DEEP LEARNING (SS2020)
SEMINAR 2

Assignment 1. Let us consider a fully connected recurrent network with n binary neurons. Denoting their outputs by $x_i = \pm 1$, the corresponding dynamic system reads

$$x_i(t+1) = \operatorname{sign}\big(\sum_{j \neq i} w_{ij} x_j(t)\big),$$

We will assume sequential updates in some fixed order over the neurons. Let us further assume that matrix of weights is symmetric, i.e. $w_{ij} = w_{ji}$.

Prove that the function

$$\mathcal{H}(x) = -\frac{1}{2} \sum_{i,j} x_i w_{ij} x_j = -\frac{1}{2} x^T W x$$

is a Ljapunov function for this dynamical system, i.e. it can decrease only:

$$\mathcal{H}(x(t+1)) \leq \mathcal{H}(x(t)).$$

Conclude that the network will eventually reach a fixpoint configuration.
Hint: express the change of $\mathcal{H}(x)$ for an update of a single neuron x_i.

Assignment 2 (Softmax)**.** Show the following properties of the function

$$\operatorname{softmax}\colon \mathbb{R}^n \to \mathbb{R}^n_+ \colon x \mapsto \frac{e^{x_i}}{\sum_j e^{x_j}}$$

(a) $\operatorname{softmax}$ is invariant to adding the same number to all scores x
(b) $\arg\max_i \operatorname{softmax}(x)_i = \arg\max_i \log \operatorname{softmax}(x)_i = \arg\max_i(x_i)$
(c) When all scores x are scaled by a big positive number, $\operatorname{softmax}(x)$ approaches the argmax indicator:

$$I_i(x) = \begin{cases} 1, & \text{if } \ x_i > x_j \ \forall j \neq i; \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

(So it would be more appropriate to call it "soft argmax").

Assignment 3. Let us consider the logistic regression model

$$p(k \mid x; w) = \log \mathrm{S}(k w^T x),$$

where $k = \pm 1$ is the class, $x \in \mathbb{R}^n$ is the feature vector, $w \in \mathbb{R}^n$ is a parameter vector and S denotes the logistic sigmoid function. Given training data $\mathcal{T}^m = \{(x_i, k_i) \mid i = 1 \ldots m\}$, we want to estimate w by maximising the (conditional) log-likelihood.

(a) Let us assume that the training data are linearly separable. Show that in this case the logistic regression problem has no finite optimal solution. To approach the optimal value in this case, the weight vector becomes infinitely large.
Hint: show that for any w that achieves a correct classification taking $w' = sw$ with $s > 1$ achieve a higher likelihood.
(b) Show that adding a regularizer on the weight norm $\lambda\|w\|^2$ fixes this problem.

Assignment 4 (*ML with noisy labels)*. Suppose that the class label $k = \pm 1$ given an observation x follows the logistic model with conditional distribution $q(k|x; w)$ as in the previous assignment. Suppose you have training pairs (x_i, t_i) where t_i might have been incorrectly labelled by the person who annotated the data, which happens with probability ε. That is, $t_i = -k_i$ with probability ε and $t_i = k_i$ with probability $1 - \varepsilon$, where k_i is the true label which is not available.

(a) Formulate conditional maximum likelihood learning of parameters w.
Hint: the conditional likelihood of the training data sample (x_i, t_i) is given by marginalizing over the unknown true label:
$p(t_i|x_i) = \sum_{k\in\{-1,1\}} p(t_i|k)q(k|x_i; w)$, where $p(t|k)$ is the labeling noise model.
(b) How is the maximum likelihood related to minimizing the cross-entropy

$$\sum_i \sum_k p_i(k) \log q(k \mid x_i; w)$$

where $p_i(k)$ are the "softened 1-hot labels": $p_i(k) = 1 - \varepsilon$ for $k = t_i$ and ε otherwise?
Hint: show that cross-entropy is an upper bound on the negative log likelihood using Jensen's inequality for log.

Assignment 5 (Backprop of scan)**.** The *inclusive cumulative sum* or for brevity *scan* operation is defined as follows: Given the input vector $x \in \mathbb{R}^n$ the output $y \in \mathbb{R}^n$ has components:

$$y_i = \sum_{j\le i} x_j.$$

Compute the backprop of scan, i.e. given $\nabla_y L$ compute $\nabla_x L$.