Learnable Explicit Density for Continuous Latent Space and Variational Inference

Chin-Wei Huang, Ahmed Touati, Laurent Dinh, Michal Drozdzal, Mohammad Havaei, Laurent Charlin & Aaron Courville

MILA, Univesité de Montreal, Imagia Inc., HEC Montréal, CIFAR

cw.huang427@gmail.com

Abstract

In this paper, we study two aspects of the variational autoencoder (VAE): the prior distribution over the latent variables and its corresponding posterior. First, we decompose the learning of VAEs into layerwise density estimation, and argue that having a flexible prior is beneficial to both sample generation and inference. Second, we analyze the family of inverse autoregressive flows (inverse AF) and show that with further improvement, inverse AF could be used as universal approximation to any complicated posterior. Our analysis results in a unified approach to parameterizing a VAE, without the need to restrict ourselves to use factorial Gaussians in the latent real space.

Introduction

VAEs [1, 2] can be interpreted as an infinite mixture model $p(x) = \int_z p(x|z)p(z)dz$ where the parameters of the class conditional distribution $p(x|z)$ are functions of the latent variable z (which is thought of as class here), and there are infinitely many classes. Such models should theoretically have enough flexibility to capture highly complex distributions such as image manifolds, but in practice its samples are found to be quite blurry. We make two explanations:

1. **The model is not calibrated for sampling from the prior**; i.e. $q(z) \neq p(z)$.
2. **Maximizing the variation lower bound introduces bias in the model**:
$$\max \log p(x) \neq \max \log p(x) - \mathcal{KL}(q(z|x)||p(z|x))$$

In this paper, we make two main contributions. First, we analyze the effect of making the prior learnable. We show that training with the variational lower bound under some limit conditions matches the marginal approximate posterior with the prior, which is desirable from the generative model point of view. We then decompose the lower bound, and show that updating the prior alone brings the prior closer to the marginal approximate posterior, suggesting that having the prior trainable is beneficial to both sample generation and inference. Our second contribution is to prove that by using the family of inverse AF [3], one can universally approximate any posterior. This theoretically justifies the use of inverse AF to improve variational inference. We unified the two aspects and propose to use invertible functionals [4] and [3] to parameterize explicit densities for both the prior and approximate posterior.

Marginal Matching Prior

- Goal: to train an auto-encoder as a generative model by keeping $q(z) = \int_x p_{\mathcal{D}}(x)q(z|x)dx$ close to the prior.

$$\textit{1. } q(z|x) \to p(z|x) \quad \forall x \sim p_{\mathcal{D}}(x) \qquad \textit{2. } p(x) \to p_{\mathcal{D}}(x)$$

$$\begin{aligned} q(z) &= \int_x p_{\mathcal{D}}(x)q_\phi(z|x)dx \\ &\xrightarrow{1.} \int_x p_{\mathcal{D}}(x)p_{\theta,\pi}(z|x)dx \\ &\xrightarrow{2.} \int_x p_{\theta,\pi}(x)p_{\theta,\pi}(z|x)dx = p_\pi(z) \end{aligned} \tag{1}$$

We can cast it as an optimization problem by minimizing the KL-divergences:

$$\begin{aligned} &\min \mathbf{E}_{p_{\mathcal{D}}(x)}[\mathcal{KL}(q(z|x)||p(z|x))] + \mathcal{KL}(p_{\mathcal{D}}(x)||p(x)) \\ &= \max \mathbf{E}_{x\sim p_{\mathcal{D}}(x)}[\underbrace{\mathbf{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z) + \log p_\pi(z) - \log q_\phi(z|x)\right]}_{\mathcal{L}(\theta,\phi,\pi;x)}] \end{aligned} \tag{2}$$

- Training of VAEs can be decomposed by coordinate-ascent into layer-wise density estimation.

$$\max_\pi \mathbf{E}[\mathcal{L}] = \min_\pi \mathcal{KL}(\mathbf{E}_{x\sim p_{\mathcal{D}}(x)}[q(z|x)]||p_\pi(z)) \tag{3}$$

- Variational distribution could be improved by having a better prior which simplifies the true posterior.

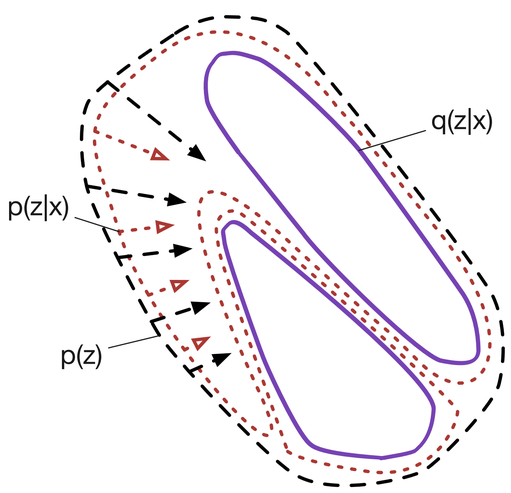

Figure 1: Effect of prior on posterior. Matching the prior $p(z)$ with the marginal approximate posterior $\mathbf{E}_x[q(z|x)]$ makes the true posterior $p(z|x)$ easier to model, since it pushes the true posterior to be closer to the approximate posterior.

Inverse Autoregressive Flows as Universal Posterior Approximator

Lemma 1. ***Existence of solution to a nonlinear independent component analysis problem [5].*** *Given a random vector $X = (X_i)_{i=1\ldots m} \in \mathbf{R}^m$, there always exists a mapping g from $\mathbf{R}^m$ to $\mathbf{R}^m$ such that the components of the random vector $Y = f(X)$ are statistically independent.*

Proposition 1. ***Inverse autoregressive transformation as universal approximator of any density.*** *Let X be a random vector in an open set $\mathcal{U} \subset \mathbf{R}^m$. We assume that X has a positive and continuous probability density distribution. There exists a sequence of mappings $(G_n)_{n\geq 0}$ from $(0,1)^m$ to $\mathbf{R}^m$ parametrized by autoregressive neural networks such that the sequence $X_n = G_n(Y)$ where $Y \sim Unif((0,1)^m)$ converges in distribution to X.*

Proposed Method

$$\begin{aligned} \mathcal{L} = &\mathbf{E}_{q(z'|x)}[\log p(x|g(z'))] + \\ &\mathbf{E}_{q(z'|x)}\left[\log p(h^{-1}\circ g(z')) + \log\left|\frac{\partial h^{-1}}{\partial z}(g(z'))\right|\right] - \\ &\mathbf{E}_{q(z'|x)}\left[\log q(z'|x) - \log\left|\frac{\partial g}{\partial z'}(z')\right|\right] \end{aligned} \tag{4}$$

Experiments

Mixture of Bivariate Gaussians

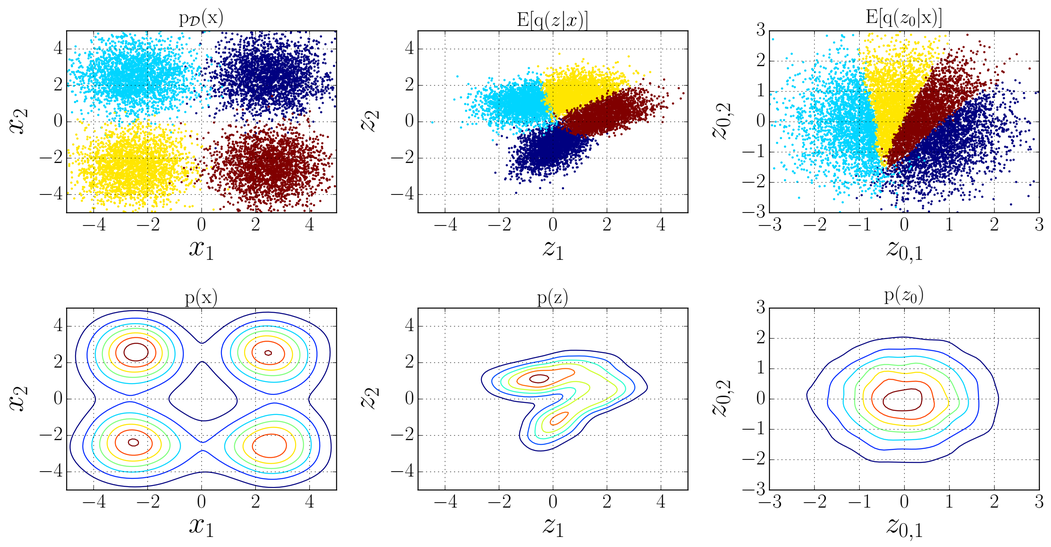

Figure 2: Fitting a Gaussian mixture distribution. $E[\cdot]$ indicates marginalization over the data $x \sim p_{\mathcal{D}}(x)$. Clockwise from top left: projection of data distribution $p_{\mathcal{D}}(x)$ onto the prior space $\mathbf{E}(q(z|x))$, and the base distribution space $\mathbf{E}(q(z_0|x))$; density maps of the base distribution $p(z_0)$, the transformed prior $p(z)$ and and marginal model distribution $p(x)$.

MNIST Density Estimation

Table 1: Effect of prior.

MLP		MLP		ResConv	
L_{post}	NLL	L_{prior}	NLL	L_{prior}	NLL
0	90.78	0	90.78	0	83.11
4	88.89	4	88.07	4	81.87
8	88.71	8	87.47	8	81.70
12	88.70	12	86.59	12	81.44

Table 2: Effect of prior and posterior.

ResConv		
L_{prior}	L_{post}	NLL
4 NVP	4 NVP	81.81
8 NVP	8 NVP	81.55
8 NVP	8 MADE	80.81
16 NVP	16 MADE	80.60

References

[1] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. 2014.

[2] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. 2014.

[3] Diederik P Kingma, Tim Salimans, and Max Welling. Improving Variational Inference with Inverse Autoregressive Flow. 2016.

[4] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. 2016.

[5] Aapo Hyvarjnen and Petteri Pajunen. Nonlinear Independent Component Analysis: Existence and Uniqueness Results. 1998.