arXiv:1907.08956v1 [cs.LG] 21 Jul 2019

Tutorial: Deriving the Standard Variational Autoencoder (VAE) Loss Function

Stephen G. Odaibo*
(1) Department of Machine Learning Research
RETINA-AI Health, Inc.
(2) Department of Head & Neck Surgery
Ophthalmology Section
MD Anderson Cancer Center
stephen.odaibo@retina-ai.com

Abstract

In Bayesian machine learning, the posterior distribution is typically computationally intractable, hence variational inference is often required. In this approach, an evidence lower bound on the log likelihood of data is maximized during training. Variational Autoencoders (VAE) are one important example where variational inference is utilized. In this tutorial, we derive the variational lower bound loss function of the standard variational autoencoder. We do so in the instance of a gaussian latent prior and gaussian approximate posterior, under which assumptions the Kullback-Leibler term in the variational lower bound has a closed form solution. We derive essentially everything we use along the way; everything from Bayes' theorem to the Kullback-Leibler divergence.

Bayes Theorem

Bayes theorem is a way to update one's belief as new evidence comes into view. The probability of a hypothesis, z, given some new data x, is denoted, $p(z|x)$, and is given by

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}, \tag{1}$$

where $p(x)$ is the probability of the data x, $p(x|z)$ is the probability of the data given a hypothesis z, and $p(z)$ is the probability of that hypothesis z. While Bayes theorem by itself can appear non-intuitive or at least difficult to intuit, the key to understanding it is to derive it. It arises directly out of the conditional probability axiom, which itself arises out of the definition of the joint probability. The probability of an event X and an event Y occurring jointly is,

$$p(X \cap Y) = p(X|Y)p(Y) \tag{2}$$

And since the 'AND' is commutative, we have,

$$p(X \cap Y) = p(Y \cap X) = p(Y|X)p(X) \tag{3}$$

$$p(X|Y)p(Y) = p(Y|X)p(X) \tag{4}$$

*Correspondence: stephen.odaibo@retina-ai.com

Table 1: **Bayesian Statistics Glossary**

Symbol	Name
z	Latent variable
x	Evidence or Data
$p(x)$	Evidence probability
$p(z)$	Prior probability
$p(z\|x)$	Posterior probability
$p(x\|z)$	Likelihood probability

Dividing both sides of Equation (4) by $p(Y)$ yields Bayes theorem,

$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)} \tag{5}$$

Kullback-Leibler Divergence

When comparing two distributions as we often do in density estimation, the central task of generative models, we need a measure of similarity between both distributions. The Kullback-Leibler divergence is a commonly used similarity measure for this purpose. It is the expectation of the information difference between both distributions. But first, what is information?

To understand what information is and to see its definition, consider the following: The higher the probability of an event, the lower its information content. This makes intuitive sense in that if someone tells us something 'obvious' i.e. highly probable i.e. something we and almost everyone else already knew, then that informant has not increased the amount of information we have. Hence the information content of highly probably event is low. Another way to say this is that the information is inversely related to the probability of an event. And since $\log(p(x)$ is directly related to $p(x)$, it follows that $-\log(p(x))$ is inversely related to $p(x)$, and is how we model information:

$$\text{Information content of event x wrt p} = I_p(x) = -\log p(x) \tag{6}$$

$$\text{Information content of event x wrt q} = I_q(x) = -\log q(x) \tag{7}$$

The difference of information between $q(x)$ and $p(x)$ is therefore:

$$\Delta I = I_p - I_q = -\log p(x) + \log q(x) = \log\left(\frac{q(x)}{p(x)}\right) \tag{8}$$

And the Kullback-Leibler is the expectation of the above difference, and is given by,

$$D_{KL}(q(x)||p(x)) := E_{\sim q}[\Delta I] = \int (\Delta I) q(x) dx = \int q(x) \log\left(\frac{q(x)}{p(x)}\right) dx \tag{9}$$

Similarly

$$D_{KL}(p(x)||q(x)) := E_{\sim p}[\Delta I] = \int (\Delta I) p(x) dx = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \tag{10}$$

Note that the Kullback-Leibler (KL) is not symmetric, i.e,

$$D_{KL}(q(x)||p(x)) \neq D_{KL}(p(x)||q(x)) \tag{11}$$

In $D_{KL}(q(x)||p(x))$, we are taking the expectation of the information difference with respect to $q(x)$ distribution, while in $D_{KL}(p(x)||q(x))$, we are taking the expectation with respect to the $p(x)$ distribution.

Hence the Kullback-Leibler is called a 'divergence' and not a 'metric' as metrics must be symmetric. There recently have been a number of symmetrization devices proposed for KL which have been shown to improve its generative fidelity [Pu et al. (2017)][Chen et al. (2017)] [Arjovsky et al. (2017)].

Note the KL divergence is always non-negative, i.e.,

$$D_{KL}(q(x)||p(x)) = -\int q(x) \log\left(\frac{p(x)}{q(x)}\right) dx \geq 0 \tag{12}$$

To see this, note that as depicted in Figure (1),

$$\log t \leq t - 1 \tag{13}$$

Therefore

$$\begin{aligned} -D_{KL}(q(x)||p(x)) = \int q(x) \log\left(\frac{p(x)}{q(x)}\right) dx \leq \\ \int q(x) \left(\frac{p(x)}{q(x)} - 1\right) dx = \\ \int q(x) \frac{p(x)}{q(x)} dx - \int q(x) dx = \\ \int p(x) dx - \int q(x) dx = \\ 1 - 1 = 0 \end{aligned} \tag{14}$$

We have just shown,

$$-D_{KL}(q(x)||p(x)) \leq 0 \tag{15}$$

which implies,

$$D_{KL}(q(x)||p(x)) \geq 0 \tag{16}$$

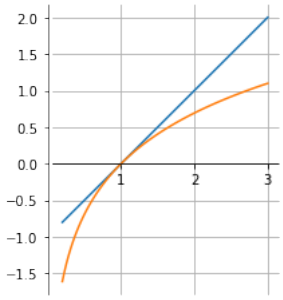

Figure 1: $\log(t) \leq t - 1$

VAE Objective

Consider variational autoencoders [Kingma et al. (2013)]. They have many applications including for finer characterization of disease [Odaibo (2019)]. The encoder portion of a VAE yields an approximate posterior distribution $q(z|x)$, and is parametrized on a neural network by weights collectively denoted θ. Hence we more properly write the encoder as $q_\theta(z|x)$. Similarly, the decoder portion of the

VAE yields a likelihood distribution $p(x|z)$, and is parametrized on a neural network by weights collectively denoted ϕ. Hence we more properly denote the decoder portion of the VAE as $p_\phi(x|z)$. The output of the encoder are parameters of the latent distribution, which is sampled to yield the input into the decoder. A VAE schematic is shown in Figure (2).

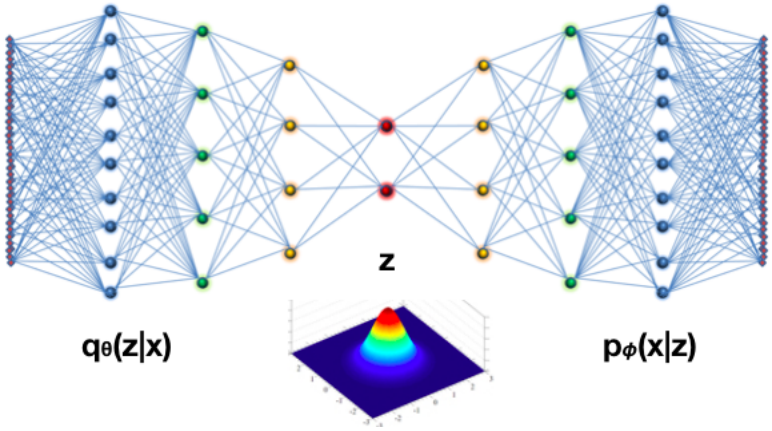

Figure 2: VAE

The KL divergence between the approximate and the real posterior distributions is given by,

$$D_{KL}\left(q_\theta(z|x_i)||p(z|x_i)\right) = -\int q_\theta(z|x_i)\log\left(\frac{p(z|x_i)}{q_\theta(z|x_i)}\right)dz \geq 0 \tag{17}$$

Applying Bayes' theorem to the above equation yields,

$$D_{KL}\left(q_\theta(z|x_i)||p(z|x_i)\right) = -\int q_\theta(z|x_i)\log\left(\frac{p_\phi(x_i|z)p(z)}{q_\theta(z|x_i)p(x_i)}\right)dz \geq 0 \tag{18}$$

This can be broken down using laws of logarithms, yielding,

$$D_{KL}\left(q_\theta(z|x_i)||p(z|x_i)\right) = -\int q_\theta(z|x_i)\left[\log\left(\frac{p_\phi(x_i|z)p(z)}{q_\theta(z|x_i)}\right) - \log p(x_i)\right]dz \geq 0 \tag{19}$$

Distributing the integrand then yields,

$$-\int q_\theta(z|x_i)\log\left(\frac{p_\phi(x_i|z)p(z)}{q_\theta(z|x_i)}\right)dz + \int q_\theta(z|x_i)\log p(x_i)dz \geq 0 \tag{20}$$

In the above, we note that $\log(p(x_i))$ is a constant and can therefore be pulled out of the second integral above, yielding,

$$-\int q_\theta(z|x_i)\log\left(\frac{p_\phi(x_i|z)p(z)}{q_\theta(z|x_i)}\right)dz + \log p(x_i)\int q_\theta(z|x_i)dz \geq 0 \tag{21}$$

And since $q_\theta(z|x_i)$ is a probability distribution it integrates to 1 in the above equation, yielding,

$$-\int q_\theta(z|x_i)\log\left(\frac{p_\phi(x_i|z)p(z)}{q_\theta(z|x_i)}\right)dz + \log p(x_i) \geq 0. \tag{22}$$

Then carrying the integral over to the other side of the inequality, we get,

$$\log p(x_i) \geq \int q_\theta(z|x_i)\log\left(\frac{p_\phi(x_i|z)p(z)}{q_\theta(z|x_i)}\right)dz. \tag{23}$$

Applying rules of logarithms, we get,

$$\log p(x_i) \geq \int q_\theta(z|x_i) \Big[\log p_\phi(x_i|z) + \log p(z) - \log q_\theta(z|x_i) \Big] dz. \tag{24}$$

Recognizing the right hand side of the above inequality as Expectation, we write,

$$\log p(x_i) \geq E_{\sim q_\theta(z|x_i)} \Big[\log p_\phi(x_i|z) + \log p(z) - \log q_\theta(z|x_i) \Big] \tag{25}$$

$$\log p(x_i) \geq E_{\sim q_\theta(z|x_i)} \Big[\log p(x_i, z) - \log q_\theta(z|x_i) \Big] \tag{26}$$

From Equation (23) it also follows that:

$$\log p(x_i) \geq \int q_\theta(z|x_i) \log \left(\frac{p(z)}{q_\theta(z|x_i)} \right) dz + \int q_\theta(z|x_i) \log p_\phi(x_i|z) dz \tag{27}$$

$$\log p(x_i) \geq -D_{KL}(q_\theta(z|x_i)||p(z)) + E_{\sim q_\theta(z|x_i)}[\log p_\phi(x_i|z)] \tag{28}$$

The right hand side of the above equation is the Evidence Lower Bound (ELBO) also known as the variational lower bound. It is so termed because it bounds the likelihood of the data which is the term we seek to maximize. Therefore maximizing the ELBO maximizes the log probability of our data by proxy. This is the core idea of variational inference, since maximization of the log probability directly is typically computationally intractable. The Kullback-Leibler term in the ELBO is a regularizer because it is a constraint on the form of the approximate posterior. The second term is called a reconstruction term because it is a measure of the likelihood of the reconstructed data output at the decoder.

Notably, we have some liberty to choose some structure for our latent variables. We can obtain a closed form for the loss function if we choose a gaussian representation for the latent prior $p(z)$ and the approximate posterior, $q_\theta(z|x_i)$. In addition to yielding a closed form loss function, the gaussian model enforces a form of regularization in which the approximate posterior have variation or spread (like a gaussian).

Closed form VAE Loss: Gaussian Latents

Say we choose:

$$p(z) \rightarrow \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(-\frac{(x-\mu_p)^2}{2\sigma_p^2} \right) \tag{29}$$

and

$$q_\theta(z|x_i) \rightarrow \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2} \right) \tag{30}$$

,

then the KL or regularization term in the ELBO becomes:

$$-D_{KL}(q_\theta(z|x_i)||p(z)) =$$

$$\int \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2} \right) \log \left(\frac{\frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(-\frac{(x-\mu_p)^2}{2\sigma_p^2} \right)}{\frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2} \right)} \right) dz \tag{31}$$

Evaluating the term in the logarithm simplifies the above into,

$$\int \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2}\right) \times \tag{32}$$
$$\left\{-\frac{1}{2}\log(2\pi) - \log(\sigma_p) - \frac{(x-\mu_p)^2}{2\sigma_p^2} + \frac{1}{2}\log(2\pi) + \log(\sigma_q) + \frac{(x-\mu_q)^2}{2\sigma_q^2}\right\} dz.$$

This further simplifies into,

$$\frac{1}{\sqrt{2\pi\sigma_q^2}} \int \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2}\right) \left\{-\log(\sigma_p) - \frac{(x-\mu_p)^2}{2\sigma_p^2} + \log(\sigma_q) + \frac{(x-\mu_q)^2}{2\sigma_q^2}\right\} dz, \tag{33}$$

which further simplifies into,

$$\frac{1}{\sqrt{2\pi\sigma_q^2}} \int \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2}\right) \left\{\log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{(x-\mu_p)^2}{2\sigma_p^2} + \frac{(x-\mu_q)^2}{2\sigma_q^2}\right\} dz. \tag{34}$$

Expressing the above as an Expectation we get,

$$\begin{aligned} -D_{KL}(q_\theta(z|x_i)||p(z)) &= E_q\left\{\log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{(x-\mu_p)^2}{2\sigma_p^2} + \frac{(x-\mu_q)^2}{2\sigma_q^2}\right\} \\ &= \log\left(\frac{\sigma_q}{\sigma_p}\right) + E_q\left\{-\frac{(x-\mu_p)^2}{2\sigma_p^2} + \frac{(x-\mu_q)^2}{2\sigma_q^2}\right\} \\ &= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}E_q\left\{(x-\mu_p)^2\right\} + \frac{1}{2\sigma_q^2}E_q\left\{(x-\mu_q)^2\right\} \end{aligned} \tag{35}$$

And since the variance σ^2 is the expectation of the squared distance from the mean, i.e.,

$$\sigma_q^2 = E_q\left\{(x-\mu_q)^2\right\}, \tag{36}$$

it follows that,

$$\begin{aligned} -D_{KL}(q_\theta(z|x_i)||p(z)) &= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}E_q\left\{(x-\mu_p)^2\right\} + \frac{\sigma_q^2}{2\sigma_q^2} \\ &= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}E_q\left\{(x-\mu_p)^2\right\} + \frac{1}{2} \\ &= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}E_q\left\{(x-\mu_q+\mu_q-\mu_p)^2\right\} + \frac{1}{2} \\ &= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}E_q\left\{(\underbrace{x-\mu_q}_{a} + \underbrace{\mu_q-\mu_p}_{b})^2\right\} + \frac{1}{2} \end{aligned} \tag{37}$$

Recall that,

$$(a+b)^2 = a^2 + 2ab + b^2, \tag{38}$$

therefore,

$$-D_{KL}(q_\theta(z|x_i)||p(z)) = \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}E_q\left\{(\underbrace{x-\mu_q}_{a}+\underbrace{\mu_q-\mu_p}_{b})^2\right\} + \frac{1}{2} \tag{39}$$

$$= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}E_q\left\{(x-\mu_q)^2 + 2(x-\mu_q)(\mu_q-\mu_p) + (\mu_q-\mu_p)^2\right\} + \frac{1}{2}$$

$$= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}E_q\left\{(x-\mu_q)^2 + 2(x-\mu_q)(\mu_q-\mu_p) + (\mu_q-\mu_p)^2\right\} + \frac{1}{2}$$

$$= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}\left[E_q\left\{(x-\mu_q)^2\right\} + 2E_q\left\{(x-\mu_q)(\mu_q-\mu_p)\right\} + E_q\left\{(\mu_q-\mu_p)^2\right\}\right] + \frac{1}{2}$$

$$= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}\left[\sigma_q^2 + 2*0*(\mu_q-\mu_p) + (\mu_q-\mu_p)^2\right] + \frac{1}{2}$$

$$= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{\sigma_q^2+(\mu_q-\mu_p)^2}{2\sigma_p^2} + \frac{1}{2}$$

And when we take $\sigma_p = 1$ and $\mu_p = 0$, we get,

$$-D_{KL}(q_\theta(z|x_i)||p(z)) = \log(\sigma_q) - \frac{\sigma_q^2+\mu_q^2}{2} + \frac{1}{2} \tag{40}$$

$$= \frac{1}{2}\log\left(\sigma_q^2\right) - \frac{\sigma_q^2+\mu_q^2}{2} + \frac{1}{2}$$

$$= \frac{1}{2}\left[1 + \log\left(\sigma_q^2\right) - \sigma_q^2 - \mu_q^2\right]$$

Recall the ELBO, Equation (28),

$$\log p(x_i) \geq -D_{KL}(q_\theta(z|x_i)||p(z)) + E_{\sim q_\theta(z|x_i)}\Big[\log p_\phi(x_i|z)\Big]$$

From which it follows that the contribution from a given datum x_i and a single stochastic draw towards the objective to be *maximized* is,

$$\frac{1}{2}\left[1 + \log\left(\sigma_j^2\right) - \sigma_j^2 - \mu_j^2\right] + E_{\sim q_\theta(z|x_i)}\Big[\log p_\phi(x_i|z)\Big] \tag{41}$$

where σ_j^2 and μ_j are parameters into the approximate distribution, q, and j is an index into the latent vector z. For a batch, the objective function is therefore given by,

$$\mathcal{G} = \sum_{j=1}^{J}\frac{1}{2}\left[1 + \log\left(\sigma_i^2\right) - \sigma_i^2 - \mu_i^2\right] + \frac{1}{L}\sum_{l}E_{\sim q_\theta(z|x_i)}\Big[\log p(x_i|z^{(i,l)})\Big] \tag{42}$$

where J is the dimension of the latent vector z, and L is the number of samples stochastically drawn according to re-parametrization trick.

Because the objective function we obtain in Equation (42) is to be maximized during training, we can think of it as a 'gain' function as opposed to a loss function. To obtain the loss function, we simply take the negative of $\mathcal{G}$:

$$\mathcal{L} = -\sum_{j=1}^{J} \frac{1}{2}\left[1 + \log\left(\sigma_i^2\right) - \sigma_i^2 - \mu_i^2\right] - \frac{1}{L}\sum_{l} E_{\sim q_\theta(z|x_i)}\Big[\log p(x_i|z^{(i,l)})\Big] \tag{43}$$

Therefore to train the VAE is to seek the optimal network parameters (θ^*, ϕ^*) that minimize $\mathcal{L}$:

$$(\theta^*, \phi^*) = argmin_{(\theta,\phi)}\mathcal{L}(\theta, \phi) \tag{44}$$

Conclusion

We have done a step-by-step derivation of the VAE loss function. We illustrated the essence of variational inference along the way, and have derived the closed form loss in the special case of gaussian latent.

Acknowledgement

The author thanks Larry Carin for helpful discussion on consequences of Kullback-Leibler divergence asymmetry, and on KL symmetrization approach.

References

Odaibo SG. retina-VAE: Variationally Decoding the Spectrum of Macular Disease. arXiv:1907.05195. 2019 Jul 11.

Kingma DP, Welling M. Autoencoding Variational Bayes. arXiv preprint arXiv:1312.6114. 2013 Dec 20.

Pu Y, Wang W, Henao R, Chen L, Gan Z, Li C, Carin L. Adversarial Symmetric Variational Autoencoder. InAdvances in Neural Information Processing Systems. 2017 (pp. 4330-4339).

Chen L, Dai S, Pu Y, Li C, Su Q, Carin L. Symmetric Variational Autoencoder and Connections to Adversarial Learning. arXiv preprint arXiv:1709.01846. 2017 Sep 6.

Arjovsky M, Bottou L. Towards Principled Methods for Training Generative Adversarial Networks. arXiv preprint arXiv:1701.04862. 2017 Jan 17.