Exercises for the course
Deep Learning 1
Winter Semester 2024/25

Machine Learning Group
Faculty IV – Electrical Engineering and Computer Science
Technische Universität Berlin

Exercise Sheet 5

Exercise 1: Neural Network Regularization (5×20 P)

For a neural network to generalize from limited data, it is desirable to make it sufficiently invariant to small local perturbations. This can be done by limiting the gradient norm $\|\partial f / \partial \boldsymbol{x}\|$ for all $\boldsymbol{x}$ in the input domain. As the input domain can be high-dimensional, it is impractical to minimize the gradient norm directly. Instead, we can minimize an upper-bound of it that depends only on the model parameters.

We consider a two-layer neural network with d input neurons, h hidden neurons, and one output neuron. Let W be a weight matrix of size $d \times h$, and $(b_j)_{j=1}^h$ a collection of biases. We denote by $W_{i,:}$ the ith row of the weight matrix and by $W_{:,j}$ its jth column. The neural network computes:

$$a_j = \max(0, W_{:,j}^\top \boldsymbol{x} + b_j) \qquad \text{(layer 1)}$$
$$f(\boldsymbol{x}) = \textstyle\sum_j a_j \qquad \text{(layer 2)}$$

The first layer detects patterns of the input data, and the second layer performs a pooling operation over these detected patterns.

(a) *Show* that the gradient norm of the network can be upper-bounded as:

$$\left\| \frac{\partial f}{\partial \boldsymbol{x}} \right\| \leq \sqrt{h} \cdot \|W\|_F$$

Hint: Use the Cauchy-Schwarz inequality.

(b) *Show* that the well-known weight decay procedure ($W^{(t+1)} \leftarrow (1-\gamma) \cdot W^{(t)}$ for some $\gamma > 0$) can be interpreted as a gradient descent of $\|W\|_F$ or some related quantity.

(c) Let $\|W\|_{\text{Mix}} = \sqrt{\sum_i \|W_{i,:}\|_1^2}$ be a ℓ_1/ℓ_2 mixed matrix norm. *Show* that the gradient norm of the network can be upper-bounded by it as:

$$\left\| \frac{\partial f}{\partial \boldsymbol{x}} \right\| \leq \|W\|_{\text{Mix}}$$

(d) *Show* that the bound is tighter than the one based on the Frobenius norm, i.e. show that $\|W\|_{\text{Mix}} \leq \sqrt{h} \cdot \|W\|_F$.

(e) *Show* that the gradient of the squared mixed norm is given by

$$\frac{\partial}{\partial W_{ij}} \|W\|_{\text{Mix}}^2 = 2 \cdot \|W_{i,:}\|_1 \cdot \text{sign}(W_{ij}).$$