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:

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

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:

$$\Big\|\frac{\partial f}{\partial \boldsymbol{x}}\Big\| \le \sqrt{h} \cdot \|W\|_F$$

Hint: Use the Cauchy-Schwarz inequality.

$$
\begin{aligned}
\Big\|\frac{\partial f}{\partial \boldsymbol{x}}\Big\|^2 = \sum_{i=1}^d \Big(\frac{\partial f}{\partial x_i}\Big)^2 &= \sum_{i=1}^d \Big(\sum_{j=1}^h 1_{a_j>0} W_{ij}\Big)^2\\
&\le \sum_{i=1}^d \Big(\sum_{j=1}^h (1_{a_j>0})^2 \sum_{j=1}^h W_{ij}^2\Big) \le \sum_{i=1}^d \Big(h \sum_{j=1}^h W_{ij}^2\Big) = h \cdot \|W\|_F^2
\end{aligned}
$$

(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.
Descending $\|W\|_F^2$ with a learning rate $\gamma/2$, we get:

$$W^{(t+1)} \leftarrow W^{(t)} - \frac{\gamma}{2} \cdot \frac{\partial \|W^{(t)}\|_F^2}{\partial W^{(t)}} = W^{(t)} - \frac{\gamma}{2} \cdot 2W^{(t)} = (1-\gamma)\cdot W^{(t)}$$

(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:

$$\Big\|\frac{\partial f}{\partial \boldsymbol{x}}\Big\| \le \|W\|_{\text{Mix}}$$

$$
\begin{aligned}
\Big\|\frac{\partial f}{\partial \boldsymbol{x}}\Big\|^2 = \sum_{i=1}^d \Big(\frac{\partial f}{\partial x_i}\Big)^2 &= \sum_{i=1}^d \Big(\sum_{j=1}^h 1_{a_j>0} W_{ij}\Big)^2\\
&\le \sum_{i=1}^d \Big(\sum_{j=1}^h |W_{ij}|\Big)^2 = \sum_{i=1}^d \big(\|W_{i,:}\|_1\big)^2 = \|W\|_{\text{Mix}}^2
\end{aligned}
$$

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

$$\|W\|^2_{\text{Mix}} = \sum_{i=1}^{d} \Big(\sum_{j=1}^{h} \big|W_{ij}\big| \Big)^2 \leq \sum_{i=1}^{d} \Big(\sum_{j=1}^{h} (1)^2 \sum_{j=1}^{h} |W_{ij}|^2 \Big) = \sum_{i=1}^{d} \Big(h \sum_{j=1}^{h} W_{ij}^2 \Big) = h \cdot \|W\|^2_F$$

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

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

$$\begin{aligned}
\frac{\partial}{\partial W_{ij}} \|W\|^2_{\text{Mix}} &= \frac{\partial}{\partial W_{ij}} \sum_{i} \|W_{i,:}\|_1^2 \\
&= \frac{\partial}{\partial W_{ij}} \sum_{i=1}^{d} \Big(\sum_{j=1}^{h} \big|W_{ij}\big| \Big)^2 \\
&= 2 \cdot \Big(\sum_{j=1}^{h} \big|W_{ij}\big| \Big) \cdot \frac{\partial}{\partial W_{ij}} \sum_{j=1}^{h} \big|W_{ij}\big| = 2 \cdot \|W_{i,:}\|_1 \cdot \text{sign}(W_{ij})
\end{aligned}$$