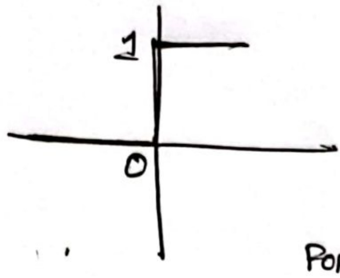
Different Activation Functions

① Binary step fn:



In backward calculation,
For loss function step fn won't be a good choice.

For
$\geq 0 : 1$ (+) positive val
$< 0 : 0$ (−) neg "

* multilevel output possible না
* Gradient zero.

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial a} \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial w}$$

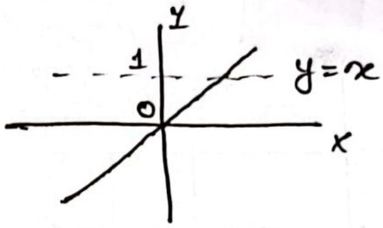
$\frac{\partial a}{\partial z} \hookrightarrow$ zero

$$\frac{dy}{dx} = \frac{d}{dx} 1 = 0$$

$$\frac{dy}{dx} = \frac{d}{dx} 0 = 0.$$

* weight update হবে না।

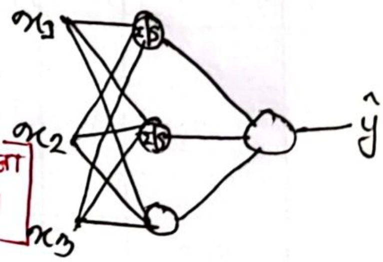
② Linear Activation fn:



* Multilevel output possible
* Not influential (Gradient র সাথে input র সম্পর্ক নেই)
* Gradient not zero but ↑

Q: All Layers of the NN collapse into one in linear funct

Ans:



Layer-1: $z^{[1]} = w^{[1]} x + b^{[1]}$

$a^{[1]} = \sigma(z^{[1]})$

$a^{[1]} = z^{[1]}$ [Sigmoid-এ change হবে না যা x তাই y বা output পাবো]

$$z^{[2]} = W^{[2]} a^{[1]} + b^{[2]}$$
$$= W^{[2]} z^{[1]} + b^{[2]}$$
$$= W^{[2]}(w^{[1]} x + b^{[1]}) + b^{[2]}$$
$$= \underbrace{W^{[2]} W^{[1]}}_{m} x + \underbrace{W^{[2]} b^{[1]} + b^{[2]}}_{c}$$
$$= mx + c = \text{linear output}$$

Why do we need nonlinear activation functions?

Ans:

1. NN with linear activation fn result in a linear combination of inputs.

$$z^{[1]} = W^{[1]}x + b^{[1]} = a^{[1]}$$

$$z^{[2]} = \omega^{[2]}a^{[1]} + b^{[2]}$$

$$= \omega^{[2]}\left(W^{[1]}x + b^{[1]}\right) + b^{[2]} = \underbrace{\omega^{[2]}\omega^{[1]}}_{m}x + \underbrace{\omega^{[2]}b^{[1]} + b^{[2]}}_{c}$$

$$= \boxed{mx + c}$$

So inputs = outputs.

Collapse in one NN.

2. Linear function are only single grade polynomial $mx+c$ so in gradient decent calculation soon it becomes zero. For single and sigmoid so becom. non convex

3. Multi layer deep NN with nonlinear activation functions can learn hierarchical and abstract representations of features of data.

4. Real world data such as images videos, text often contains nonlinear relationships and high dimentionality. So Nonlinear activation fn allow NN to capture and learn these intricate patterns enabling better generailization to unseen data.

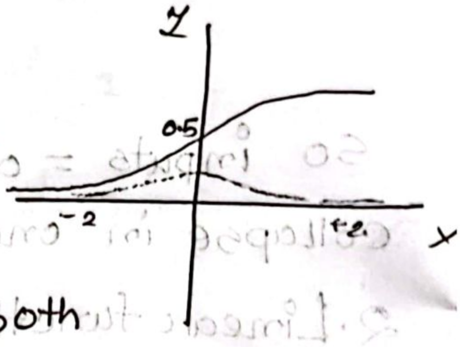
Backpropagation:

$$y = x$$

$$\frac{dy}{dx} = \frac{d}{dx}x = 1$$

$$\frac{d^2}{dx^2} = 0 \quad [\text{Soon becomes zero}]$$

9. Sigmoid:



→ Multi level output ~~(2 label)~~

→ Gradient influencial with input

→ 0.5 centered → ওটা problem.

কারণ output positive always (+, -) both

কারণ center 0 না।

→ vanishing gradient problem [Gradien ক্রমশ zero হয়ে:

→ Computionally expensive $\sigma = \frac{1}{1+e^{-x}}$

Computing Loss function: (এখানে total backpropagation এর calculation দেখানো হয়েছে with sigmoid activation fn)

To calculate Back Propagation:

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial a}\frac{\partial a}{\partial z}\frac{\partial z}{\partial w_1}$$

$$\frac{\partial L}{\partial a} = \frac{\partial}{\partial a} - \left[y \log_2 a + (1-y) \log_2 (1-a) \right]$$

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

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

$$= - \left[y/a + \frac{1-y}{1-a} (-1) \right]$$

$$= -\frac{y}{a} + \frac{1-y}{1-a}$$

$$\frac{\partial a}{\partial z} = \frac{\partial}{\partial z} \left(\frac{1}{1+e^{-z}} \right)$$

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

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

$$= -1 (1+e^{-z})^{-2} \quad e^{-z} \frac{\partial}{\partial z} (-z)$$

$$= -1 (1+e^{-z})^{-2} \quad e^{-z} (-1)$$

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

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

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

$$= \left(\frac{1+e^{-z}}{1+e^{-z}} - \frac{1}{1+e^{-z}} \right) \cdot \frac{1}{1-e^{-z}}$$

$$= \left(1 - \frac{1}{1+e^{-z}} \right) \cdot \frac{1}{1-e^{-z}}$$

$$= (1-a)\, a$$

So sigmoid-র ফিরে আসছে but in different terms.

Now,

$$\frac{\partial z}{\partial w_1} = \frac{\partial}{\partial w_1}\left(w_1 x_1 + w_2 x_2 + b\right)$$

$$= x_1$$

$$dw_1 = \frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial a} \frac{\partial a}{\partial z} \frac{\partial z}{\partial w_1}$$

$$= \frac{-y}{a} + \frac{1-y}{1-a} \times a(1-a) \times x_1$$

$$= \frac{-y + ay + a - ay}{a(\cancel{1-a})} \times a(\cancel{1-a}) \times x_1$$

$$= (a-y)\, x_1$$

for bias, $b = (a-y)$

$$dw_2 = (a-y)x_2$$

$$dw_3 = (a-y)x_3$$

- - - - - -

- - - - - -

*** এখন পুরো back propagation টাই same, আমাদের just activation fn র derivation change হবে যে একটা activa fn র জন্য।

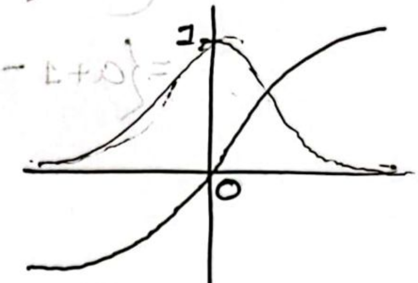
4. Tanh Activation fn:

→ Center zero

→ Gradient steep.

→ Vanishing Gradient problem.

→ complex computation $\dfrac{e^{z}-e^{-z}}{e^{z}+e^{-z}}$



$$\frac{\partial a}{\partial z} = \frac{\partial}{\partial z}\left(\frac{e^{z}-e^{-z}}{e^{z}+e^{-z}}\right) \qquad \frac{u}{v} = \frac{vu' - uv'}{v^2}$$

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

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

$$= \left(\frac{e^{z}+e^{-z}}{e^{z}+e^{-z}}\right)^2 - \left(\frac{e^{z}-e^{-z}}{e^{z}+e^{-z}}\right)^2$$

$$= 1 - a^2$$

~~5. ReLU (Rectified Linear Unit)~~

$$dw_1 = \frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial a}\,\frac{\partial a}{\partial z}\,\frac{\partial z}{\partial w_1}$$

$$= -\frac{y}{a} + \frac{1-y}{1-a} \times (1-a^2) \times x_1.$$

$$= \frac{-y+ay+a-ay}{a(1-a)} \times (1-a^2) \times x_1$$

$$= \frac{a-y}{a(1-a)} \times (1-a^2)\, x_1.$$

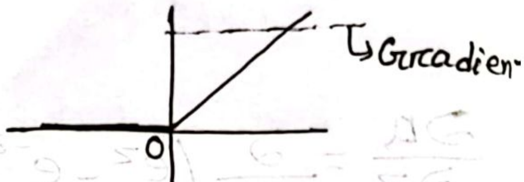
$$= \frac{(a-y)(1+a)}{a(1-a^2)} \times (1-a^2)\, x_1.$$

$$= \frac{a^2-ay+a-y}{a} \times x_1$$

$$= \left(a - y + 1 - y/a\right) x_1$$

$$= \left\{a + 1 - y\left(1 + \frac{1}{a}\right)\right\} x_1$$

4. ReLU: Rectified Linear Unit



So As activation fn:

$x \geq 0 \quad y = x$

$$\frac{dy}{dx} = \frac{d}{dx} x = 1$$

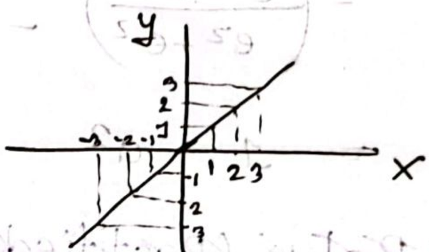
$x \geq 0 \quad y = x$

$x < 0 \quad y = 0$

$x<0$, $y=0$; $\frac{dy}{dx} = 0$

Q: Proves that ReLU looks like a linear but actions like nonlinear.

Linear fn ~~y≥0 x≥0~~ $y = x$



& if $x < 0$ then $y < 0$

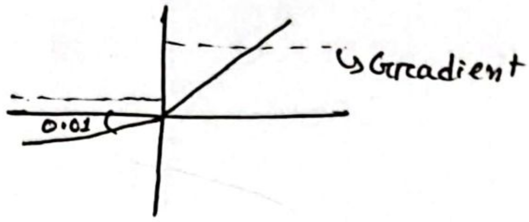
if $x \geq 0$ then $y \geq 0$

but here for negative value Relu gives zero value always. So ReLU is not linear.

*** যেখানে negative value র জন্য Gradient হঠাৎ নাই হয়ে যায় এটা die ReLU problem. যেহেতু neg =0 তাই কখনো কিছু layer-এ কিছু node/neuron থেকে output পাবেই না। ReLU automatic dropout করে যেখানে neg=0 তা

→ Computationally efficient.

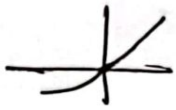
5. Leaky ReLU



→ Parameterized Relu

এখানে neg(-) ভ্যালু manually দেয়া হয় (0.02 or 0.01)

→ Exponential Relu



neg এর slop কিছু curved হয় তাই।

$x \geq 0 \quad y = x \quad \frac{dy}{dx} = 1$

$x < 0 \quad y = 0.01x \quad \frac{dy}{dx} = 0.01$

[যাদের derivative ২টা Loss বের করার সময় দুইটা করা লাগবে।]

~~So~~ the derivation turns into $(1-\tanh^2)$ where except tanh no other variable is existed. Hence it is visiable that derivation of tanh only depends on itself.

Q: "The derivative of the hyperbolic tangent function is more steep than the sigmoid function" - Justify the statement with proper evidance.

Ans: We know, for sigmoid activation function if large value is assigned then the gradient becomes zero. Same goes for very small value.

Now the sigmoid function, $g(z) = \frac{1}{1+e^{-z}}$

And the derivation of sigmoid is,

$$g'(z) = g(z).(1-g(z))$$

Let's, if $z=10$; $g(z) = \frac{1}{1+e^{-10}} = 0.999 \approx 1$

So $g'(z) = 1(1-1) = 0$ ✓

Again, if $z=-10$; $g(z) = \frac{1}{1+e^{-(-10)}} = 0.000045 \approx 0$

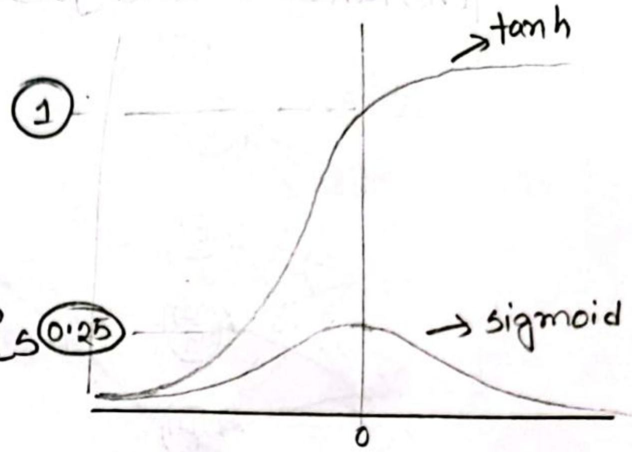
So, $g'(z) = 0(1-0) = 0$ ✓

For $z=0$; $g(z) = \frac{1}{1+e^{-0}} = \frac{1}{2}$

So, $g'(z) = \frac{1}{2}\left(1-\frac{1}{2}\right) = \frac{1}{4} = 0.25$ ✓

So it is proved that for large and low value sigmoid shows vanishing gradient problem and

It's steepness is about 0.25 unit long if it is drawn on graphically.



Now for tanh,

tanh, $g(z) = \dfrac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$

The derivation of tanh(z) is

$$g'(z) = 1 - (g(z))^2$$

Let's if $z = 10$; $g(z) = \dfrac{e^{10} - e^{-10}}{e^{10} + e^{-10}} \approx 0.9999 \approx 1$.

so $g'(z) = 1 - (1)^2 = 0$

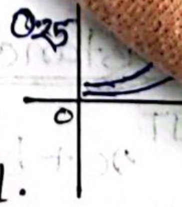
Again if $z = -10$; $g(z) = \dfrac{e^{-10} - e^{-(-10)}}{e^{-10} + e^{-(-10)}} \approx -1$

so $g'(z) = 1 - (-1)^2 = 0$

Now, $z = 0$; $g(z) = \dfrac{e^{0} - e^{-0}}{e^{0} + e^{-0}} = 0$

so $g'(z) = 1 - 0^2 = 1$

Tanh also suffers from vanishing G. problem but it's steepness much higher than sigmoid about 1.



Therefore the statement is justified.

Fig 1: Representation of steepness of tanh 4 sigmoid