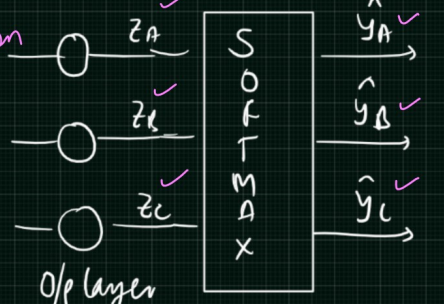
SOFTMAX ACTIVATION FUNCTION:-

no. of classes ≥ 2

At o/p layer multiclass classification



$$\Sigma e = e^{z_A} + e^{z_B} + e^{z_C} \quad -(1)$$

$$\frac{e^{z_A}}{e^{z_A} + e^{z_B} + e^{z_C}}$$

$$\hat{y}_A + \hat{y}_B + \hat{y}_C = \frac{e^{z_A} + e^{z_B} + e^{z_C}}{\Sigma e} = \frac{\Sigma e}{\Sigma e} = 1$$

Outcome: Probability distribution

$$\Sigma y = 1$$

$\hat{y}_A = 0.7 \rightarrow$ 70% of chance is the prediction belongs to class A

$\hat{y}_B = 0.3 \rightarrow$

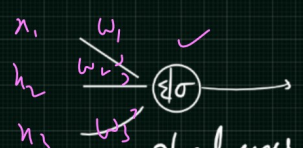
$\hat{y}_C = 0.0$

Case Study:-

Case 1: Binary classification. { sigmoid }

60% $0.6 \rightarrow$ A

$0 \rightarrow 0.5$ A, $0.5 \rightarrow 1$ B



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

$$z = w_1x_1 + w_2x_2 + w_3x_3 = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}^T \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$w^T \cdot x = \theta^T \cdot x$$

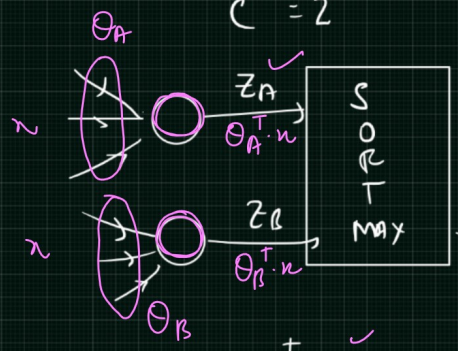
$$(w_1\ w_2\ w_3) \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$p(y=1|x) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-\theta^T x}}$$

$$p(y=0|x) = 1 - p(y=1|x) = 1 - \frac{1}{1+e^{-z}} = \frac{e^{-z}}{1+e^{-z}}$$

$$p(y=0|x) = \frac{e^{-\theta^T x}}{1+e^{-\theta^T x}}$$

Case 2: act fn: softmax Binary classification



$$\hat{y}_A \quad \frac{e^{z_A}}{e^{z_A} + e^{z_B}} = p(y=1|x) = \frac{e^{\theta_A^T \cdot x}}{e^{\theta_A^T \cdot x} + e^{\theta_B^T \cdot x}}$$

$$\hat{y}_B \quad \frac{e^{z_B}}{e^{z_A} + e^{z_B}} \quad \therefore p(y=0|x) =$$

$$p(y=1|x) = \frac{e^{\theta_A^T x}}{e^{\theta_A^T x} + e^{\theta_B^T \cdot x}} = \frac{e^{\theta_A^T \cdot x} / e^{\theta_A^T x}}{e^{\theta_A^T x} / e^{\theta_A^T x} + e^{\theta_B^T \cdot x} / e^{\theta_A^T x}}$$

$$= \frac{1}{1 + e^{\theta_B^T \cdot x - \theta_A^T \cdot x}}$$

$$= \frac{1}{1 + e^{-(\theta_A^T - \theta_B^T) x}}$$

Assume $\theta_A^T - \theta_B^T = \theta^T$

$$p(y=1|x) = \frac{1}{1 + e^{-\theta^T \cdot x}}$$

$$p(y=0|x) = \frac{e^{\theta_B^T \cdot x}}{e^{\theta_A^T \cdot x} + e^{\theta_B^T \cdot x}}$$

$$= \frac{e^{\theta_B^T \cdot x - \theta_A^T \cdot x}}{1 + e^{\theta_B^T x - \theta_A^T \cdot x}}$$

$$= \frac{e^{-(\theta_A^T - \theta_B^T) x}}{1 + e^{-(\theta_A^T - \theta_B^T) x}}$$

$$p(y=0|x) = \frac{e^{\theta^T x}}{1 + e^{\theta^T x}}$$

Conclusion: For binary classification Softmax is equivalent to sigmoid

Drawback :-

i) It contains exponential terms ✓

⇓

computationally intensive ✓

Softmax is always used in output layer

2) ELU (Exponential Linear Units)

$$elu(z,\alpha) = \begin{cases} z & z \geq 0 \\ \alpha\cdot(e^z - 1) & z < 0 \end{cases}$$

forward pass

domain $z \in (-\infty, \infty)$

Range $elu(z,\alpha) \in (-\alpha, \infty)$

α = scalar value | it's a slope of -ve section

at $z = 0$, let's see continuity of elu ✓

$\xrightarrow{0^-} \; 0 \; \xleftarrow{0^+}$

$$\begin{cases} LHL = \lim\limits_{z\to 0^-} elu(z,\alpha) = \lim\limits_{z\to 0^-} \alpha(e^z - 1) \simeq \alpha(e^0 - 1) \\ \quad = \alpha(1-1) = 0 \end{cases}$$ ✓

$$\begin{cases} RHL = \lim\limits_{z\to 0^+} elu(z,\alpha) = \lim\limits_{z\to 0^+} z = 0 \end{cases}$$ ✓

$elu(z=0,\alpha) = 0$

$\therefore$ LHL = RHL = $elu(z=0,\alpha)$

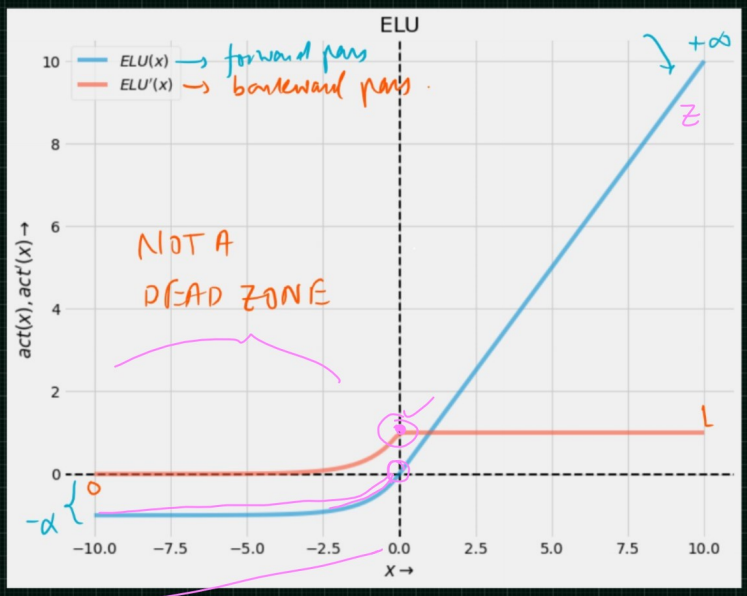
$\Rightarrow$ elu is continuous. ✓

Backward pass

$$elu'(z,\alpha) = \begin{cases} 1 & z \geq 0 \\ \alpha e^z & z < 0 \end{cases}$$

domain $z \in (-\infty, \infty)$

range $elu'(z,\alpha) \in (0, 1]$



Advantage :-

i) No dead ReLU issues
ii) Approx zero centered

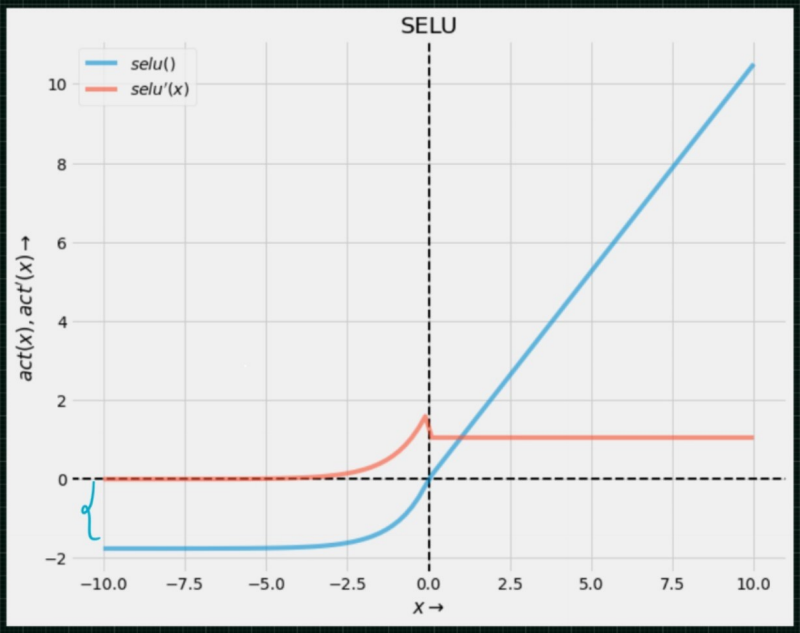
Disadvantage :-

i) exponential term ⇒ compute intensive

3) SELU → Scaled ELU

$$selu(z, \alpha) = scale \times elu(z, \alpha)$$

$$= k \times elu(z, \alpha)$$



4) Swish

$$\text{swish}(z) = z \cdot \sigma(\beta \cdot z) = \frac{z}{1 + e^{-z\beta}}$$

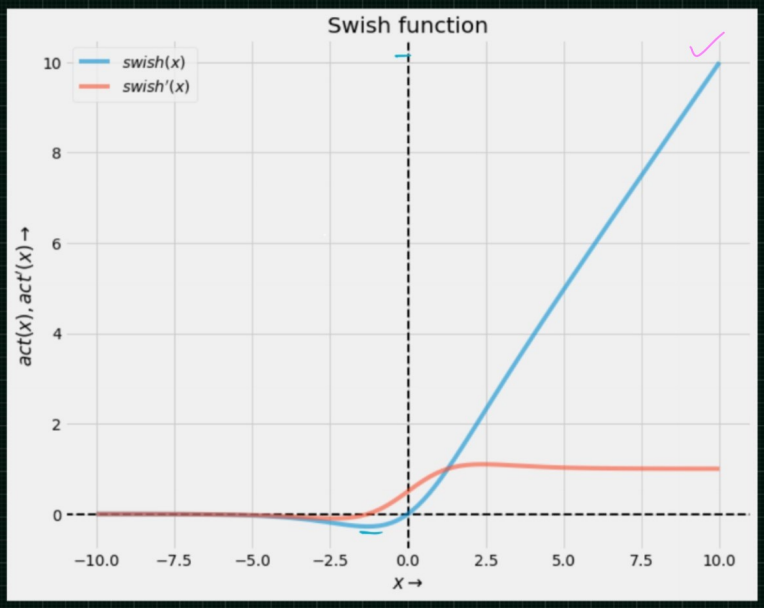
domain $z \in (-\infty, \infty)$

range $\in (0, \infty)$

β can be a learnable parameter

for $\beta = 1$.

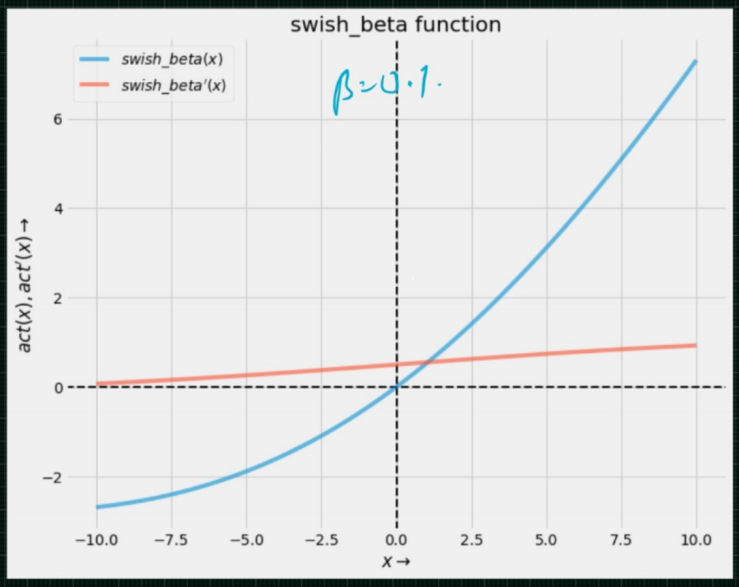
$$\text{swish} - 1 \Rightarrow \text{swish}(z) = z \cdot \sigma(z) = \frac{z}{1 + e^{-z}}$$



derivative of swish

domain $z \in (-\infty, \infty)$

range $\in (0, 1)$



Drawbacks :-

i) Presence of e^x term

↓

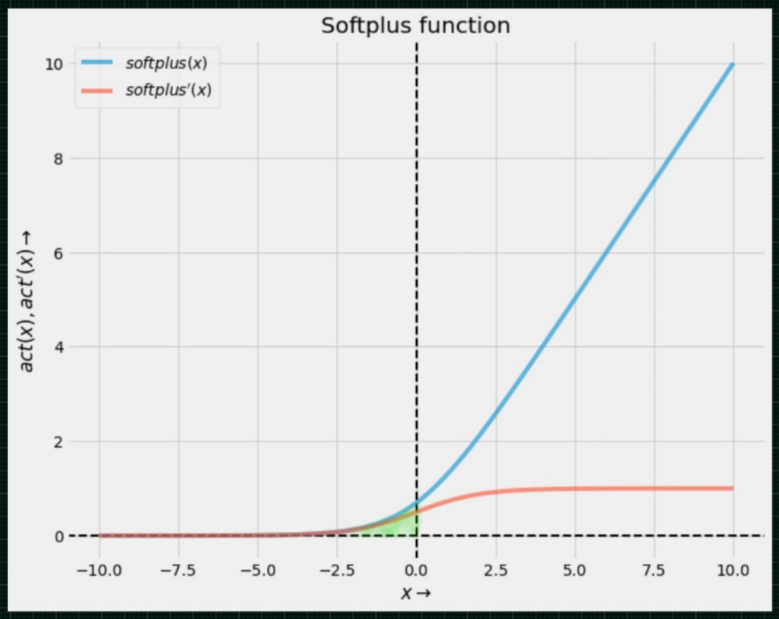
compute intensive

Advantages :-

i) Its unique the origin

SOFTPLUS

$$\text{softplus}(z) = \log(1 + e^z)$$



Disadvantage :-

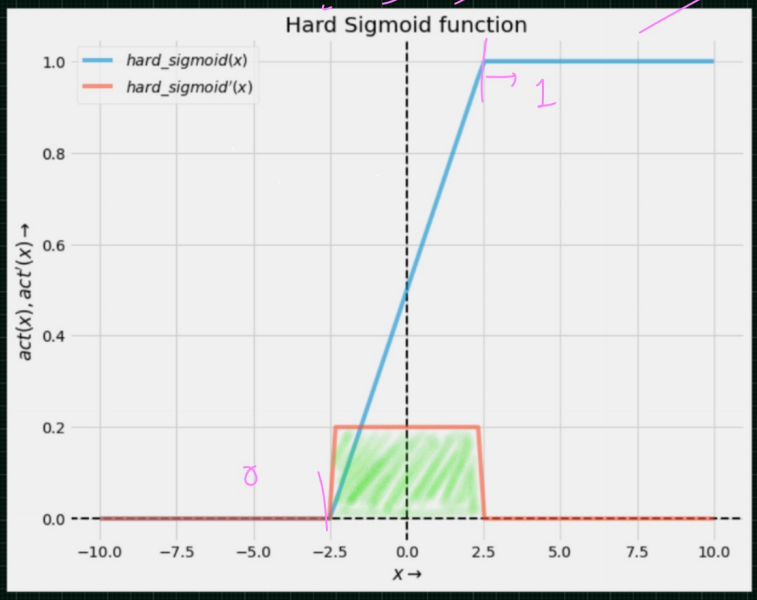
i) Presence of e^z & log ⇒ computation intensive

Advantage :-

i) Is differentiable everywhere

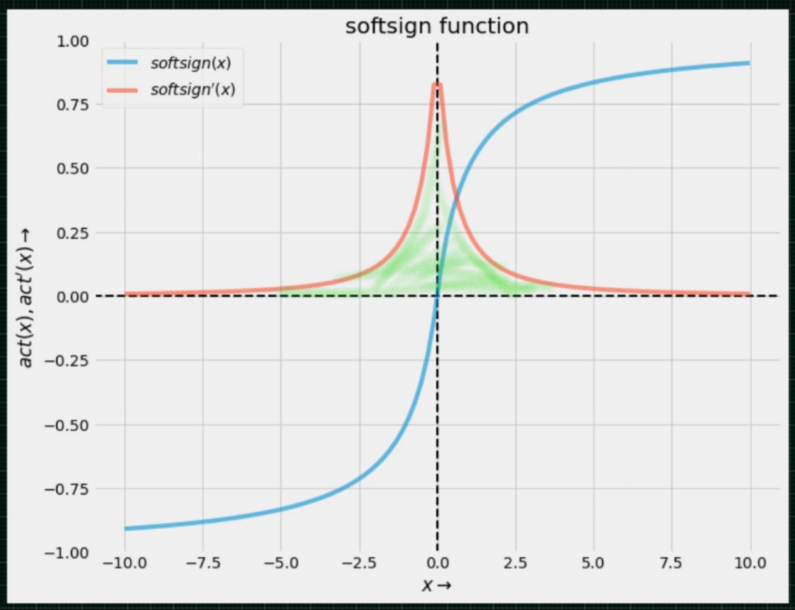
Hard Sigmoid

$$\phi(z) = \begin{cases} 0 & z < -2.5 \\ 1 & z > 2.5 \\ 0.2z + 0.5 & -2.5 \leq z \leq 2.5 \end{cases}$$ if else

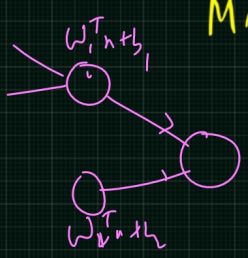


SOFTSIGN

$$SS(z) = \frac{z}{(|z|+1)} = \begin{cases} \frac{z}{z+1} & z \geq 0 \\ \frac{z}{-z+1} & z < 0 \end{cases}$$



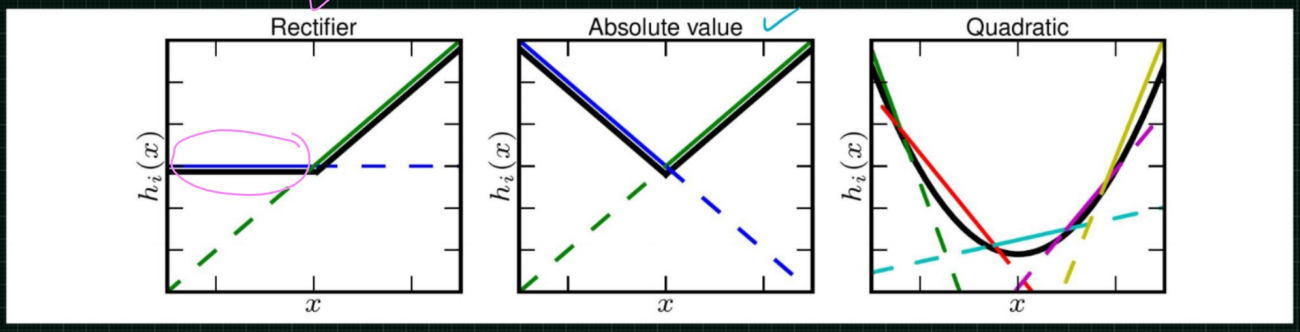
MAXOUT



Generalization of ReLU & Leaky ReLU

$$f(x) = \max\left(W_1^T x + b_1, W_2^T x + b_2\right)$$

$$\left(0, (\ )\right)$$



Drawback :-

i) It is computationally intensive as it doubles the number of parameters.

→ How to start / begin selection of activation function?

for any problem,

Go with ReLU for hidden layers (90%)

hid — ReLU, hid variant ReLU (Leaky ReLU, Parametric ReLU, ELU)

o/p — softmax

versions of relu { as a future trend }

for classification :-

binary :-

sigmoid / softmax at o/p layer

o/p unit = 1 o/p unit = 2

Multiclassification

softmax only at o/p layer

for Regression :- o/p

No activation function

Remark

DL → Architecture → CNN → Std data

Imagenet, CIFAR } previous