a new class of objectives (InfoVAE) that mitigate these problems. We show that our model can significantly improve the quality of the variational posterior and can make effective use of the latent features regardless of the flexibility of the decoding distribution. Through extensive qualitative and quantitative analyses, we demonstrate that our models outperform competing approaches on multiple performance metrics.

Introduction

Generative models have shown great promise in modeling complex distributions such as natural images and text (Radford, Metz, and Chintala 2015; Zhu et al. 2017; Yang et al. 2017; Li, Song, and Ermon 2017). These are directed graphical models which represent the joint distribution between the data and a set of hidden variables (features) capturing latent factors of variation. The joint is factored as the product of a prior over the latent variables and a conditional distribution of the visible variables given the latent ones. Usually a simple prior distribution is provided for the latent variables, while the distribution of the input conditioned on latent variables is complex and modeled with a deep network.

However, variational autoencoders have several problems. First, the approximate inference distribution is often significantly different from the true posterior. Previous methods have resorted to using more flexible variational families to better approximate the true posterior distribution (Kingma, Salimans, and Welling 2016). However we find that the problem is rooted in the ELBO training objective itself. In fact, we show that the ELBO objective favors fitting the data distribution over performing correct amortized inference. When the two goals are conflicting (e.g., because of limited capacity), the ELBO objective tends to sacrifice correct inference to better fit (or potentially overfit) the training data.

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Another problem that has been observed is that when the conditional distribution is sufficiently expressive, the latent variables are often ignored (Chen et al. 2016). That is, the model only uses a single conditional distribution component to model the data, effectively ignoring the latent variables and fails to take advantage of the mixture modeling capability of the VAE. In addition, one goal of unsupervised learning is to learn meaningful latent representations, but this obviously fails if the latent variables are ignored. Some solutions have been proposed in (Chen et al. 2016) by limiting the capacity of the conditional distribution, but this requires manual and problem-specific design of the features we would like to extract.

In this paper we propose a novel solution by framing both problems as explicit modeling choices: we introduce new training objectives where it is possible to weight the preference between correct inference and fitting the data distribution, and specify a preference on how much the model should rely on the latent variables. This choice is only implicitly made in the traditional ELBO objective. We make this choice explicit and generalize the ELBO objective by adding additional terms that allow users to select their preferences. Despite of the addition of seemingly intractable terms, we find an equivalent form that can still be efficiently optimized.

Our new family also generalizes known models including the β-VAE (Higgins et al. 2017) and Adversarial Autoencoders (Makhzani et al. 2015). In addition to deriving these models as special cases, we provide generic principles for hyper-parameter selection that work well in all the experimental settings we considered. Finally we perform extensive experiments to evaluate our newly introduced model family, and compare with existing models on multiple metrics of performance such as log-likelihood, sampling quality, and semi-supervised performance. An instantiation of our general framework called MMD-VAE achieves better or on-par performance on all metrics we considered. We further observe that our model can lead to better amortized inference, and utilize the latent variables even in the presence of a very flexible decoder.

Variational Autoencoders

A latent variable generative model defines a joint distribution between (latent) features $z \in \mathcal{Z}$, and inputs $x \in \mathcal{X}$. Usually we assume a simple prior distribution $p(z)$ over the features, such as Gaussian or uniform, and model the data distribution with a complex conditional distribution $p_\theta(x|z)$, where $p_\theta(x|z)$ is often parameterized with a neural network. Suppose the true underlying distribution is $p_{\mathcal{D}}(x)$ (that is approximated by a training set), then a natural training objective is maximum (marginal) likelihood

$$\mathbb{E}_{p_{\mathcal{D}}(x)}[\log p_\theta(x)] = \mathbb{E}_{p_{\mathcal{D}}(x)}[\log \mathbb{E}_{p(z)}[p_\theta(x|z)]] \quad (1)$$

However direct optimization of the likelihood is intractable because computing $p_\theta(x) = \int_z p_\theta(x|z)p(z)\mathrm{d}z$ requires integration. A classic approach (Kingma and Welling 2013) is to define an amortized inference distribution $q_\phi(z|x)$ and jointly optimize a lower bound to the log-likelihood

$$\begin{aligned}\mathcal{L}_{\mathrm{ELBO}}(x) &= -D_{\mathrm{KL}}(q_\phi(z|x)||p(z)) + \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] \\ &\leq \log p_\theta(x)\end{aligned}$$

We further average this over the data distribution $p_{\mathcal{D}}(x)$ to obtain the final optimization objective

$$\mathcal{L}_{\mathrm{ELBO}} = \mathbb{E}_{p_{\mathcal{D}}(x)}[\mathcal{L}_{\mathrm{ELBO}}(x)] \leq \mathbb{E}_{p_{\mathcal{D}}(x)}[\log p_\theta(x)]$$

Equivalent Forms of the ELBO Objective

There are several ways to equivalently rewrite the ELBO objective that will become useful in our following analysis. We define the joint generative distribution as

$$p_\theta(x, z) \equiv p(z)p_\theta(x|z)$$

In fact we can correspondingly define a joint "inference distribution"

$$q_\phi(x, z) \equiv p_{\mathcal{D}}(x)q_\phi(z|x)$$

Note that the two definitions are symmetrical. In the former case we start from a known distribution $p(z)$ and learn the conditional distribution on $\mathcal{X}$, in the latter we start from a known (empirical) distribution $p_{\mathcal{D}}(x)$ and learn the conditional distribution on $\mathcal{Z}$ given x. We also correspondingly define any conditional and marginal distributions as follows:

$$\begin{aligned}p_\theta(x) &= \int_z p_\theta(x, z)dz && \text{Marginal of } p_\theta(x, z) \text{ on } x \\ p_\theta(z|x) &\propto p_\theta(x, z) && \text{Posterior of } p_\theta(x|z) \\ q_\phi(z) &= \int_x q_\phi(x, z)dx && \text{Marginal of } q_\phi(x, z) \text{ on } z \\ q_\phi(x|z) &\propto q_\phi(x, z) && \text{Posterior of } q_\phi(z|x)\end{aligned}$$

For the purposes of optimization, the ELBO objective can be written equivalently (up to an additive constant) as

$$\begin{aligned}\mathcal{L}_{\mathrm{ELBO}} &\equiv -D_{\mathrm{KL}}(q_\phi(x, z)\|p_\theta(x, z)) && (2) \\ &= -D_{\mathrm{KL}}(p_{\mathcal{D}}(x)\|p_\theta(x)) && (3) \\ &\qquad - \mathbb{E}_{p_{\mathcal{D}}(x)}[D_{\mathrm{KL}}(q_\phi(z|x)\|p_\theta(z|x))] \\ &= -D_{\mathrm{KL}}(q_\phi(z)\|p(z)) \\ &\qquad - \mathbb{E}_{q_\phi(z)}[D_{\mathrm{KL}}(q_\phi(x|z)\|p_\theta(x|z))] && (4)\end{aligned}$$

We prove the first equivalence in the appendix. The second and third equivalence are simple applications of the additive property of KL divergence. All three forms of ELBO in Eqns. (2),(3),(4) are useful in our analysis.

Two Problems of Variational Autoencoders

Amortized Inference Failures

Under ideal conditions, optimizing the ELBO objective using sufficiently flexible model families for $p_\theta(x|z)$ and $q_\phi(z|x)$ over θ, ϕ will achieve both goals of correctly capturing $p_{\mathcal{D}}(x)$ and performing correct amortized inference. This can be seen by examining Eq. (3). This form indicates that the ELBO objective is minimizing the KL divergence between the data distribution $p_{\mathcal{D}}(x)$ and the (marginal) model distribution $p_\theta(x)$, as well as the KL divergence between the variational posterior $q_\phi(z|x)$ and the true posterior $p_\theta(z|x)$. However, with finite model capacity the two goals can be conflicting and subtle tradeoffs and failure modes can emerge from optimizing the ELBO objective.

In particular, one limitation of the ELBO objective is that it might fail to learn an amortized inference distribution $q_\phi(z|x)$ that approximates the true posterior $p_\theta(z|x)$. This can happen for two different reasons:

Inherent properties of the ELBO objective: the ELBO objective can be maximized (even to $+\infty$ in pathological cases) even with a very inaccurate variational posterior $q_\phi(z|x)$.

Implicit modeling bias: common modeling choices (such as the high dimensionality of $\mathcal{X}$ compared to $\mathcal{Z}$) tend to sacrifice variational inference vs. data fit when modeling capacity is not sufficient to achieve both.

We will explain in turn why these failures happen.

Good ELBO Values Do Not Imply Accurate Inference We first provide some intuition to this phenomena, then formally prove the result for a pathological case of continuous spaces and Gaussian distributions. Finally we justify in the experiments section that this happens in realistic settings on real datasets (in both continuous and discrete spaces).

The ELBO objective in original form has two components, a log likelihood (reconstruction) term $\mathcal{L}_{\mathrm{AE}}$ and a regularization term $\mathcal{L}_{\mathrm{REG}}$:

$$\begin{aligned}&\mathcal{L}_{\mathrm{ELBO}}(x) = \mathcal{L}_{\mathrm{AE}}(x) + \mathcal{L}_{\mathrm{REG}}(x) \\ &\equiv \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{\mathrm{KL}}(q_\phi(z|x)||p(z))\end{aligned}$$

Let us first consider what happens if we only optimize $\mathcal{L}_{\mathrm{AE}}$ and not $\mathcal{L}_{\mathrm{REG}}$. The first term maximizes the log likelihood of observing data point x given its inferred latent variables $z \sim q_\phi(z|x)$. Consider a finite dataset $\{x_1, \cdots, x_N\}$. Let q_ϕ be such that for $x_i \neq x_j$, $q_\phi(z|x_i)$ and $q_\phi(z|x_j)$ are distributions with disjoint supports. Then we can learn a $p_\theta(x|z)$ mapping the support of each $q_\phi(z|x_i)$ to a distribution concentrated on x_i, leading to very large $\mathcal{L}_{\mathrm{AE}}$ (for continuous distributions $p_\theta(x|z)$ may even tend to a Dirac delta distribution and $\mathcal{L}_{\mathrm{AE}}$ tends to $+\infty$). Intuitively, the $\mathcal{L}_{\mathrm{AE}}$ component will encourage choosing $q_\phi(z|x_i)$ with disjoint support when $x_i \neq x_j$.

In almost all practical cases, the variational distribution family for q_ϕ is supported on the entire space $\mathcal{Z}$ (e.g., it is a Gaussian with non-zero variance, or IAF posterior (Kingma, Salimans, and Welling 2016)), preventing disjoint supports. However, attempting to learn disjoint supports for $q_\phi(z|x_i), x_i \neq x_j$ will "push" the mass of the

distributions away from each other. For example, for continuous distributions, if q_ϕ maps each x_i to a Gaussian $\mathcal{N}(\mu_i, \sigma_i)$, the $\mathcal{L}_{\text{AE}}$ term will encourage $\mu_i \to \infty, \sigma_i \to 0^+$.

This undesirable result may be prevented if the $\mathcal{L}_{\text{REG}}$ term can counter-balance this tendency. However, we show that the regularization term $\mathcal{L}_{\text{REG}}$ is not always sufficiently strong to prevent this issue. We first prove this fact in the simple case of a mixture of two Gaussians. We will then evaluate this finding empirically on realistic datasets in the experiments section .

Proposition 1. *Let $\mathcal{X}, \mathcal{Z} = \mathbb{R}$, and $\mathcal{D}$ be a dataset with two samples $\{-1, 1\}$; let $p_\theta(x|z)$ be selected from the family of all functions $\mu_\theta^p, \sigma_\theta^p$ that map $z \in \mathcal{Z}$ to a Gaussian $\mathcal{N}(\mu_\theta^p(z), \sigma_\theta^p(z))$ on $\mathcal{X}$, and $q_\phi(z|x)$ be selected from the family of all functions $\mu_\phi^q, \sigma_\phi^q$ that map $x \in \mathcal{X}$ to a Gaussian $\mathcal{N}(\mu_\phi^q(z), \sigma_\phi^q(z))$ on $\mathcal{Z}$. Then $\mathcal{L}_{ELBO}$ can be maximized to $+\infty$ when*

$$\mu_\phi^q(x=1) \to +\infty \qquad \mu_\phi^q(x=-1) \to -\infty$$
$$\sigma_\phi^q(x=1) \to 0^+ \qquad \sigma_\phi^q(x=-1) \to 0^+$$

and θ is optimally selected given ϕ. In addition the variational gap $D_{\text{KL}}(q_\phi(z|x)\|p_\theta(z|x)) \to +\infty$ for all $x \in \mathcal{D}$.

A proof can be found the in Appendix[1]. This means that amortized inference has completely failed, even though the ELBO objective can be made arbitrarily large. The model learns an inference distribution $q_\phi(z|x)$ that pushes all probability mass to ∞. This will become infinitely far from the true posterior $p_\theta(z|x)$ as measured by D_{KL}.

Modeling Bias In the above example we indicated a potential problem with the ELBO objective where the model tends to push the probability mass of $q_\phi(z|x)$ too far from 0. This tendency is a property of the ELBO objective and true for any $\mathcal{X}$ and $\mathcal{Z}$. However this is made worse by the fact that $\mathcal{X}$ is often higher dimensional compared to $\mathcal{Z}$, so any error in fitting $\mathcal{X}$ will be magnified compared to $\mathcal{Z}$.

For example, consider fitting an n dimensional distribution $\mathcal{N}(0, I)$ with $\mathcal{N}(\epsilon, I)$ using KL divergence, then

$$D_{\text{KL}}(\mathcal{N}(0, I), \mathcal{N}(\epsilon, I)) = n\epsilon^2/2$$

As n increases with some fixed ϵ, the Euclidean distance between the means of the two distributions is $\Theta(\sqrt{n})$, yet the corresponding D_{KL} becomes $\Theta(n)$. For natural images, the dimensionality of $\mathcal{X}$ is often orders of magnitude larger than the dimensionality of $\mathcal{Z}$. Recall in Eq.(4) that ELBO is optimizing both $D_{\text{KL}}(q_\phi(z)\|p(z))$ and $D_{\text{KL}}(q_\phi(x|z)\|p_\theta(x|z))$. Because the same per-dimensional modeling error incurs a much larger loss in $\mathcal{X}$ space than $\mathcal{Z}$ space, when the two objectives are conflicting (e.g., because of limited modeling capacity), the model will tend to sacrifice divergences on $\mathcal{Z}$ and focus on minimizing divergences on $\mathcal{X}$.

Regardless of the cause (properties of ELBO or modeling choices), this is generally an undesirable phenomenon for two reasons:

1) One may care about accurate inference more than generating sharp samples. For example, generative models are often used for down stream tasks such as semi supervised learning.

2) Overfitting: Because $p_\mathcal{D}$ is an empirical (finite) distribution in practice, matching it too closely can lead to poor generalization (Shu et al. 2018).

Both issues are observed in the experiments section.

[1]Appendix in technical report (Zhao, Song, and Ermon 2017)

The Information Preference Property

Using a complex decoding distribution $p_\theta(x|z)$ such as PixelRNN/PixelCNN (van den Oord, Kalchbrenner, and Kavukcuoglu 2016; Gulrajani et al. 2016) has been shown to significantly improve sample quality on complex natural image datasets. However, this approach suffers from a new problem: it tends to neglect the latent variables z altogether, that is, the mutual information between z and x becomes vanishingly small. Intuitively, the reason is that the learned $p_\theta(x|z)$ is the same for all $z \in \mathcal{Z}$, implying that the z is not dependent on the input x. This is undesirable because a major goal of unsupervised learning is to learn meaningful latent features which *should* depend on the inputs.

This effect, which we shall refer to as the *information preference* problem, was studied in (Chen et al. 2016) with a coding efficiency argument. Here we provide an alternative interpretation and a novel solution to this problem.

We inspect the ELBO in the form of Eq.(3), and consider the two terms respectively. We show that both can be optimized to 0 *without utilizing the latent variables*.

$D_{\text{KL}}(p_\mathcal{D}(x)||p_\theta(x))$: Suppose the model family $\{p_\theta(\cdot|z), \theta \in \Theta\}$ is sufficiently flexible and there exists a θ^* such that for every $z \in \mathcal{Z}$, $p_{\theta^*}(\cdot|z)$ and $p_\mathcal{D}(\cdot)$ are identical. Then we select this θ^* and the marginal $p_{\theta^*}(x) = \int_z p(z)p_{\theta^*}(x|z)dz = p_\mathcal{D}(x)$, hence this divergence $D_{\text{KL}}(p_\mathcal{D}(x)||p_\theta(x)) = 0$ which is optimal.

$\mathbb{E}_{p_\mathcal{D}(x)}[D_{\text{KL}}(q_\phi(z|x)||p_\theta(z|x))]$: Because $p_{\theta^*}(\cdot|z)$ is the same for every z (x is independent from z) we have $p_{\theta^*}(z|x) = p(z)$. Because $p(z)$ is usually a simple distribution, if it is possible to choose ϕ such that $q_\phi(z|x) = p(z), \forall x \in \mathcal{X}$, this divergence will also achieve the optimal value of 0.

Because $\mathcal{L}_{\text{ELBO}}$ is the sum of the above divergences, when both are 0, this is a global optimum. There is no incentive for the model to learn otherwise, undermining our purpose of learning a latent variable model.

The InfoVAE Model Family

In order to remedy these two problems we define a new training objective that will learn both the correct model and amortized inference distributions. We begin with the form of ELBO in Eq. (4)

$$\mathcal{L}_{\text{ELBO}} = -D_{\text{KL}}(q_\phi(z)\|p(z)) - \mathbb{E}_{p(z)}[D_{\text{KL}}(q_\phi(x|z)\|p_\theta(x|z))]$$

First we add a scaling parameter λ to the divergence between $q_\phi(z)$ and $p(z)$ to increase its weight and counter-act the imbalance between $\mathcal{X}$ and $\mathcal{Z}$ (see previous discussion). Next we add a mutual information maximization term to encourage high mutual information between x and z. This encourages the model to use the latent code and avoids the

information preference problem. We arrive at the following objective

$$\mathcal{L}_{\text{InfoVAE}} = -\lambda D_{\text{KL}}(q_\phi(z)\|p(z)) - \mathbb{E}_{q(z)}[D_{\text{KL}}(q_\phi(x|z)\|p_\theta(x|z))] + \alpha I_q(x;z) \quad (5)$$

where $I_q(x;z)$ is the mutual information between x and z under the distribution $q_\phi(x,z)$.

Even though we cannot directly optimize this objective, we can rewrite this into an equivalent form that we can optimize more efficiently (we prove this in the Appendix)

$$\mathcal{L}_{\text{InfoVAE}} \equiv \mathbb{E}_{p_\mathcal{D}(x)}\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - (1-\alpha)\mathbb{E}_{p_\mathcal{D}(x)}D_{\text{KL}}(q_\phi(z|x)||p(z)) - (\alpha+\lambda-1)D_{\text{KL}}(q_\phi(z)\|p(z)) \quad (6)$$

The first two terms can be optimized by the reparameterization trick as in the original ELBO objective. The last term $D_{\text{KL}}(q_\phi(z)\|p(z))$ is not easy to compute because we cannot tractably evaluate $\log q_\phi(z)$. However we can obtain unbiased samples from it by first sampling $x \sim p_\mathcal{D}$, then $z \sim q_\phi(z|x)$, so we can optimize it by likelihood free optimization techniques (Goodfellow et al. 2014; Nowozin, Cseke, and Tomioka 2016; Arjovsky, Chintala, and Bottou 2017; Gretton et al. 2007). In fact we may replace the term $D_{\text{KL}}(q_\phi(z)\|p(z))$ with anther divergence $D(q_\phi(z)\|p(z))$ that we can efficiently optimize over. Changing the divergence may alter the empirical behavior of the model but we show in the following theorem that replacing D_{KL} with any (strict) divergence is still correct (a divergence is strict if $D(q_\phi(z)\|p(z)) = 0$ iff $q_\phi(z) = p(z)$). Let $\hat{\mathcal{L}}_{\text{InfoVAE}}$ be the objective where we replace $D_{\text{KL}}(q_\phi(z)\|p(z))$ with any strict divergence $D(q_\phi(z)\|p(z))$.

Proposition 2. *Let $\mathcal{X}$ and $\mathcal{Z}$ be continuous spaces, and $\alpha < 1$, $\lambda > 0$. Given fixed mutual information I_0, among the set of q_ϕ that satisfy $I_q(x;z) = I_0$, $\hat{\mathcal{L}}_{\text{InfoVAE}}$ is globally optimized if $p_\theta(x) = p_\mathcal{D}(x)$ and $q_\phi(z|x) = p_\theta(z|x), \forall z$.*

Proof of Proposition 2. See Appendix. □

Note that in the proposition we have the additional requirement that the mutual information $I_q(x;z)$ is bounded. This is inevitable because if $\alpha > 0$ the objective can be optimized to $+\infty$ by simply increasing the mutual information infinitely. In our experiments simply ensuring that $q_\phi(z|x)$ does not have vanishing variance is sufficient to regularize the behavior of the model.

Relation to VAE and β-VAE: This model family generalizes several previous models. When $\alpha = 0$ and $\lambda = 1$ we get back the original ELBO objective. When $\lambda > 0$ is freely chosen, while $\alpha + \lambda - 1 = 0$, and we use the D_{KL} divergences, we get the β-VAE (Higgins et al. 2017) model family. However, β-VAE models cannot effectively trade-off weighing of $\mathcal{X}$ and $\mathcal{Z}$ and information preference. In particular, for every λ there is a unique value of $\alpha = 1 - \lambda$ that we can choose. For example, if we choose a large value of $\lambda \gg 1$ to balance the importance of observed and latent spaces ($\mathcal{X}$ and $\mathcal{Z}$), we must also choose $\alpha \ll 0$, which forces the model to penalize mutual information. This in turn can lead to under-fitting or ignoring the latent variables.

Relation to Adversarial Autoencoders (AAE): When $\alpha = 1$, $\lambda = 1$ and D is chosen to be the Jensen Shannon divergence we get the adversarial autoencoders in (Makhzani et al. 2015). This paper generalizes AAE, but more importantly we provide a deeper understanding of the correctness and desirable properties of AAE. Furthermore, we characterize settings when AAE is preferable compared to VAE (i.e. when we would like to have $\alpha = 1$).

Relation to Information Theoretical Analysis: Several other lines of work analyze mutual information in latent variable generative models from different perspectives, including (Zhao, Song, and Ermon 2018; Alemi et al. 2017).

Our generalization introduces new parameters, but the meaning and effect of the various parameter choices is clear. We always select λ to a value that makes the loss on $\mathcal{X}$ approximately the same as the loss on $\mathcal{Z}$. We also recommend setting $\alpha = 0$ when $p_\theta(x|z)$ is a simple distribution, and $\alpha = 1$ when $p_\theta(x|z)$ is a complex distribution and information preference is a concern. The final degree of freedom is the divergence $D(q_\phi(z)\|p(z))$ to use. We will explore this topic in the next section.

Divergences Families

We consider and compare three divergences in this paper.

Adversarial Training: Adversarial autoencoders (AAE) proposed in (Makhzani et al. 2015) use an adversarial discriminator to approximately minimize the Jensen-Shannon divergence (Goodfellow et al. 2014) between $q_\phi(z)$ and $p(z)$. However, when $p(z)$ is a simple distribution such as Gaussian, there are preferable alternatives. In fact, adversarial training can be unstable and slow even when we apply recent techniques for stabilizing GAN training (Arjovsky, Chintala, and Bottou 2017; Gulrajani et al. 2017).

Stein Variational Gradient: The Stein variational gradient (Liu and Wang 2016) is a simple and effective framework for matching a distribution q to p by computing effectively $\nabla_\phi D_{\text{KL}}(q_\phi(z)||p(z))$ which we can use for gradient descent minimization of $D_{\text{KL}}(q_\phi(z)||p(z))$. However a weakness of these methods is that they are difficult to apply efficiently in high dimensions. We give a detailed overview of this method in the Appendix.

Maximum-Mean Discrepancy: Maximum-Mean Discrepancy (MMD) (Gretton et al. 2007; Li, Swersky, and Zemel 2015; Dziugaite, Roy, and Ghahramani 2015) is a framework to quantify the distance between two distributions by comparing all of their moments. It can be efficiently implemented using the kernel trick. Letting $k(\cdot,\cdot)$ be any positive definite kernel, the MMD between p and q is

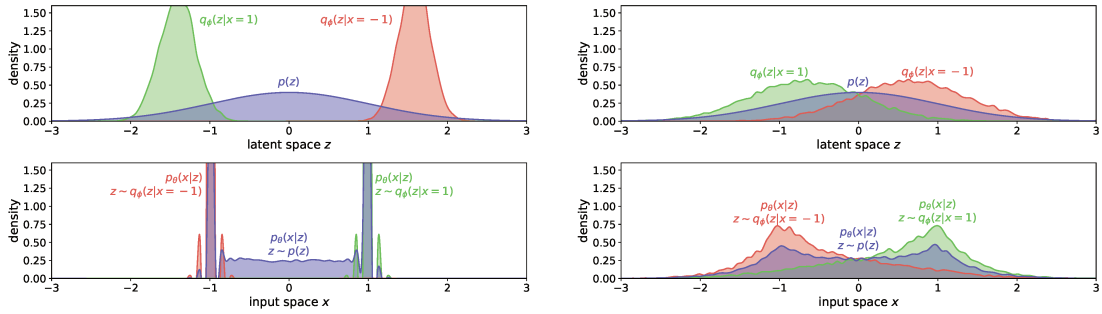
$$D_{\text{MMD}}(q\|p) = \mathbb{E}_{p(z),p(z')}[k(z,z')] - 2\mathbb{E}_{q(z),p(z')}[k(z,z')] + \mathbb{E}_{q(z),q(z')}[k(z,z')]$$

$D_{\text{MMD}} = 0$ if and only if $p = q$.

Experiments

Variance Overestimation with ELBO training

We first perform some simple experiments on toy data and MNIST to demonstrate that ELBO suffers from inaccurate inference in practice, and adding the scaling term λ in Eq.(5)



ELBO InfoVAE ($\lambda = 500$)

Figure 1: Verification of Proposition 1 where the dataset only contains two examples $\{-1, 1\}$. **Top:** density of the distributions $q_\phi(z|x)$ when $x = 1$ (red) and $x = -1$ (green) compared with the true prior $p(z)$ (purple). **Bottom:** The "reconstruction" $p_\theta(x|z)$ when z is sampled from $q_\phi(z|x = 1)$ (green) and $q_\phi(z|x = -1)$ (red). Also plotted is $p_\theta(x|z)$ when z is sampled from the true prior $p(z)$ (purple). When the dataset consists of only two data points, ELBO (**left**) will push the density in latent space Z away from 0, while InfoVAE (**right**) does not suffer from this problem.

can correct for this. Next, we will perform a comprehensive set of experiments to carefully compare different models on multiple performance metrics.

Mixture of Gaussian We verify the conclusions in Proposition 1 by using the same setting in that proposition. We use a three layer deep network with 200 hidden units in each layer to simulate the highly flexible function family. For InfoVAE we choose the scaling coefficient $\lambda = 500$, information preference $\alpha = 0$, and divergence optimized by MMD.

The results are shown in Figure 1. It can be observed that the predictions of the theory are reflected in the experiments: ELBO training leads to poor inference and a significantly over-estimated $q_\phi(z)$, while InfoVAE demonstrates a more stable behavior.

MNIST We demonstrate the problem on a real world dataset. We train ELBO and InfoVAE (with MMD regularization) on binarized MNIST with different training set sizes ranging from 500 to 50000 images; We use the DCGAN architecture (Radford, Metz, and Chintala 2015) for both models. For InfoVAE, we use the scaling coefficient $\lambda = 1000$, and information preference $\alpha = 0$. We choose the number of latent features $\text{dimension}(\mathcal{Z}) = 2$ to plot the latent space, and 10 for all other experiments.

First we verify that in real datasets ELBO over-estimates the variance of $q_\phi(z)$, while InfoVAE does not (with recommended choice of λ). In Figure 2 we plot estimates for the log determinant of the covariance matrix of $q_\phi(z)$, denoted as $\log\det(\text{Cov}[q_\phi(z)])$ as a function of the size of the training set. For standard factored Gaussian prior $p(z)$, $\text{Cov}[p(z)] = I$, so $\log\det(\text{Cov}[p(z)]) = 0$. Values above or below zero give us an estimate of over or under-estimation of the variance of $q_\phi(z)$, which should in theory match the prior $p(z)$. It can be observed that the for ELBO, variance of $q_\phi(z)$ is significantly over-estimated. This is especially severe when the training set is small. On the other hand, when we use a large value for λ, InfoVAE can avoid this problem.

To make this more intuitive we plot in Figure 3 the contour plot of $q_\phi(z)$ when training on 500 examples. It can be

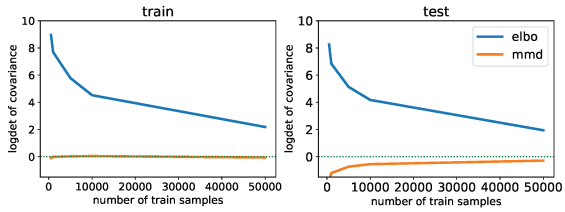


Figure 2: $\log\det(\text{Cov}[q_\phi(z)])$ for ELBO vs. MMD-VAE under different training set sizes. The correct prior $p(z)$ has value 0 on this metric, and values above or below 0 correspond to over-estimation and under-estimation of the variance respectively. ELBO (blue curve) shows consistent over-estimation while InfoVAE does not.

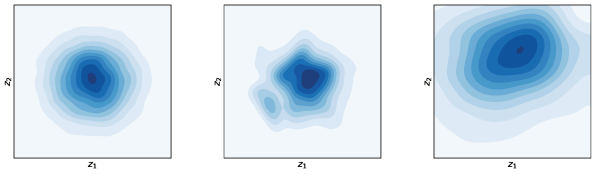
seen that with ELBO $q_\phi(z)$ matches $p(z)$ very poorly, while InfoVAE matches significantly better.

To verify that ELBO trains inaccurate amortized inference we plot in Figure 4 the comparison between samples from the approximate posterior $q_\phi(z|x)$ and samples from the true posterior $p_\theta(z|x)$ computed by rejection sampling. The same trend can be observed. ELBO consistently gives very poor approximate posterior, while the InfoVAE posterior is mostly accurate.

Finally show the samples generated by the two models in Figure 5. ELBO generates very sharp reconstructions, but very poor samples when sampled ancestrally $x \sim p(z)p_\theta(x|z)$. InfoVAE, on the other hand, generates samples of consistent quality, and in fact, produces samples of reasonable quality after only training on a dataset of 500 examples. This reflect InfoVAE's ability to control overfitting and demonstrate consistent training time and testing time behavior.

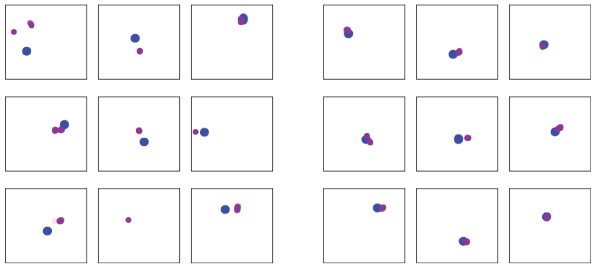
Comprehensive Comparison

In this section, we perform extensive qualitative and quantitative experiments on the binarized MNIST dataset to eval-



Prior $p(z)$ InfoVAE $q_\phi(z)$ ELBO $q_\phi(z)$

Figure 3: Comparing the prior, InfoVAE $q_\phi(z)$ and ELBO $q_\phi(z)$. InfoVAE $q_\phi(z)$ is almost identical to the true prior, while the ELBO $q_\phi(z)$ is very far off. $q_\phi(z)$ for both models is computed by $\int_x p_{\mathcal{D}} q_\phi(z|x)$ where $p_{\mathcal{D}}$ is the test set.



ELBO Posterior InfoVAE Posterior

Figure 4: Comparing true posterior $p_\theta(z|x)$ (green, generated by importance sampling) with approximate posterior $q_\phi(z|x)$ (blue) on testing data after training on 500 samples. For ELBO the approximate posterior is generally further from the true posterior compared to InfoVAE. (All plots are drawn with the same scale)

uate the performance of different models. We would like to answer these questions:

1) Compare the models on a comprehensive set of numerical metrics of performance. Also compare the stability and training speed of different models.

2) Evaluate and compare the possible types of divergences (Adversarial, Stein, MMD).

For all two questions, we find InfoVAE with MMD regularization to perform better in almost all metrics of performance and demonstrate the best stability and training speed. The details are presented in the following sections.

For models we use ELBO, Adversarial autoencoders, InfoVAE with Stein variational gradient, and InfoVAE with MMD ($\alpha = 1$ because information preference is a concern, $\lambda = 1000$ which can put the loss on $\mathcal{X}$ and $\mathcal{Z}$ on the same order of magnitude). In this setting we also use a highly flexible PixelCNN as the decoder $p_\theta(x|z)$ so that information preference is also a concern. Detailed experimental setup is explained in the Appendix.

We consider multiple quantitative evaluations, including the quality of the samples generated, the training speed and stability, the use of latent features for semi-supervised learning, and log-likelihoods on samples from a separate test set.

Distance between $q_\phi(z)$ and $p(z)$: To measure how well $q_\phi(z)$ approximates $p(z)$, we use two numerical metrics. The first is the full batch MMD statistic over the full data.

reconstruction generation

Figure 5: Samples generated by ELBO vs. MMD InfoVAE ($\lambda = 1000$) after training on 500 samples (plotting mean of $p_\theta(x|z)$). **Top:** Samples generated by ELBO. Even though ELBO generates very sharp reconstruction for samples on the training set, model samples $p(z)p_\theta(x|z)$ is very poor, and differ significantly from the reconstruction samples, indicating over-fitting, and mismatch between $q_\phi(z)$ and $p(z)$. **Bottom:** Samples generated by InfoVAE. The reconstructed samples and model samples look similar in quality and appearance, suggesting better generalization in the latent space.

Table 1: Log likelihood estimates for different models on the MNIST dataset. MMD-VAE achieves the best results, even though it is not explicitly optimizing a lower bound to the true log likelihood.

	Log likelihood estimate
ELBO	82.75
MMD-VAE	**80.76**
Stein-VAE	81.47
Adversarial VAE	82.21

MMD is also used during training of MMD-VAE but only estimated from mini-batches; during evaluation we can use the full dataset to obtain more accurate estimates. The second is the log determinant of the covariance matrix of $q_\phi(z)$. Ideally when $p(z)$ is the standard Gaussian Σ_{q_ϕ} should be the identity matrix, so $\log\det(\Sigma_{q_\phi}) = 0$. In our experiments we plot the log determinant divided by the dimensionality of the latent space. This measures the average under/over estimation per dimension of the learned covariance.

The results are plotted in Figure 6 (A,B). This is different from the experiments in Figure 2 because in this case the decoder is a highly complex pixel recurrent model and the concern that we highlight is failure to use latent features rather than inaccurate posterior. MMD achieves the best performance except for ELBO. Even though ELBO achieves extremely low error, this is trivial because for this experimental setting of flexible decoders, ELBO learns a latent code z that does not contain any information about x, and $q_\phi(z|x) \approx p(z)$ for all z.

Sample distribution: If the generative model $p(z)p_\theta(x|z)$ has true marginal $p_{data}(x)$, then the dis-

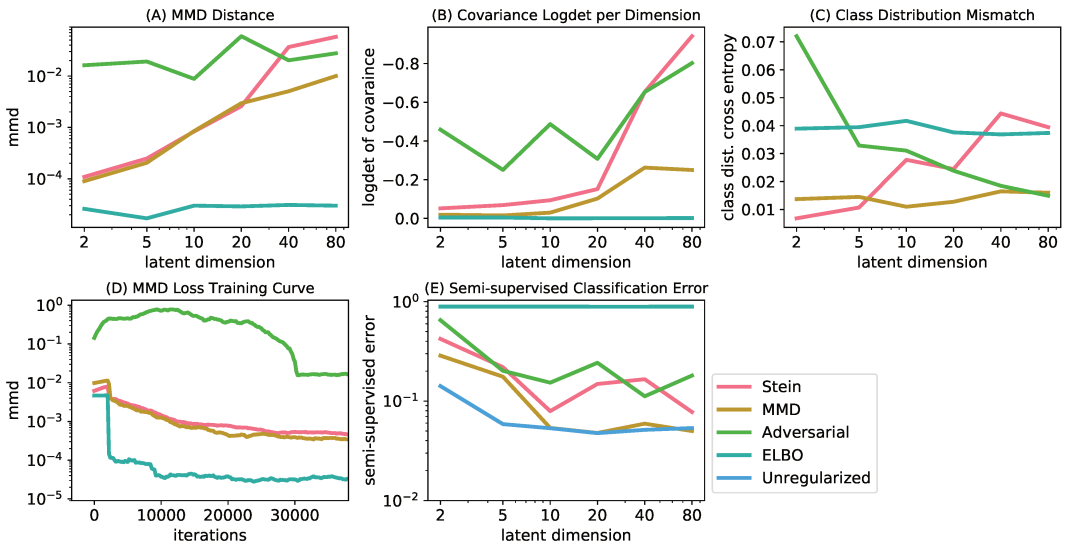


Figure 6: Comparison of numerical performance. We evaluate MMD, log determinant of sample covariance, cross entropy with correct class distribution, and semi-supervised learning performance. 'Stein', 'MMD', 'Adversarial' and 'ELBO' corresponds to the VAE where the latent code are regularized with the respective methods, and 'Unregularized' corresponds to the vanilla autoencoder without regularization over the latent dimensions.

tribution of different object classes should be identical to the original dataset. On the other hand, an incorrect generative model is unlikely to generate a correct object class distribution. Let $\boldsymbol{c}$ denote the class distribution in the real dataset, and $\hat{\boldsymbol{c}}$ denote the class distribution of the generated images, computed by a pretrained classifier. We use cross entropy loss $l_{ce}(\boldsymbol{c}, \hat{\boldsymbol{c}}) = -\boldsymbol{c}^T(\log \hat{\boldsymbol{c}} - \log \boldsymbol{c})$ to measure the deviation from the true class distribution.

The results are plotted in Figure 6 (C). Stein regularization performs well on low dimensional latent space (2 dimensions), whereas adversarial regularization performs better on higher dimensions; MMD regularization performs well in all the cases.

Training Speed and Stability: We prefer a model that is stable, trains quickly and requires little hyperparameter tuning. In Figure 6 (D) we plot the change of MMD statistic vs. the number of iterations. In this respect, adversarial autoencoder becomes less desirable because it takes longer to converge, and sometimes converges to poor results even with WGAN-GP discriminators (Gulrajani et al. 2017).

Semi-supervised Learning: To evaluate the quality of the learned features for downstream tasks such as semi-supervised learning, we train a SVM directly on the learned MNIST latent features. We use the M1+TSVM setting in (Kingma et al. 2014), and train with 1000 samples. Lower classification error would suggest that the learned features z contain more information about the data x. The results are shown in Figure 6 (E). We observe that an unregularized autoencoder (which does not use any regularization $\mathcal{L}_{\text{REG}}$) is superior when the latent dimension is low and MMD catches up when it is high. Furthermore, the latent code learned by ELBO contains almost no information about the input: the classification error is no better than random guess.

Log likelihood: To be consistent with previous results, we use the stochastically binarized MNIST (Salakhutdinov and Murray 2008). Estimation of log likelihood is achieved by importance sampling. We use 5-dimensional latent features in our log likelihood experiments. The values are shown in Table 1. All the objectives perform on-par or superior compared to our ELBO baseline. We do not explicitly optimize a lower bound to the true log likelihood, but still achieve good log likelihood because of Proposition 2

Conclusion

Despite the recent success of variational autoencoders, they can fail to perform amortized inference, or learn meaningful latent features. We trace both issues back to the ELBO learning criterion, and modify the ELBO objective to propose a new model family that can fix both problems. We perform extensive experiments to verify the effectiveness of our approach. Our experiments show that a particular subset of our model family, MMD-VAEs perform on-par or better than all other approaches on multiple metrics of performance.

Acknowledgements

This research was supported by Intel, TRI, NSF (#1651565, #1522054, #1733686), AFOSR, and ONR.

References

Alemi, A. A.; Poole, B.; Fischer, I.; Dillon, J. V.; Saurous, R. A.; and Murphy, K. 2017. An information-theoretic analysis of deep latent-variable models. *arXiv preprint arXiv:1711.00464*.

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

Chen, X.; Kingma, D. P.; Salimans, T.; Duan, Y.; Dhariwal, P.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*.

Dziugaite, G. K.; Roy, D. M.; and Ghahramani, Z. 2015. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2672–2680.

Gretton, A.; Borgwardt, K. M.; Rasch, M.; Schölkopf, B.; and Smola, A. J. 2007. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, 513–520.

Gulrajani, I.; Kumar, K.; Ahmed, F.; Taiga, A. A.; Visin, F.; Vázquez, D.; and Courville, A. C. 2016. Pixelvae: A latent variable model for natural images. *CoRR* abs/1611.05013.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. 2017. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*.

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

Kingma, D. P.; Rezende, D. J.; Mohamed, S.; and Welling, M. 2014. Semi-supervised learning with deep generative models. *CoRR* abs/1406.5298.

Kingma, D. P.; Salimans, T.; and Welling, M. 2016. Improving variational inference with inverse autoregressive flow. *arXiv preprint arXiv:1606.04934*.

Li, Y.; Song, J.; and Ermon, S. 2017. Infogail: Interpretable imitation learning from visual demonstrations. In *Advances in Neural Information Processing Systems*, 3812–3822.

Li, Y.; Swersky, K.; and Zemel, R. 2015. Generative moment matching networks. In *International Conference on Machine Learning*, 1718–1727.

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

Makhzani, A.; Shlens, J.; Jaitly, N.; and Goodfellow, I. 2015. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

Nowozin, S.; Cseke, B.; and Tomioka, R. 2016. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, 271–279.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Salakhutdinov, R., and Murray, I. 2008. On the quantitative analysis of deep belief networks. In *Proceedings of the 25th international conference on Machine learning*, 872–879. ACM.

Shu, R.; Bui, H. H.; Zhao, S.; Kochenderfer, M. J.; and Ermon, S. 2018. Amortized inference regularization. *NIPS*.

van den Oord, A.; Kalchbrenner, N.; and Kavukcuoglu, K. 2016. Pixel recurrent neural networks. *CoRR* abs/1601.06759.

Yang, Z.; Hu, Z.; Salakhutdinov, R.; and Berg-Kirkpatrick, T. 2017. Improved variational autoencoders for text modeling using dilated convolutions. *CoRR* abs/1702.08139.

Zhao, S.; Song, J.; and Ermon, S. 2017. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*.

Zhao, S.; Song, J.; and Ermon, S. 2018. The information autoencoding family: A lagrangian perspective on latent variable generative models. *arXiv preprint arXiv:1806.06514*.

Zhu, J.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR* abs/1703.10593.