Khoa Nguyen

Homework 1 - Theory

DS-GA 1008
Deep Learning
October 19, 2021

Following is my answers for the theory part of the homework 01 of the course Deep Learning at NYU, 2021 [LC21]. These answers are not verified to be correct. Please read with a big grain of salt.

1 Instructions

- Every vector is treated as a column vector
- Use numerator-layout notation[1] for matrix calculus.
- Only use vector and matrix (no tensor)
- Missing transpose are considered wrong.

2 Two-Layer Neural Networks

We have the following neural net:

$$\boldsymbol{x} \to \text{Linear}_1 \to f \to \text{Linear}_2 \to g \to \hat{\boldsymbol{y}},$$

where $\text{Linear}_i(x) = \boldsymbol{W^{(i)}}\boldsymbol{x} + \boldsymbol{b^{(i)}}$, and f, g are element-wise nonlinear activation functions. $\boldsymbol{x} \in \mathbb{R}^n$, $\hat{\boldsymbol{y}} \in \mathbb{R}^K$.

3 Regression Task

Choose $f = \text{ReLU}$, and g to be an identity function. We choose mean square error (MSE) as the loss function: $l_{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2$.

3.1 Training Overview

Task

Name and mathematically describe the 5 programming steps you would take to train this model with PyTorch using SGD on a single batch of data.

- Step 1: set all the gradients to zeros as

$$\frac{\partial l_{MSE}}{\partial \boldsymbol{W^{(i)}}} = 0, \frac{\partial l_{MSE}}{\partial \boldsymbol{b^{(i)}}} = 0$$

[1] https://en.wikipedia.org/wiki/Matrix_calculus#Numerator-layout_notation

Homework 1 - Theory

- Step 2: do the forward pass, put $\boldsymbol{x}$ through the network and get $\hat{\boldsymbol{y}}$ (see more detailed mathematical explanations in 3.2).
- Step 3: calculate the loss $l_{MSE}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2$
- Step 4: calculate the gradients of the loss function with respect to the weights

$$\frac{\partial l_{MSE}}{\partial \boldsymbol{W}^{(i)}}, \frac{\partial l_{MSE}}{\partial \boldsymbol{b}^{(i)}}$$

- Step 5: update the weights according to the calculated gradients as

$$\boldsymbol{W}^{(i)} \leftarrow \boldsymbol{W}^{(i)} - \gamma \frac{\partial l_{MSE}}{\partial \boldsymbol{W}^{(i)}}, \tag{1}$$

$$\boldsymbol{b}^{(i)} \leftarrow \boldsymbol{b}^{(i)} - \gamma \frac{\partial l_{MSE}}{\partial \boldsymbol{b}^{(i)}} \tag{2}$$

3.2 Forward Pass

Task

For a single data point $(\boldsymbol{x}, \boldsymbol{y})$, write down all inputs and outputs for forward pass of each layer. You can only use variable $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)}, \boldsymbol{W}^{(2)}, \boldsymbol{b}^{(2)}$ in your answer (note that $\text{Linear}_i(x) = \boldsymbol{W}^{(i)}\boldsymbol{x} + \boldsymbol{b}^{(i)}$).

Layer	Input	Output
Linear_1	$\boldsymbol{x}$	$\boldsymbol{z}_1 = \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$
f	$\boldsymbol{z}_1$	$\boldsymbol{z}_2 = \text{ReLU}(\boldsymbol{z}_1)$
Linear_2	$\boldsymbol{z}_2$	$\boldsymbol{z}_3 = \boldsymbol{W}^{(2)}\boldsymbol{z}_2 + \boldsymbol{b}^{(2)}$
g	$\boldsymbol{z}_3$	$\hat{\boldsymbol{y}} = \boldsymbol{z}_3$
Loss	$\hat{\boldsymbol{y}}, \boldsymbol{y}$	$\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2$

We have the input and output data as

$$\boldsymbol{x}_{n\times 1} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n. \quad \boldsymbol{y}_{K\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{bmatrix} \in \mathbb{R}^K. \tag{3}$$

The weight matrix and the bias vector of the first linear layer is

$$\boldsymbol{W}^{(1)}_{h\times n} = \begin{bmatrix} w^{(1)}_{11} & w^{(1)}_{12} & \dots & w^{(1)}_{1n} \\ w^{(1)}_{21} & w^{(1)}_{22} & \dots & w^{(1)}_{2n} \\ \vdots & & \ddots & \vdots \\ w^{(1)}_{h1} & w^{(1)}_{h2} & \dots & w^{(1)}_{hn} \end{bmatrix} \in \mathbb{R}^{h\times n}; \quad \boldsymbol{b}^{(1)}_{h\times 1} = \begin{bmatrix} b^{(1)}_1 \\ b^{(1)}_2 \\ \vdots \\ b^{(1)}_n \end{bmatrix} \in \mathbb{R}^h, \tag{4}$$

Homework 1 - Theory

where h is the hidden dimension.

The output of the first linear layer is

$$\boldsymbol{z}_1 = \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)} = \begin{bmatrix} w_{11}^{(1)}x_1 + w_{12}^{(1)}x_2 + \cdots + w_{1n}^{(1)}x_n + b_1^{(1)} \\ w_{21}^{(1)}x_1 + w_{22}^{(1)}x_2 + \cdots + w_{2n}^{(1)}x_n + b_2^{(1)} \\ \vdots \\ w_{h1}^{(1)}x_1 + w_{h2}^{(1)}x_2 + \ldots + w_{hn}^{(1)}x_n + b_h^{(1)} \end{bmatrix} = \begin{bmatrix} z_1^{(1)} \\ z_2^{(1)} \\ \vdots \\ z_h^{(1)} \end{bmatrix} \in \mathbb{R}^h. \tag{5}$$

Applying non-linear function, we get

$$\boldsymbol{z}_2 = \text{ReLU}\left(\boldsymbol{z}_1\right) = \begin{bmatrix} z_1^{(2)} \\ z_2^{(2)} \\ \vdots \\ z_h^{(2)} \end{bmatrix} \in \mathbb{R}^h, \text{ where } z_i^{(2)} = \begin{cases} z_i^{(1)} & \text{if } z_i^{(1)} \geqslant 0 \\ 0 & \text{if } z_i^{(1)} < 0 \end{cases} \tag{6}$$

For the second linear layer, the weight matrix and the bias vector are

$$\boldsymbol{W}_{K\times h}^{(2)} = \begin{bmatrix} w_{11}^{(2)} & w_{12}^{(2)} & \cdots & w_{1h}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{k1}^{(2)} & w_{k2}^{(2)} & \cdots & w_{Kh}^{(2)} \end{bmatrix} \in \mathbb{R}^{K\times h}; \quad \boldsymbol{b}_{K\times 1}^{(2)} = \begin{bmatrix} b_1^{(2)} \\ b_2^{(2)} \\ \vdots \\ b_K^{(2)} \end{bmatrix} \in \mathbb{R}^K. \tag{7}$$

As we assumed that g is an identity function, or $\hat{\boldsymbol{y}} = \boldsymbol{z}_3$, therefore

$$\hat{\boldsymbol{y}}_{K\times 1} = \boldsymbol{z}_3 = \boldsymbol{W}^{(2)}\boldsymbol{z}_2 + \boldsymbol{b}^{(2)} = \begin{bmatrix} w_{11}^{(2)}z_1^{(2)} + w_{12}^{(2)}z_2^{(2)} + \cdots + w_{1n}^{(2)}z_n^{(2)} + b_1^{(2)} \\ w_{21}^{(2)}z_1^{(2)} + w_{22}^{(2)}z_2^{(2)} + \cdots + w_{2n}^{(2)}z_n^{(2)} + b_2^{(2)} \\ \vdots \\ w_{K1}^{(2)}z_1^{(2)} + w_{K2}^{(2)}z_2^{(2)} + \ldots + w_{Kn}^{(2)}z_n^{(2)} + b_K^{(2)} \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_K \end{bmatrix} \in \mathbb{R}^K. \tag{8}$$

The loss function (MSE) is

$$l = \frac{1}{k}[(\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 + \ldots + (\hat{y}_K - y_K)^2]. \tag{9}$$

3.3 Backward Pass

Task

Write down the gradient calculated from the backward pass. You can only use the following variables: $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)}, \boldsymbol{W}^{(2)}, \boldsymbol{b}^{(2)}, \frac{\partial l}{\partial \hat{\boldsymbol{y}}}, \frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1}, \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3}$ in your answer.

Homework 1 - Theory

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{b}^{(2)}} &= \frac{\partial l}{\partial \hat{\boldsymbol{y}}} \cdot \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3} \cdot \frac{\partial \boldsymbol{z}_3}{\partial \boldsymbol{b}^{(2)}} \\ &= \frac{\partial l}{\partial \hat{\boldsymbol{y}}} \cdot \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3} \cdot 1,\end{aligned} \tag{10}$$

Furthermore, as g is the identity function, $\hat{\boldsymbol{y}} = \boldsymbol{z}_3 \Rightarrow \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3} = \mathbb{I}$, where $\mathbb{I}$ is the identity matrix. The simplified versions of the above equation would be:

$$\frac{\partial l}{\partial \boldsymbol{b}^{(2)}} = \frac{\partial l}{\partial \hat{\boldsymbol{y}}}. \tag{11}$$

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{W}^{(2)}} &= \frac{\partial l}{\partial \hat{\boldsymbol{y}}} \cdot \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3} \cdot \frac{\partial \boldsymbol{z}_3}{\partial \boldsymbol{W}^{(2)}} \\ &= \frac{\partial l}{\partial \hat{\boldsymbol{y}}} \cdot \mathbb{I} \cdot \boldsymbol{z}_2^\intercal. \\ &= \frac{\partial l}{\partial \hat{\boldsymbol{y}}} \cdot [\text{ReLU}(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})]^\intercal.\end{aligned} \tag{12}$$

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{W}^{(1)}} &= \frac{\partial l}{\partial \hat{\boldsymbol{y}}} \cdot \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3} \cdot \frac{\partial \boldsymbol{z}_3}{\partial \boldsymbol{z}_2} \cdot \frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1} \cdot \frac{\partial \boldsymbol{z}_1}{\partial \boldsymbol{W}^{(1)}} \\ &= \frac{\partial l}{\partial \hat{\boldsymbol{y}}} \cdot \quad \boldsymbol{W}^{(2)^\intercal} \cdot \frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1} \cdot \boldsymbol{x}^\intercal.\end{aligned} \tag{13}$$

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{b}^{(1)}} &= \frac{\partial l}{\partial \hat{\boldsymbol{y}}} \cdot \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3} \cdot \frac{\partial \boldsymbol{z}_3}{\partial \boldsymbol{z}_2} \cdot \frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1} \cdot \frac{\partial \boldsymbol{z}_1}{\partial \boldsymbol{b}^{(1)}} \\ &= \frac{\partial l}{\partial \hat{\boldsymbol{y}}} \cdot \quad \boldsymbol{W}^{(2)^\intercal} \cdot \frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1}.\end{aligned} \tag{14}$$

3.4 Elements of The Derivatives

Task

Show the elements of $\frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1}$, $\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3}$, $\frac{\partial l}{\partial \hat{\boldsymbol{y}}}$ (be careful about the dimensionality)?

Note that the derivatives here follow the numerator-layout notation.

The derivative of the loss with respect to (w.r.t) $\hat{\boldsymbol{y}}$ is

$$\begin{aligned}\frac{\partial l}{\partial \hat{\boldsymbol{y}}} &= \frac{1}{k}\begin{bmatrix}\frac{\partial l}{\partial \hat{y}_1} & \frac{\partial l}{\partial \hat{y}_2} & \frac{\partial l}{\partial \hat{y}_3} & \cdots \frac{\partial l}{\partial \hat{y}_k}\end{bmatrix} \\ &= \frac{1}{k}\begin{bmatrix}2(\hat{y} - y_1) & 2(\hat{y}_2 - y_2) & \ldots & 2(\hat{y}_k - y_k)\end{bmatrix} \in \mathbb{R}^K.\end{aligned} \tag{15}$$

Homework 1 - Theory

The derivative of $\hat{\boldsymbol{y}}$ w.r.t $\boldsymbol{z}_3$ is

$$\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3} = \frac{\partial \hat{\boldsymbol{y}}}{\partial \hat{\boldsymbol{y}}} = \begin{bmatrix} \frac{\partial \hat{y}_1}{\partial \hat{y}_1} & \frac{\partial \hat{y}_1}{\partial \hat{y}_2} & \cdots & \frac{\partial \hat{y}_1}{\partial \hat{y}_k} \\ \frac{\partial \hat{y}_2}{\partial \hat{y}_1} & \frac{\partial \hat{y}_2}{\partial \hat{y}_2} & \cdots & \frac{\partial \hat{y}_2}{\partial \hat{y}_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \hat{y}_k}{\partial \hat{y}_1} & \frac{\partial \hat{y}_k}{\partial \hat{y}_2} & \cdots & \frac{\partial \hat{y}_k}{\partial \hat{y}_k} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = \mathbb{I}_{K \times K}. \tag{16}$$

The derivative of $\boldsymbol{z}_2$ w.r.t $\boldsymbol{z}_1$ is

$$\frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1} = \begin{bmatrix} \frac{\partial z_1^{(2)}}{\partial z_1^{(1)}} & \frac{\partial z_1^{(2)}}{\partial z_2^{(1)}} & \cdots & \frac{\partial z_1^{(2)}}{\partial z_h^{(1)}} \\ \frac{\partial z_2^{(2)}}{\partial z_1^{(1)}} & \frac{\partial z_2^{(2)}}{\partial z_2^{(1)}} & \cdots & \frac{\partial z_2^{(2)}}{\partial z_h^{(1)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial z_h^{(2)}}{\partial z_1^{(1)}} & \frac{\partial z_h^{(2)}}{\partial z_2^{(1)}} & \cdots & \frac{\partial z_h^{(2)}}{\partial z_h^{(1)}} \end{bmatrix} = \begin{bmatrix} \frac{\partial z_1^{(2)}}{\partial z_1^{(1)}} & 0 & \dots & 0 \\ 0 & \frac{\partial z_2^{(2)}}{\partial z_2^{(1)}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{\partial z_h^{(2)}}{\partial z_h^{(1)}} \end{bmatrix} \in \mathbb{R}^{h \times h}. \tag{17}$$

We have $\frac{\partial z_i^{(2)}}{\partial z_i^{(1)}} = 1$ if $z_i^{(1)} \geq 0$, otherwise $\frac{\partial z_i^{(2)}}{\partial z_i^{(1)}} = 0$.

4 Classification Task

We would like to perform multi-class classification task, so we set both $f, g = \sigma$, where σ is the logistic sigmoid function: $\sigma(z) = \frac{1}{1+e^{-z}}$.

4.1 Using Sigmoid instead of ReLU

Task

If you want to train this network, what do you need to change in the equations of the forward pass, backward pass, and elements of the derivatives in section 3, assuming we are using the same MSE loss function.

4.1.1 Forward Pass

Layer	Input	Output
Linear_1	$\boldsymbol{x}$	$\boldsymbol{z}_1 = \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$
f	$\boldsymbol{z}_1$	$\boldsymbol{z}_2 = \sigma(\boldsymbol{z}_1)$
Linear_2	$\boldsymbol{z}_2$	$\boldsymbol{z}_3 = \boldsymbol{W}^{(2)}\boldsymbol{z}_2 + \boldsymbol{b}^{(2)}$
g	$\boldsymbol{z}_3$	$\hat{\boldsymbol{y}} = \sigma(\boldsymbol{z}_3)$
Loss	$\hat{\boldsymbol{y}}, \boldsymbol{y}$	$\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2$

Homework 1 - Theory

In the forward pass, the equation for $\boldsymbol{z}_2$ changes to

$$\boldsymbol{z}_2 = \sigma\left(\boldsymbol{z}_1\right) = \begin{bmatrix} z_1^{(2)} \\ z_2^{(2)} \\ \vdots \\ z_h^{(2)} \end{bmatrix} \in \mathbb{R}^h, \text{ where } z_i^{(2)} = \frac{1}{1+e^{-z_i^{(1)}}}. \tag{18}$$

Similarly, for each element of $\hat{\boldsymbol{y}}$ we have $\hat{y}_i = \frac{1}{1+e^{-z_i^{(3)}}}$, where

$$\boldsymbol{z}_3 = \boldsymbol{W}^{(2)}\boldsymbol{z}_2 + \boldsymbol{b}^{(2)} = \begin{bmatrix} w_{11}^{(2)} z_1^{(2)} + w_{12}^{(2)} z_2^{(2)} + \cdots + w_{1n}^{(2)} z_n^{(2)} + b_1^{(2)} \\ w_{21}^{(2)} z_1^{(2)} + w_{22}^{(2)} z_2^{(2)} + \cdots + w_{2n}^{(2)} z_n^{(2)} + b_2^{(2)} \\ \vdots \\ w_{K1}^{(2)} z_1^{(2)} + w_{K2}^{(2)} z_2^{(2)} + \ldots + w_{Kn}^{(2)} z_n^{(2)} + b_K^{(2)} \end{bmatrix} = \begin{bmatrix} z_1^{(3)} \\ z_2^{(3)} \\ \vdots \\ z_K^{(3)} \end{bmatrix} \in \mathbb{R}^K.$$

4.1.2 Backward Pass and The Elements of the Derivatives

First, let's derive the derivative of the sigmoid function. The sigmoid function is

$$y = \sigma(z) = \frac{1}{1+e^{-z}},$$

where y and z are two scalar values. Let $u(z) = 1 + e^{-z}$, we have $y = \frac{1}{u}$. Using the chain rule for the derivative we have

$$\begin{aligned}
\frac{\partial y}{\partial z} &= \frac{\partial y}{\partial u} * \frac{\partial u}{\partial z} \\
&= -\frac{1}{u^2} * (-e^{-z}) \\
&= -\frac{1}{(1+e^{-z})^2} * (-e^{-z}) \\
&= \frac{e^{-z}}{(1+e^{-z})^2} \\
&= \frac{(1+e^{-z})-1}{(1+e^{-z})^2} \\
&= \frac{1}{1+e^{-z}} - \frac{1}{(1+e^{-z})^2} \\
&= \frac{1}{1+e^{-z}} * (1 - \frac{1}{1+e^{-z}}) \\
&= y(1-y)
\end{aligned}$$

Homework 1 - Theory

The derivative of $\hat{\boldsymbol{y}}$ w.r.t $\boldsymbol{z}_3$ is

$$\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3} = \begin{bmatrix} \frac{\partial \hat{y}_1}{\partial z_1^{(3)}} & \frac{\partial \hat{y}_1}{\partial z_2^{(3)}} & \cdots & \frac{\partial \hat{y}_1}{\partial z_k^{(3)}} \\ \frac{\partial \hat{y}_2}{\partial z_1^{(3)}} & \frac{\partial \hat{y}_2}{\partial z_2^{(3)}} & \cdots & \frac{\partial \hat{y}_2}{\partial z_k^{(3)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \hat{y}_k}{\partial z_1^{(3)}} & \frac{\partial \hat{y}_k}{\partial z_2^{(3)}} & \cdots & \frac{\partial \hat{y}_k}{\partial z_k^{(3)}} \end{bmatrix} = \begin{bmatrix} \hat{y}_1(1-\hat{y}_1) & \hat{y}_1(1-\hat{y}_1) & \cdots & \hat{y}_1(1-\hat{y}_1) \\ \hat{y}_2(1-\hat{y}_2) & \hat{y}_2(1-\hat{y}_2) & \cdots & \hat{y}_2(1-\hat{y}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{y}_k(1-\hat{y}_k) & \hat{y}_k(1-\hat{y}_k) & \cdots & \hat{y}_k(1-\hat{y}_k) \end{bmatrix}. \tag{19}$$

Similarly, the derivative of $\boldsymbol{z}_2$ w.r.t $\boldsymbol{z}_1$ is

$$\frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1} = \begin{bmatrix} \frac{\partial z_1^{(2)}}{\partial z_1^{(1)}} & \frac{\partial z_1^{(2)}}{\partial z_2^{(1)}} & \cdots & \frac{\partial z_1^{(2)}}{\partial z_h^{(1)}} \\ \frac{\partial z_2^{(2)}}{\partial z_1^{(1)}} & \frac{\partial z_2^{(2)}}{\partial z_2^{(1)}} & \cdots & \frac{\partial z_2^{(2)}}{\partial z_h^{(1)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial z_h^{(2)}}{\partial z_1^{(1)}} & \frac{\partial z_h^{(2)}}{\partial z_2^{(1)}} & \cdots & \frac{\partial z_h^{(2)}}{\partial z_h^{(1)}} \end{bmatrix} = \begin{bmatrix} z_1^{(2)}(1-z_1^{(2)}) & z_1^{(2)}(1-z_1^{(2)}) & \cdots & z_1^{(2)}(1-z_1^{(2)}) \\ z_2^{(2)}(1-z_2^{(2)}) & z_2^{(2)}(1-z_2^{(2)}) & \cdots & z_2^{(2)}(1-z_2^{(2)}) \\ \vdots & \vdots & \ddots & \vdots \\ z_h^{(2)}(1-z_h^{(2)}) & z_h^{(2)}(1-z_h^{(2)}) & \cdots & z_h^{(2)}(1-z_h^{(2)}) \end{bmatrix}. \tag{20}$$

Now, we can plug these information into the backpropagation

$$\frac{\partial l}{\partial \boldsymbol{b}^{(2)}} = \frac{\partial l}{\partial \hat{\boldsymbol{y}}} \cdot \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3} \tag{21}$$

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{W}^{(2)}} &= \frac{\partial l}{\partial \hat{\boldsymbol{y}}} \cdot \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3} \cdot \frac{\partial \boldsymbol{z}_3}{\partial \boldsymbol{W}^{(2)}} \\ &= \frac{\partial l}{\partial \hat{\boldsymbol{y}}} \cdot \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3} \cdot \boldsymbol{z}_2^\intercal \end{aligned} \tag{22}$$

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{W}^{(1)}} &= \frac{\partial l}{\partial \hat{\boldsymbol{y}}} \cdot \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3} \cdot \frac{\partial \boldsymbol{z}_3}{\partial \boldsymbol{z}_2} \cdot \frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1} \cdot \frac{\partial \boldsymbol{z}_1}{\partial \boldsymbol{W}^{(1)}} \\ &= \frac{\partial l}{\partial \hat{\boldsymbol{y}}} \cdot \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3} \cdot \boldsymbol{W}^{(2)\intercal} \cdot \frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1} \cdot \boldsymbol{x}^\intercal. \end{aligned} \tag{23}$$

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{b}^{(1)}} &= \frac{\partial l}{\partial \hat{\boldsymbol{y}}} \cdot \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3} \cdot \frac{\partial \boldsymbol{z}_3}{\partial \boldsymbol{z}_2} \cdot \frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1} \cdot \frac{\partial \boldsymbol{z}_1}{\partial \boldsymbol{b}^{(1)}} \\ &= \frac{\partial l}{\partial \hat{\boldsymbol{y}}} \cdot \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}_3} \cdot \boldsymbol{W}^{(2)\intercal} \cdot \frac{\partial \boldsymbol{z}_2}{\partial \boldsymbol{z}_1}. \end{aligned} \tag{24}$$

4.2 Using Binary Cross Entropy (BCE) Loss

Now you think you can do a better job by using BCE:

$$l_{BCE}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{1}{K} \sum_{i=1}^{K} -[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{25}$$

Homework 1 - Theory

Task

What do you need to change in the equations of the forward pass, backward pass, and in the elements of the derivatives?

4.2.1 Forward Pass

Layer	Input	Output
Linear_1	$\boldsymbol{x}$	$\boldsymbol{z}_1 = \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$
f	$\boldsymbol{z}_1$	$\boldsymbol{z}_2 = \sigma(\boldsymbol{z}_1)$
Linear_2	$\boldsymbol{z}_2$	$\boldsymbol{z}_3 = \boldsymbol{W}^{(2)}\boldsymbol{z}_2 + \boldsymbol{b}^{(2)}$
g	$\boldsymbol{z}_3$	$\hat{\boldsymbol{y}} = \sigma(\boldsymbol{z}_3)$
Loss	$\hat{\boldsymbol{y}}, \boldsymbol{y}$	$\frac{1}{K}\sum_{i=1}^{K} -[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]$

4.2.2 Backward Pass and the Elements of the Derivatives

The only thing that changes now is the derivative of the loss function with respect to $\hat{\boldsymbol{y}}$, $\frac{\partial l}{\partial \hat{\boldsymbol{y}}}$. We have l as a scalar, and $\hat{\boldsymbol{y}}_{K\times 1}$ is a column vector. According to the numerator-layout notation, $\frac{\partial l}{\partial \hat{\boldsymbol{y}}}$ will be a row vector of length K as

$$\frac{\partial l}{\partial \hat{\boldsymbol{y}}} = \begin{bmatrix} \frac{\partial l}{\partial \hat{y}_1} & \frac{\partial l}{\partial \hat{y}_2} \cdots \frac{\partial l}{\partial \hat{y}_K} \end{bmatrix} \tag{26}$$

with each element as

$$\begin{aligned}
\frac{\partial l}{\partial \hat{y}_i} &= -\frac{1}{K}\left[\frac{\partial y_i \log(\hat{y}_i)}{\partial \hat{y}_i} + \frac{\partial (1-y_i)\log(1-\hat{y}_i)}{\partial \hat{y}_i}\right] \\
&= -\frac{1}{K}\left[y_i \frac{\partial \log(\hat{y}_i)}{\partial \hat{y}_i} + (1-y_i)\frac{\partial \log(1-\hat{y}_i)}{\partial \hat{y}_i}\right] \\
&= -\frac{1}{K}\left[y_i \frac{1}{\hat{y}_i} + (1-y_i)\frac{-1}{1-\hat{y}_i}\right] \\
&= -\frac{1}{K}\left(\frac{y_i}{\hat{y}_i} - \frac{1}{1-\hat{y}_i} + \frac{y_i}{1-\hat{y}_i}\right) \\
&= -\frac{1}{K}\frac{y_i(1-\hat{y}_i) - \hat{y}_i + y_i\hat{y}_i}{\hat{y}_i(1-\hat{y}_i)} \\
&= \frac{1}{K}\frac{\hat{y}_i - y_i}{\hat{y}_i(1-\hat{y}_i)}
\end{aligned} \tag{27}$$

Homework 1 - Theory

4.3 Using one ReLU and one Sigmoid

Task

You realize that not all hidden activations need to be a soft version of binary (output of sigmoid). You decide to use $f(\cdot) = (\cdot)^+ = \text{ReLU}(\cdot)$, but keep g as σ. Explain why this choice of f is beneficial for training (deeper) neural networks?

With these changes, we now have the final neural network as

Layer	Input	Output
Linear_1	$\boldsymbol{x}$	$\boldsymbol{z}_1 = \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$
f	$\boldsymbol{z}_1$	$\boldsymbol{z}_2 = \text{ReLU}\left(\boldsymbol{z}_1\right)$
Linear_2	$\boldsymbol{z}_2$	$\boldsymbol{z}_3 = \boldsymbol{W}^{(2)}\boldsymbol{z}_2 + \boldsymbol{b}^{(2)}$
g	$\boldsymbol{z}_3$	$\hat{\boldsymbol{y}} = \sigma(\boldsymbol{z}_3)$
Loss	$\hat{\boldsymbol{y}}, \boldsymbol{y}$	$\frac{1}{K}\sum_{i=1}^{K} -[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]$

At the end of the neural network, where we often need to have $\hat{\boldsymbol{y}}$ as a vector of values between 0 and 1 (probabilities), using the sigmoid (or softmax) functions is a good choice. However, for the activation maps in the hidden layers, we do not need to squash their values to be in the range $(0, 1)$. The ReLU function is a non-linear function that keeps the positive values of its input, therefore it is scale equivariant and is a better choice for f to train deeper networks.

References

[LC21] Yann LeCun and Alfredo Canziani. "DS-GA 1008: Deep Learning". In: (2021). URL: https://atcold.github.io/NYU-DLSP21/.