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.

(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: 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:

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

where $W_{(l)}$ is the weight matrix of layer l and $b^{(l)}$ is the bias vector of layer l.
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 $z_{(l)}^{(i)} = \sum_j W_{(l)}^{(ij)} a_{(l-1)}^{(j)} + b_{(l)}^{(i)}$ for all layers $l \in \{2, ..., L\}$ stays constant, i.e. $\mathrm{Var}(\sum_j W_{(l+1)}^{(ij)} a_{(l)}^{(i)} + b_{(l+1)}^{(i)}) = \mathrm{Var}(\sum_j W_{(l)}^{(ij)} a_{(l-1)}^{(j)} + b_{(l)}^{(i)})$.

(b) How do you need to choose the initialization for tanh if you assume the following?

- Around 0, we have $\tanh(x) \approx x$.

Show your work.

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.

$\frac{\partial J}{\partial w_i} = \lambda \mathbb{E}_{\hat{p}}\left[\frac{1}{2}(w^T x - t)\cdot x_i\right]$ $H = \frac{\partial}{\partial w w^T}\mathbb{E}\left[\frac{1}{2}(w^T x - t)^2\right]$

$\frac{\partial J}{\partial w_j \partial w_i} = \mathbb{E}[x_j x_i]$ $= \frac{1}{2}(w^T x)(x^T w)$

$= \frac{\partial}{\partial w w^T}\mathbb{E}\left[\boxed{\frac{1}{2}(w^T x)(w^T x)} + \text{lin.} + \text{const}\right]$

$= \frac{\partial}{\partial w w^T}\mathbb{E}\left[\frac{1}{2}(w^T x)^2\right] = \mathbb{E}[x x^T]$

$= \text{Cov}(x, x^T) + \mathbb{E}[x]\mathbb{E}[x^T]$

$= \delta^2 I + \mu \cdot \mu^T$

$H = \mathbb{E}[x x^T]$

$= (\text{Cov}(X, X^T) + \mathbb{E}[x]\mathbb{E}[x^T])$

$= \delta^2 I + \mu \cdot \mu^T$

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

condition number: $\frac{\lambda_{max}}{\lambda_{min}}$

Assume $\vec{v}$ is the eigenvalue of λ

$H\vec{v} = \lambda\vec{v}$

$v^T H \vec{v} = \lambda v^T \vec{v} = \lambda (v_1\ v_2\ \cdots\ v_d)\begin{pmatrix} v_1 \\ v_2 \\ i \\ v_d \end{pmatrix} = \lambda \sum_{i=1}^{d} v_i^2 = \lambda \|\vec{v}\|^2$

$\lambda_1 = \lambda_{max} = \frac{v^T H \vec{v}}{\|\vec{v}\|} = \frac{1}{\|\vec{v}\|^2} v^T(\delta^2 I + \mu\mu^T)v$

$= \frac{\delta^2}{\|\vec{v}\|^2} v^T v + \frac{1}{\|\vec{v}\|^2} v^T \mu\mu^T v$

$= \delta^2 + \frac{1}{\|\vec{v}\|^2} v^T\mu(v^T\mu)^T$

$= \delta^2 + \frac{1}{\|\vec{v}\|^2}\cdot\|v^T\mu\|$

$= \delta^2 + \frac{\sum_i (v_i\mu_i)^2}{\sum_i (v_i)^2}$

$= \delta^2 + \sum_i \mu_i^2$

$= \delta^2 + \|\mu\|^2$

$\lambda_2 = \lambda_{min} = \delta^2$ (if $\|\mu\| = 0$)

$\lambda_1 = \lambda_{max} = \max_{\|v\|=1} v^T H v$

$= \max_{\|v\|=1} v^T(\delta^2 I + \mu\cdot\mu^T)v$

$= \max_{\|v\|=1} v^T v\cdot\delta^2 + v^T\mu\cdot\mu^T v$

$= \max_{\|v\|=1} \|v\|^2\delta^2 + (\mu^T v)^T\mu^T v$

$= \max_{\|v\|=1} \delta^2 + \|\mu^T v\|^2$

$= \max_{\|v\|=1} \delta^2 - \sum_i (\mu_i v_i)^2$

$= \delta^2 - \sum_i (\mu_i)^2$ ($\forall v_i,\ \mu_i = v_i$, i.e. $\mu = \vec{v}$)

$= \delta^2 - \|\mu\|^2$

$\lambda_d = \lambda_{min}$

$= \min_{\|v\|=1} \delta^2 - \sum_i (\mu_i v_i)^2$

$= \delta^2 - 0 = \delta^2$

RML Sheet 9

$$\frac{\lambda_1}{\lambda_2} = \frac{\delta^2 + \|\mu\|^2}{\delta^2} = 1 + \frac{\|\mu\|^2}{\delta^2}$$

(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.

Pro: centering makes $\|\mu\| = 0$ so that $\frac{\lambda_1}{\lambda_2} = 1 + \frac{0}{\delta^2} = 1$. The condition number will be lower and therefore it converge faster.

Centering makes condition number $\frac{\lambda_1}{\lambda_2}$ lower: $1 + \frac{0}{\delta^2} < 1 + \frac{\|\mu\|^2}{\delta^2}$ therfore, convergence is faster

Con: Centering may lead to a higher error cause the error function is influenced by the means and covariance of the data ($\mathcal{E}(w) = w^T \mathbb{E}[(x-\mu)(x-\mu)^T + \mu\mu^T + \lambda I]w$ + linear + const)

Model performance will likely decrease due to wrong functional form

Model based on centered data $f(x) = w^T(x - \mathbb{E}[x])$ doesn't contain the ground truth $f(x) = v^T x$

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:

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

where $W_{(l)}$ is the weight matrix of layer l and $b^{(l)}$ is the bias vector of layer l.
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 $z_{(l)}^{(i)} = \sum_j W_{(l)}^{(ij)} a_{(l-1)}^{(j)} + b_{(l)}^{(i)}$ for all layers $l \in \{2, ..., L\}$ stays constant, i.e. $\mathrm{Var}(\sum_j W_{(l+1)}^{(ij)} a_{(l)}^{(i)} + b_{(l+1)}^{(i)}) = \mathrm{Var}(\sum_j W_{(l)}^{(ij)} a_{(l-1)}^{(j)} + b_{(l)}^{(i)})$.

R solutions 要都表 $l=1$ 的情况

$$\mathbb{V}[z_l^i] = \mathbb{V}\left[\sum_j W_l^{ij} a_{l-1}^j + b_l^i\right] = \mathbb{V}\left[\sum_j W_l^{ij} a_{l-1}^j\right]$$
$$= \mathbb{E}\left[\left(\sum_j W_l^{ij} a_{l-1}^j\right)^2\right] - \mathbb{E}\left[\sum_j W_l^{ij} a_{lj}^j\right]^2$$

$$\mathbb{E}\left[\sum_j W_l^{ij} a_{lj}^j\right]^2 = \left(n \underbrace{\mathbb{E}[W_l^{ij}]}_{0} \cdot \mathbb{E}[a_{lj}^j]\right)^2 = 0$$

$$\mathbb{E}\left[\left(\sum_j W_l^{ij} a_{l-1}^j\right)^2\right] = \mathbb{E}\left[\sum_j \left((W_l^{ij} a_{l-1}^j)^2\right) + \sum_{j \neq m} W_l^{ij} a_{l-1}^j W_l^{im} a_{l-1}^m\right]$$

$$= \mathbb{E}\Big[\sum_j (w_\ell^{ij})^2 (a_{\ell-1}^j)^2\Big] + 0 \qquad \left(w_\ell^{ij}, a_{\ell-1}^j \text{ are independent, so } \sum_{j\neq m} \mathbb{E}[w_\ell^{ij} a_{\ell-1}^j w_\ell^{im} a_{\ell-1}^m] = 0\right)$$

$$= \sum_j \mathbb{E}[(w_\ell^{ij})^2] \cdot \mathbb{E}[(a_{\ell-1}^j)^2]$$

$$= \sum_j \Big(\mathbb{V}[(w_\ell^{ij})] + \mathbb{E}^2[w_\ell^{ij}]\Big) \cdot \mathbb{E}[(a_{\ell-1}^j)^2]$$

$$= \sum_j \left(\frac{2}{n} + 0\right) \cdot \mathbb{E}[(a_{\ell-1}^j)^2]$$

$$= \sum_j \frac{2}{n} \cdot \mathbb{E}[\mathrm{ReLu}(z_{\ell-1}^j)^2]$$

$$= \sum_j \frac{2}{n} \cdot \frac{1}{2} \mathbb{V}[z_{\ell-1}^j]$$

$$= n \cdot \frac{1}{n} \mathbb{V}[z_{\ell-1}^j]$$

$$= \mathbb{V}[z_{\ell-1}^j]$$

$$\mathbb{V}[z_\ell^i] = \mathbb{V}\Big[\sum_j w_\ell^{ij} a_{\ell-1}^j + b_\ell^i\Big] = \mathbb{E}\Big[\Big(\sum_j w_\ell^{ij} a_{\ell-1}^j\Big)^2\Big] - \mathbb{E}\Big[\sum_j w_\ell^{ij} a_{\ell j}^j\Big]^2$$

$$= \mathbb{V}[z_{\ell-j}^i]$$

(b) Now assume instead of ReLU you choose tanh as an activation function. How do you need to choose network parameter initialization if you want to achieve the result from (a) on constant variance of the latent variables? Hint: Around 0, we have $\tanh(x) \approx x$. Use this to approximate an expectation value. *Show* your work.

$$\tanh(z_\ell^i) = a_\ell^j$$

To satisfy: $\mathbb{V}[z_\ell^i] = \mathbb{V}[a_j^i]$

$$\mathbb{V}[z_\ell^i] = \mathbb{V}\Big[\sum_j w_\ell^{ij} a_{\ell-1}^j\Big] = \sum_j \Big(\mathbb{E}[(w_\ell^{ij} a_{\ell-1}^j)^2] - \underbrace{\mathbb{E}[w_\ell^{ij} a_{\ell-1}^j]^2}_{0}\Big)$$

$$= \sum_j \left(\mathbb{E}[(w_\ell^{ij} a_{\ell-1}^j)^2]\right)$$

$$= \mathbb{E}\Big[\sum_j \Big((w_\ell^{ij} a_{\ell-1}^j)^2\Big) + \sum_{j\neq m} w_\ell^{ij} a_{\ell-1}^j w_\ell^{im} a_{\ell-1}^m\Big]$$

$$= \mathbb{E}\Big[\sum_j (w_\ell^{ij})^2 (a_{\ell-1}^j)^2\Big] + 0$$

$$= \sum_j \mathbb{E}[(w_\ell^{ij})^2] \cdot \mathbb{E}[(a_{\ell-1}^j)^2]$$

$$= \sum_j \Big(\mathbb{V}[(w_\ell^{ij})] + \underbrace{\mathbb{E}^2[w_\ell^{ij}]}_{0}\Big) \cdot \mathbb{E}[(a_{\ell-1}^j)^2]$$

$$= \sum_j \Big(\mathbb{V}[w_\ell^{ij}]\Big) \cdot \mathbb{E}[(a_{\ell-1}^j)^2]$$

$$= \sum_j \Big(\mathbb{V}[w_\ell^{ij}]\Big) \cdot \mathbb{E}[\tanh(z_{\ell-1}^j)^2]$$

$$\approx \sum_j \Big(\mathbb{V}[w_\ell^{ij}]\Big) \cdot \mathbb{E}[(z_{\ell-1}^j)^2]$$

$$\overset{!}{=} \mathbb{V}[(z^i_{l-1})]$$

$$\Leftrightarrow \sum_j \left(\mathbb{V}[w^{ij}_l]\right) \cdot \mathbb{E}[(z^j_{l-1})^2] = \mathbb{V}[z^i_{l-1}] = \mathbb{E}[(z^i_{l-1})^2] - \mathbb{E}[z^i_{l-1}]^2$$

$$n\left(\mathbb{V}[w^{ij}_l]\right) \cdot \mathbb{E}[(z^j_{l-1})^2] = \mathbb{E}[(z^i_{l-1})^2] - \mathbb{E}[z^i_{l-1}]^2$$

$$\mathbb{V}[w^{ij}_l] = \frac{1}{n} - \frac{\mathbb{E}[z^i_{l-1}]^2 = 0}{n\,\mathbb{E}[(z^i_{l-1})^2]}$$

$$= \frac{1}{n}$$

$$\left(\text{Because of } \mathbb{E}[z^i_j] = \mathbb{E}\left[\sum_j w^{ij}_l a^j_{lj}\right]^2 = \left(n \underbrace{\mathbb{E}[w^{ij}_l]}_{0} \cdot \mathbb{E}[a^j_{lj}]\right)^2 = 0 \right)$$

$$\Rightarrow w^{ij}_l \sim \mathcal{N}\left(0, \frac{1}{n}\right)$$

$$b^i_l = 0$$

We make approximation $\mathbb{E}[\tanh(t^i_l)^2] = \mathbb{E}[z^{i\,2}_l]$

from b we have $\mathbb{V}(z^i_l) = \sum_i \frac{2}{n} \mathbb{E}[\tanh(t^i_l)^2]$

$$= \sum_i \frac{2}{n} \mathbb{E}[z^{i\,2}_l]$$

$$= \sum_i \frac{2}{n} \mathbb{E}\left[\left(\sum_j w^{ij}_l a^j_{l-1}\right)^2\right]$$

$$= \sum_i \frac{2}{n} \left(\mathbb{V}\left[\sum_j w^{ij}_l a^j_{l-1}\right] + 0\right)$$