Exercise 9-1 Variational Autoencoder

Remember the Kullback-Leibler divergence between two probability distributions \( q \) and \( p \) is given by

\[
KL(q \parallel p) = - \int q(x) \log \frac{p(x)}{q(x)} \, dx
= - \int q(x) \log p(x) \, dx + \int q(x) \log q(x) \, dx
\]

(a) Show that the KL-divergence between two normal distributions \( q = \mathcal{N}(\mu_1, \sigma_1^2) \) and \( p = \mathcal{N}(\mu_2, \sigma_2^2) \) is:

\[
KL(q \parallel p) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}.
\]

*Hint:* Use that \( \mathbb{E}_p[f(x)] = \int p(x)f(x) \, dx \) and \( \sigma^2[x] = \mathbb{E}[(x - \mathbb{E}[x])^2] = \mathbb{E}[x^2] - (\mathbb{E}[x])^2 \).

(b) Let \( p \) be the standard normal distribution. How does the above term simplify?

(c) As you have learned in tutorial 06, the ELBO lower bound in variational inference is defined as

\[
L = \int q(z) \log \frac{p(z, x)}{q(z)} \, dz.
\]

Rearrange the above formula such that one can see that it is the expectation of \( \log p(x|z) \) with respect to \( q(z) \) minus the KL-divergence of \( q(z) \) w.r.t. \( p(z) \).

(d) Explain intuitively why \( \mathbb{E}_{q(z)} [\log P(x|z)] \) is called reconstruction loss.

(e) Download the Jupyter notebook for this exercise from the lecture web-page. It contains an implementation of a variational autoencoder which is applied to the MNIST dataset. Read and understand the content of this notebook.

Exercise 9-2 Generative Adversarial Networks (GANs)

(a) Explain the role of the generator \( G \) and the discriminator \( D \) in GANs.

(b) The loss for \( D \) is given as:

\[
J^{(D)} = -\frac{1}{2} \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \frac{1}{2} \mathbb{E}_{z \sim p_{z}(z)} [\log (1 - D(G(z))].
\]

Explain the terms in this loss!
(c) The generator tries to fool the discriminator, so its loss can be defined as:

\[ J^{(G)} = -J^{(D)} \]

Write down this optimization problem as a minimax game!

(d) Why might the usage of \( J^{(G)} = -J^{(D)} \) as loss for the generator lead to slow learning?

*Hint:* What happens to the gradient of the losses if \( D(G(z)) \) is small? (No calculation required)