DEEP LEARNING (SS2021)
SEMINAR 6

Assignment 1 (ML with noisy labels)**.** We want to learn a binary classifier $q(k \mid x; \theta)$ with classes $k = \pm 1$. It is defined as a neural network with parameters θ and with the sigmoid logistic distribution in the output.

The true labels k_i of the images x_i are however unknown. Instead we are given training pairs (x_i, t_i) with "noisy labels" t_i. They might have been incorrectly assigned by the person who annotated the data. More specifically, let us assume that the label t_i is correct ($t_i = k_i$) with probability $1 - \varepsilon$ and incorrect ($t_i = -k_i$) with probability ε.

a) Formulate the conditional maximum likelihood learning of the parameters θ.
Hint: the conditional likelihood of the training data sample (x_i, t_i) is obtained by marginalizing over the unknown true label

$$p(t_i \mid x_i) = \sum_{k \in \{-1,1\}} p(t_i \mid k) q(k \mid x_i; \theta),$$

where $p(t \mid k)$ is the labelling noise model.

b) A popular practical solution is to minimize the cross-entropy loss

$$-\sum_i \sum_k p_i(k) \log q(k \mid x_i; w), \tag{1}$$

where $p_i(k)$ denote "softened 1-hot labels": $p_i(k) = 1 - \varepsilon$ for $k = t_i$ and ε otherwise. Prove that the negative cross-entropy (1) is a lower bound of the log likelihood in a). Use Jensen's inequality for $\log$.

Assignment 2. Let $q(x)$ and $p(x)$ be two factorising probability distributions for random vectors $x \in \mathbb{R}^n$, i.e.

$$p(x) = \prod_{i=1}^{n} p(x_i) \text{ and } q(x) = \prod_{i=1}^{n} q(x_i).$$

Prove that their KL-divergence decomposes into a sum of KL-divergences for the components, i.e.

$$D_{KL}(q(x) \parallel p(x)) = \sum_{i=1}^{n} D_{KL}(q(x_i) \parallel p(x_i))$$

Assignment 3. Compute the KL-divergence of two univariate normal distributions.

Assignment 4 (Bernoulli VAE)**.** Let us consider a VAE with binary valued latent variables $z \in \mathcal{Z} = \{0,1\}^n$. Training such VAEs by maximising the ELBO criterion requires computation of the gradient of the data term w.r.t. encoder parameters φ

$$\nabla_\varphi \mathbb{E}_{q_\varphi(z \,|\, x)} \log p_\theta(x \,|\, z) = \nabla_\varphi \sum_{z \in \mathcal{Z}} q_\varphi(z \,|\, x) \log p_\theta(x \,|\, z). \tag{2}$$

a) we can explicitly sum over $z \in \mathcal{Z}$ if the dimension of the latent space is small. This is however not possible for high dimensional latent spaces.

b) (Score function, log-trick) Prove the following equality

$$\nabla_\varphi \sum_{z \in \mathcal{Z}} q_\varphi(z \,|\, x) \log p_\theta(x \,|\, z) = \sum_{z \in \mathcal{Z}} q_\varphi(z \,|\, x) \nabla_\varphi \log q_\varphi(z \,|\, x) \log p_\theta(x \,|\, z)$$

Conclude that the following procedure implements an unbiased stochastic estimator of the required gradient (2):

Sample $z \sim q_\varphi(z \,|\, x)$ and compute $\nabla_\varphi \log q_\varphi(z \,|\, x) \log p_\theta(x \,|\, z)$