1 Introduction

ELBO loss function is a function that defines the loss of a variational autoencoder network. In this document, we will derive it from first principles, explain why each choice is made, and end with a closed-form formula for the loss.

This derivation is very much adapted and taken from the derivation by Priyam Mazumdar, I have attempted to add my own interpretation on his derivation.

1.1 Why ElBO?

ELBO is a quantity used as an approximation of the lower bound of the marginalized log probability of the dataset. It naturally occurs when we employ the Jensen inequality on the KL divergence of the variational approximation of the posterior with the target posterior.

1.2 The task of ELBO

Before we go any further in the derivation, I will quickly define the terms we will be using in this derivation:

1. $p(x)$ – The marginal probability density function of x.
2. $p(z|x)$ – This is the true posterior distribution, which is intractable.
3. $q(z)$ – The latent distribution of z, which is distributed as a standard normal
4. $q(z|x)$ – The posterior mapping function, given some data point x, what is the probability that the latent variable z was generated from that datapoint. This is an approximation of $P(z|x)$ using a $\mathcal{N}(\mu, \sigma^2)$ the best parameters of the approximation are found with our encoder.
5. $p(x|z)$ – The likelihood that we can reconstruct x from given latent variable z.

2 Derivation

The idea behind the derivation begins with, given some parameter θ, our loss should minimize the distance between $q(z|x)$ and $p(z|x)$,between the approximation of the posterior and the target posterior. We can find the distance between the two with KL-Divergence.

$$D_{KL}(q_\theta(z|x)||p(z/x)) \tag{1}$$

$$\sum_{z\in Z} q_\theta(z|x) log(\frac{q_\theta(z|x)}{p(z|x)}) \tag{2}$$

The equation above is the KL-Divergence, which I like to think of as a kind of expectation, between the log of the ratio. This equation gives a measure of distance between two distributions. Using the rule of conditional probabilities, we can first switch $p(z|x)$ with the fraction of the joint probability density function and the marginalized probability density function of x, then apply log rules to separate the expressions and finally break the summation apart. ’

$$\sum_{z\in Z} q_\theta(z|x) log(\frac{q_\theta(z|x)}{p(z,x)}) + \sum_{z\in Z} q_\theta(z|x) log(p(x) \tag{3}$$

$$\sum_{z\in Z} q_\theta(z|x) log(\frac{q_\theta(z|x)}{p(z,x)}) + log(p(x) \sum_{z\in Z} q_\theta(z|x) \tag{4}$$

$$\sum_{z\in Z} q_\theta(z|x) = 1 \tag{5}$$

$$\sum_{z\in Z} q_\theta(z|x)log(\frac{q_\theta(z|x)}{p(z,x)}) + log(p(x) \tag{6}$$

By Jensen's inequality, we understand the above expression is greater then or equal to zero.

$$\sum_{z\in Z} q_\theta(z|x)log(\frac{q_\theta(z|x)}{p(z,x)}) + log(p(x) \geq 0 \tag{7}$$

The expression above can be used to find the lower bound of the marginalized log probability, since log is monotonic, maximizing the left side will give us an elegant solution to our problem.

$$-\sum_{z\in Z} q_\theta(z|x)log(\frac{p(z,x)}{q_\theta(z|x)}) \leq log(p(x) \tag{8}$$

$$\sum_{z\in Z} q_\theta(z|x)log(\frac{p(z,x)}{q_\theta(z|x)}) \leq log(p(x) \tag{9}$$

$$log(p(x)) \geq ELBO \tag{10}$$

Now we have one main issue: for a neural network, not every term in the above is easily calculable; we must do some smart manipulation to make each term calculable.

$$log(p(x) \geq \sum_{z\in Z} q_\theta(z|x)log(\frac{p(z,x)}{q_\theta(z|x)}) \tag{11}$$

$$log(p(x) \geq \sum_{z\in Z} q_\theta(z|x)log(\frac{p(z,x)}{q_\theta(z|x)}) \tag{12}$$

$$log(p(x) \geq \sum_{z\in Z} q_\theta(z|x)log(\frac{p_\phi(x|z)p(z)}{q_\theta(z|x)}) \tag{13}$$

$$log(p(x) \geq \sum_{z\in Z} q_\theta(z|x)log(\frac{p(z)}{q_\theta(z|x)}) + \sum_{z\in Z} q_\theta(z|x)log(p_\phi(x|z)) \tag{14}$$

Notice that by the statistician's rule, the second part can be written as an expectation and the second is almost the kl divergence.

$$log(p(x) \geq \sum_{z\in Z} q_\theta(z|x)log(\frac{p(z)}{q_\theta(z|x)}) + E_{q_\theta(z|x)}(log(p(_\phi(x|z)) \tag{15}$$

$$log(p(x) \geq \sum_{z\in Z} q_\theta(z|x)log(\frac{p(z)}{q_\theta(z|x)}) + E_{q_\theta(z|x)}(log(p(_\phi(x|z)) \tag{16}$$