Homework 1

Guilherme Albertini

September 23, 2022

Theory

Let $Linear_1 \to f \to Linear_2 \to g$ be a be a two-layer neural net architecture whereby $Linear_i(x) = \boldsymbol{W}^{(i)}\boldsymbol{x} + \boldsymbol{b}^{(i)}$ is the i^{th} affine transformation and f, g are element-wise nonlinear activation functions (else must use transposes and/or Hadamard products). When an input $\boldsymbol{x} \in \mathbb{R}^n$ is fed into the network, $\hat{\boldsymbol{y}} \in \mathbb{R}^K$ is obtained as output.

Problem 1: Regression Task

We would like to perform regression task. We choose $f(\cdot) = 5(\cdot)^+ = 5ReLU(\cdot)$ and g to be the identity function. To train, we choose MSE loss function, $\ell_{MSE}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = ||(\hat{\boldsymbol{y}} - \boldsymbol{y})||^2$.

1. 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.

   (a) First compute a prediction from the model (the forward pass); $\tilde{y} = model(x)$

   (b) Second compute the loss through computation of the energy

   (c) Zero the gradient parameters; $optimiser.zero\_grad()$

   (d) Compute and accumulate gradient parameters; $L.backward()$

   (e) Finally step in the opposite direction of the gradient; $optimiser.step()$

2. For a single data point (x, y), write down all inputs and outputs for forward pass of each layer. You can only use variables and mechanics specified prior in your answer.

—$Linear_1$—
Input: $\boldsymbol{x}$
Output (z_1): $\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$
—f—
Input: $Linear_1(\boldsymbol{x}) = \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$
Output (z_2): $f(Linear_1(\boldsymbol{x})) = 5ReLU(Linear_1(\boldsymbol{x}))$
$= 5ReLU(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})$
$= 5\max(0, \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})$
—$Linear_2$—
Input: $f(Linear_1(\boldsymbol{x}))$
Output (z_3): $Linear_2(f(Linear_1(\boldsymbol{x}))) = \boldsymbol{W}^{(2)}\boldsymbol{f(Linear_1(x))} + \boldsymbol{b}^{(2)}$
$= 5\boldsymbol{W}^{(2)}((\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})) + \boldsymbol{b}^{(2)}$
—g—
Input: $Linear_2(f(Linear_1(\boldsymbol{x})))$
Output: $g(Linear_2(f(Linear_1(\boldsymbol{x})))) =$
$5(\boldsymbol{W}^{(2)}(\max(0, \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})) + \boldsymbol{b}^{(2)})\boldsymbol{I} =$
$5(\boldsymbol{W}^{(2)}(\max(0, \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})) + \boldsymbol{b}^{(2)}) = \hat{\boldsymbol{y}}$
—Loss—
Input: $\hat{\boldsymbol{y}}$
Output: $||(5(\boldsymbol{W}^{(2)}(\max(0, \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})) + \boldsymbol{b}^{(2)}) - \boldsymbol{y})||^2$

3. Write down the gradients calculated from the backward pass. You can only use the following variables: $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{W^{(1)}}, \boldsymbol{b^{(1)}}, \boldsymbol{W^{(2)}}, \boldsymbol{b^{(2)}}, \frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}, \frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}, \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}$, where $\boldsymbol{z_1}, \boldsymbol{z_2}, \boldsymbol{z_3}, \hat{\boldsymbol{y}}$ are outputs of $Linear_1, f, Linear_2, g$, respectively.

$$
\begin{aligned}
\frac{\partial \boldsymbol{z_{3i}}}{\partial \boldsymbol{W^{(2)}}} &= \frac{\partial(5\boldsymbol{W^{(2)}}\,\mathbf{max}(\mathbf{0}, \boldsymbol{W^{(1)}}\boldsymbol{x} + \boldsymbol{b^{(1)}}) + \boldsymbol{b^{(2)}})}{\partial \boldsymbol{W^{(2)}}} \\
&= 5\,\mathbf{max}(\mathbf{0}, \boldsymbol{W^{(1)}}\boldsymbol{x} + \boldsymbol{b^{(1)}})^T, \text{ else } 0^T \\
\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{b^{(2)}}} &= \frac{\partial(5\boldsymbol{W^{(2)}}\,\mathbf{max}(\mathbf{0}, \boldsymbol{W^{(1)}}\boldsymbol{x} + \boldsymbol{b^{(1)}}) + \boldsymbol{b^{(2)}})}{\partial \boldsymbol{b^{(2)}}} \\
&= diag(1) \\
\frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{W^{(2)}}} &= \frac{\partial \boldsymbol{\ell}}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}} \frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{W^{(2)}}} \\
&= 5\frac{\partial \boldsymbol{\ell}}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}\,\mathbf{max}(\mathbf{0}, \boldsymbol{W^{(1)}}\boldsymbol{x} + \boldsymbol{b^{(1)}})^T \\
\frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{b^{(2)}}} &= \frac{\partial \boldsymbol{\ell}}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}} \frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{b^{(2)}}} \\
&= \frac{\partial \boldsymbol{\ell}}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{z_2}} &= \frac{\partial(\boldsymbol{W^{(2)}}5\max(0, \boldsymbol{W^{(1)}}\boldsymbol{x} + \boldsymbol{b^{(1)}}))}{\partial 5\max(0, \boldsymbol{W^{(1)}}\boldsymbol{x} + \boldsymbol{b^{(1)}})} = \boldsymbol{W^{(2)}} \\
\frac{\partial \boldsymbol{z_{1i}}}{\partial \boldsymbol{W^{(1)}}} &= \frac{\partial(\boldsymbol{W^{(1)}}\boldsymbol{x} + \boldsymbol{b^{(1)}})}{\boldsymbol{W^{(1)}}} = x^T, \text{ else } 0^T \\
\frac{\partial \boldsymbol{z_1}}{\partial \boldsymbol{b^{(1)}}} &= \frac{\partial(\boldsymbol{W^{(1)}}\boldsymbol{x} + \boldsymbol{b^{(1)}})}{\boldsymbol{b^{(1)}}} = diag(1) \\
\frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{W^{(1)}}} &= \frac{\partial \boldsymbol{\ell}}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}} \frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{z_2}} \frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}} \frac{\partial \boldsymbol{z_1}}{\partial \boldsymbol{W^{(1)}}} \\
&= \frac{\partial \boldsymbol{\ell}}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}} \boldsymbol{W^{(2)}} \frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}} \boldsymbol{x^T} \\
\frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{b^{(1)}}} &= \frac{\partial \boldsymbol{\ell}}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}} \frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{z_2}} \frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}} \frac{\partial \boldsymbol{z_1}}{\partial \boldsymbol{b^{(1)}}} \\
&= \frac{\partial \boldsymbol{\ell}}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}} \boldsymbol{W^{(2)}} \frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}
\end{aligned}
$$

4. Show the elements of $\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}$, $\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}$, $\frac{\partial \boldsymbol{\ell}}{\partial \hat{\boldsymbol{y}}}$. Be careful about dimensionality.

Note: $i \in \{1, \ldots, m\}$ used throughout.

$$\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}} = \frac{\partial(5 \max(0, \boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}}))}{\partial(\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}})}$$

$$\left(\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}\right)_{ii} = \begin{cases} 0, & z_{1i} < 0 \\ 5, & z_{1i} > 0 \\ \text{undefined (or assigned a value 0 in code)}, & z_{1i} = 0 \end{cases}$$

Note: $\boldsymbol{z_1} = \boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}}$ and the above is a diagonal matrix

$\left(\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}\right)_{ii} = 1$ (and 0 elsewhere, off of diagonal; an identity matrix)

$$\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}} = \frac{\partial(||\hat{\boldsymbol{y}} - \boldsymbol{y}||^2)}{\partial \hat{\boldsymbol{y}}} = 2(\hat{\boldsymbol{y}} - \boldsymbol{y})^T \text{ (A vector)}$$

Problem 2: Classification Task

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

1. If you want to train this network, what do you need to change in the equations of (1.2), (1.3) and (1.4), assuming we are using the same MSE loss function.

—f—
Input: $Linear_1(\boldsymbol{x}) = \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$
Output (z_2): $f(Linear_1(\boldsymbol{x})) = \tanh(Linear_1(\boldsymbol{x}))$
$= \tanh(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})$
$= \frac{\exp(\boldsymbol{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)})-\exp(-\boldsymbol{W}^{(1)}\boldsymbol{x}-\boldsymbol{b}^{(1)})}{\exp(\boldsymbol{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)})+\exp(-\boldsymbol{W}^{(1)}\boldsymbol{x}-\boldsymbol{b}^{(1)})}$
—$Linear_2$—
Input: $f(Linear_1(\boldsymbol{x})) = \tanh(Linear_1(x))$
Output (z_3): $Linear_2(f(Linear_1(\boldsymbol{x}))) = \boldsymbol{W}^{(2)}\boldsymbol{f(Linear_1(x))} + \boldsymbol{b}^{(2)}$
$= \boldsymbol{W^{(2)}\tanh\left(W^{(1)}x + b^{(1)}\right) + b^{(2)}}$
—g—
Input: $Linear_2(f(Linear_1(\boldsymbol{x})))$
Output: $g(Linear_2(f(Linear_1(\boldsymbol{x})))) = \frac{1}{1+\exp(-z_3)} = \frac{1}{1+\exp(-(\boldsymbol{W^{(2)}\tanh\left(W^{(1)}x+b^{(1)}\right)+b^{(2)}}))} = \hat{\boldsymbol{y}}$
—Loss—
Input: $\hat{\boldsymbol{y}}$
Output: $||(\frac{1}{1+\exp(-(\boldsymbol{W^{(2)}\tanh\left(W^{(1)}x+b^{(1)}\right)+b^{(2)}}))} - \boldsymbol{y})||^2$

$$
\begin{aligned}
\frac{\partial \boldsymbol{z_{3i}}}{\partial \boldsymbol{W^{(2)}}} &= \frac{\partial(\boldsymbol{W^{(2)}}\tanh\left(\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}}\right) + \boldsymbol{b^{(2)}})}{\partial \boldsymbol{W^{(2)}}} \\
&= \tanh(\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}})^T \text{ , else } 0^T \\
\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{b^{(2)}}} &= \frac{\partial(\boldsymbol{W^{(2)}}\tanh\left(\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}}\right) + \boldsymbol{b^{(2)}})}{\partial \boldsymbol{b^{(2)}}} \\
&= diag(1) \\
\frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{W^{(2)}}} &= \frac{\partial \boldsymbol{\ell}}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{W^{(2)}}} \\
&= \frac{\partial \boldsymbol{\ell}}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}\tanh(\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}})^T \\
\frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{b^{(2)}}} &= \frac{\partial \boldsymbol{\ell}}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{b^{(2)}}} \\
&= \frac{\partial \boldsymbol{\ell}}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{z_2}} &= \frac{\partial(\boldsymbol{W^{(2)}}\tanh\left(\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}}\right) + \boldsymbol{b^{(2)}})}{\partial \tanh(\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}})} = \boldsymbol{W^{(2)}} \\
\left(\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}\right)_{ii} &= \frac{\partial(\tanh(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}))}{\partial(\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}})} = \boldsymbol{1} - \tanh(\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}})^2 \\
\left(\frac{\partial \hat{\boldsymbol{y}}}{\partial z_3}\right)_{ii} &= \frac{\partial(1 + \exp(-(\boldsymbol{W^{(2)}}\tanh\left(\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}}\right) + \boldsymbol{b^{(2)}})))^{-1}}{\partial(\tanh(W^{(1)}x + b^{(1)}) + b^{(2)})} \\
&= \sigma(z_3)(1 - \sigma(z_3))_i, \\
&\quad \text{Note the two partials above represent diagonal matrices} \\
\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}} &= \frac{\partial(||\hat{\boldsymbol{y}} - \boldsymbol{y}||^{\boldsymbol{2}})}{\partial \hat{\boldsymbol{y}}} = 2\boldsymbol{(\hat{y} - y)^T} \text{ (A vector)}
\end{aligned}
$$

2. Now you think you can do a better job by using a Binary Cross Entropy (BCE) loss function $\ell_{BCE}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{1}{K}\sum_{i=1}^{K} -[y_i \log(\hat{y_i}) + (1 - y_i)\log(1 - \hat{\hat{y}}_i)]$. What needs to change from the previous?

—Loss—
Input: $\hat{\boldsymbol{y}}$
Output: $\ell_{BCE}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{1}{K}\sum_{i=1}^{K} -[y_i \log(\hat{y_i}) + (1-y_i)\log(1-\hat{\hat{y}_i})]$,
$\hat{y_i} = \frac{1}{1+\exp(-z_{3i})}$
$\frac{\partial \ell}{\partial \hat{y}} = (\frac{-y_i}{\hat{y_i}} + \frac{1-y_i}{1-\hat{y_i}})^T$

3. 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 σ. Explain why this choice of f can be beneficial for training a (deeper) network.

Sigmoid function is more computationally intensive to compute compared to ReLU due to exponential operation; the latter produces sparsity in matrices which encourages numerical optimization techniques taking advantage of this time and space complexity reduction.

ReLU also avoids the vanishing gradient problem, whereby the number of parameters receive very small updates such that the nodes deviate greatly from their optimal value. As the gradient is constant for ReLU compared to the sigmoid gradient always being smaller than 1, successive operations will start to prohibit learning.

Problem 3: Conceptual Questions

1. Why is softmax actually softargmax?

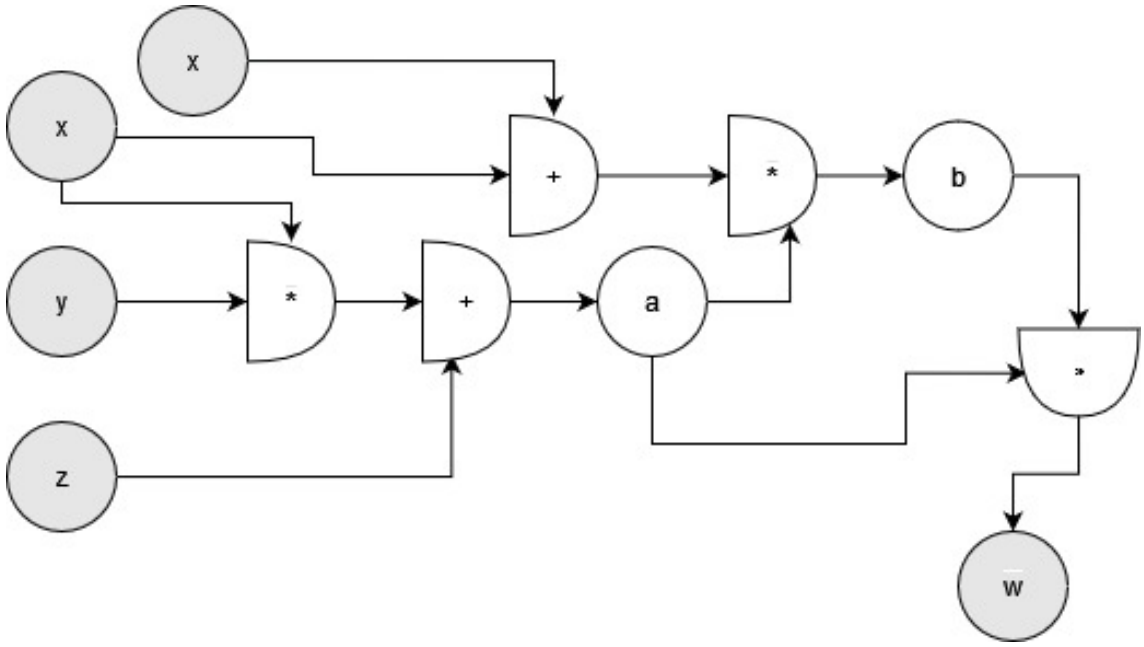
The "softmax" is not a smooth approximation to the maximum function – it is actually an approximation to the arg max function whose value is the index that has the maximum. Accordingly, some prefer the "softargmax" terminolgy to emphasize this distinction.

2. 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, w, and operators $+, \star$ in your solution. Be sure to use the correct shape for symbols and operators as shown in class.

   (a) $a = x \star y + z$

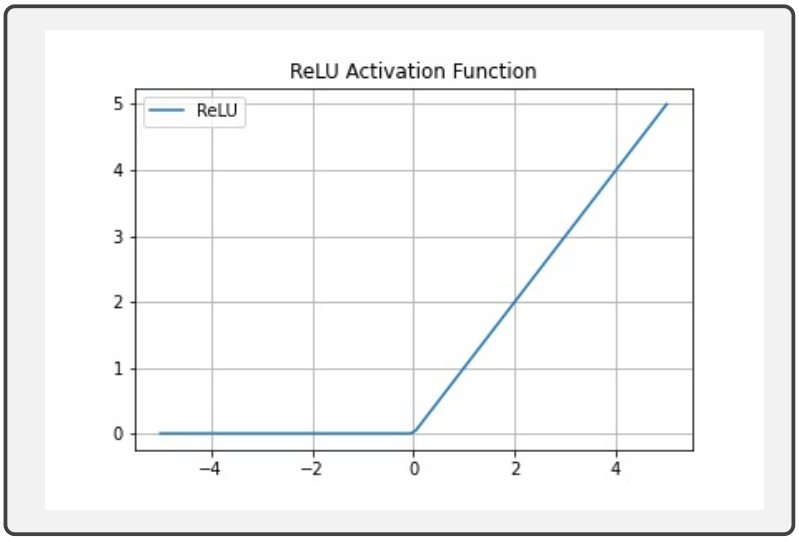
   (b) $b = (x + x) \star a$

   (c) $w = a \star b$

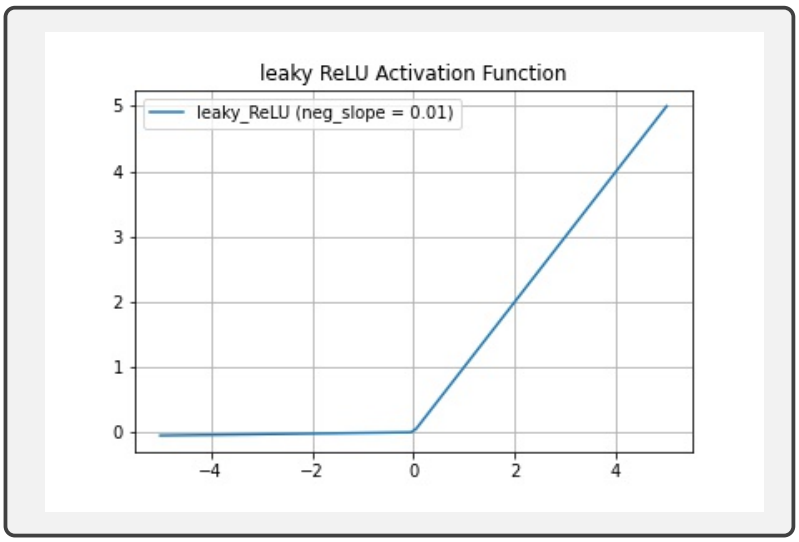


3. Draw the graph of the following:

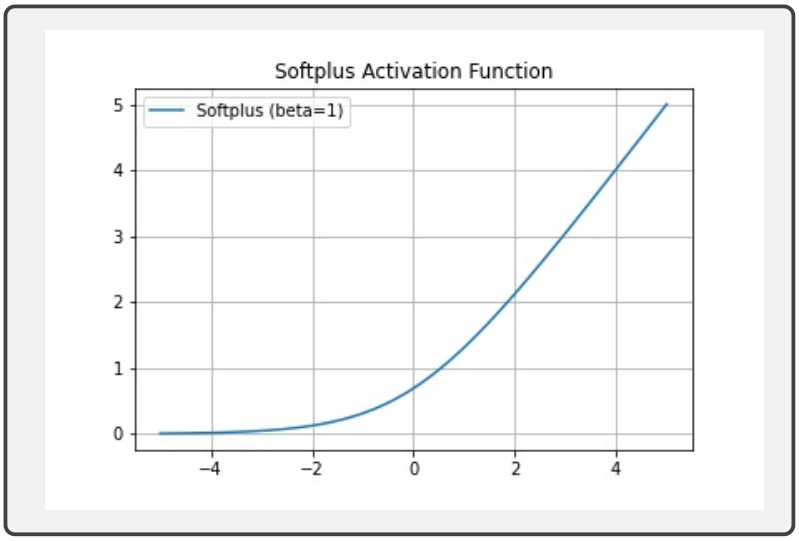
   (a) ReLU()



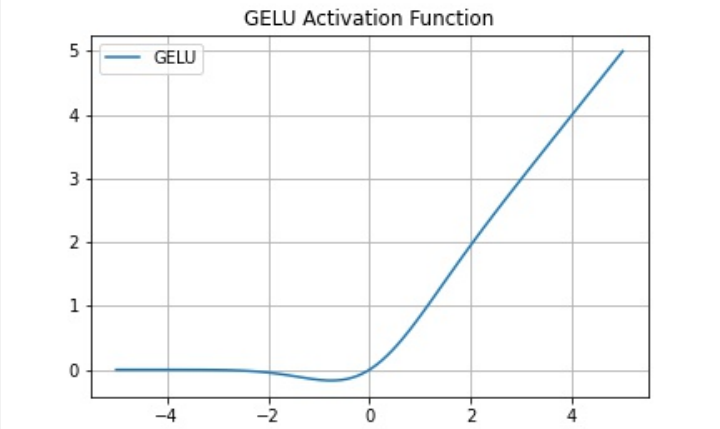
   (b) LeakyReLU(neg slope is 0.01)



(c) Softplus(beta is 1)



(d) GELU



4. What are 4 different types of linear transformations? What is the role of linear transformation and non linear transformation in a neural network?

   Reflection, rotation, projection, and dilation are the main types of linear transformations. In neural networks, each output unit produces the linear combination of the inputs and the connection weights. Once we allow translation, these essentially become affine transformations. To introduce nonlinearity to the model, activation functions ingest these transformations through a nonlinear function and treat that as the unit output.

5. Given a neural network F parameterized by parameters θ, denoted F_θ, dataset $D = x_1, x_2, \ldots, x_N$, and labels $Y = y_1, y_2, \ldots, y_N$, write down the mathematical definition of training a neural network with the MSE loss function $\ell_{MSE}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = ||\hat{y} - y||^2$

   $$\theta_i = \theta_i - \alpha \frac{\partial \ell}{\partial \theta_i}$$

Section 2.2

Problem 6

Blah

Problem 7

Blah

Problem 10

Blah