Deep Learning (SS2021)
Seminar 5

April 16, 2021

Assignment 1 (BN with Weight Decay)
In the lecture we discussed that combining Batch Normalisation with weight decay regularisation leads to an ill posed optimisation problem. Let us consider this in a simplified scenario for a single neuron. Its output is given by $y = \frac{w^\top x}{\|w\|}$, 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) Compute the gradient of $R(w)$ and show that a step towards decreasing it can be interpreted as "weight decay". Suppose that w_0 is optimal for the non-regularized loss L. What will gradient descent on $\tilde{L}$ do if started at w_0?

b) Consider a point w_0 such that $\|w_0\| = 1$. Assume that $g = \nabla_w L(y)$ is non-zero. Show that g is orthogonal to w and hence also to $\nabla_w R(w)$. Draw these vectors and the sphere $\|w\| = 1$.

c) Let $\|g\| = a$ and $\|\nabla_w R(w)\| = \lambda$ at w_0 and $\|w_0\| = 1$. Consider a single step of gradient descent with step length α. For which $\alpha > 0$, the norm $\|w\|$ will decrease?

Assignment 2 (Trust Region Problem with Box Constraints, FGSM)
Let us consider a loss function $L(\theta)$ and denote its gradient at θ^t by $g^t = \nabla_{\theta^t} L$.

a) Solve the following trust region problem:

$$\begin{aligned} &\arg\min_{\theta} \big[L(\theta^t) + \langle g, \theta - \theta^t \rangle\big], \\ &\text{s.t. } \|\theta_i - \theta_i^t\| \leq \varepsilon \; \forall i \end{aligned}$$

by using the technique of Lagrange multipliers.
Hint: Make a substitution of variables $\Delta\theta = \theta - \theta^t$. Square the constraints to make their derivative simpler. Note that the Lagrange multipliers for inequality constraints must be non-negative.

b) Show that the fast sign gradient attack solves a similar constrained optimization problem (formulate this problem).

Assignment 3 (Proximal Problem with Regularization)
Consider the problem of minimizing the training loss of a neural network with a weight regularization $\lambda\|\theta\|^2$. Since the weight regularization is known and given in closed form, we do not need to approximate it linearly by computing its gradient. Derive an SGD like optimization algorithm for solving the composite proximal step problem

$$\min_{\theta} \Big[\langle g, \theta - \theta_t \rangle + \lambda \|\theta\|^2 + \frac{1}{\varepsilon} \|\theta - \theta_t\|^2 \Big], \tag{1}$$

where θ_t is the current parameter vector, g is the (stochastic) gradient at θ_t, λ is the regularization strength and ε is the learning rate.