CMSC 472: Introduction to Deep Learning **Released: Jan-30. Due Feb-06 5pm.**

Assignment 1

Name: James Zhang *UID:* 118843940

Instructions:

- Submit the assignment on Gradescope. You **must assign** pages to corresponding questions to receive full credit.
- Assignments have to be formatted in LaTeX. You can use overleaf for writing your assignments.
- Submit only the compiled PDF version of the assignment.
- Refer to policies (collaboration, late days, etc.) on the course website.

1 Probability

1. **Density function**. Let p be a Gaussian distribution with zero mean and variance of 0.2. Compute the density of p at 0.

   Sol:

   Recall that a Gaussian pdf p is defined as

   $$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2)$$

   Plugging in known values $\mu = 0, \sigma = \sqrt{0.2}$, we get

   $$p(x) = \frac{1}{\sqrt{0.4\pi}} \exp(-\frac{1}{2}(\frac{x}{\sqrt{0.2}})^2)$$

   Now to find the density at 0, plug in 0 for x and we get

   $$p(x) = \frac{1}{\sqrt{0.4\pi}} \approx 0.8921$$

2. **Bayes rule**. Consider the probability distribution of you getting sick given the weather in the table below.

Sick?	Weather			
	sunny	rainy	cloudy	snow
yes	0.144	0.02	0.016	0.02
no	0.576	0.08	0.064	0.08

Compute P(sick = no | weather = cloudy).

Sol:

By the definition of conditional probability

$$\mathbb{P}(\text{sick=no} \mid \text{weather=cloudy}) = \frac{\mathbb{P}(\text{sick=no} \cap \text{weather=cloudy})}{\mathbb{P}(\text{weather=cloudy})}$$

Looking at the given graph, we plug in

$$\mathbb{P}(\text{sick=no} \mid \text{weather=cloudy}) = \frac{0.064}{0.08} = 0.8$$

3. **Conditional probability**. A student is taking a one-hour-time-limit makeup examination. Suppose the probability that the student will finish the exam in less than x hours is $x/2$, ∀x ∈ $[0, 1]$. Given that the student is still working after 0.75 hours, what is the conditional probability that the full hour will be used?

   Sol:

   We're told that $\mathbb{P}(X < x) = \frac{x}{2}\ \forall\ x \in [0, 1]$. We seek to find

   $$\mathbb{P}(\text{takes exactly 1 hour} \mid \text{still working after 0.75 hours})$$

   For simplicity, let us denote A as the event the student takes exactly 1 hour, and let B denote the event that the student is still working after 0.75 hours. Once more, by the definition of condiitonal probability, this is equal to

   $$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

   To find $\mathbb{P}(B)$, we compute

   $$\mathbb{P}(B) = 1 - \frac{0.75}{2} = 0.625$$

   For $\mathbb{P}(A \cap B)$, note that in order for the student to take the full hour, they must still be working after 0.75 hours, so $\mathbb{P}(A \cap B) = \mathbb{P}(A) = 1 - \frac{1}{2} = \frac{1}{2}$. This stems from the Law of Total Probability, which states

   $$\mathbb{P}(A) = \sum_n \mathbb{P}(A \cap B_n)$$

   for any events A, B. Therefore, the answer is

   $$\mathbb{P}(A|B) = \frac{0.5}{0.625} = 0.8$$

2 Calculus and Linear Algebra

For each of the following questions, we expect to see all the steps for reaching the solution.

1. Compute the derivative of the function $f(z)$ with respect to z $\left(\text{i.e., } \frac{df}{dz}\right)$, where

$$f(z) = \frac{1}{1 + e^{-z}}$$

Sol:

By the chain rule, we get

$$\frac{df}{dz} = -\frac{-e^{-z}}{(1 + e^{-z})^2} = \frac{e^{-z}}{(1 + e^{-z})^2} = \frac{1}{1 + e^{-z}}(1 - \frac{1}{1 + e^{-z}})$$

2. Compute the derivative of the function $f(w)$ with respect to w_i, where $w, x \in \mathbb{R}^{\mathrm{D}}$ and

$$f(w) = \frac{1}{1 + e^{-w^T x}}$$

Sol:

Note that we can rewrite the inner product as a summation so we get

$$w^T x = -\sum_j w_j x_j,\ j \in [0, D-1] \implies \frac{\partial}{\partial w_i}(-w^T x) = -x_i$$

With respect to a w_i,

$$\frac{df}{dw_i} = \frac{x_i e^{-w^T x}}{(1 + e^{-w^T x})^2}$$

by the chain rule twice.

3. Compute the derivative of the loss function $J(w)$ with respect to w, where

$$J(w) = \frac{1}{2}\sum_{i=1}^{m}\left|w^T x^{(i)} - y^{(i)}\right|$$

Sol:

Note that the derivative of a scalar loss function with respect to w, a column vector, is a resulting row vector. Further note that $\frac{d}{dx}(|x|) = \frac{x}{|x|}$. Thus, with respect to an arbitrary i

$$\frac{d}{dw}(|w^T x^{(i)} - y^{(i)}|) = \frac{(w^T x^{(i)} - y^{(i)})x^{(i)T}}{|w^T x^{(i)} - y^{(i)}|}$$

Therefore,

$$\frac{dJ}{dw} = \frac{1}{2}\sum_{i=1}^{m}\frac{w^T x^{(i)} - y^{(i)}}{|w^T x^{(i)} - y^{(i)}|}(x^{(i)T})$$

which is a row vector as desired.

4. Compute the derivative of the loss function $J(w)$ with respect to w, where

$$J(w) = \frac{1}{2}\left[\sum_{i=1}^{m}\left(w^T x^{(i)} - y^{(i)}\right)^2\right] + \lambda\|w\|_2^2$$

Sol:

Note that this is a loss function with a regression penalty but we can split them up

$$\frac{dJ}{dw} = \frac{d}{dw}(\frac{1}{2}\sum_{i=1}^{m}(w^T x^{(i)} - y^{(i)})^2 + \lambda||w||_2^2) = \frac{1}{2}\frac{d}{dw}(\sum_{i=1}^{m}(w^T x^{(i)} - y^{(i)})^2) + \frac{d}{dw}(\lambda||w||_2^2)$$

Consider an arbitrary i, such that we consider

$$\frac{d}{dw}((w^T x^{(i)} - y^{(i)})^2) = 2(w^T x^{(i)} - y^{(i)})(x^{(i)T})$$

Thus,

$$\frac{1}{2}\frac{d}{dw}(\sum_{i=1}^{m}(w^T x^{(i)} - y^{(i)})^2) = \sum_{i=1}^{m}(w^T x^{(i)} - y^{(i)})(x^{(i)T})$$

Now for the derivative of the ridge penalty, note that

$$\frac{d}{dw}(\lambda||w||_2^2) = \lambda\frac{d}{dw}(||w||_2^2) = \lambda\frac{d}{dw}(w^T w) = 2\lambda w^T$$

Therefore,

$$\frac{dJ}{dw} = 2\lambda w^T + \sum_{i=1}^{m}(w^T x^{(i)} - y^{(i)})(x^{(i)T})$$

5. Compute the derivative of the loss function $J(w)$ with respect to w, where

$$J(w) = \sum_{i=1}^{m} \left[y^{(i)} \log\left(\frac{1}{1+e^{-w^T x^{(i)}}}\right) + \left(1-y^{(i)}\right) \log\left(1 - \frac{1}{1+e^{-w^T x^{(i)}}}\right)\right]$$

Sol:

Recall that we previously calculated that

$$\frac{d}{dw}\left(\frac{1}{1+e^{-w^T x^{(i)}}}\right) = \frac{e^{-w^T x^{(i)}} x^{(i)T}}{(1+e^{-w^T x^{(i)}})^2}$$

Further note that $\frac{d}{dx}(\log x) = \frac{1}{x}$ plus the chain rule if necessary. Therefore,

$$\frac{d}{dw}(\log(\frac{1}{1+e^{-w^T x^{(i)}}})) = (1+e^{-w^T x^{(i)}})(\frac{e^{-w^T x^{(i)}} x^{(i)T}}{(1+e^{-w^T x^{(i)}})^2}) = \frac{e^{-w^T x^{(i)}} x^{(i)T}}{1+e^{-w^T x^{(i)}}}$$

Similarly the derivative of $\log(1 - \frac{1}{1+e^{-w^T x^{(i)}}}) = \log(\frac{e^{-w^T x^{(i)}}}{1+e^{-w^T x^{(i)}}})$ to be

$$\frac{1+e^{-w^T x^{(i)}}}{e^{-w^T x^{(i)}}}(\frac{e^{-w^T x^{(i)}} x^{(i)T}}{(1+e^{-w^T x^{(i)}})^2}) = -\frac{x^{(i)T}}{1+e^{-w^T x^{(i)}}}$$

which means the derivative over all terms of the summation is

$$\frac{dJ}{dw} = \sum_{i=1}^{m} \left(y^{(i)} \frac{x^{(i)^T} e^{-w^T x^{(i)}}}{1+e^{-w^T x^{(i)}}} + (1-y^{(i)}) \frac{-x^{(i)^T}}{1+e^{-w^T x^{(i)}}} \right)$$

$$\frac{dJ}{dw} = \sum_{i=1}^{m} \left(\frac{y^{(i)} e^{-w^T x^{(i)}} - 1 + y^{(i)}}{1+e^{-w^T x^{(i)}}} \right) \left(x^{(i)^T} \right)$$

6. Compute $\nabla_w f$, where $f(w) = \tanh\left[w^T x\right]$.

Sol:

$$\tanh(w^T x) = \frac{e^{w^T x} - e^{-w^T x}}{e^{w^T x} + e^{-w^T x}}$$

The gradient is a column vector given as $\begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial w} \end{bmatrix}^T$

$$\frac{\partial f}{\partial x} = \frac{(e^{w^T x} + e^{-w^T x})(w^T e^{w^T x} + w^T e^{-w^T x}) - (e^{w^T x} - e^{-w^T x})(w^T e^{w^T x} - w^T e^{w^T x})}{(e^{w^T x} + e^{-w^T x})^2}$$

$$\frac{\partial f}{\partial x} = w^T(1 - \tanh(w^T x))$$

$$\frac{\partial f}{\partial w} = \frac{(e^{x^T x} + e^{-x^T x})(x^T e^{w^T x} + x^T e^{-w^T x}) - (e^{w^T x} - e^{-w^T x})(x^T e^{w^T x} - x^T e^{w^T x})}{(e^{w^T x} + e^{-w^T x})^2}$$

$$\frac{\partial f}{\partial w} = x^T(1 - \tanh(w^T x))$$

Therefore, the gradient is

$$\nabla f = \begin{bmatrix} \dfrac{\partial f}{\partial x} \\ \\ \dfrac{\partial f}{\partial w} \end{bmatrix} = \begin{bmatrix} w^T(1 - \tanh(w^T x)) \\ x^T(1 - \tanh(w^T x)) \end{bmatrix}$$

7. Find the solution to the system of linear equations given by Ax=b, where

$$A = \begin{pmatrix} 2 & 1 & -1 \\ -3 & -1 & 2 \\ -2 & 1 & 2 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 8 \\ -11 \\ -3 \end{pmatrix}$$

Sol:

Let row reduce our the appropriate augmented matrix

$$\left(\begin{array}{ccc|c} 2 & 1 & -1 & 8 \\ -3 & -1 & 2 & -11 \\ -2 & 1 & 2 & -3 \end{array}\right) \underset{\longrightarrow}{R_2 + \frac{3}{2}R_1, R_3 + R_1} \left(\begin{array}{ccc|c} 2 & 1 & -1 & 8 \\ 0 & \frac{1}{2} & \frac{1}{2} & 1 \\ 0 & 2 & 1 & 5 \end{array}\right)$$

$$\underset{\longrightarrow}{R_1 - 2R_2, R_3 - 4R_2} \left(\begin{array}{ccc|c} 2 & 0 & -2 & 6 \\ 0 & \frac{1}{2} & \frac{1}{2} & 1 \\ 0 & 0 & -1 & 1 \end{array}\right) \underset{\longrightarrow}{R_1 - 2R_3, R_2 + \frac{1}{2}R_3} \left(\begin{array}{ccc|c} 2 & 0 & 0 & 4 \\ 0 & \frac{1}{2} & 0 & \frac{3}{2} \\ 0 & 0 & -1 & 1 \end{array}\right)$$

$$\implies x = \begin{pmatrix} 2 \\ 3 \\ -1 \end{pmatrix}$$

8. Find the eigenvalues and associated eigenvectors of the matrix:

$$A = \begin{bmatrix} 7 & 0 & -3 \\ -9 & -2 & 3 \\ 18 & 0 & -8 \end{bmatrix}$$

Sol:

To find the eigenvalues of A, let us solve $\det(A - \lambda I)$

$$\det \begin{pmatrix} 7-\lambda & 0 & -3 \\ -9 & -2-\lambda & 3 \\ 18 & 0 & -8-\lambda \end{pmatrix} = (-2-\lambda)((7-\lambda)(-8-\lambda)+54)$$

$$= (-2-\lambda)(\lambda^2+\lambda-56+54) = (-2-\lambda)(\lambda^2+\lambda-2) = (-2-\lambda)(\lambda+2)(\lambda-1)$$

$$\implies \lambda = -2 \text{ with multiplciity } 2, 1$$

The associated eigenvectors solve the equation $(A - \lambda_i I)v_i = 0$.

For the eigenvalue -2,

$$\begin{pmatrix} 9 & 0 & -3 \\ -9 & 0 & 3 \\ 18 & 0 & -6 \end{pmatrix} v = 0 \implies v = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{1}{3} \\ 0 \\ 1 \end{pmatrix}$$

For the eigenvalue 1,

$$\begin{pmatrix} 6 & 0 & -3 \\ -9 & -3 & 3 \\ 18 & 0 & -9 \end{pmatrix} v = 0 \implies \begin{pmatrix} 2 & 0 & -1 \\ -3 & -1 & 1 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \underset{\longrightarrow}{R_2 + \frac{3}{2}R_1} = \begin{pmatrix} 2 & 0 & -1 \\ 0 & -1 & -\frac{1}{2} \\ 0 & 0 & 0 \end{pmatrix} v = 0$$

$$\implies v = \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}$$

Therefore, the eigenvalue/eigenvalue pairs are

$$\{2, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}\}, \{2, \begin{pmatrix} \frac{1}{3} \\ 0 \\ 1 \end{pmatrix}\}, \{1, \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}\}$$

3 Activation functions

For each of the following activation functions, write their equations and their derivatives. Plot the functions and derivatives, with $x \in [-5, 5]$ and $y \in [-10, 10]$ plot limits. (No need to submit the code for plots.)

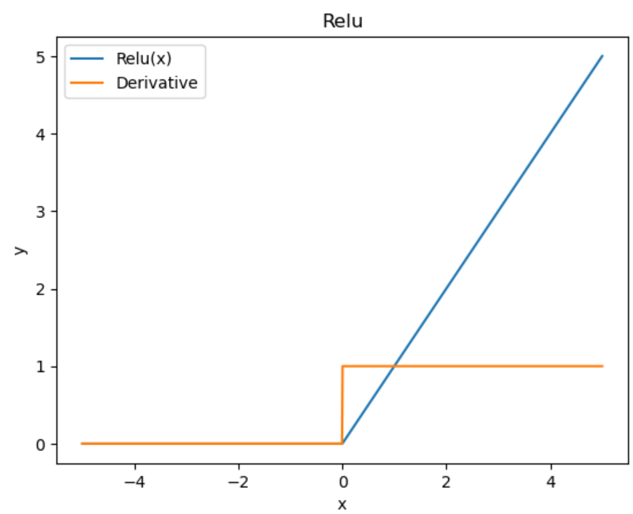
1. Relu

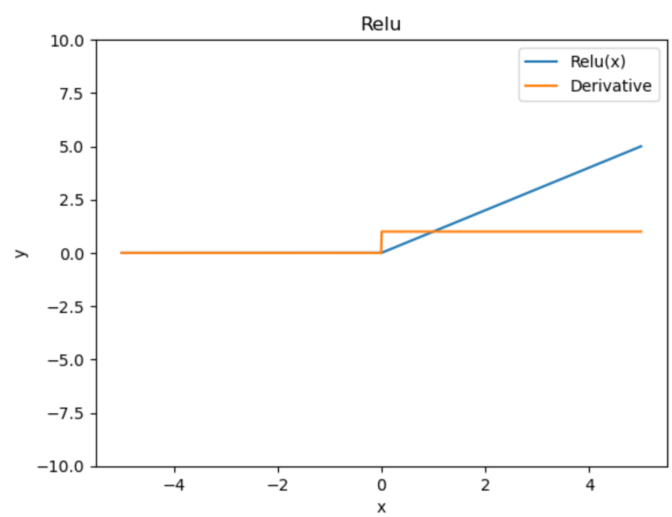
   Sol:

   In all of the following plots, I plot two graphs: the two graphs contains a zoomed in plot of the function and its derivative, and the bottom plot contains a zoomed out plot such that $x \in [-5, 5], y \in [-10, 10]$, since I was confused with the directions.

$$\text{Relu}(x) = \begin{cases} 0 \text{ if } x < 0 \\ x \text{ otherwise} \end{cases}$$

$$\text{Derivative of Relu}(x) = \begin{cases} 0 \text{ if } x < 0 \\ 1 \text{ otherwise} \end{cases}$$





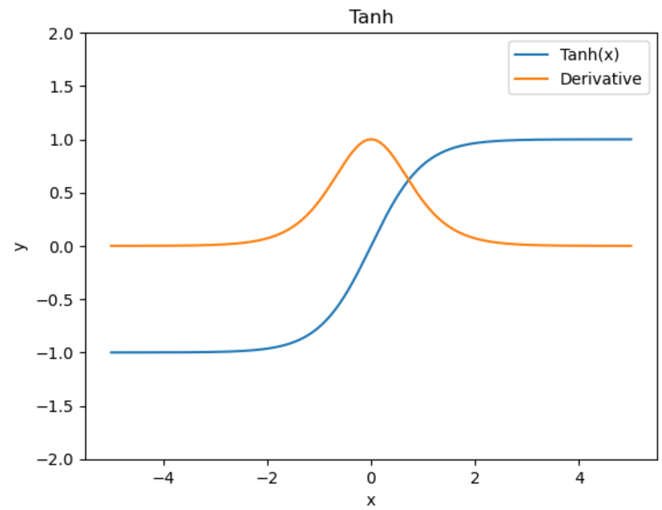
2. Tanh

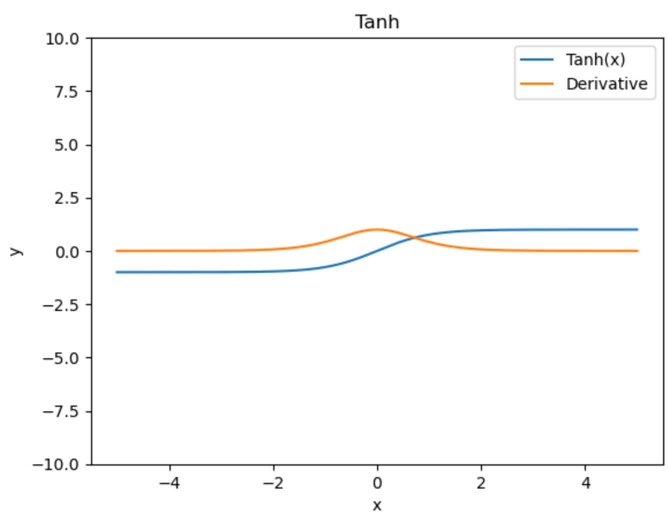
Sol:

$$\text{Tanh}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Using the quotient rule we can determine the derivative

$$\frac{(e^x + e^{-x})(e^x + e^{-x}) - (e^x - e^{-x})(e^x - e^{-x})}{(e^x + e^{-x})^2} = 1 - \frac{(e^x - e^{-x})^2}{(e^x + e^{-x})^2} = 1 - \tanh^2(x)$$



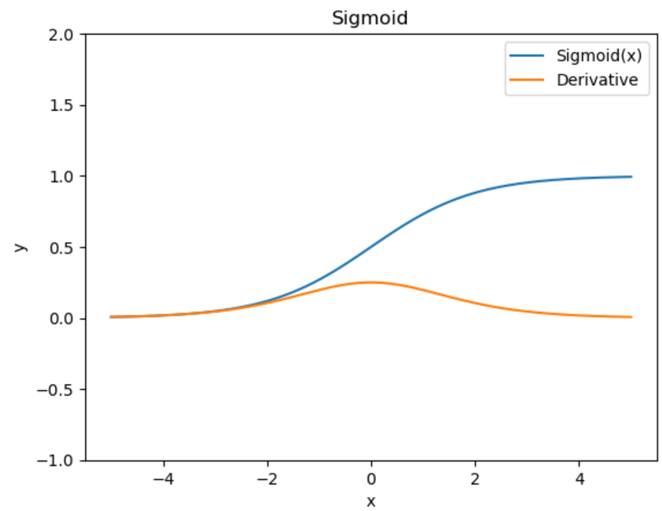


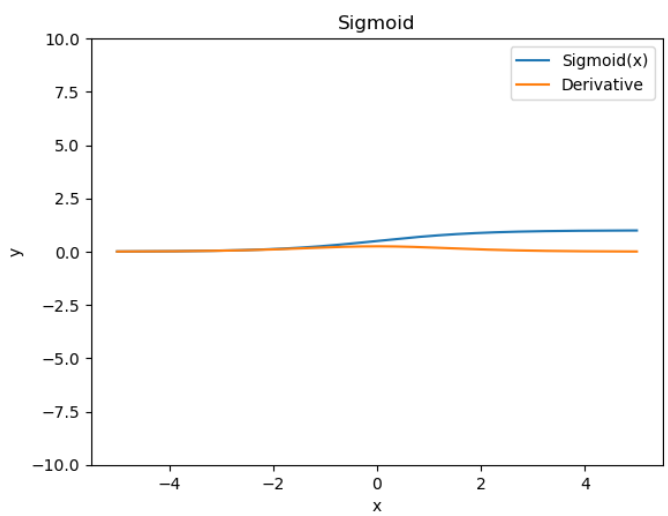
3. Sigmoid

Sol:

$$\text{Sigmoid}(x) = \frac{1}{1+e^{-x}}$$

$$\text{Derivative of Sigmoid}(x) = \frac{e^{-x}}{(1+e^{-x})^2} = (\frac{1}{1+e^{-x}})(1-\frac{1}{1+e^{-x}})$$



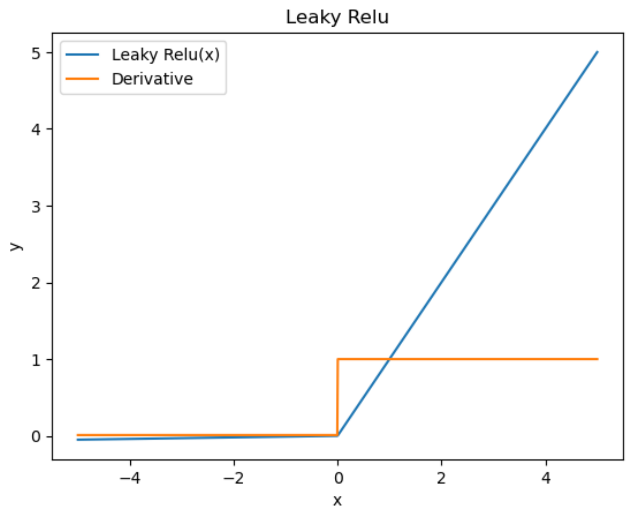


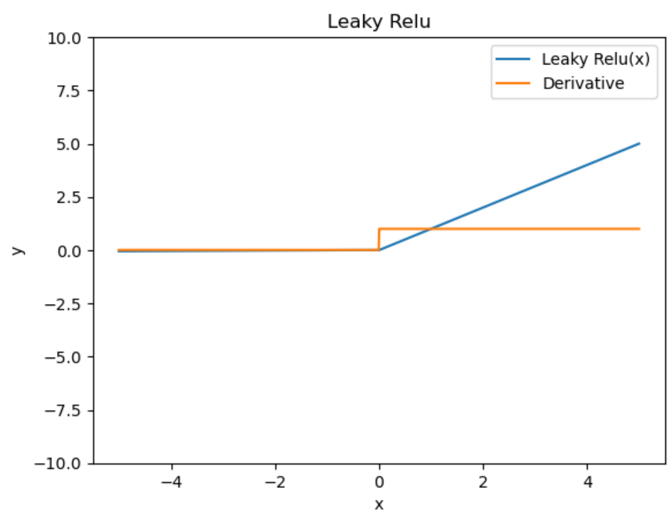
4. Leaky ReLU

Sol:

$$\text{Leaky Relu}(x) = \begin{cases} 0.01x \text{ if } x < 0 \\ x \text{ otherwise} \end{cases}$$

$$\text{Derivative of Leaky Relu}(x) = \begin{cases} 0.01 \text{ if } x < 0 \\ 1 \text{ otherwise} \end{cases}$$



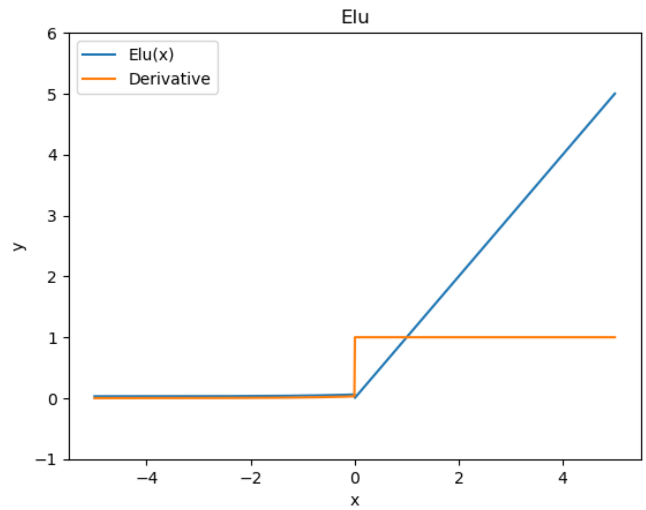


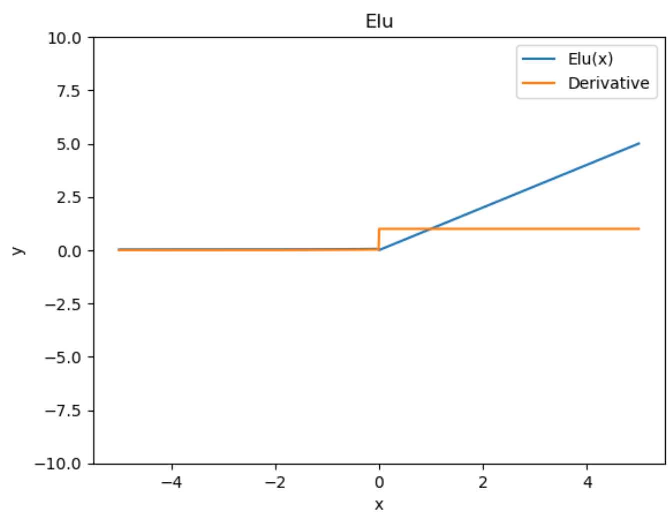
5. ELU (plot with $\alpha = 0.3$)

Sol:

$$ELU(x) = \begin{cases} \alpha(e^x - 1) \text{ if } x < 0 \\ x \text{ otherwise} \end{cases}$$

$$\text{Derivative of ELU(x)} = \begin{cases} \alpha e^x \text{ if } x < 0 \\ 1 \text{ otherwise} \end{cases}$$





6. Sinc

Sol:

$$\text{Sinc}(x) = \frac{\sin(x)}{x}$$

$$\text{Derivative of Sinc}(x) = \frac{\cos(x)}{x} - \frac{\sin(x)}{x^2}$$