$$\begin{aligned}
D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z} \mid \mathbf{x})) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{q_\phi(\mathbf{z} \mid \mathbf{x})}{p_\theta(\mathbf{z} \mid \mathbf{x})}\right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{q_\phi(\mathbf{z} \mid \mathbf{x})p_\theta(\mathbf{x})}{p_\theta(\mathbf{x},\mathbf{z})}\right] \quad \text{since } p_\theta(\mathbf{z} \mid \mathbf{x}) = \frac{p_\theta(\mathbf{x},\mathbf{z})}{p_\theta(\mathbf{x})} \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{q_\phi(\mathbf{z} \mid \mathbf{x})}{p_\theta(\mathbf{x},\mathbf{z})}\right] + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x})\right] \\
&= -\underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})}\right]}_{\mathcal{L}_\theta,\phi(\mathbf{x})} + \log p_\theta(\mathbf{x})
\end{aligned}$$

Now in a simplified form we can say that

$$D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z} \mid \mathbf{x})) = -\mathcal{L}_{\theta,\phi}(\mathbf{x}) + \log p_\theta(\mathbf{x}) \tag{1}$$

Some observations about the 3 terms, starting off with the KL divergence D_{KL}, this is a distance so we know that $D \geqslant 0$.

The last term represents the *evidence* since its a log of a probability the result will be negative, that is, $\log p_\theta(\mathbf{x}) < 0$.

From the two facts above we know that to minimze the KL divergence we want to maximize the evidence lower bound (ELBO) $\mathcal{L}_{\theta,\phi}(\mathbf{x})$. Expressed formally we want to find the parameters such that

$$\theta^*, \phi^* = \arg\max_{\theta,\phi} \mathcal{L}_{\theta,\phi}(\mathbf{x})$$

To understand what the ELBO is we can decompose the term into two more meaningful parts using the following derivation

$$\begin{aligned}
\mathcal{L}_{\theta,\phi}(\mathbf{x}) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})}\right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z}) \frac{p_\theta(\mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})}\right] \quad \text{(bayes rule)} \\
&= \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right]}_{\text{reconstruction error}} + \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})}\right]}_{\text{regularization term}}
\end{aligned}$$

The first term is the reconstruction error, remember that $p_\theta(\mathbf{x} \mid \mathbf{z})$ is just a parameterized distribution, such as a normal distribution, so for example we have that

$$p_\theta(\mathbf{x} \mid \mathbf{z}) = N(\mathbf{x} \mid \mu_\theta(\mathbf{x}'), \boldsymbol{\Sigma}_\theta^2(\mathbf{x}'))$$

A quick sidenote, I've seen a fair bit of different notation used to describe parameterized distributions, I think all that matters is that its understood that the mean (first) term just says that the multivariate (normal) distribution is centered around whatever this term is, and the second term is just the variance in all directions, its sometimes also shown as $\eta \mathbf{I}$ where η is some scalar describing the variance and $\mathbf{I}$ is the identity matrix, so its describing the variance in all directions.

Now to answer the main question, how is that first term the reconstruction loss, or more accurately the expected reconstruction loss. We can create the following derivation.

$$\begin{aligned}\ln p_\theta(\mathbf{x} \mid \mathbf{z}) &= \ln\left(\frac{\exp[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2]}{\sigma\sqrt{2\pi}}\right)\\ &= -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 - \ln(\sigma\sqrt{2\pi})\\ &= -\frac{1}{2\sigma}(x-\mu)^2 - \ln(\sigma\sqrt{2\pi})\end{aligned}$$

One thing you might see now is that this term represents the l2 loss, or the squared loss, which is a similar idea to the basic mean square error in that we are computed the error just as the difference of the model output and the input. So intuitively we want to minimize this term to get the best reconstruction.

Moving on we still have the regularization term to understand. It might look a bit weird at first but its simply the KL divergence between the prior, that is the *assumed distribution* of the latent space and the approximated distribution of the latent space for a given input $q_\phi(\mathbf{z} \mid \mathbf{x})$. So we can say

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\frac{p_\theta(\mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})}\right] = D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z}))$$

A common choice for the assumed prior is the standard MVN, so $N(\mathbf{0}, \mathbf{I})$, where $\mathbf{0}$ is the zero vector and $\mathbf{I}$ is the identity matrix. Meaning we rewrite the above KL divergence as

$$D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel N(\mathbf{0}, \mathbf{I}))$$

To quickly recap, the objective when training a VAE is to maximize the ELBO (or equivelantly minimize the negative ELBO). If we phrase this in terms of a loss function we can say that we want to minimize the following loss function

$$\text{loss}(q_\phi, p_\theta) = -\mathcal{L}_{\theta,\phi}(\mathbf{x}) = -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z}')\right] + D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel N(\mathbf{0}, \mathbf{I}))$$

Remember that all we've done here is just restate the ELBO as a loss function to minimize, nothing else.

Now that we have a loss function we have to naturally ask how do we find the gradient of this loss function with respect to the parameters θ and ϕ, since these are the ones we intend to optimze. Firstly we can rewrite our ELBO in a form that makes it easier to differentiate, that is

$$\mathcal{L}_{\theta,\phi}(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z} \mid \mathbf{x})\right]$$

Now taking the gradient with respect to θ we get

$$\begin{aligned}\nabla_\theta \mathcal{L}_{\theta,\phi}(\mathbf{x}) &= \nabla \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z} \mid \mathbf{x})\right]\\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\nabla_\theta(\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z} \mid \mathbf{x}))\ \right]\\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\nabla_\theta \log p_\theta(\mathbf{x}, \mathbf{z})\right] \quad \text{(constant rule)}\\ &= \int \nabla_\theta \log p_\theta(\mathbf{x}, \mathbf{z}) q_\phi(\mathbf{z} \mid \mathbf{x}) d\mathbf{z}\end{aligned}$$

Since we generally can't compute this integral easily we tend to approximate it using for example Monte Carlo sampling, that is to say we just approximate it by literally sampling $\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{x})$ and then computing the gradient at that point. Which gives the following approximation

$$\nabla_\theta \mathcal{L}_{\theta,\phi}(\mathbf{x}) \approx \frac{1}{L} \sum_{l=1}^{L} \nabla_\theta \log p_\theta(\mathbf{x}, \mathbf{z}^{(l)})$$

So now that we have a way to compute that we can move on to the gradient with respect to ϕ. It starts out same as before

$$\nabla_\phi \mathcal{L}_{\theta,\phi}(\mathbf{x}) = \nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z} \mid \mathbf{x})\right]$$

The issue now is, we were computing the gradient with respect to ϕ and we have $q_\phi(\mathbf{z} \mid \mathbf{x})$ as an 'argument' to the expectation, putting the gradient inside the expectation is simply not equivalent due to the fact that we end up not capturing the gradient for the probability distribution we are computing the expectation over. Or more formally

$$\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z} \mid \mathbf{x})\right] = \nabla_\phi \left(\int \left[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z} \mid \mathbf{x})\right] q_\phi(\mathbf{z} \mid \mathbf{x}) d\mathbf{z} \right)$$

Whereas if we put the gradient inside the expectation we get

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\nabla_\phi (\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z} \mid \mathbf{x}))\right] = \int \nabla_\phi \left[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z} \mid \mathbf{x})\right] q_\phi(\mathbf{z} \mid \mathbf{x}) d\mathbf{z}$$

Put more simply, due to the dependence of the expectation on ϕ the complexity of computing the gradient is higher because we must compute the gradient with respect to the entire expectation, as opposed to the case with θ where we could just move the gradient inside the expectation.

One key thing we want to do is get rid of this dependence on ϕ one way of doing this is to use the reparametrization trick. The idea is to rewrite the expectation in a way that we can move the gradient inside the expectation. We do this by first separating out the random component that generates $q_\phi(\mathbf{z} \mid \mathbf{x})$ from the deterministic components that depend on $\mathbf{z}$, formally we can say

$$\begin{aligned} q_\theta(\mathbf{z} \mid \mathbf{x}) &= N(\mathbf{z} \mid \mu_\phi(\mathbf{x}), \boldsymbol{\Sigma}_\phi(\mathbf{x})) \\ &= \mu_\phi(\mathbf{x}) + \boldsymbol{\Sigma}_\phi(\mathbf{x})\mathbf{e} \quad \text{where } \mathbf{e} \sim N(\mathbf{0}, \mathbf{I}) \\ &= g_\phi(\mathbf{x}, \mathbf{e}) \\ &= \mathbf{z} \end{aligned}$$

With this form we can rewrite the ELBO as

$$\mathcal{L}_{\theta,\phi}(\mathbf{x}) = \mathbb{E}_{p(\mathbf{e})} \left[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z} \mid \mathbf{x})\right]$$

A sample estimator of this ELBO is then

$$\tilde{\mathcal{L}}_{\theta,\phi} = \log_\theta(\mathbf{x}, \mathbf{z}^{(i)}) - \log q_\phi(\mathbf{z}^{(i)} \mid \mathbf{x}) \quad \text{where } \mathbf{z}^{(i)} = g_\phi(\mathbf{x}, \mathbf{e}^{(i)}) \text{ and } \mathbf{e}^{(i)} \sim N(\mathbf{0}, \mathbf{I})$$

Meaning we can rewrite our ELBO as

$$\mathbb{E}_{p(\mathbf{e})} \left[\tilde{\mathcal{L}}_{\theta,\phi}(\mathbf{x})\right] = \mathcal{L}_{\theta,\phi}(\mathbf{x})$$

Thus if we want the gradient we can now use the same approach as before, since we know how to compute the gradient at specific samples, and since we know that we can use the expectation of the these sample gradients to compute the gradient of the ELBO we can finally derive the following

$$\mathbb{E}_{p(\mathbf{e})}\left[\nabla_{\theta,\phi}\tilde{\mathcal{L}}_{\theta,\phi}(\mathbf{x})\right] = \nabla_{\theta,\phi}\mathcal{L}_{\theta,\phi}(\mathbf{x})$$

With this reparametrization trick we can now rewrite our loss as

$$\text{loss}(q_\phi, p_\theta) = D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel N(\mathbf{0}, \mathbf{I})) - \mathbb{E}_{p(\mathbf{e})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{e} \times \sigma_\phi(\mathbf{x}) + \mu_\phi(\mathbf{x}))\right]$$

Some questions I had were for example, how exactly do we have

$$\mathbb{E}_{p(\mathbf{e})}[f(\mathbf{z})] = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[f(\mathbf{z})]$$

So lets prove this equality

$$\begin{aligned}
\mathbb{E}_{p(\mathbf{e})}[f(\mathbf{z})] &= \int f(\mathbf{z}) p(\mathbf{e}) d\mathbf{e} \\
&= \int f(\mu_\phi(\mathbf{x}) + \mathbf{\Sigma}_\phi(\mathbf{x})\mathbf{e}) N(\mathbf{0}, \mathbf{I}) d\mathbf{e} \\
u = \mathbf{\Sigma}_\phi(\mathbf{x})\mathbf{e} + \mu_\phi(\mathbf{x}) \quad du &= \frac{1}{\mathbf{\Sigma}\phi(\mathbf{x})} d\mathbf{e} \\
&= \int f(u) \frac{1}{\mathbf{\Sigma}\phi(\mathbf{x})\sqrt{2\pi}} \exp\left(-\frac{(u - \mu_\phi(\mathbf{x}))^2}{2\Sigma\phi(\mathbf{x})}\right) du \\
&= \int f(u) N(u \mid \mu_\phi(\mathbf{x}), \mathbf{\Sigma}_\phi(\mathbf{x})) du \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[f(\mathbf{z})]
\end{aligned}$$