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 3

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

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

(b) *Show* that the condition number of the Hessian is given by: $\dfrac{\lambda_1}{\lambda_d} = 1 + \dfrac{\|\boldsymbol{\mu}\|^2}{\sigma^2}$.
Symmetric real-valued matrices have an eigendecomposition into orthonormal eigenvectors with real eigenvalues. Hence we can deduct:

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

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 : $1 + \|\boldsymbol{0}\|^2/\sigma^2 < 1 + \|\boldsymbol{\mu}\|^2/\sigma^2$, 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: Initialization (35 + 10 P)

Consider a deep neural network with L layers with width n and a ReLU activation function. Assume the dataset X, which consists of samples $x \in \mathbb{R}^n$ which are iid.. X is centered and whitened, i.e., $\mathbb{E}[x^{(i)}] = 0$ and $\mathrm{Var}[x^{(i)}] = 1 \forall i \in \{1, \ldots, n\}$, $\mathrm{Cov}(x^{(i)}, x^{(j)}) = 0 \forall i \neq j$ where i, j indicate the dimensions.
The He-initialization is defined as follows:

$$W_{(l)}^{(ij)} \sim \mathcal{N}\left(0, \frac{2}{n}\right)$$
$$b_{(l)}^{i} = 0\,,$$

where $W_{(l)}$ is the weight matrix of layer l and $b^{(l)}$ is the bias vector of layer l. Also, the $W_{ij}^{(l)}$ are mutually independent.
You may use the following assumptions/hints:

- For a random variable Y centered around 0, i.e. $\mathbb{E}[Y] = 0$, we assume $\mathbb{E}[\mathrm{ReLU}(Y)^2] = \frac{1}{2}\mathrm{Var}(Y)$.
- For mutually independent random variables a, b, we have $\mathbb{E}[ab] = \mathbb{E}[a]\mathbb{E}[b]$.
- $\mathbb{E}[\sum_i Y^{(i)}] = \sum_i \mathbb{E}[Y^{(i)}] = n\mathbb{E}[Y]$.
- a_0 is the input to the neural network.

(a) *Show* by induction that, when using the initialization scheme of He et al. (2015), the variance of the latent variables $z_{(l)}^{(j)} = \sum_i W_{(l)}^{(ij)} a_{(l-1)}^{(i)} + b_{(l)}^{(j)}$ for all layers $l \in \{1, ..., L\}$ stays constant, i.e. $\mathrm{Var}(\sum_i W_{(l+1)}^{(ij)} a_{(l)}^{(i)} + b_{(l+1)}^{(j)}) = \mathrm{Var}(\sum_i W_{(l)}^{(ij)} a_{(l-1)}^{i} + b_{(l)}^{(j)})$.

Solution 1 (short)

$a_i = \max(0, z_i)$
$z_j = \sum_i a_i w_{ij} + b_j$

$$\begin{aligned}
\mathrm{Var}(\overbrace{\sum_i a_i w_{ij}}^{z_j}) &= \sum_i \mathrm{Var}(a_i w_{ij}) && \text{(variance of a sum of decorrelated variables)} \\
&= \sum_i \mathbb{E}[a_i^2 w_{ij}^2] - \mathbb{E}[a_i]^2 \mathbb{E}[w_{ij}]^2 && \text{(variance of product of indep. variables)} \\
&= \sum_i \mathbb{E}[(a_i w_{ij})^2] && \text{(weights have mean zero)} \\
&= \sum_i \mathbb{E}[a_i^2]\mathbb{E}[w_{ij}^2] && \text{(expectation of product of indep. variables)} \\
&= \sum_i \frac{1}{2}\mathbb{E}[z_i^2]\mathrm{Var}(w_{ij}) && \text{(relu slashes squared expectation by two, weights have mean zero)} \\
&= \sum_i \frac{1}{2}\mathbb{E}[z_i^2]\frac{2}{N} && \text{(set value for variance of weights)} \\
&= \frac{1}{N}\sum_i \mathbb{E}[z_i^2] && \text{(reorganize, use induction base and hypothesis)} \\
&= \mathrm{Var}[z_i^2] && (z_i\text{s are identically distributed, pre-activations are centered})
\end{aligned}$$

Induction: requires pre-activations are centered, iid., then verifies that same holds for next layer, etc.

Solution 2(extended)

We show this in multiple steps:

1. For all $l \in \{1, \dots, L\}, l' \in \{1, \dots, l-1\}$ we have that $W_{(l)}^{(ij)}$ and $a_{(l')}^{(i)}$ are independent: $l' = 0$ follows from definition of $W_{(l)}^{(ij)}$ and $a_0^{(i)} = x^{(i)}$. For $l' > 0$ we see that

$$a_{(l')}^{(i)} = \text{ReLU}(\sum_k W_{(l')}^{(ki)} a_{(l'-1)}^{(k)} + b_{(l')}^{(i)}).$$

Sum, multiplication and ReLU are measurable maps, and the $W_{l'}^{(ki)}, a_{(l'-1)}^{(k)}, b_{(l')}^{(i)}$ are independent of $W_{(l)}^{(ij)}$ by definition of induction hypothesis. Hence $a_{(l')}^{(i)}$ is independent or $W_{(l)}^{(ij)}$ too.

2. For all $l \in \{1, \dots, L\}$ we have :

$$\mathbb{E}[\sum_i W_{(l)}^{(ij)} a_{l-1)}^{(i)}] \overset{1, \text{ hint}}{=} \sum_i \underbrace{\mathbb{E}[W_{(l)}^{(ij)}]}_{=0} \mathbb{E}[a_{(l-1)}^{(i)}] = n \cdot 0 \cdot \mathbb{E}[a_{(l-1)}^{(i)}] = 0$$

$$\mathbb{E}[(W_{(l)}^{(ij)})^2] = \text{Var}(W_{(l)}^{(ij)}) + \mathbb{E}[W_{(l)}^{(ij)}]^2 = \text{Var}(W_{(l)}^{(ij)})$$

3. We show that for all $l \in \{1, \dots, L\}$ it holds $\text{Var}(z_{(l)}^{(i)}) = 2$.

   (a) $l = 1$:

$$\begin{aligned}\text{Var}(z_1^i) &= \text{Var}(\sum_k W_{(1)}^{(ki)} a_{(0)}^{(k)} + b_{(l)}^i) = \sum_k \text{Var}(W_{(1)}^{(ki)} x^{(k)}) \\ &\overset{1}{=} \sum_k \text{Var}(W_1^{(ik)}) \text{Var}(x^{(k)}) = \sum_k \frac{2}{n} \cdot 1 = 2.\end{aligned}$$

   (b) $l > 1$:

$$\begin{aligned}\text{Var}(z_{(l)}^{(i)}) =& \text{Var}(\sum_i W_{(l)}^{(ij)} a_{(l-1)}^{(i)} + b_{(l)}^{(i)})) \\ =& \text{Var}(\sum_i W_{(l)}^{(ij)} a_{(l-1)}^{(i)}) \\ =& \mathbb{E}[(\sum_i W_{(l)}^{(ij)} a_{(l)-1}^{(i)})^2] - \mathbb{E}[\sum_i W_{(l)}^{(ij)} a_{(l-1)}^{(i)}]^2 \\ =& \mathbb{E}[\sum_i W_{(l)}^{(ij)} a_{(l-1)}^{(i)} \sum_k W_{(l)}^{(kj)} a_{(l-1)}^{(k)}] - (0)^2 \\ =& \sum_i \mathbb{E}[(W_{(l)}^{(ij)} a_{(l-1)}^{(i)})^2] + \sum_i \sum_{k \neq i} \mathbb{E}[W_{(l)}^{(ij)} a_{(l-1)}^{(i)} W_{(l)}^{(kj)} a_{(l-1)}^{(k)}] \\ =& \sum_i \text{Var}(W_{(l)}^{(ij)}) \mathbb{E}[(a_{(l-1)}^{(i)})^2] \\ =& \sum_i \frac{2}{n} \mathbb{E}[(a_{(l-1)}^{(i)})^2] \\ =& \sum_i \frac{2}{n} \mathbb{E}[(\text{ReLU}(z_{(l-1)}^{(i)})^2] \\ =& \sum_i \frac{2}{n} \frac{1}{2} \text{Var}(z_{(l-1)}^{(i)}) \\ =& \frac{1}{n} \sum_i \text{Var}(z_{(l-1)}^{(i)}) \\ \overset{\text{induction hypothesis}}{=}& \frac{1}{n} \sum_i 2 = 2\end{aligned}$$

(b) Now assume instead of ReLU you choose tanh as an activation function. Consider what needs to be taken care of with this activation function and explain how to choose network parameter initialization accordingly.

Hint: Around 0, we have $\tanh(x) \approx x$. You can use this to approximate an expectation value that mights to be computed along the way.
Show your work.
We want $z_l^{(i)}$ to be in the linear regime of tanh for all l and i. This is achieve by following the proof from 2a) with the assumptions

$$W_{ij}^{(l)} \sim \mathcal{N}\left(0, \frac{1}{n}\right)$$
$$b_i^{(l)} = 0$$

In the adapted proof one then just has to make the approximation $\mathbb{E}[\tanh(z_{(l)}^{(i)})^2] = \mathbb{E}[z_{(l)}^{(i)2}]$.