Machine Learning Group
Department of Electrical Engineering and Computer Science
Fakultät IV, Technische Universität Berlin

Exercise Sheet 2

Exercise 1: Neural Network Optimization (10 + 10 + 10 P)

Consider the one-layer neural network

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$$

applied to data points $\boldsymbol{x} \in \mathbb{R}^d$, and where $\boldsymbol{w} \in \mathbb{R}^d$ is the parameter of the model. We would like to optimize the mean square error objective:

$$J(\boldsymbol{w}) = \mathbb{E}_{\hat{p}}\Big[\frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{x} - t)^2\Big],$$

where the expectation is computed over an empirical approximation $\hat{p}$ of the true joint distribution $p(\boldsymbol{x}, t)$. The ground truth is known to be of type: $t|\boldsymbol{x} = \boldsymbol{v}^\top \boldsymbol{x} + \varepsilon$, with the parameter $\boldsymbol{v}$ unknown, and where ε is some small i.i.d. Gaussian noise. The input data follows the distribution $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I)$ where $\boldsymbol{\mu}$ and σ^2 are the mean and variance.

(a) *Compute* the Hessian of the objective function J at the current location $\boldsymbol{w}$ in the parameter space, and as a function of the parameters $\boldsymbol{\mu}$ and σ of the data.

(b) *Show* that the condition number of the Hessian is given by: $\dfrac{\lambda_1}{\lambda_d} = 1 + \dfrac{\|\boldsymbol{\mu}\|^2}{\sigma^2}$.

(c) *Explain* for this particular problem what would be the advantages and disadvantages of centering the data before training. You answer could include the following aspects: (1) condition number and speed of convergence, (2) ability to reach a low prediction error.

Exercise 2: Neural Network Regularization (10 + 10 + 10 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:

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

and *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.

(b) 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:

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

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

(c) *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 3: Programming (40 P)

Download the programming files on ISIS and follow the instructions.