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:

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

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:

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

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

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

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.

$$\frac{\partial f}{\partial x_i} = \sum_{j=1}^{h} \frac{\partial f}{\partial a_j} \cdot \frac{\partial a_j}{\partial x_i} = \sum_{j=1}^{h} 1_{a_j > 0} \cdot \frac{\partial (W_{:,j}^\top x + b_j)}{\partial x_i}$$

$$= \sum_{j=1}^{h} 1_{a_j>0} \cdot \frac{\partial \sum_{i=1}^{d} W_{i,j} \cdot x_i}{\partial x_i}$$

$$= \sum_{j=1}^{h} 1_{a_j>0} \cdot W_{i,j}$$

$$\left(\sum a_i b_i\right)^2 \leq \left(\sum a_i^2\right)\left(\sum b_i^2\right)$$

$$\left\| \frac{\partial f}{\partial x} \right\| = \sqrt{\sum_{i=1}^{d} \left(\frac{\partial f}{\partial x_i}\right)^2} = \sqrt{\sum_{i=1}^{d} \left(\sum_{j=1}^{h} 1_{a_j>0} \cdot W_{i,j}\right)^2}$$

$$\leq \sqrt{\sum_{i=1}^{d} \left(\sum_{j=1}^{h} 1_{a_j>0}^2\right)\left(\sum_{j=1}^{h} W_{i,j}^2\right)}$$

$$\leq \sqrt{\sum_{i=1}^{d} h \cdot \left(\sum_{j=1}^{h} W_{i,j}^2\right)}$$

$$= \sqrt{h} \sqrt{\sum_{i=1}^{d} \sum_{j=1}^{h} W_{i,j}^2}$$

$$= \sqrt{h} \cdot \|W\|$$

(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.

$$W^{(t+1)} \leftarrow (1-\gamma) W^{(t)} = W^{(t)} - \gamma W^{(t)}$$

Descending $\|W\|_F^2$ with a learning rate $\frac{\gamma}{2}$

$$= W^{(t)} - \frac{\gamma}{2} \cdot 2 W^{(t)}$$

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

(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:

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

from 1.a) we have showed: $\left\|\frac{\partial f}{\partial x}\right\| = \sqrt{\sum_{i=1}^{d}\left(\sum_{j=1}^{h} 1_{a_j>0} \cdot W_{i,j}\right)^2}$

then we have: $\leq \sqrt{\sum_{i=1}^{d}\left(\sum_{j=1}^{h} |W_{i,j}|\right)^2}$

$$= \sqrt{\sum_{i=1}^{d} \|W_{i,:}\|_1^2}$$

$$= \|W\|_{mix}$$

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

$$\|W\|_{mix} = \sqrt{\sum_{i=1}^{d}\left(\sum_{j=1}^{h} W_{i,j}\right)^2}$$

$$\leq \sqrt{\sum_{i=1}^{d}\left(\sum_{j=1}^{h} 1^2 \cdot \sum_{j=1}^{h} W_{i,j}\right)}$$ Cauchy-Schwarz inequality

$$\leq \sqrt{\sum_{i=1}^{d} h \cdot \sum_{j=1}^{h} W_{i,j}}$$

$$= \sqrt{h} \cdot \|W\|_F$$

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

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

$$\|W\|^2_{Mix} = \sum_{i=1}^{d} \|W_{i,:}\|_1^2$$

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

$$= 2\|W_{i,:}\|_1 \cdot \frac{\partial}{\partial W_{ij}}\left(\sqrt{\sum_{j=1}^{h} |W_{i,j}|}\right) = 2\|W_{i,:}\|_1 \cdot \frac{\partial}{\partial W_{ij}}\left(\sum_{j=1}^{h} |W_{ij}|\right)$$

$$= 2\|W_{i,:}\|_1 \cdot \text{sign}(W_{i,j})$$