VAE ELBO Loss Derivation

Alina Selega

March 2022

1 ELBO loss

In a VAE, we want to learn a variational posterior distribution $q(z|x)$ that is as close as possible to the target distribution $p(z|x)$. Let's recall our terminology:

- $p(x|z)$: decoder
- $p(z|x)$: encoder
- $p(z) \sim N(0,1)$: prior

We typically learn $q(z|x) \sim N(g(x), h(x))$ with the help of neural networks approximating functions g, h. So we express our optimization problem over g, h as functions minimizing the KL divergence between variational and target distributions:

$$\begin{aligned}
KL(q(z|x), p(z|x)) &= \mathbb{E}_{q(z|x)}\left[\log \frac{q(z|x)}{p(z|x)}\right] \\
&= \mathbb{E}_{q(z|x)}\left[\log q(z|x) - \log \frac{p(x|z)p(z)}{p(x)}\right] \\
&= \mathbb{E}_{q(z|x)}\left[\log q(z|x) - \log p(x|z) - \log p(z) + \log p(x)\right] \\
&= \mathbb{E}_{q(z|x)}\left[\log \frac{q(z|x)}{p(z)} - \log p(x|z) + \log p(x)\right] \\
&= KL(q(z|x), p(z)) - \mathbb{E}_{q(z|x)}\left[\log p(x|z)\right] + \mathbb{E}_{q(z|x)}\left[\log p(x)\right] \\
&= KL(q(z|x), p(z)) - \mathbb{E}_{q(z|x)}\left[\log p(x|z)\right] + \log p(x)
\end{aligned}$$

Recall that we want to minimise the KL divergence above w.r.t. g, h. As $p(x)$ is independent of g, h we can drop it from the optimization, which becomes:

$$\min KL(q(z|x), p(z)) - \mathbb{E}_{q(z|x)}\left[\log p(x|z)\right]$$

When modelling Gaussian likelihood $p(x|z) \sim N(f(z), cI)$, we can drop the constant and get:

$$\min KL(q(z|x), p(z)) - \mathbb{E}_{q(z|x)}\left[-\frac{(x - f(z))^2}{2c}\right]$$
$$= \min KL(q(z|x), p(z)) + \mathbb{E}_{q(z|x)}\left[\frac{(x - f(z))^2}{2c}\right]$$

The second term is the reconstruction error and, for a Gaussian likelihood, is typically written as $C(x - \hat{x})^2$ for input x and its decoding $\hat{x}$, where $C = \frac{1}{2c}$ represents a balancing term between minimising the reconstruction error and the KL divergence between the variational posterior and prior.

Now let's derive the KL divergence term.

2 KL divergence between two Gaussians

This StackExchange answer gives the general form for KL divergence between two Gaussians. Below I derive a more detailed working.

Let $p_1(x) \sim N(\mu_1, \sigma_1^2)$ and $p_2(x) \sim N(\mu_2, \sigma_2^2)$.

$$\begin{aligned} KL(p_1(x), p_2(x)) &= \int p_1(x) \log \frac{p_1(x)}{p_2(x)} dx \\ &= \int p_1(x) \log p_1(x) dx - \int p_1(x) \log p_2(x) dx \end{aligned}$$

2.1 First term

$$\begin{aligned} \int p_1(x) \log p_1(x) dx &= \int p_1(x) \left(\log \frac{1}{Z_1} - \frac{(x - \mu_1)^2}{2\sigma_1^2} \right) dx \\ &= \log \frac{1}{Z_1} \int p_1(x) dx - \int p_1(x) \frac{(x - \mu_1)^2}{2\sigma_1^2} dx \\ &= -\log Z_1 - \frac{1}{2\sigma_1^2} \int p_1(x)(x^2 - 2x\mu_1 + \mu_1^2) dx \\ &= -\log Z_1 - \frac{1}{2\sigma_1^2} \left((\sigma_1^2 + \mu_1^2) - 2\mu_1^2 + \mu_1^2 \right) \end{aligned}$$

The last step is achieved by computing $\int p_1(x) x^2 dx = \mathbb{E}_{p_1(x)}[x^2] = \sigma_1^2 + \mu_1^2$ and similarly, computing expected values of $-2x\mu_1$ and μ_1^2 under $p_1(x)$.

Finally, we get

$$\int p_1(x) \log p_1(x) dx = -\log Z_1 - \frac{1}{2} = -\frac{1}{2}(\log 2\pi\sigma_1^2 + 1)$$

2.2 Second term

$$
\begin{aligned}
-\int p_1(x)\log p_2(x)dx &= -\int p_1(x)\left(\log\frac{1}{Z_2} - \frac{(x-\mu_2)^2}{2\sigma_2^2}\right)dx \\
&= \log Z_2 + \frac{1}{2\sigma_2^2}\int p_1(x)(x^2 - 2x\mu_2 + \mu_2^2)dx \\
&= \log Z_2 + \frac{1}{2\sigma_2^2}((\sigma_1^2 + \mu_1^2) - 2\mu_2\mu_1 + \mu_2^2) \\
&= \frac{1}{2}\log 2\pi\sigma_2^2 + \frac{1}{2\sigma_2^2}(\sigma_1^2 + (\mu_1 - \mu_2)^2)
\end{aligned}
$$

2.3 Final KL

$$
\begin{aligned}
KL(p_1(x), p_2(x)) &= -\frac{1}{2}(\log 2\pi\sigma_1^2 + 1) + \frac{1}{2}\log 2\pi\sigma_2^2 + \frac{1}{2\sigma_2^2}(\sigma_1^2 + (\mu_1 - \mu_2)^2) \\
&= \log\frac{\sigma_2}{\sigma_1} + \frac{1}{2\sigma_2^2}(\sigma_1^2 + (\mu_1 - \mu_2)^2) - \frac{1}{2}
\end{aligned}
$$

3 Final ELBO loss

In the ELBO loss, the KL divergence term we are interested is between $q(z|x) \sim N(g(x), h(x))$ and $p(z) \sim N(0, 1)$. Using the result from the previous section, we get:

$$
\begin{aligned}
KL(q(z|x), p(z)) &= \log\frac{1}{\sqrt{h(x)}} + \frac{1}{2}\left(h(x) + g(x)^2\right) - \frac{1}{2} \\
&= \frac{1}{2}\left(h(x) + g(x)^2 - \log h(x) - 1\right)
\end{aligned}
$$

Finally, the ELBO with Gaussian likelihood is then given by:

$$
C(x - \hat{x})^2 + \frac{1}{2}\left(h(x) + g(x)^2 - \log h(x) - 1\right)
$$