# CSCI-GA 2572 Deep Learning
# Homework 1: Backpropagation

Chuanyang Jin

February 12, 2023

## 1.1 Two-Layer Neural Nets

You are given the following neural net architecture:

$$\text{Linear}_1 \to f \to \text{Linear}_2 \to g$$

where $\text{Linear}_i(x) = \boldsymbol{W}^{(i)}\boldsymbol{x} + \boldsymbol{b}^{(i)}$ is the i-th affine transformation, and $f, g$ are element-wise nonlinear activation functions. When an input $\mathbf{x} \in \mathbb{R}^n$ is fed to the network, $\tilde{\boldsymbol{y}} \in \mathbb{R}^K$ is obtained as the output.

## 1.2 Regression Task

We would like to perform regression task. We choose $f(\cdot) = 3(\cdot)^+ = 3\text{ReLU}(\cdot)$ and $g$ to be the identity function. To train this network, we want to minimize the energy loss $L$ and this is computed via the squared Euclidean distance cost C, such that $L(\boldsymbol{w}, \boldsymbol{x}, \boldsymbol{y}) = F(\boldsymbol{x}, \boldsymbol{y}) = C(\boldsymbol{y}, \tilde{\boldsymbol{y}}) = ||\tilde{\boldsymbol{y}} - \boldsymbol{y}||^2$, where $\mathbf{y}$ is the output target.

(a) (1pt) 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.

*Solution.*

(1) Generate a prediction: $\boldsymbol{y} = \text{model}(\boldsymbol{x})$

(2) Compute the loss: $L(\boldsymbol{w}, \boldsymbol{x}, \boldsymbol{y}) = F(\boldsymbol{x}, \boldsymbol{y}) = C(\boldsymbol{y}, \tilde{\boldsymbol{y}}) = ||\tilde{\boldsymbol{y}} - \boldsymbol{y}||^2$

(3) Clear the gradients: optimizer.zero_grad()

(4) Back propagation to compute and accumulate the gradients: $L$.backward()

(5) Update the parameters: optimizer.step()

□

(b) (4pt) For a single data point (x, y), write down all inputs and outputs for forward pass of each layer. You can only use variable $\boldsymbol{x}, \boldsymbol{y}, \mathbf{W}^{(1)}, \boldsymbol{b}^{(1)}, \boldsymbol{W}^{(2)}, \boldsymbol{b}^{(2)}$ in your answer. (note that $\text{Linear}_i(x) = \boldsymbol{W}^{(i)}\boldsymbol{x} + \boldsymbol{b}^{(i)}$).

*Solution.*

(1) $\text{Linear}_1 : \boldsymbol{x} \to \boldsymbol{s}_1,\ \boldsymbol{s}_1 = \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$

(2) $f : \boldsymbol{s}_1 \to \boldsymbol{a}_1,\ \boldsymbol{a}_1 = 3\text{ReLU}(\boldsymbol{s}_1) = 3\text{ReLU}(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})$

(3) $\text{Linear}_2 : \boldsymbol{a}_1 \to \boldsymbol{s}_2,\ \boldsymbol{s}_2 = \boldsymbol{W}^{(2)}\boldsymbol{a}_1 + \boldsymbol{b}^{(2)} = 3\boldsymbol{W}^{(2)}\text{ReLU}(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}) + \boldsymbol{b}^{(2)}$

(4) $g : \boldsymbol{s}_2 \to \tilde{\boldsymbol{y}},\ \tilde{\boldsymbol{y}} = s^{(2)} = 3\boldsymbol{W}^{(2)}\text{ReLU}(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}) + \boldsymbol{b}^{(2)}$

□

(c) (6pt) Write down the gradients calculated from the backward pass. You can only use the following variables: $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)}, \boldsymbol{W}^{(2)}, \boldsymbol{b}^{(2)}, \frac{\partial C}{\partial \tilde{\boldsymbol{y}}}, \frac{\partial \boldsymbol{a}_1}{\partial \boldsymbol{s}_1}, \frac{\partial \tilde{\boldsymbol{y}}}{\partial \boldsymbol{s}_2}$ in your answer, where $\boldsymbol{s}_1, \boldsymbol{a}_1, \boldsymbol{s}_2, \tilde{\boldsymbol{y}}$ are the outputs of $\text{Linear}_1, f, \text{Linear}_2, g$.

*Solution.*

(1)

$$\frac{\partial C}{\partial \tilde{\boldsymbol{y}}} \text{ as given}$$

(2)

$$\frac{\partial C}{\partial \boldsymbol{s}_2} = \frac{\partial C}{\partial \tilde{\boldsymbol{y}}} \frac{\partial \tilde{\boldsymbol{y}}}{\partial \boldsymbol{s}_2}$$

$$\frac{\partial C}{\partial \boldsymbol{W}^{(2)}} = \frac{\partial C}{\partial \boldsymbol{s}_2} \frac{\partial \boldsymbol{s}_2}{\partial \boldsymbol{W}^{(2)}} = \boldsymbol{a}_1 \frac{\partial C}{\partial \tilde{\boldsymbol{y}}} \frac{\partial \tilde{\boldsymbol{y}}}{\partial \boldsymbol{s}_2} = 3\text{ReLU}(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}) \frac{\partial C}{\partial \tilde{\boldsymbol{y}}} \frac{\partial \tilde{\boldsymbol{y}}}{\partial \boldsymbol{s}_2}$$

$$\frac{\partial C}{\partial \boldsymbol{b}^{(2)}} = \frac{\partial C}{\partial \boldsymbol{s}_2} \frac{\partial \boldsymbol{s}_2}{\partial \boldsymbol{b}^{(2)}} = \frac{\partial C}{\partial \tilde{\boldsymbol{y}}} \frac{\partial \tilde{\boldsymbol{y}}}{\partial \boldsymbol{s}_2}$$

(3)

$$\frac{\partial C}{\partial \boldsymbol{a}_1} = \frac{\partial C}{\partial \boldsymbol{s}_2} \frac{\partial \boldsymbol{s}_2}{\partial \boldsymbol{a}_1} = \frac{\partial C}{\partial \tilde{\boldsymbol{y}}} \frac{\partial \tilde{\boldsymbol{y}}}{\partial \boldsymbol{s}_2} \boldsymbol{W}^{(2)}$$

(4)

$$\frac{\partial C}{\partial \boldsymbol{s}_1} = \frac{\partial C}{\partial \boldsymbol{a}_1} \frac{\partial \boldsymbol{a}_1}{\partial \boldsymbol{s}_1} = \frac{\partial C}{\partial \tilde{\boldsymbol{y}}} \frac{\partial \tilde{\boldsymbol{y}}}{\partial \boldsymbol{s}_2} \boldsymbol{W}^{(2)} \cdot \frac{\partial \boldsymbol{a}_1}{\partial \boldsymbol{s}_1}$$

$$\frac{\partial C}{\partial \boldsymbol{W}^{(1)}} = \frac{\partial C}{\partial \boldsymbol{s}_1} \frac{\partial \boldsymbol{s}_1}{\partial \boldsymbol{W}^{(1)}} = \boldsymbol{x} \frac{\partial C}{\partial \tilde{\boldsymbol{y}}} \frac{\partial \tilde{\boldsymbol{y}}}{\partial \boldsymbol{s}_2} \boldsymbol{W}^{(2)} \frac{\partial \boldsymbol{a}_1}{\partial \boldsymbol{s}_1}$$

$$\frac{\partial C}{\partial \boldsymbol{b}^{(1)}} = \frac{\partial C}{\partial \boldsymbol{s}_1} \frac{\partial \boldsymbol{s}_1}{\partial \boldsymbol{b}^{(1)}} = \frac{\partial C}{\partial \tilde{\boldsymbol{y}}} \frac{\partial \tilde{\boldsymbol{y}}}{\partial \boldsymbol{s}_2} \boldsymbol{W}^{(2)} \frac{\partial \boldsymbol{a}_1}{\partial \boldsymbol{s}_1}$$

□

(d) (2pt) Show us the elements of $\frac{\partial \boldsymbol{a}_1}{\partial \boldsymbol{s}_1}$, $\frac{\partial \tilde{\boldsymbol{y}}}{\partial \boldsymbol{s}_2}$ and $\frac{\partial C}{\partial \tilde{\boldsymbol{y}}}$ (be careful about the dimensionality)?

*Solution.*

(1) $\frac{\partial \boldsymbol{a}_1}{\partial \boldsymbol{s}_1} = M$ where $M$ is a diagonal matrix such that

$$M[i][j] = \begin{cases} 3 & i = j, \boldsymbol{s}_1[i] \geq 0 \\ 0 & i = j, \boldsymbol{s}_1[i] < 0 \\ 0 & i \neq j \end{cases}$$

(2) $\frac{\partial \tilde{\boldsymbol{y}}}{\partial \boldsymbol{s}_2} = I$, where $I$ is a $K \times K$ identity matrix.

(3) $\frac{\partial C}{\partial \tilde{\boldsymbol{y}}} = 2(\tilde{\boldsymbol{y}} - \boldsymbol{y})^T$, which is a row vector of length $K$.

□

## 1.3 Classification Task

We would like to perform multi-class classification task, so we set $f = tanh$ and $g = \sigma$, the logistic sigmoid function $\sigma(x) = (1 + \exp(-x))^{-1}$.

(a) (4pt + 6pt + 2pt) If you want to train this network, what do you need to change in the equations of (b), (c) and (d), assuming we are using the same squared Euclidean distance loss function.

*Solution.*

We need to change in the equations of (b) that:

(1) $f : \boldsymbol{s}_1 \to \boldsymbol{a}_1$, $\boldsymbol{a}_1 = \tanh(\boldsymbol{s}_1) = \tanh(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})$

(2) The form of $\text{Linear}_2$ doesn't change. Only the value of $\boldsymbol{a}_1$ in it changes.
$\text{Linear}_2 : \boldsymbol{a}_1 \to \boldsymbol{s}_2$, $\boldsymbol{s}_2 = \boldsymbol{W}^{(2)}\boldsymbol{a}_1 + \boldsymbol{b}^{(2)} = \boldsymbol{W}^{(2)}\tanh(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}) + \boldsymbol{b}^{(2)}$

(3) $g : \boldsymbol{s}_2 \to \tilde{\boldsymbol{y}}$, $\tilde{\boldsymbol{y}} = \sigma(s^{(2)}) = (1 + \exp(-\boldsymbol{W}^{(2)}\tanh(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}) - \boldsymbol{b}^{(2)}))^{-1}$

The forms of the equations of (c) doesn't change.
We need to change in (d) that:

(1) $\frac{\partial \boldsymbol{a}_1}{\partial \boldsymbol{s}_1} = M'$ where $M'$ is a diagonal matrix such that

$$M[i][j] = \begin{cases} 1 - \tanh^2(\boldsymbol{s}_1[i]) & i = j \\ 0 & i \neq j \end{cases}$$

(2) $\frac{\partial \tilde{\boldsymbol{y}}}{\partial \boldsymbol{s}_2} = N'$ where $N$ is a diagonal matrix such that

$$N[i][j] = \begin{cases} \sigma((\boldsymbol{s}_2[i])^2)(1 - \sigma((\boldsymbol{s}_2[i])^2)) & i = j \\ 0 & i \neq j \end{cases}$$

□

(b) (4pt + 6pt + 2pt) Now you think you can do a better job by using a Binary Cross Entropy (BCE) loss function $D_{\text{BCE}}(\boldsymbol{y}, \tilde{\boldsymbol{y}}) = \frac{1}{K}\sum_{i=1}^{K} -[y_i \log(\tilde{y}_i) + (1 - y_i)\log(1 - \tilde{y}_i)]$. What do you need to change in the equations of (b), (c) and (d)?

*Solution.*

We still need to make all the changes in 1.3 (a). Besides that, we need to change in (d) that:

(1)

$$\frac{\partial C}{\partial \tilde{\boldsymbol{y}}} = -\frac{1}{K}\frac{\boldsymbol{y}^T}{\tilde{\boldsymbol{y}}^T} + \frac{1}{K}\frac{1 - \boldsymbol{y}^T}{1 - \tilde{\boldsymbol{y}}^T}$$

□

(c) (1pt) Things are getting better. You realize that not all intermediate hidden activations need to be binary (or soft version of binary). You decide to use $f(\cdot) = (\cdot)^+$ but keep $g$ as tanh. Explain why this choice of $f$ can be beneficial for training a (deeper) network.

*Solution.*

This choice of $f$ can be beneficial for training a (deeper) network since

(1) the ReLU activation avoids saturation, which slows down or stops the learning process, by only saturating for negative inputs while the sigmoid activation saturates for both large positive and negative inputs;

(2) ReLU activation is computationally more efficient.

□

## 1.4 Conceptual Questions

(a) (1pt) Why is softmax actually softargmax?

*Solution.*

$$\text{softmax}_\beta(e) = \frac{1}{\beta} \log \frac{1}{N} \sum_{n=1}^{N} \exp(\beta e_n)$$

This real softmax can produce a probability distribution over multiple classes.

$$\text{softargmax}_\beta(e) = \frac{\exp(\beta e)}{\sum_{n=1}^{N} \exp(\beta e_n)}$$

The softmax we use as the activation function for the output layer is actually the softargmax. It can select one class with the maximum probability. □

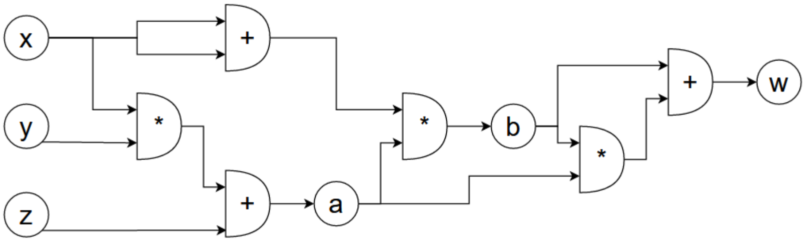
(b) (3pt) Draw the computational graph defined by this function, with inputs $x, y, z \in \mathbb{R}$ and output $w \in \mathbb{R}$. You make use symbols $x$, $y$, $z$, $o$, and operators $*$, $+$ in your solution. Be sure to use the correct shape for symbols and operators as shown in class.

$$a = xy + z$$

$$b = a(x + x)$$

$$w = ab + b$$

*Solution.*

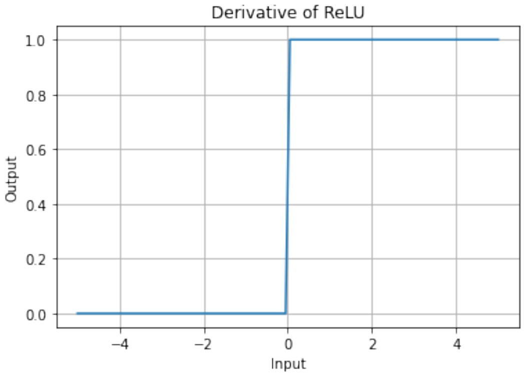


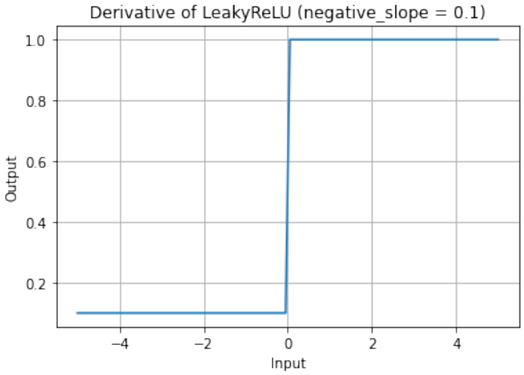
□

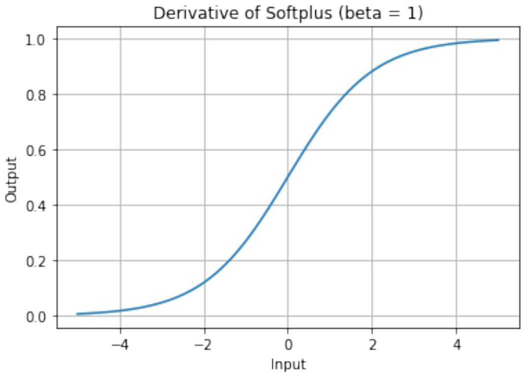
(c) (2pt) Draw the graph of the derivative for the following functions?

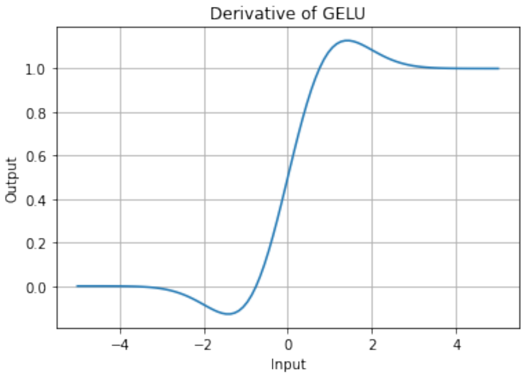
- ReLU()
- LeakyReLU(negative_slope = 0.1)
- Softplus(beta = 1)
- GELU()

*Solution.*


Derivative of ReLU
1.0
0.8
0.6
0.4
0.2
0.0
Output
−4
−2
0
2
4
Input


Derivative of LeakyReLU (negative_slope = 0.1)
1.0
0.8
0.6
0.4
0.2
Output
−4
−2
0
2
4
Input


Derivative of Softplus (beta = 1)
1.0
0.8
0.6
0.4
0.2
0.0
Output
−4
−2
0
2
4
Input


Derivative of GELU
1.0
0.8
0.6
0.4
0.2
0.0
Output
−4
−2
0
2
4
Input

□

(d) (3pt) Given function $f(x) = \boldsymbol{W}_1\boldsymbol{x}$ with $\boldsymbol{W}_1 \in \mathbb{R}^{b\times a}$ and $g(\boldsymbol{x}) = \boldsymbol{W}_2\boldsymbol{x}$ with $W_2 \in \mathbb{R}^{b\times a}$

(a) What is the Jacobian matrix of $f$ and $g$

(b) What is the Jacobian matrix of $h(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x})$

(c) What is the Jacobian matrix of $h(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x})$ if $\boldsymbol{W}_1 = \boldsymbol{W}_2$

*Solution.*

(a) The Jacobian matrix of $f$ is $\boldsymbol{W}_1$ and the Jacobian matrix of $g$ is $\boldsymbol{W}_2$.

(b) The Jacobian matrix of $h(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x})$ is $\boldsymbol{W}_1 + \boldsymbol{W}_2$.

(c) If $\boldsymbol{W}_1 = \boldsymbol{W}_2$, the Jacobian matrix of $h(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x})$ is $2\boldsymbol{W}_1$.

□

(e) (3pt) Given function $f(\boldsymbol{x}) = \boldsymbol{W}_1\boldsymbol{x}$ with $\boldsymbol{W}_1 \in \mathbb{R}^{b\times a}$ and $g(\boldsymbol{x}) = \boldsymbol{W}_2\boldsymbol{x}$ with $\boldsymbol{W}_2 \in \mathbb{R}^{c\times b}$

(a) What is the Jacobian matrix of $f$ and $g$

(b) What is the Jacobian matrix of $h(\boldsymbol{x}) = g(f(\boldsymbol{x})) = (g \circ f)(\boldsymbol{x})$

(c) What is the Jacobian matrix of $h(\boldsymbol{x})$ if $\boldsymbol{W}_1 = \boldsymbol{W}_2$ (so $a = b = c$)

*Solution.*

(a) The Jacobian matrix of $f$ is $\boldsymbol{W}_1$ and the Jacobian matrix of $g$ is $\boldsymbol{W}_2$.

(b) The Jacobian matrix of $h(\boldsymbol{x}) = g(f(\boldsymbol{x})) = (g \circ f)(\boldsymbol{x})$ is $\boldsymbol{W}_2\boldsymbol{W}_1$.

(c) If $\boldsymbol{W}_1 = \boldsymbol{W}_2$, the Jacobian matrix of $h(\boldsymbol{x})$ is $\boldsymbol{W}_1\boldsymbol{W}_1$.

□