

# VAE Review Notes

Sina Hajimiri  
SUT, CE-40719

Spring 2020

## 1 Introduction

Variational Autoencoder is a deep generative model, working explicitly with the density function.

Suppose we have a distribution on  $X$  named  $p_D(x)$ . We want to model this distribution with a neural network ( $\theta$ ), to be able to take new samples from it.

What is an ordinary way to train the network and learn  $\theta$ ?

A natural training objective for learning  $\theta$  is maximum likelihood:

$$\max_{\theta} \mathbb{E}_{p_D(x)} [\log p_{\theta}(x)] \quad (1)$$

Assuming that every data point is generated from an underlying latent representation  $z$ , we can write  $p_{\theta}(x)$  as  $p_{\theta}(x, z)$  marginalized on  $z$ . We can rewrite equation 1 as follows:

$$\mathbb{E}_{p_D(x)} [\log p_{\theta}(x)] = \mathbb{E}_{p_D(x)} [\log \int_z p_{\theta}(x|z)p(z)dz] \quad (2)$$

This equation is intractable and can not be optimized efficiently.

## 2 How to overcome intractability?

Until now we have a network ( $\theta$ ) that maps from  $z$  to  $x$ . This network provides us with a joint distribution on  $X$  and  $Z$  (Which is called a generative distribution):

$$p_{\theta}(x, z) = p_{\theta}(x|z)p(z) \quad (3)$$

We create another network ( $\phi$ ) to map from  $x$  to  $z$ . This network too, models a joint distribution on  $X$  and  $Z$  (Which is called an inference distribution):

$$q_{\phi}(x, z) = q_{\phi}(z|x)p_D(x) \quad (4)$$

The auxiliary  $q_{\phi}(z|x)$  distribution can help us overcome intractability.

We can now rewrite the intractable  $\log \int_z p_{\theta}(x|z)p(z)dz$  in equation 2 as

$$\log \int_z p_\theta(x|z)p(z)dz = \log \int_z \frac{q_\phi(z|x)}{q_\phi(z|x)} p_\theta(x|z)p(z)dz \quad (5)$$

$$= \log \int_z q_\phi(z|x) \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)} dz \quad (6)$$

$$= \log \mathbb{E}_{q_\phi(z|x)} \left[ \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)} \right] \quad (7)$$

$$\geq \mathbb{E}_{q_\phi(z|x)} \left[ \log \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)} \right] \quad (8)$$

$$= \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)] + \mathbb{E}_{q_\phi(z|x)} \left[ \log \frac{p(z)}{q_\phi(z|x)} \right] \quad (9)$$

$$= \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)] - \text{KL}(q_\phi(z|x) || p(z)) \quad (10)$$

in which equation 8 is based on [Jensen's inequality](#).

### 3 VAE Objective Function

We define the objective function for a specific  $x$  as

$$\begin{aligned} \mathcal{L}_{\text{ELBO}}(x) &= \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)] - \text{KL}(q_\phi(z|x) || p(z)) \\ &\leq \log \int_z p_\theta(x|z)p(z)dz = \log p_\theta(x) \end{aligned} \quad (11)$$

which, as shown in the equation, is a lower bound on log of likelihood. We also define the final objective function of the VAE as:

$$\max_{\phi, \theta} \mathcal{L}_{\text{ELBO}} = \mathbb{E}_{p_D(x)} \left[ \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)] - \text{KL}(q_\phi(z|x) || p(z)) \right] \quad (12)$$

### 4 How can we optimize $\mathcal{L}_{\text{ELBO}}$ ?

We have 3 different kinds of elements in equation 12, which we rewrite in turns.

- Expectations: We know that we can estimate an expectation  $\mathbb{E}_{f(u)}[g(u)]$  as  $\frac{1}{N} \sum_{n=1}^N g(u_n)$  which  $u_1, \dots, u_N$  are  $N$  samples taken from the  $f(u)$  distribution. Since we can take samples from both  $p_D(x)$  and  $q_\phi(z|x)$ , we can compute both of the expectations of equation 12.
- $\log p_\theta(x|z)$ : We can compute  $\log p_\theta(x|z)$  analytically since the distribution is set to be either multivariate Bernoulli or multivariate Gaussian.
  - Multivariate Bernoulli case: The output of the decoder (after a sig-

moid layer) is denoted as  $a$ . We have:

$$\begin{aligned}
p_\theta(x|z) &= \text{Bern}(x; a) = \prod_{k=1}^d a_k^{x_k} (1 - a_k)^{(1-x_k)} \\
\Rightarrow \log p_\theta(x|z) &= \sum_{k=1}^d x_k \log(a_k) + (1 - x_k) \log(1 - a_k) \\
&= -\text{BCE}(x, a)
\end{aligned} \tag{13}$$

- Multivariate Gaussian case: The output of the decoder is denoted as  $\mu$ . Also,  $\sigma^2$  is used in equations but in practice, it is usually set to 1. We have:

$$\begin{aligned}
p_\theta(x|z) &= \mathcal{N}(x; \mu, \sigma^2 I) \\
&= (2\pi)^{-\frac{d}{2}} \det(\sigma^2 I)^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^\top (\sigma^2 I)^{-1} (x-\mu)} \\
&= (2\pi\sigma^2)^{-\frac{d}{2}} e^{-\frac{1}{2\sigma^2} \|x-\mu\|^2} \\
\Rightarrow \log p_\theta(x|z) &= -\frac{d}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|x-\mu\|^2 \\
&= -\frac{1}{2\sigma^2} \|x-\mu\|^2 + \text{const} \\
&\equiv -\frac{1}{2\sigma^2} \text{MSE}(x, \mu)
\end{aligned} \tag{14}$$

- $\text{KL}(q_\phi(z|x) \parallel p(z))$ : This term can be computed analytically too, since both distributions are set to be Gaussian. We know that for two multivariate Gaussian distributions as

$$\begin{aligned}
q(z) &= \mathcal{N}(z; \mu_1, \Sigma_1) \\
p(z) &= \mathcal{N}(z; \mu_2, \Sigma_2)
\end{aligned} \tag{15}$$

their KL divergence can be computed as following:

$$\begin{aligned}
\text{KL}(q(z) \parallel p(z)) &= \frac{1}{2} \left[ \log \frac{|\Sigma_2|}{|\Sigma_1|} - d + \text{tr}\{\Sigma_2^{-1} \Sigma_1\} \right. \\
&\quad \left. + (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) \right]
\end{aligned} \tag{16}$$

By setting

$$\begin{aligned}
p(z) &= \mathcal{N}(z; 0, I) \\
q(z) &= q_\phi(z|x) = \mathcal{N}(z; \mu, \Sigma)
\end{aligned} \tag{17}$$

where  $\mu = (\mu_1, \dots, \mu_d)$  and  $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$ , the KL term of equation

12 can be computed as following:

$$\begin{aligned}
\text{KL}(q_\phi(z|x) \parallel p(z)) &= \frac{1}{2} \left[ \log\left(\frac{|I|}{|\Sigma|}\right) - d + \text{tr}\{I^{-1}\Sigma\} \right. \\
&\quad \left. + (0 - \mu)^T I^{-1} (0 - \mu) \right] \\
&= \frac{1}{2} \left[ -\log(|\Sigma|) - d + \text{tr}\{\Sigma\} + \mu^T \mu \right] \quad (18) \\
&= \frac{1}{2} \left[ -\log\left(\prod_i \sigma_i^2\right) - d + \sum_i \sigma_i^2 + \sum_i \mu_i^2 \right] \\
&= \frac{1}{2} \left[ -\sum_i \log(\sigma_i^2) - d + \sum_i \sigma_i^2 + \sum_i \mu_i^2 \right]
\end{aligned}$$

So, all of the terms in equation 12 can be computed efficiently, and we can optimize  $\phi$  and  $\theta$  to maximize the equation.

## 5 Equivalent Forms of the ELBO

$\mathcal{L}_{\text{ELBO}}$  has different forms, which may not be directly optimizable, but are useful in theoretical analyses. You may see these alternative forms in papers extending VAE framework.

$$\mathcal{L}_{\text{ELBO}} \equiv -\text{KL}(q_\phi(x, z) \parallel p_\theta(x, z)) \quad (19)$$

$$= -\text{KL}(p_D(x) \parallel p_\theta(x)) - \mathbb{E}_{p_D(x)}[\text{KL}(q_\phi(z|x) \parallel p_\theta(z|x))] \quad (20)$$

$$= -\text{KL}(q_\phi(z) \parallel p(z)) - \mathbb{E}_{q_\phi(z)}[\text{KL}(q_\phi(x|z) \parallel p_\theta(x|z))] \quad (21)$$

Proof of equation 19:

$$\begin{aligned}
\mathcal{L}_{\text{ELBO}} &= \mathbb{E}_{p_D(x)} \left[ \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)] - \text{KL}(q_\phi(z|x) \parallel p(z)) \right] \\
&= \mathbb{E}_{p_D(x)} \left[ \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)] - \mathbb{E}_{q_\phi(z|x)} \left[ \frac{q_\phi(z|x)}{p(z)} \right] \right] \\
&= \mathbb{E}_{q_\phi(x, z)} [\log p_\theta(x|z) + \log p(z) - \log q_\phi(z|x)] \\
&= \mathbb{E}_{q_\phi(x, z)} [\log p_\theta(x, z) - \log q_\phi(z|x) - \log p_D(x) + \log p_D(x)] \quad (22) \\
&= \mathbb{E}_{q_\phi(x, z)} [\log p_\theta(x, z) - \log q_\phi(x, z) + \log p_D(x)] \\
&= \mathbb{E}_{q_\phi(x, z)} \left[ \log \frac{p_\theta(x, z)}{q_\phi(x, z)} + \log p_D(x) \right] \\
&= -\text{KL}(q_\phi(x, z) \parallel p_\theta(x, z)) + \mathbb{E}_{p_D(x)} [\log p_D(x)]
\end{aligned}$$

Since  $\mathbb{E}_{p_D(x)} [\log p_D(x)]$  is a constant (entropy of a fixed distribution), optimizing  $-\text{KL}(q_\phi(x, z) \parallel p_\theta(x, z))$  is equivalent to optimizing  $\mathcal{L}_{\text{ELBO}}$ .

Proofs of the equivalence of optimizing equations 20 and 21 to equation 12 are left for exercise.

## References

- [1] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *CoRR*, vol. abs/1312.6114, 2014.
- [2] S. Zhao, J. Song, and S. Ermon, “Infovae: Balancing learning and inference in variational autoencoders,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 5885–5892, 2019.
- [3] M. Soleymani, “Deep learning course slides.” University Lecture, May 2020. Computer Engineering Department of Sharif University of Technology.