Exercises for the course
Deep Learning 1
Winter Semester 2022/23

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

Exercise Sheet 3

Exercise 1: Neural Network Optimization (20 + 15 + 15 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.

$$\begin{aligned}
H &= \frac{\partial}{\partial \boldsymbol{w}\boldsymbol{w}^\top}\mathbb{E}\Big[\frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{x} - t)^2\Big] \\
&= \frac{\partial}{\partial \boldsymbol{w}\boldsymbol{w}^\top}\mathbb{E}\Big[\frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{x})(\boldsymbol{x}^\top \boldsymbol{w}) + \text{lin.} + \text{const.}\Big] \\
&= \mathbb{E}\big[\boldsymbol{x}\boldsymbol{x}^\top\big] = \text{Cov}(\boldsymbol{x}) + \mathbb{E}[\boldsymbol{x}]\,\mathbb{E}[\boldsymbol{x}]^\top = \sigma^2 I + \boldsymbol{\mu}\boldsymbol{\mu}^\top
\end{aligned}$$

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

$$\begin{aligned}
\lambda_1 &= \max_{\|\boldsymbol{v}\|=1} \boldsymbol{v}^\top H \boldsymbol{v} \\
&= \max_{\|\boldsymbol{v}\|=1} \boldsymbol{v}^\top (\sigma^2 I + \boldsymbol{\mu}\boldsymbol{\mu}^\top)\boldsymbol{v} \\
&= \max_{\|\boldsymbol{v}\|=1} \sigma^2 + \|\boldsymbol{v}^\top \boldsymbol{\mu}\|^2 \\
&= \sigma^2 + \Big\|\frac{\boldsymbol{\mu}^\top}{\|\boldsymbol{\mu}\|}\boldsymbol{\mu}\Big\|^2 \\
&= \sigma^2 + \|\boldsymbol{\mu}\|^2 \\
\lambda_2 &= \max_{\binom{\|\boldsymbol{v}\|=1}{\boldsymbol{v}^\top \boldsymbol{\mu}=0}} \boldsymbol{v}^\top (\sigma^2 I + \boldsymbol{\mu}\boldsymbol{\mu}^\top)\boldsymbol{v} = \sigma^2 \\
\lambda_3, \dots, \lambda_d &= \sigma^2
\end{aligned}$$

Therefore, $\lambda_1/\lambda_d = (\sigma^2 + \|\boldsymbol{\mu}\|^2)/\sigma^2 = 1 + \|\boldsymbol{\mu}\|^2/\sigma^2$

(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.
Advantage: centering makes λ_1/λ_d lower : $1 + \|\mathbf{0}\|^2/\sigma^2 < 1 + \|\boldsymbol{\mu}\|^2/\sigma^2$, therefore, convergence is faster. Disadvantage: The set of homogeneous models based on centered data $f(\boldsymbol{x}) = \boldsymbol{w}^\top(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])$ does not contain the ground truth $f(\boldsymbol{x}) = \boldsymbol{v}^\top \boldsymbol{x}$.

Exercise 2: Programming (50 P)

Download the programming files on ISIS and follow the instructions.