Homework 1

1.1 Regression Task

(a)
- Forward the data through the model
- Compute loss
- Zero gradients
- Step in the opposite direction of the gradient

(b) Forward pass. Given f is ReLU and g is the identity function.

Layer	**Input**	**Output**
Linear_1	x	$W^{(1)}x + b^{(1)}$
f	$W^{(1)}x + b^{(1)}$	$\left[W^{(1)}x + b^{(1)}\right]^+$
Linear_2	$\left[W^{(1)}x + b^{(1)}\right]^+$	$W^{(2)}\left[W^{(1)}x + b^{(1)}\right]^+ + b^{(2)}$
g	$W^{(2)}\left[W^{(1)}x + b^{(1)}\right]^+ + b^{(2)}$	$W^{(2)}\left[W^{(1)}x + b^{(1)}\right]^+ + b^{(2)}$
Loss	$W^{(2)}\left[W^{(1)}x + b^{(1)}\right]^+ + b^{(2)},\ y$	$\frac{1}{2}\left(W^{(2)}\left[W^{(1)}x + b^{(1)}\right]^+ + b^{(2)} - y\right)^2$

(c) Backward pass.

Parameter	**Gradient**
$W^{(1)}$	$\frac{\partial l}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_3}W^{(2)}\frac{\partial z_2}{\partial z_1}x$
$b^{(1)}$	$\frac{\partial l}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_3}W^{(2)}\frac{\partial z_2}{\partial z_1}$
$W^{(2)}$	$\frac{\partial l}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_3}\,z_2 = \frac{\partial l}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_3}\left[W^{(1)}x + b^{(1)}\right]^+$
$b^{(2)}$	$\frac{\partial l}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_3}$

(d)
$$\frac{\partial z_2}{\partial z_1} = \begin{cases} 1, & \text{if } z_1 > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\frac{\partial \hat{y}}{\partial z_3} = 1$$

$$\frac{\partial l}{\partial \hat{y}} = \hat{y} - y$$

1.2 Classification Task

(a) Forward Pass

Layer	Input	Output
Linear_1	x	$W^{(1)}x + b^{(1)}$
f	$W^{(1)}x + b^{(1)}$	$\left[1 + \exp\left(-W^{(1)}x - b^{(1)}\right)\right]^{-1}$
Linear_2	$\left[1 + \exp\left(-W^{(1)}x - b^{(1)}\right)\right]^{-1}$	$W^{(2)}\left[1 + \exp\left(-W^{(1)}x - b^{(1)}\right)\right]^{-1} + b^{(2)}$
g	$W^{(2)}\left[1 + \exp\left(-W^{(1)}x - b^{(1)}\right)\right]^{-1} + b^{(2)}$	$\left[1 + \exp\left(-W^{(2)}\left[1 + \exp\left(-W^{(1)}x - b^{(1)}\right)\right]^{-1} - b^{(2)}\right)\right]^{-1}$
Loss	$\left[1 + \exp\left(-W^{(2)}\left[1 + \exp\left(-W^{(1)}x - b^{(1)}\right)\right]^{-1} - b^{(2)}\right)\right]^{-1}, \ y$	$\frac{1}{2}\left(\left[1 + \exp\left(-W^{(2)}\left[1 + \exp\left(-W^{(1)}x - b^{(1)}\right)\right]^{-1} - b^{(2)}\right)\right]^{-1} - y\right)^2$

Backward Pass (Need to change $W^{(2)}$)

Parameter	Gradient
$W^{(1)}$	$\frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} W^{(2)} \frac{\partial z_2}{\partial z_1} x$
$b^{(1)}$	$\frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} W^{(2)} \frac{\partial z_2}{\partial z_1}$
$W^{(2)}$	$\frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} \left[1 + \exp\left(-W^{(1)}x - b^{(1)}\right)\right]^{-1}$
$b^{(2)}$	$\frac{\partial l}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3}$

Elements of $\frac{\partial l}{\partial \hat{y}}$, $\frac{\partial \hat{y}}{\partial z_3}$, $\frac{\partial z_2}{\partial z_1}$

$$\frac{\partial l}{\partial \hat{y}} = \hat{y} - y \qquad \text{(Same as 1.1)}$$

$$\frac{\partial \hat{y}}{\partial z_3} = \frac{e^{-z_3}}{\left(1 + e^{(-z_3)}\right)^2}$$

$$\frac{\partial z_2}{\partial z_1} = \frac{e^{-z_1}}{\left(1 + e^{(-z_1)}\right)^2}$$

(b) Forward Pass

Layer	Input	Output
Linear_1	x	$W^{(1)}x + b^{(1)}$
f	$W^{(1)}x + b^{(1)}$	$\left[1+\exp\left(-W^{(1)}x - b^{(1)}\right)\right]^{-1}$
Linear_2	$\left[1+\exp\left(-W^{(1)}x - b^{(1)}\right)\right]^{-1}$	$W^{(2)}\left[1+\exp\left(-W^{(1)}x - b^{(1)}\right)\right]^{-1} + b^{(2)}$
g	$W^{(2)}\left[1+\exp\left(-W^{(1)}x - b^{(1)}\right)\right]^{-1} + b^{(2)}$	$\left[1+\exp\left(-W^{(2)}\left[1+\exp\left(-W^{(1)}x - b^{(1)}\right)\right]^{-1} - b^{(2)}\right)\right]^{-1}$
Loss	$\left[1+\exp\left(-W^{(2)}\left[1+\exp\left(-W^{(1)}x - b^{(1)}\right)\right]^{-1} - b^{(2)}\right)\right]^{-1},\ y$	$-y\log\left(\left[1+\exp\left(-W^{(2)}\left[1+\exp\left(-W^{(1)}x - b^{(1)}\right)\right]^{-1} - b^{(2)}\right)\right]^{-1}\right)$ $-(1-y)\log\left(1-\left[1+\exp\left(-W^{(2)}\left[1+\exp\left(-W^{(1)}x - b^{(1)}\right)\right]^{-1} - b^{(2)}\right)\right]^{-1}\right)$

Backward Pass

Parameter	Gradient
$W^{(1)}$	$\frac{\partial l}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_3}W^{(2)}\frac{\partial z_2}{\partial z_1}x$
$b^{(1)}$	$\frac{\partial l}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_3}W^{(2)}\frac{\partial z_2}{\partial z_1}$
$W^{(2)}$	$\frac{\partial l}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_3}\left[1+\exp\left(-W^{(1)}x - b^{(1)}\right)\right]^{-1}$
$b^{(2)}$	$\frac{\partial l}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial z_3}$

Elements of $\frac{\partial l}{\partial \hat{y}}$, $\frac{\partial \hat{y}}{\partial z_3}$, $\frac{\partial z_2}{\partial z_1}$

$$\frac{\partial l}{\partial \hat{y}} = -\left[\frac{y}{\hat{y}} + \left(-\frac{1-y}{1-\hat{y}}\right)\right] = \frac{1-y}{1-\hat{y}} - \frac{y}{\hat{y}} = \frac{\hat{y}-y}{\hat{y}(1-\hat{y})}$$

$$\frac{\partial \hat{y}}{\partial z_3} = \frac{e^{-z_3}}{\left(1+e^{(-z_3)}\right)^2} \qquad \text{(Same as 1.2 (a))}$$

$$\frac{\partial z_2}{\partial z_1} = \frac{e^{-z_1}}{\left(1 + e^{(-z_1)}\right)^2} \qquad \text{(Same as 1.2 (a))}$$

(c) Benefites of using ReLU:

- Using ReLU can prevent the problem of vanishing gradients, which can happen if using Sigmoid.
- Gradient of ReLU is easier to interpret and faster to compute than the gradient of Sigmoid.