Ryan Ott
14862565
ryan.ott@student.uva.nl

1 Generative Part of the VAE

1.1 Steps needed to sample an image from the decoder

1. Sample a latent vector $\boldsymbol{z}_n$ from the prior distribution $p(\boldsymbol{z}_n)$. In our case $\boldsymbol{z}_n \sim \mathcal{N}(0, \boldsymbol{I}_D)$
2. Pass $\boldsymbol{z}_n$ through the decoder network f_θ to yield the parameters of the distribution over $\boldsymbol{x}_n$
3. Sample a vector $\boldsymbol{x}_n$ from this distribution parameterized by $f_\theta(\boldsymbol{z}_n)$. In our case because we are using a categorial distribution, $f_\theta(\boldsymbol{z}_n)$ will be a vector of probabilities of pixel intensities. This $\boldsymbol{x}_n$ vector is the sampled image.

1.2 Why Monte-Carlo sampling is inefficient for VAE training

Monte-Carlo sampling requires a high number of samples to get an accurate estimate of the expectation due to the high variance of the estimator. This is for example shown in Figure 2 where we see the sampled points $z \sim p(z)$ not being representative of the true probability given our data $p(z|x)$ which is actually mostly concentrated around (1,1). Most samples will not contribute meaningfully to the integral we are trying to estimate. This problem gets exponentially worse as the dimensionality of the latent space increases. This is because the volume of the space grows exponentially with the dimensionality of the latent variable. It means that the probability mass of the distribution is spread out over a large volume, requiring even more samples to get an accurate estimate of the expectation.

1.3 Why ELBO is a lower bound on the log-likelihood

The right hand side of equation 10 is our ELBO. We can see that it is a lower bound on the log-likelihood because we are subtracting the KL-divergence between the variational distribution and the prior distribution. The KL-divergence is always non-negative, so the ELBO is always less than or equal to the log-probability $\log p(\boldsymbol{x}_n)$.

1.4 What happens to left hand terms when when the lower bound is pushed up

A higher ELBO implies that the difference between our variational distribution and true posterior is smaller, meaning that our latent space representation is more accurate.
Maximising the ELBO also means we are increasing the likelihood of the data under the model. This means that the model is more accurate at reconstructing the input data. The model parameters are more likely to have generated the data we are seeing.

1.5 Why are the names reconstruction loss and regularization loss appropriate

The first term can be seen as the reconstruction loss, because it measures how well the model predicts (or "reconstructs") an observation given a sampled latent variable from the variational posterior.
The second term measures the amount of information about the observation that is lost when we compress it into the latent variable. It is a regularization term because it penalizes the model for having a variational posterior that is too different from the prior distribution.

1.6 Why does sampling prevent gradient computation and how does reparameterization solve this

Sampling prevents us from computing $\nabla_\phi \mathcal{L}$ because the gradient of the sampling operation is not defined. Sampling is stochastic and so non-differentiable meaning we cannot establish how a small change in ϕ affects the loss.
Reparameterization solves this problem by decoupling the sampling operation from the parameters ϕ, instead introducing an auxiliary independent variable ϵ that is sampled from a fixed distribution $p(\epsilon)$, often a standard multivariate normal that we shift and scale to match the desired distribution of the latent variable. The

sampled latent variable $\boldsymbol{z}$ is then a deterministic differentiable function of ϵ and ϕ: $\boldsymbol{z} = g(\epsilon, \phi)$. This allows us to express our expectation in terms of ϵ instead, where the stochasticity is now encapsulated, and move the expectation outside of the gradient operation, allowing us to compute the gradient w.r.t. ϕ.