Homework 1: Theory

Yuxiang Chai

September 27, 2022

1.1 Two-Layer Neural Nets

1.2 Regression Task

(a) i. $\hat{\boldsymbol{y}} = \text{model}(\boldsymbol{x})$. First we need to generate the prediction of the given model and input.

ii. $\ell = \ell_{\text{MSE}} = \|\hat{\boldsymbol{y}} - \boldsymbol{y}\|^2$. Second we compute the loss between the prediction and the ground truth.

iii. `optimizer.zero_grad()`. Then we need to set the optimizer with zero gradient. The optimizer can be SGD, Adam or others.

iv. ℓ.backward(). Then we compute the gradient.

v. `optimizer.step()`. Last we update the weight of each parameter.

(b) i. For $\texttt{Linear}_1$, the input is $\boldsymbol{x}$, and the output is $\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}}$.

ii. For f, the input is $\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}}$, and the output is $5 \cdot (\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}})^+$.

iii. For $\texttt{Linear}_2$, the input is $5 \cdot (\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}})^+$, and the output is $5 \cdot \boldsymbol{W^{(2)}}(\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}})^+ + \boldsymbol{b^{(2)}}$

iv. For g, the input is $5 \cdot \boldsymbol{W^{(2)}}(\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}})^+ + \boldsymbol{b^{(2)}}$, and the output is $5 \cdot \boldsymbol{W^{(2)}}(\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}})^+ + \boldsymbol{b^{(2)}}$

(c) i.
$$\frac{\partial \ell}{\partial \boldsymbol{b^{(2)}}} = \frac{\partial \ell}{\partial \boldsymbol{z_3}} \frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{b^{(2)}}} = \frac{\partial \ell}{\partial \boldsymbol{z_3}} = \frac{\partial \ell}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}$$

ii.
$$\frac{\partial \ell}{\partial \boldsymbol{W^{(2)}}} = \frac{\partial \ell}{\partial \boldsymbol{z_3}} \frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{W^{(2)}}} = \frac{\partial \ell}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}} \frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{W^{(2)}}} = 5 \cdot \frac{\partial \ell}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}} ((\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}})^+)^\top$$

iii.
$$\frac{\partial \ell}{\partial \boldsymbol{b^{(1)}}} = \frac{\partial \ell}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}} \frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{z_2}} \frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}} \frac{\partial \boldsymbol{z_1}}{\partial \boldsymbol{b^{(1)}}} = \frac{\partial \ell}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}} \boldsymbol{W^{(2)}} \frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}$$

iv.
$$\frac{\partial \ell}{\partial \boldsymbol{W^{(1)}}} = \frac{\partial \ell}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}} \frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{z_2}} \frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}} \frac{\partial \boldsymbol{z_1}}{\partial \boldsymbol{W^{(1)}}} = \frac{\partial \ell}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}} \boldsymbol{W^{(2)}} \frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}} \boldsymbol{x}^\top$$

(d) i. $\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}$ is a $m \times m$ matrix $\boldsymbol{M}$ if $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$ is $\mathbb{R}^m$, where $\boldsymbol{M}_{ij} = 0$ if $i \neq j$ and $\boldsymbol{M}_{ii} = 5$ if $\boldsymbol{z_1}[i] > 0$. Assume that `ReLU()` gradient at 0 is 0

ii. $\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}$ is a $K \times K$ identity matrix $\boldsymbol{I}$.

iii. $\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}$ is a K dimension row vector $\boldsymbol{V}$, where $\boldsymbol{V} = 2(\hat{\boldsymbol{y}} - \boldsymbol{y})^\top$

1.3 Classification Task

(a) (b) i. For $\texttt{Linear}_1$, the input is $\boldsymbol{x}$, and the output is $\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}}$.

ii. For f, the input is $\boldsymbol{W^{(1)}x}+\boldsymbol{b^{(1)}}$, and the output is $\tanh(\boldsymbol{W^{(1)}x}+\boldsymbol{b^{(1)}})$.

iii. For $\texttt{Linear}_2$, the input is $\tanh(\boldsymbol{W^{(1)}x}+\boldsymbol{b^{(1)}})$, and the output is $\boldsymbol{W^{(2)}}\tanh(\boldsymbol{W^{(1)}x}+\boldsymbol{b^{(1)}})+\boldsymbol{b^{(2)}}$

iv. For g, the input is $\boldsymbol{W^{(2)}}\tanh(\boldsymbol{W^{(1)}x}+\boldsymbol{b^{(1)}})+\boldsymbol{b^{(2)}}$, and the output is $\sigma(\boldsymbol{W^{(2)}}\tanh(\boldsymbol{W^{(1)}x}+\boldsymbol{b^{(1)}})+\boldsymbol{b^{(2)}})$

(c) i.
$$\frac{\partial \ell}{\partial \boldsymbol{b^{(2)}}}=\frac{\partial \ell}{\partial \boldsymbol{z_3}}\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{b^{(2)}}}=\frac{\partial \ell}{\partial \boldsymbol{z_3}}=\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}$$

ii.
$$\frac{\partial \ell}{\partial \boldsymbol{W^{(2)}}}=\frac{\partial \ell}{\partial \boldsymbol{z_3}}\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{W^{(2)}}}=\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{W^{(2)}}}=\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}\tanh(\boldsymbol{W^{(1)}x}+\boldsymbol{b^{(1)}})^\top$$

iii.
$$\frac{\partial \ell}{\partial \boldsymbol{b^{(1)}}}=\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{z_2}}\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}\frac{\partial \boldsymbol{z_1}}{\partial \boldsymbol{b^{(1)}}}=\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}\boldsymbol{W^{(2)}}\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}$$

iv.
$$\frac{\partial \ell}{\partial \boldsymbol{W^{(1)}}}=\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{z_2}}\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}\frac{\partial \boldsymbol{z_1}}{\partial \boldsymbol{W^{(1)}}}=\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}\boldsymbol{W^{(2)}}\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}\boldsymbol{x}^\top$$

(d) i. $\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}$ is a $m\times m$ matrix $\boldsymbol{M}$ if $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$ is $\mathbb{R}^m$, where $\boldsymbol{M}_{ij}=0$ if $i\neq j$ and $\boldsymbol{M}_{ii}=1-\boldsymbol{z_2}_i^2$.

ii. $\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}$ is a $K\times K$ matrix $\boldsymbol{P}$, where $\boldsymbol{P}_{ij}=0$ if $i\neq j$ and $\boldsymbol{P}_{ii}=\hat{\boldsymbol{y}}_i(1-\hat{\boldsymbol{y}}_i)$.

iii. $\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}$ is a K dimension row vector $\boldsymbol{V}$, where $\boldsymbol{V}=2(\hat{\boldsymbol{y}}-\boldsymbol{y})^\top$

(b) (b) i. For $\texttt{Linear}_1$, the input is $\boldsymbol{x}$, and the output is $\boldsymbol{W^{(1)}x}+\boldsymbol{b^{(1)}}$.

ii. For f, the input is $\boldsymbol{W^{(1)}x}+\boldsymbol{b^{(1)}}$, and the output is $\tanh(\boldsymbol{W^{(1)}x}+\boldsymbol{b^{(1)}})$.

iii. For $\texttt{Linear}_2$, the input is $\tanh(\boldsymbol{W^{(1)}x}+\boldsymbol{b^{(1)}})$, and the output is $\boldsymbol{W^{(2)}}\tanh(\boldsymbol{W^{(1)}x}+\boldsymbol{b^{(1)}})+\boldsymbol{b^{(2)}}$

iv. For g, the input is $\boldsymbol{W^{(2)}}\tanh(\boldsymbol{W^{(1)}x}+\boldsymbol{b^{(1)}})+\boldsymbol{b^{(2)}}$, and the output is $\sigma(\boldsymbol{W^{(2)}}\tanh(\boldsymbol{W^{(1)}x}+\boldsymbol{b^{(1)}})+\boldsymbol{b^{(2)}})$

(c) i.
$$\frac{\partial \ell}{\partial \boldsymbol{b^{(2)}}}=\frac{\partial \ell}{\partial \boldsymbol{z_3}}\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{b^{(2)}}}=\frac{\partial \ell}{\partial \boldsymbol{z_3}}=\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}$$

ii.
$$\frac{\partial \ell}{\partial \boldsymbol{W^{(2)}}}=\frac{\partial \ell}{\partial \boldsymbol{z_3}}\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{W^{(2)}}}=\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{W^{(2)}}}=\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}\tanh(\boldsymbol{W^{(1)}x}+\boldsymbol{b^{(1)}})^\top$$

iii.
$$\frac{\partial \ell}{\partial \boldsymbol{b^{(1)}}}=\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{z_2}}\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}\frac{\partial \boldsymbol{z_1}}{\partial \boldsymbol{b^{(1)}}}=\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}\boldsymbol{W^{(2)}}\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}$$

iv.
$$\frac{\partial \ell}{\partial \boldsymbol{W^{(1)}}}=\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{z_2}}\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}\frac{\partial \boldsymbol{z_1}}{\partial \boldsymbol{W^{(1)}}}=\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}\boldsymbol{W^{(2)}}\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}\boldsymbol{x}^\top$$

(d) i. $\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}$ is a $m\times m$ matrix $\boldsymbol{M}$ if $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$ is $\mathbb{R}^m$, where $\boldsymbol{M}_{ij}=0$ if $i\neq j$ and $\boldsymbol{M}_{ii}=1-\boldsymbol{z_2}_i^2$.

ii. $\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}$ is a $K\times K$ matrix $\boldsymbol{P}$, where $\boldsymbol{P}_{ij}=0$ if $i\neq j$ and $\boldsymbol{P}_{ii}=\hat{\boldsymbol{y}}_i(1-\hat{\boldsymbol{y}}_i)$.

iii. $\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}$ is a K dimension row vector $\boldsymbol{V}$, where $\boldsymbol{V}_i=\frac{1}{K}(\frac{1-\boldsymbol{y}_i}{1-\hat{\boldsymbol{y}}_i}-\frac{\boldsymbol{y}_i}{\hat{\boldsymbol{y}}_i})$

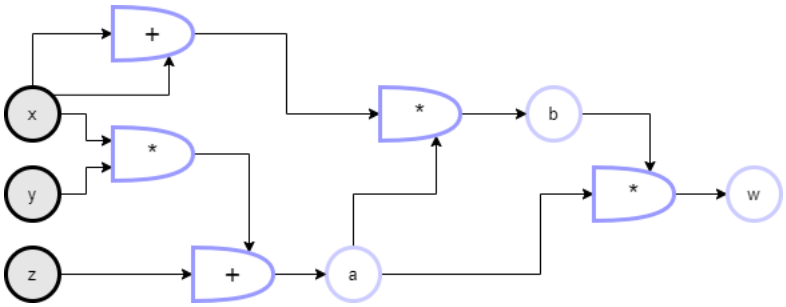
(c) First, using `ReLU` can prevent vanishing gradient problem in `sigmoid` or `tanh` because of the gradient of `ReLU` is 1 or 0.

And using `ReLU` is faster than using `sigmoid` or `tanh`. As we can see in the previous parts, the gradient computation of `ReLU` is easier than those of `sigmoid` or `tanh`.

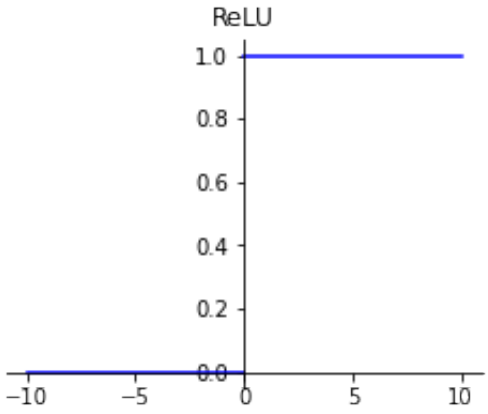
1.4 Conceptual Questions

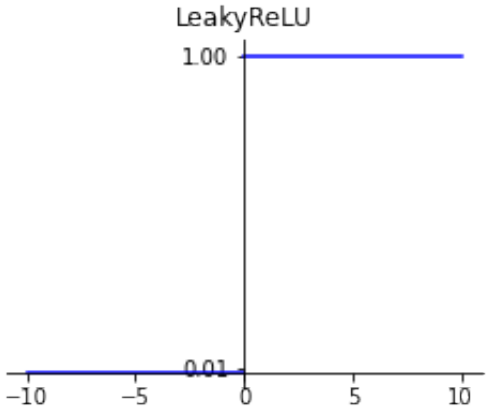
(a) Because the definition of pytorch `softmax` is $\boldsymbol{z}_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$, where the denominator for every element is the same. So the larger the numerator the larger $\boldsymbol{z}_i$. So `softmax` is actually choosing the index that has the largest probability.

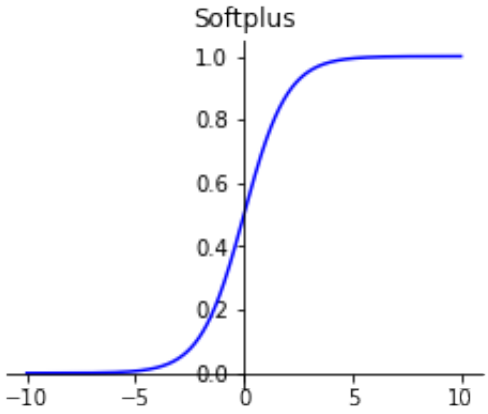
(b)

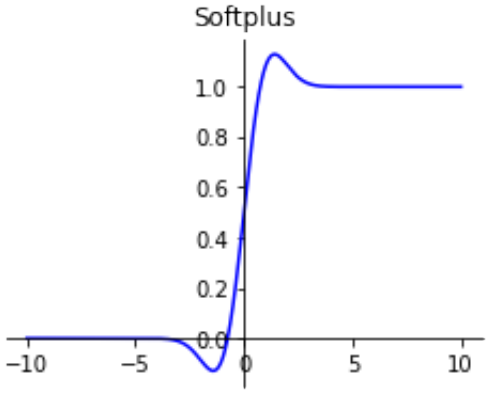


(c)









(d) Four types: Scaling, Rotation, Translation and Shearing.

Linear transformation can separate the current space.

Non-linear transformation is to prevent the model from becoming a linear transformation and non-linearly separate space.

(e) For every iteration, do the following:

i. $\hat{Y} = F_\theta(D)$, where $\hat{Y} = \hat{y}_1, \hat{y}_2, ..., \hat{y}_N$

ii. $\ell = \ell_{\text{MSE}} = \|\hat{Y} - Y\|^2$

iii. Compute gradients of θ, $\frac{\partial \ell}{\partial \theta}$

iv. Update θ, $\theta = \theta - \eta \frac{\partial \ell}{\partial \theta}$