DEEP LEARNING (SS2023)
SEMINAR 5

Assignment 1 (Trust Region Problems, FGSM)**.**
Let us consider a loss function $L(\theta)$ and denote its gradient at θ^t by $g^t = \nabla_\theta L(\theta^t)$. In this exercise θ can represent parameters of neural network, relevant for learning, or a given input image, relevant for adversarial attack. Solve the following trust region problems.

a) $\arg\min_\theta \big[L(\theta^t) + \langle g^t, \theta - \theta^t\rangle\big]$,

s.t. $\|\theta - \theta^t\|_2 \leq \varepsilon$.

Hints: Make a simplifying substitution of variables $\Delta\theta = \theta - \theta^t$. Use the method of Lagrange multipliers. The constraint can be squared to make it easier to differentiate. The linear function on a convex set attains its minimum at the boundary, so that the constraint can be replaced with equality.

b) $\arg\min_\theta \big[L(\theta^t) + \langle g^t, \theta - \theta^t\rangle\big]$,

s.t. $|\theta_i - \theta_i^t| \leq \varepsilon \quad \forall i$.

Hint: Observe that the minimization decouples over individual coordinates. Solve for a single coordinate graphically.

Assignment 2 (BN with Weight Decay)**.**
In the previous seminar, we have discussed that the output of BN layer after a linear layer is invariant to the scale of the weight vector. It would seem that applying weight decay regularization makes no sense. Nevertheless it is applied in some papers and receipts. Let's study the effect it has on optimization.

We will consider a simplified scenario for a single neuron and *weight normalization*. Its output is given by $y = \frac{w^\mathsf{T} x}{\|w\|}$, where x is the input. The regularized loss function is given by $\tilde{L}(w) = L(y(w)) + R(w)$, where $R(w) = \frac{\lambda}{2}\|w\|^2$ and $\lambda > 0$.

a) Suppose that w_0 is optimal for the non-regularized loss L. What will the gradient descent on $\tilde{L}$ do if started at w_0?

b) Consider a point w_0 on the unit sphere for which the gradient $g = \nabla_w L(y(w_0))$ is non-zero. Show that g is orthogonal to w_0 and hence also to $\nabla_w R(w_0)$. Draw these vectors and the sphere $\|w\| = 1$ in a plane.

c) In the drawing above, let $\|g\| = a$ and $\|\nabla_w R(w_0)\| = \lambda$ at w_0 with $\|w_0\| = 1$. Consider a single step of the gradient descent for $\tilde{L}$ with step length $\alpha > 0$. Give a condition on α that ensures a decreasing norm $\|w\|$.
Hint: the problem can be solved in 2D given a, λ and α.

Assignment 3 (Mirror Descent for Box Constraints)**.**
Sometimes we need to optimize a non-linear objective $f(x)$ over box constraints $x \in$

$(0, 1)$, we consider 1D case for simplicity. This is relevant for learning with constrained parameters, in a multi-step adversarial attack or adversarially robust training. In such cases it is beneficial to use gradient descent with the steps found from solving the proximal step problem

$$\min_x \langle \nabla f(x^0), x - x^0 \rangle + \frac{1}{\varepsilon} D(x, x^0),$$

where $x^0 \in (0, 1)$ and the divergence D is designed to respect constraints. For example, a suitable choice is[1]:

$$D(x, x^0) = x \log \frac{x}{x^0} + (1 - x) \log \frac{1 - x}{1 - x^0},$$

which is convex in x, has minimum at x^0 and its derivatives approach infinity at the edges of the interval $(0, 1)$.

a) Find the solution to the proximal step problem and express it using the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$ and the logit function: $\mathrm{logit}(p) = \log \frac{p}{1-p}$, which is the inverse of sigmoid.

b) Let $x^0 = \mathrm{sigmoid}(\eta^0)$. Rewrite the algorithm iterations such that $x^t = \mathrm{sigmoid}(\eta^t)$ and η^t is updated using η^{t-1}.

[1] Kullback-Leibler divergence of Bernoulli distributions with probabilities x and x_0.