Learnable Explicit Density for Continuous Latent Space and Variational Inference

Chin-Wei Huang [1] **Ahmed Touati** [1] **Laurent Dinh** [1] **Michal Drozdzal** [1 2] **Mohammad Havaei** [1 2]
Laurent Charlin [1 3] **Aaron Courville** [1 4]

Abstract

In this paper, we study two aspects of the VAE: the prior distribution over the latent variables and its corresponding posterior. 1, 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. 2, 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.

1. Introduction

Deep Gaussian Latent Models (Rezende et al., 2014), also known as VAEs (Kingma & Welling, 2014), fall within the paradigm of MLE and are often applied in computer vision problems. However, training with MLE usually leads to overestimation of the entropy of the data distribution (Minka, 2005). This is an undesirable property, as natural images are usually assumed to lie within a lower dimensional manifold, and the additional entropy (and other simplifying modeling assumptions for the purpose of explicit density estimation) often leads to a marginal likelihood with probability mass spread out in the data space where there is no support in the training data, which causes the blurriness of samples. These observations motivate the design of more flexible, complex families of model densities.

Since a clv z is introduced to the model, VAEs 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 lv 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 it is found to be overshadowed by tractable density models such as autoregressive models (Van Den Oord et al., 2016), or GANs (Goodfellow et al., 2014) in terms of sample generation quality.

It is believed that the relative poor performance in sample quality lies in the fact that the introduction of a latent representation requires approximate inference, as the model distribution is biased by simplifying posterior densities (Buntine & Jakulin, 2004); i.e. training is achieved by maximizing the vlb:

$$\mathcal{L}(\theta, \phi, \pi; x) = \mathbf{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x|z)p_\pi(z)}{q_\phi(z|x)}\right] \quad (1)$$

We discuss two aspects of training with the bound. 1, maximizing (1) with respect to ϕ amounts to minimizing $\mathcal{KL}(q_\phi(z|x)||p(z|x))$; the variational distribution, $q(z|x)$, can thus be viewed as an approximate to the true posterior, $p(z|x)$. Simplifying $q(z|x)$ (e.g. by using a factorial Gaussian as a common practice) is problematic, as the marginal log likelihood of interest $\log p(x)$ can only be optimized to the extent we are able to approximate the true posterior using the variational distribution. This motivates a direct improvement of variational inference (Rezende & Mohamed, 2015; Ranganath et al., 2015; Kingma et al., 2016).

2, during training of the VAE, only a part of the latent space is explored. When marginalizing out the input vector x, we recover the marginal $q(z) = \int_x q(z|x)p_{\mathcal{D}}(x)$, where $\mathcal{D}$ indicates the true data distribution. When the marginal approximate posterior fails to fill up the prior as the prior-contractive term requires, one would risk sampling from untrained regions in the latent space. A direct and non-parametric treatment of sampling from such regions of the prior would be to take $q(z)$ as the prior, but the integral is intractable and the data distribution is only partially speci-

[1]MILA, Université de Montréal, Canada [2]Imagia Inc., Canada [3]HEC Montréal, Canada [4]CIFAR Fellow, Canada. Correspondence to: Chin-Wei Huang <cw.huang427@gmail.com>.

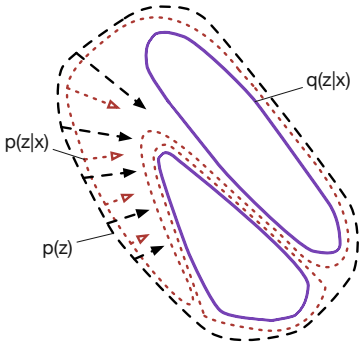

Figure 1. Effect of prior on posterior. Matching the prior $p(z)$ with the marginal approximate posterior $\mathbf{E}_x[q(z|x)]$ makes the true posterior $p(z|x)$ easier to model, since it pushes the true posterior to be closer to the approximate posterior.

fied by a limited training data. Even if we take the empirical distribution of $p_{\mathcal{D}}(x)$, we would have a mixture model of up to n components, where n is the number of training data points, which would be impractical given the scale of modern machine learning tasks. A workaround of this problem is to take a random subset of $\mathcal{D}$, or introduce a learnable set of pseudo-data of size K, and set the prior to be $p(z) = \sum_{j=1}^{K} \frac{1}{K} q(z_j|x_j)$, which is shown to be promising in the recent work done by Tomczak & Welling (2017). Another approach is to directly regularize the autoencoder by matching the aggregated posterior with the prior, as in Makhzani et al. (2015).

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 (Kingma et al., 2016), 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 Dinh et al. (2016) and Kingma et al. (2016) to parameterize explicit densities for both the prior and approximate posterior.

2. Marginal Matching Prior

We claim that maximizing the vlb explicitly matches the marginal $q(z)$ with the prior $p(z)$. By decomposing the lower bound , we then suggest using a learnable prior to improve sampling , i.e. to have a prior that matches the marginal $q(z)$ instead.

Let us define encoding and decoding distributions as $q(z|x)$ and $p(x|z)$ respectively, a prior as $p(z)$ and a data distribution as $p_{\mathcal{D}}(x)$. Our goal is 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. This can be achieved at the limits of the following two conditions (Hoffman & Johnson, 2016):

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

In words, given a perfect approximate posterior $q(z|x)$ of $p(z|x)$ and a perfect marginal likelihood $p(x)$ of $p_{\mathcal{D}}(x)$, we have the marginal $q(z)$ converge to the prior, i.e.

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

That is, to have $q(z) \to p(z)$, we need to ensure the two conditions are satisfied. We can cast it as an optimization problem by minimizing the KL-divergences:

$$\begin{aligned} &\min \mathbf{E}_{p_{\mathcal{D}}(x)}[\mathcal{KL}(q(z|x)||p(z|x))] + \mathcal{KL}(p_{\mathcal{D}}||p) \\ &= \max \mathbf{E}_{x\sim p_{\mathcal{D}}(x)}[\mathcal{L}(\theta, \phi, \pi; x)] \end{aligned} \tag{3}$$

==> maximizing the vlb brings us to the limit conditions under which marginal approximate posterior $q(z)$ should match the prior given enough flexibility in the assumed form of densities.

Now if we maximize (3) wrt π while holding θ and ϕ fixed like doing coordinate ascent, the samples $x \sim p_{\mathcal{D}}(x), z \sim q(z|x)$ can be thought of as a projected data distribution that we want to model using the prior distribution:

$$\max_\pi \mathbf{E}[\mathcal{L}] = \min_\pi \mathcal{KL}(\mathbf{E}_{x\sim p_{\mathcal{D}}(x)}[q(z|x)]||p_\pi(z)) \tag{4}$$

As a result, having a learnable prior allows us to sample from the marginal approximate posterior if the above divergence metric goes to zero.

Another advantage of a learnable prior can be visualized by the cartoon plot in Figure 1. When we fix the approximate posterior and update the prior such that it becomes closer to the marginal approximate posterior, it concentrates the probability mass in such a way that the true posterior becomes closer to the approximate posterior, as $p(z|x) \propto p(z)$. In other words, the region of high posterior density not covered by the approximate posterior will be reduced, which effectively means our proposal as variational distribution could be improved by having a better prior which simplifies the true posterior.

3. Inverse Autoregressive Flows as Universal Posterior Approximator

In Kingma et al. (2016), a powerful family of invertible functions called the Inverse Autoregressive Flows (inverse AF or IAF) were introduced, to improve variational inference. It is thus of practical and fundamental importance to understand the benefits of using inverse AF and how to improve them.

In this section, we show that normalizing flows from a base distribution (such as uniform distribution) under autoregressive assumptions are universal approximators of any density (as suggested in Goodfellow (2017)), given enough capacity when a neural network is used to parameterize non-linear dependencies.

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

Proof. See Hyvarjnen & Pajunen (1998) for the full proof. Here we point out that the transformation g used in the proof falls within the family of autoregressive functions: $f = (f_i)_{i=1\dots m}$ where $y_i = f_i(x_i, y_1, \dots, y_{i-1}) = P(x \leq x_i \mid y_1, ..., y_{i-1})$, for $i = 1 \dots m$. f_i is the conditional CDF and $Y \sim U([0,1]^m)$. Then any distribution of a rv x can be warped into an independent distribution via the CDFs, specifically by a kind of Gram-Schmidt process-like construction. □

Proposition 1. ***Inverse autoregressive transformation as universal approximator of any density.*** *Let X be a rv in an open set $\mathcal{U} \subset \mathbf{R}^m$. assume that*

X has a positive and continuous pdf. 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 $Y \sim U((0,1)^m)$ converges in distribution to X.

Proof. We consider the mapping f defined in the proof of Lemma 1. As f is autoregressive, the Jacobian of f is an upper triangular matrix whose diagonal entries are equal to the conditional densities which are positive by assumption. The determinant of the Jacobian, which is equal to the product of diagonal entries, is positive. By the *inverse function theorem*, f is locally invertible. As f is also injective (as follows from the bijectivity of CDF), f is globally invertible and let g denotes its inverse. g is an autoregressive function and by the *universal approximation theorem* (Cybenkot, 1989), we know that there exists a sequence of mappings $(G_n)_{n\geq 0}$ from $(0,1)^m$ to $\mathbf{R}^m$ parametrized by autoregressive neural networks that converge uniformly to g. Let $X_n = G_n(Y)$ where $Y \sim U((0,1)^m)$. Let h be a real-valued bounded continuous function on $\mathbf{R}^m$. The latter uniform convergence implies that since G_n converge pointwise to g, then by continuity of h, $h \circ G_n$ converges pointwise to $h \circ g$. As h is bounded, the *dominated convergence theorem* gives that $\mathbf{E}[h(X_n)] = \mathbf{E}[h(G_n(Y))]$ converges to $\mathbf{E}[h(g(Y))] = \mathbf{E}[h(X)]$. As the latter statement is valid for all bounded continuous function h, X_n converge to X in distribution. □

Note that G is usually parameterized as an invertible function, at the expense of flexibility, to have a tractable Jacobian. Special designs of such a function, other than affine transformation (Kingma et al., 2016), could be made to improve the flow; otherwise one would need to compose multiple layers of transformations to have a richer distribution family. Our proof shows that, with careful designs of approximate posteriors, VAEs could have asymptotic consistency.

4. Proposed Method

As suggested in sections 2 and 3, we propose to use one-to-one correspondence to define a learnable explicit density (LED) model for both inference and sample generation. First, inspired by (4), we found that updating the prior alone is reminiscent of MLE. One can think of data points projected onto the latent space via Monte Carlo sampling as a data distribution $q_{\mathcal{D}}(z) = \mathbf{E}_{p_{\mathcal{D}}(x)}[q(z|x)]$ in space z. A unimodal prior tends to overestimate the entropy of $q_{\mathcal{D}}(z)$. A powerful family of real non-volume preserving (Real NVP) transformations (Dinh et al., 2016) can be applied to real variables. It is thus natural to incorporate Real NVP into VAEs to jointly train an explicit density model as prior. We define the prior (and also the approximate posterior) with change of variable formula: $p(z) = p(z_0)|\frac{\partial h}{\partial z_0}(z_0)|^{-1}$ where $h : z_0 \to z$. To compute the density of the projected data distribution, we inversely (h^{-1}) transform the samples $z \sim q_{\mathcal{D}}(z)$ into the base variable z_0 with tractable density (Dinh et al., 2014). We define the posterior likewise, as in Rezende & Mohamed (2015), with $g : z' \to z$. Objective (1) ==>

$$\begin{aligned}\mathcal{L} = &\mathbf{E}_{q(z'|x)}[\log p(x|g(z'))] + \\ &\mathbf{E}_{q(z'|x)}\left[\log p(h^{-1} \circ g(z')) + \log\left|\frac{\partial h^{-1}}{\partial z}(g(z'))\right|\right] - \\ &\mathbf{E}_{q(z'|x)}\left[\log q(z'|x) - \log\left|\frac{\partial g}{\partial z'}(z')\right|\right]\end{aligned} \tag{5}$$

For permutation invariant latent variables, h is implemented with random masks. For latent variables that preserve the spatial correlation when a convolutional network is used, we choose to use a checkerboard style mask (Dinh

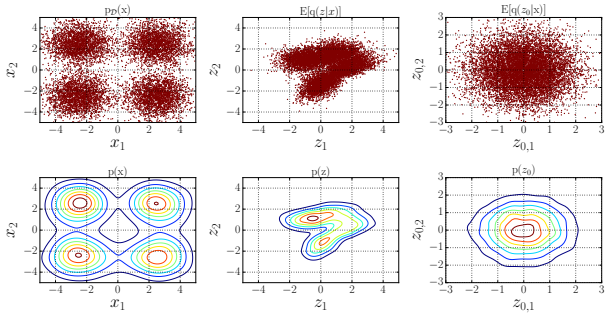

Figure 2. Fitting a Gaussian mixture distribution. $E[\cdot]$ indicates marginalization over the data $x \sim p_{\mathcal{D}}(x)$. Clockwise from top left: projection of data distribution $p_{\mathcal{D}}(x)$ onto the prior space $\mathbf{E}(q(z|x))$, and the base distribution space $\mathbf{E}(q(z_0|x))$;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

Table 1. Effect of increasing prior complexity. L_{post}: number of MADE layers used for posterior. Two hidden layers of 512 nodes were used for each layer of transformation. L_{prior}: number of NVP layers used for prior. One hidden layer of 100 nodes was used for each layer of transformation. For multi-layer perceptron, two hidden layers with 200 nodes were used and the dimension of the latent variable is 50. Rectifier is used as non-linear activation. For Residual ConvNet, we have 3 layers of residual strided convolution (He et al., 2015) with [16,32,32] feature maps, using filter of size 3×3. Before the stochastic layer a hidden layer of 450 nodes is used. The dimension of the latent variable is 32. We use exponential linear units (Clevert et al., 2015) as non-linearity.

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

et al., 2016; Agrawal & Dukkipati, 2016). Interestingly, sampling of such models is similar to block Gibbs sampling for energy based models (e.g. Ising models) that define the correlation between adjacent pixels.

Second, for the posterior distribution, we construct g by inverse AF, which is parallelizable when combined with MADE (Germain et al., 2015) or PixelCNN (Van Den Oord et al., 2016). In fact, inverse AF can be thought of as a generalization of Real NVP, as the Jacobian of the masked operation used in Real NVP is upper triangular.

5. Experiments

Mixture of Bivariate Gaussians. We experiment on a Gaussian mixture toy example, and visualize the effect of having a learnable prior in Figure 2. During training, we observe that models with flexible prior are easier to train than models with flexible posterior. Our first conjecture is that to refine the posterior density, we only draw one sample of z for each data point x, whereas refining the prior density can be viewed as modeling the projected data distribution and thus depends on as many samples as there are in the training set. Second, it might be due to the kind of transformation and the distance metrics that are used. To learn the posterior, we implicitly minimize $\mathcal{KL}(q(z|x)||p(z|x))$, which is zero forcing since samples in region that has low target density are heavily penalized. If q begins with a sharper shape, it pays a high penalty by expansion to move to another mode. It is thus easy for the distribution to be stuck in local minima if the true posterior is multimodal, while learning the prior does not have this mode seeking problem since the forward KL in (4) is zero avoiding.

Table 2. Effect of increasing both prior and posterior complexity.

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

MNIST. We also tested our proposed method on binarized MNIST (Larochelle & Murray, 2011), and report the estimated negative log likelihood as an evaluation metric.

We compare the effects of adding more invertible transformation layers on either the prior or posterior (see Table 1), or both (Table 2). From Table 1, we see that models having a flexible prior easily outperform models with a flexible posterior. Likelihood of a model with flexible prior can be further improved by using expressive posterior (Table 2) such as real NVP (81.70 → 81.55), or with MADE to introduce more autoregressive dependencies (81.55 → 80.81).

6. Discussion and Future Work

In this paper, we first reinterpret training with the variational lower bound as layer-wise density estimation. Treating the Monte Carlo samples from the approximate posterior distributions as projected data distribution suggests using a flexible prior to avoid overestimate of entropy. We leave experiments on larger datasets and sample generation as future work. Second, we showed that parameterizing inverse AF using neural networks allows us to universally approximate any posterior, which theoretically justifies the use of inverse AF. Our proof also implies using affine coupling law to autoregressively warp the distribution is limited. It is thus possible to consider designs of more flexible invertible functions to improve approximate posterior.