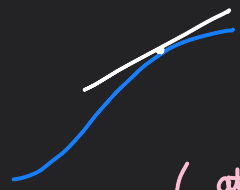
Calculus

[Maths for ML]

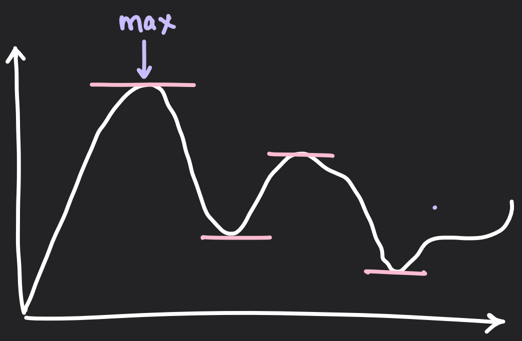
Course - Deep Learning. AI

Derivative - is the instantaneous rate of change of the function.

eg. velocity = $\frac{\Delta x}{\Delta t}$

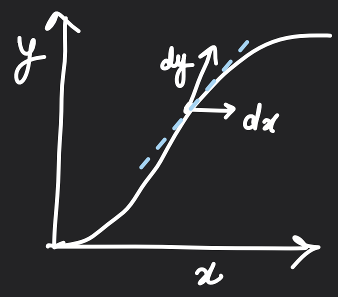


$\frac{dx}{dt}$ (inst rate of change) (at a point)



— zero slope

$$\text{Slope} = \frac{\text{Vertical change}}{\text{horizontal change}} = \frac{dy}{dx}$$



If $y = f(x)$

slope $= f'(x)$ (Lagrange's notation)

$= \frac{d}{dx} f(x)$ (Leibniz's notation)

If $y = c$, $\frac{dy}{dx} = 0$	If $y = ax + b$, $\frac{dy}{dx} = a$	If $y = x^2$, $\frac{dy}{dx} = 2x$

$f(x) = ax + b$

$$\frac{\Delta y}{\Delta x} = \frac{a(\cancel{x} + \Delta x) + b - (a\cancel{x} + b)}{\Delta x}$$

$$\boxed{\frac{\Delta f}{\Delta x} = \frac{f(x + \Delta x) - f(x)}{\Delta x}}$$

$$= \frac{a\Delta x + \cancel{b} - \cancel{b}}{\Delta x} = a\frac{\Delta x}{\Delta x} = a$$

If $y = x^3$
$f'(x) = 3x^2$

If $y = \frac{1}{x} = x^{-1}$
$f'(x) = (-1) \cdot x^{-1-1}$
$= -x^{-2}$

$$f(x) = x^n \Rightarrow f'(x) = nx^{n-1}$$

Inverse functions:

$$x \xrightarrow[(x^2)]{f} x^2 \xrightarrow[(\sqrt{x})]{g} x$$

$g(x) = f^{-1}(x)$

$g(f(x)) = x$

$\text{slope} = \frac{\Delta f}{\Delta x}$ Slope of inverse f. $= \frac{\Delta x}{\Delta f}$

so $g'(y) = \frac{1}{f'(x)}$

Derivative of trigonometric functions:

$$\frac{d}{dx}(\sin x) = \cos x$$

$$\frac{d}{dx}(\cos x) = -\sin x$$

Exponential Function :

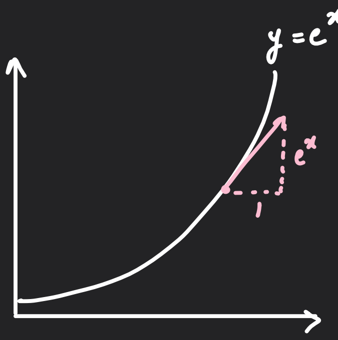
euler no. $(e) = 2.71828182\ldots$

in $\left(1+\frac{1}{n}\right)^n$ if $n \to \infty$ then $\left(1+\frac{1}{n}\right)^n = e$

example of banks

Better

Bank 1	Bank 2	Bank 3	....	Bank ∞
after 1 yr returns 100%.	after 6 mon. return 50%	after 4 mont return 33.3%.		returns $\underline{e}$



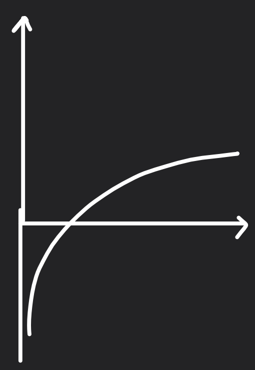
$\frac{d(e^x)}{dx} = e^x$

Logarithmic function :

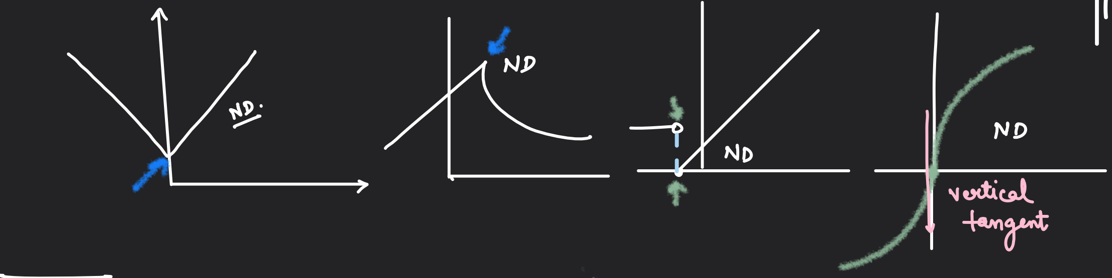
$e^{?} = 3$ ← log (3)

$f(x) = e^x \qquad f^{-1}(y) = \log(y)$

$\frac{d(\log y)}{dy} = \frac{1}{y}$



Differentiable & Non-differential (ND) fnc :



Rules :

$f = cg$
$\Downarrow$
$f' = cg'$

Sum Rule

$f = g + h$
$\Downarrow$
$f' = g' + h'$

Pdt Rule

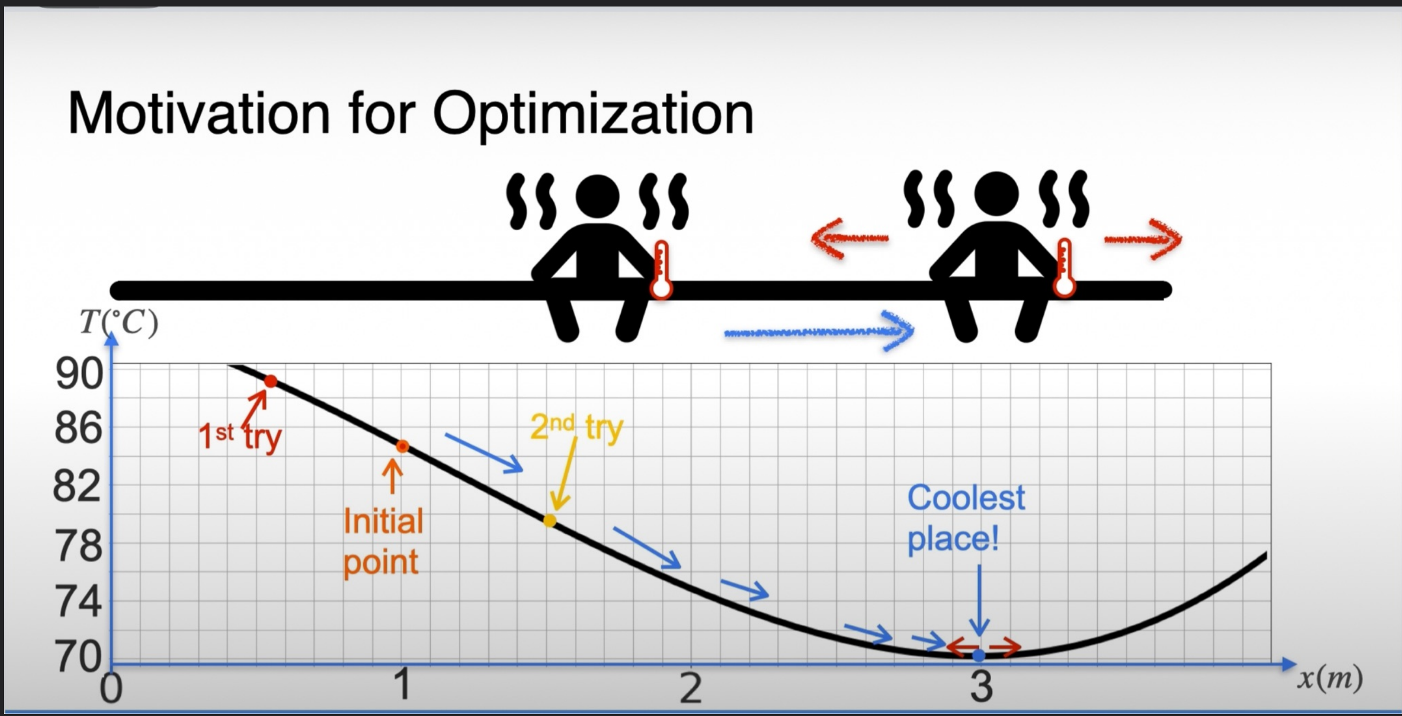
$f = gh$
$\Downarrow$
$f' = g'h + gh'$

Chain Rule

$\frac{d}{dt}\, g(h(t))$

$= \frac{dg}{dt} \cdot \frac{dh}{dt}$

Optimisation :



→ Take the derivative and put it to zero.

eg. $\frac{d}{dx}\left((x-a)^2 + (x-b)^2\right) = 0$

solve this to get the minimum point.

Log - Loss :

Let's say we win if we get 7 H and 3 T. , so what's the best coin ?

$g(p) = p^7(1-p)^3$

$\log(g(p)) = \log(p^7(1-p)^3)$

$= \log(p^7) + \log((1-p)^3)$

$= 7\log p + 3\log(1-p) = G(p)$

$\boxed{-G(p) \text{ is the log loss}}$

we minimize this

-ve bcz log p. is -ve b/w 1 and 0.

$\frac{d(G(p))}{dp} = \frac{d}{dp}(7\log p + 3\log(1-p))$

$= \frac{7}{p} + 3\frac{1}{1-p}(-1)$

$= \frac{7(1-p) + 3p}{p(p-1)} = 0 \Rightarrow \boxed{p = 0.7}$

{Less solving than actual one}

Why log ?

→ derivative of pdt is hard, but of sums is easy

→ pdt of lots of tiny things is tiny
but in log of tiny no is -ve big no {and computers can handle big no. better than smaller}

Gradients :

$$f(x, y) = x^2 + y^2 \quad (y \leftarrow \text{constant})$$

$$f_x = \frac{\partial f}{\partial x} = 2x + 0$$

$$f(x, y) = 3x^2y^3$$

$$\frac{\partial f}{\partial x} = 3(2x)y^3 = 6xy^3$$

$$\frac{\partial f}{\partial y} = 3x^2(3y^2) = 9x^2y^2$$

If $f(x, y) = x^2 + y^2$

$$\nabla f = \begin{bmatrix} 2x \\ 2y \end{bmatrix}$$ then gradient at $(2,3)$

$$= \begin{bmatrix} 2 \cdot 2 \\ 2 \cdot 3 \end{bmatrix}$$

$$= \begin{bmatrix} 4 \\ 6 \end{bmatrix}$$

[Minimum is when all slopes $= 0$]

Exercise :

$$T = f(x,y) = 85 - \frac{1}{90}x^2(x-6)y^2(y-6)$$

$$\frac{\partial f}{\partial x} = \left[-\frac{2x}{90}(x-6) + \frac{-1x^2}{90}\right]y^2(y-6)$$

$$= \frac{1}{90}(-2x^2 + 12x - x^2)y^2(y-6) = \frac{-1}{90}x(3x-12)y^2(y-6) = 0$$

$$\frac{\partial f}{\partial y} = \frac{-1}{90}x^2(x-6)\left[2y(y-6) + y^2\right]$$

$$= \frac{-1}{90}x^2(x-6)(2y^2 - 12y + y^2)$$

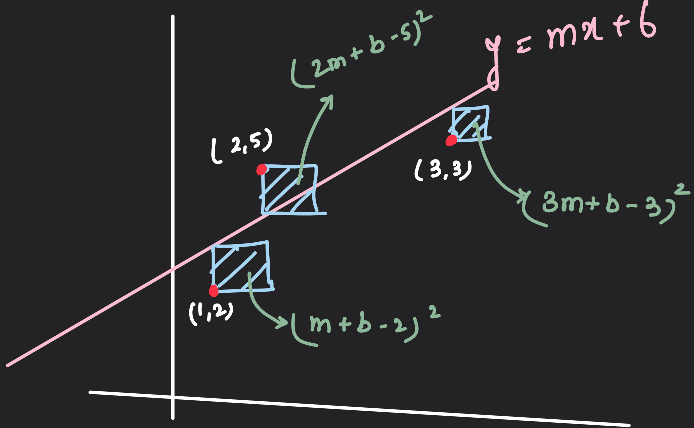
$$= \frac{-1}{90}x^2(x-6)[3y^2 - 12y] = 0$$

$x=0$ or $x=6$ or
$y=0$ or $y=4$

$x=0$ or
$x=4$ or
$y=0$ or
$y=6$

Check for each, put value in fnc or from graph.

In Linear Regression :



Goal : minimize sum of squares cost

$$(m+b-2)^2 + (2m+b-5)^2 + (3m+b-3)^2$$

$\Rightarrow$

$$E(m,b) = 14m^2 + 3b^2 + 38 + 12mb - 42m - 20b$$

Get $\frac{\partial E}{\partial m}$ and $\frac{\partial E}{\partial b}$

then solve $\frac{\partial E}{\partial m} = 0$ and $\frac{\partial E}{\partial b} = 0$ to get m and b.

Gradient Descent :

New point = old point - slope

$$x_1 = x_0 - \alpha f'(x_0) \quad ; \quad \alpha \equiv \text{Learning Rate}$$

[large slope → big step ; small slope → small step]

then $x_2 = x_1 - \alpha f'(x_1)$

⋮

Repeat until you are close enough to true minimum or when steps don't change that much.

Drawbacks of GD:

— may stuck in local minima.

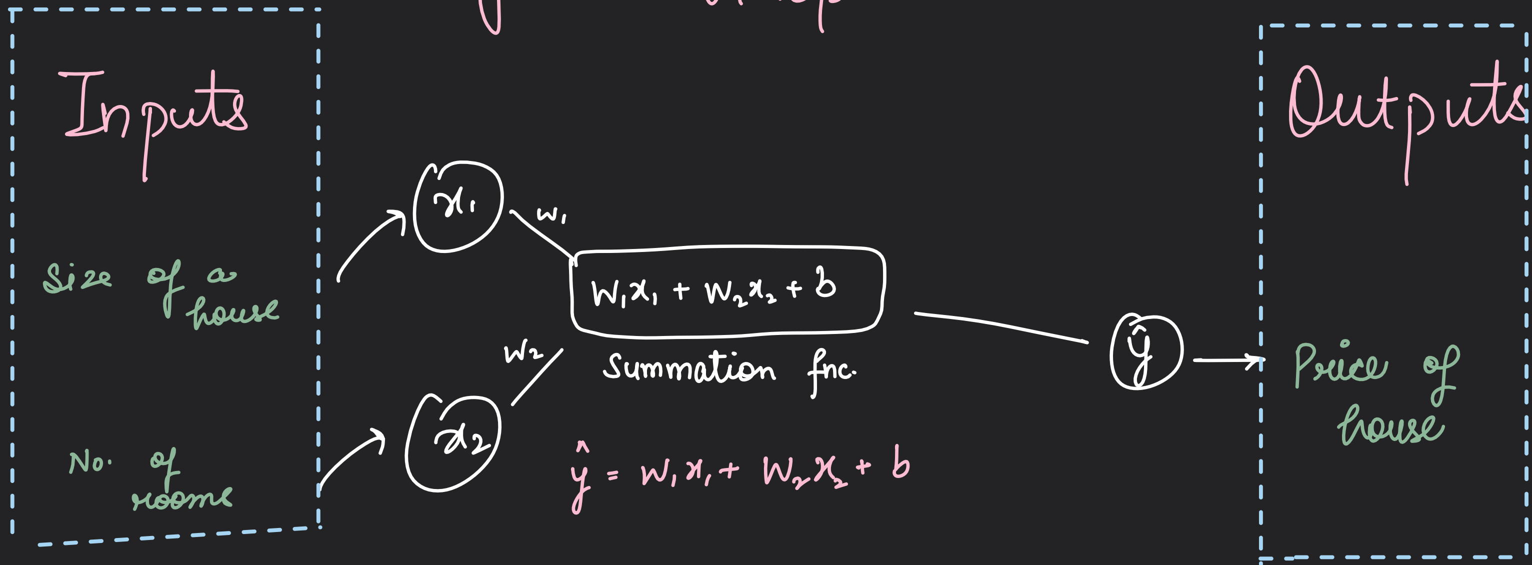
GD for 2 variables :

$$\begin{bmatrix} x_k \\ y_k \end{bmatrix} = \begin{bmatrix} x_{k-1} \\ y_{k-1} \end{bmatrix} - \alpha \nabla f(x_{k-1}, y_{k-1})$$

and repeat.

Regression with a perceptron :

Single Layer Neural Network
Perceptron



$$\hat{y} = w_1x_1 + w_2x_2 + b$$

Main Goal :

Find weights and bias that will optimise the predictions

i.e. Reduce errors in predictions.

The Loss Function

Mean Squared Error :

Loss fnc $L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$

$\frac{1}{2}$ → to cancel 2 when we take derivative

Applying Gradient D :

$$w_1 \rightarrow w_1 - \alpha \frac{\partial L}{\partial w_1}$$

$$w_2 \rightarrow w_2 - \alpha \frac{\partial L}{\partial w_2}$$

$$b \rightarrow b - \alpha \frac{\partial L}{\partial b}$$

Find these also

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b} \quad ; \quad \frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1} \quad ; \quad \frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2}$$

$$\begin{cases} \dfrac{\partial L}{\partial \hat{y}} = -(y - \hat{y}) \\[2ex] \dfrac{\partial \hat{y}}{\partial b} = 1 \qquad \dfrac{\partial \hat{y}}{\partial w_1} = x_1 \qquad \dfrac{\partial \hat{y}}{\partial w_2} = x_2 \end{cases}$$

$$\frac{\partial L}{\partial b} = -(y - \hat{y}) \qquad \frac{\partial L}{\partial w_1} = -(y - \hat{y})x_1 \qquad \frac{\partial L}{\partial w_2} = -(y - \hat{y})x_2$$

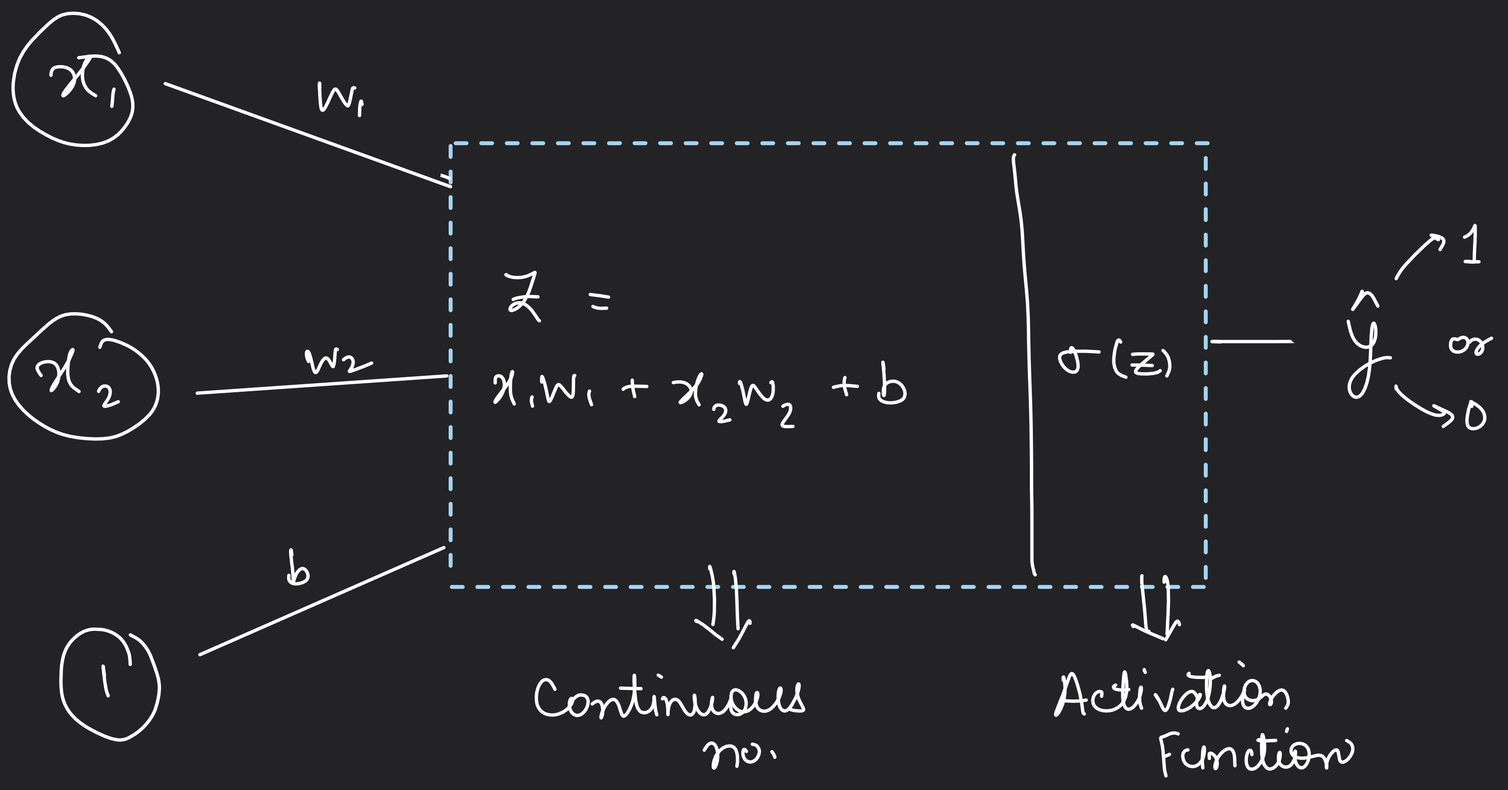
$$w_1 \longrightarrow w_1 - \alpha\left(-x_1 (y - \hat{y})\right)$$

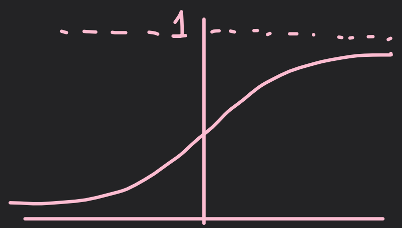
$$w_2 \longrightarrow w_2 - \alpha\left(-x_2 (y - \hat{y})\right)$$

$$b \longrightarrow b - \alpha\left(-(y - \hat{y})\right)$$

... and keep on updating

Classification with Perceptron :





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

Sigmoid Func.

Derivative of Sigmoid :

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

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

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

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

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

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

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

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

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

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

$$\frac{d}{dz}(\sigma(z)) = \sigma(z)\,[1 - \sigma(z)]$$

for classification : Log Loss (error fnc)

Pred^c Func^n : $\hat{y} = \sigma(w_1 x_1 + w_2 x_2 + b)$

Loss Func^n :

$$L(y, \hat{y}) = -y \ln(\hat{y}) - (1-y) \ln(1-\hat{y})$$

For GD:

$$w_1 \longrightarrow w_1 - \alpha \frac{\partial L}{\partial w_1}$$

$$w_2 \longrightarrow w_2 - \alpha \frac{\partial L}{\partial w_2}$$

$$b \longrightarrow b - \alpha \frac{\partial L}{\partial b}$$

Find partial derivatives

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_1} \qquad \frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_2} \qquad \frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial b}$$

$$\left\{ \frac{\partial L}{\partial \hat{y}} = \frac{-y}{\hat{y}} + \frac{1-y}{1-\hat{y}} = \frac{-y + \cancel{y\hat{y}} + \hat{y} - \cancel{y\hat{y}}}{\hat{y}(1-\hat{y})} = \frac{-(y-\hat{y})}{\hat{y}(1-\hat{y})} \right\}$$

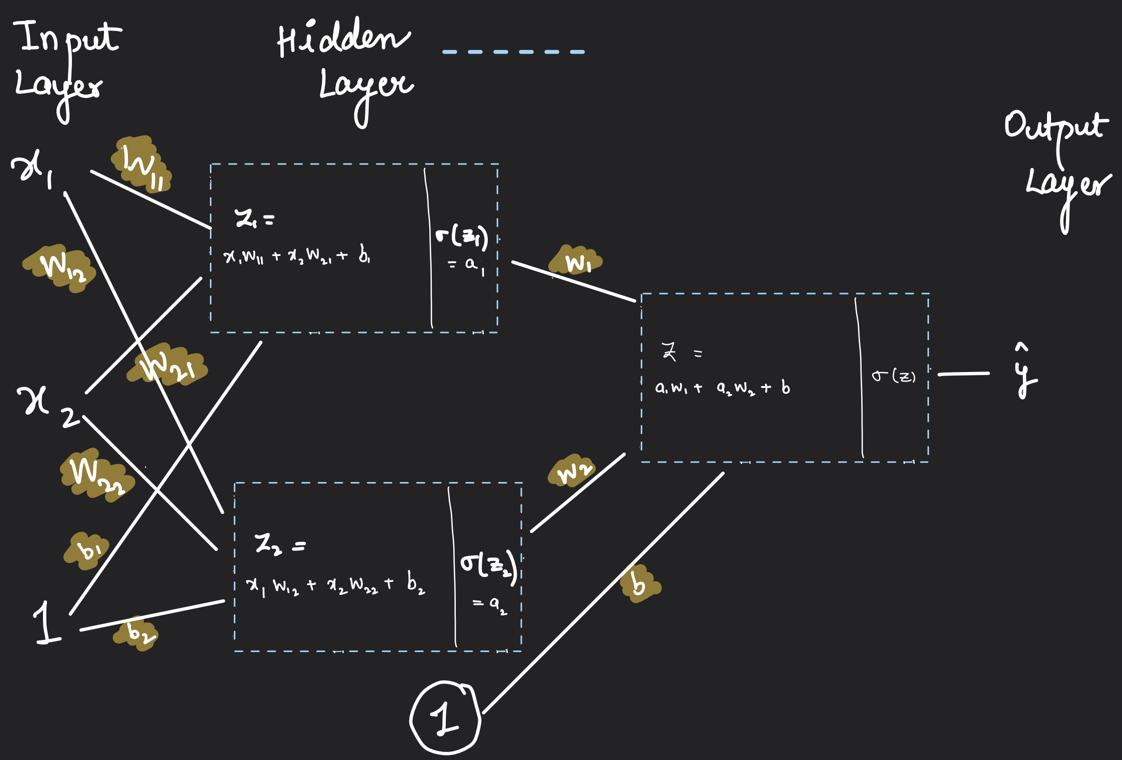
$$\frac{\partial L}{\partial w_1} = \frac{-(y-\hat{y})}{\cancel{\hat{y}(1-\hat{y})}} \cdot x_1 \cancel{\hat{y}(1-\hat{y})} = -(y-\hat{y}) \cdot x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{-(y-\hat{y})}{\cancel{\hat{y}(1-\hat{y})}} \cdot x_2 \cancel{\hat{y}(1-\hat{y})} = -(y-\hat{y}) \cdot x_2$$

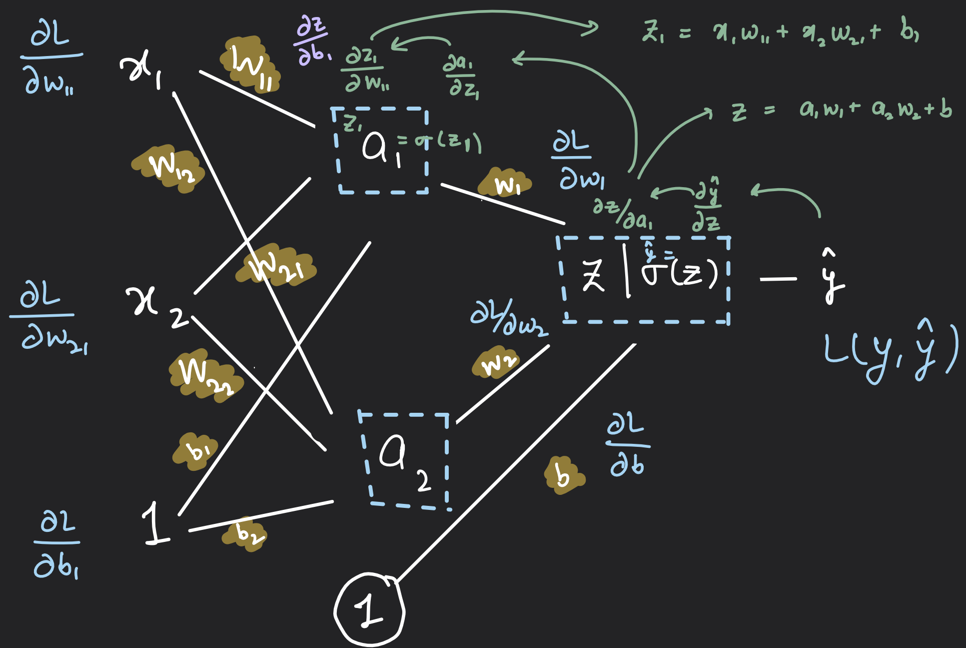
$$\frac{\partial L}{\partial b} = \frac{-(y-\hat{y})}{\cancel{\hat{y}(1-\hat{y})}} \cdot \cancel{\hat{y}(1-\hat{y})} = -(y-\hat{y})$$

Now go with gradient descent step of updating w_1, w_2, b.

Neural Network:



Goal : Adjust each of the highlighted weights and biases.



$$\frac{\partial L}{\partial w_{11}} = \frac{\partial z_1}{\partial w_{11}} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$= x_1 \cdot a_1(1-a_1) \cdot w_1 \cdot \hat{y}(1-\hat{y}) \cdot \frac{-(y-\hat{y})}{\hat{y}(1-\hat{y})}$$

$$= -x_1 w_1 a_1 (1-a_1)(y-\hat{y})$$

now perform GD on w_{11}

$$w_{11} \longrightarrow w_{11} - \alpha \cdot x_1 \, w_1 a_1 (1-a_1)(y-\hat{y})$$

For b_1

$$\frac{\partial L}{\partial b} = \frac{\partial z_1}{\partial b_1} \cdot \frac{\partial a_1}{\partial z_1} \cdot \frac{\partial z}{\partial a_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$= 1 \cdot a_1(1-a_1) \cdot w_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y-\hat{y})}{\cancel{\hat{y}(1-\hat{y})}}$$

$$= -w_1 a_1 (1-a_1)(y-\hat{y})$$

lly w_{12}.

For w_1,

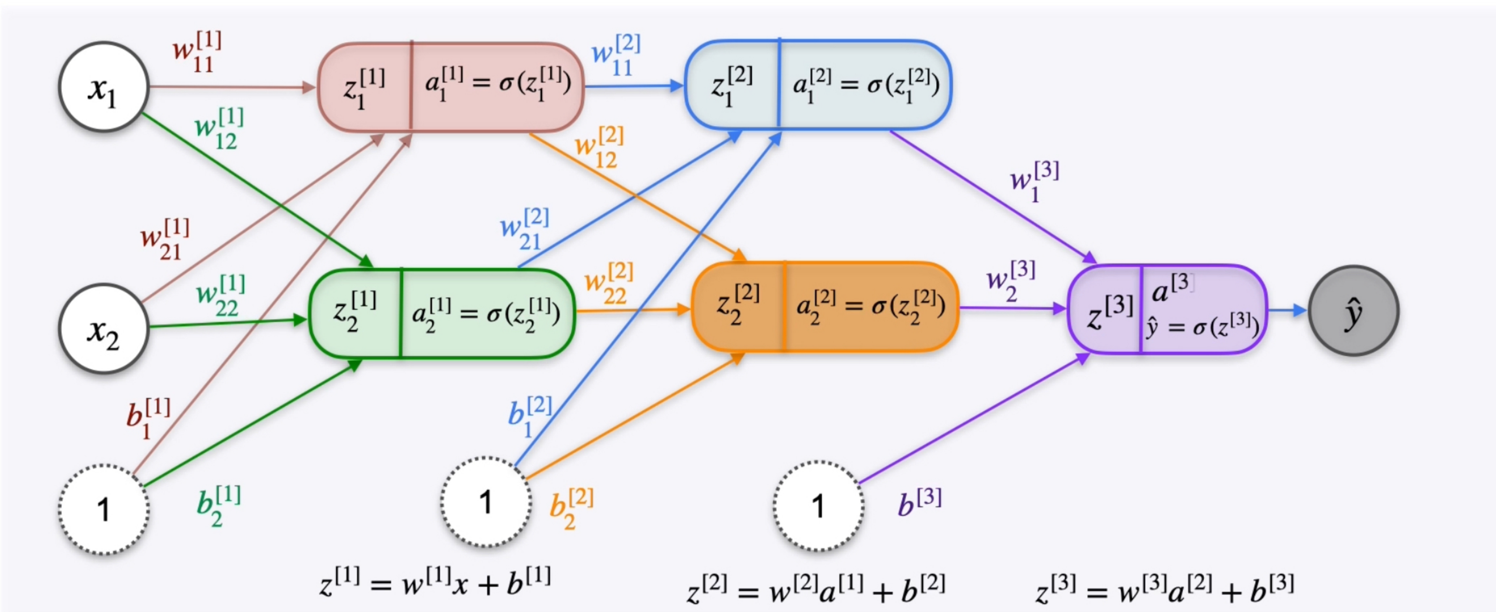
$$\frac{\partial L}{\partial w_1} = \frac{\partial z}{\partial w_1} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{dL}{\partial \hat{y}}$$

$$= a_1 \cdot \cancel{\hat{y}(1-\hat{y})} \cdot \frac{-(y-\hat{y})}{\cancel{\hat{y}(1-\hat{y})}}$$

$$= -a_1(y-\hat{y})$$

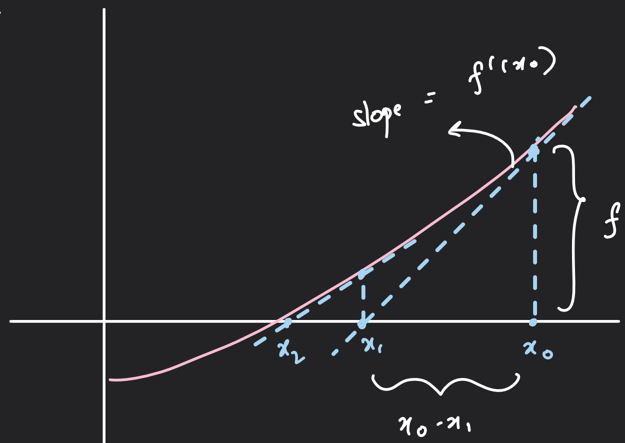
lly w_2 and b.

Back Propagation Introduction



Newton's Method : [Alternate to GD]

{ only find zero of $f(x)$ }



$$\frac{f(x_0)}{x_0 - x_1} = f'(x_0)$$

$$\frac{f(x_0)}{f'(x_0)} = x_0 - x_1$$

$$x_1 = x_0 - \frac{f(x)}{f'(x_0)}$$

NM for optimisation :

1. Start with some x_0
2. $$x_{k+1} = x_k - \frac{g'(x_k)}{(g'(x_k))'}$$
3. Repeat 2.) until you find minimum.

$f(x) \rightarrow g'(x)$

$f'(x) \rightarrow (g'(x))'$

Example :

$$g(x) = e^x - \log(x)$$

$$f(x) = g'(x) = e^x - \frac{1}{x} \qquad (\min x^* = 0.5671)$$

$$f'(x) = (g'(x))' = e^x + \frac{1}{x^2}$$

$x_0 = 0.05$

$$x_1 = x_0 - \frac{f(x)}{f'(x)} = 0.05 - \frac{\left(e^{0.05} - \frac{1}{0.05}\right)}{\left(e^{0.05} - \frac{1}{0.05^2}\right)} = 0.097$$

$$x_2 = x_1 - \frac{f(x)}{f'(x)} = 0.183$$

$x_3 = 0.320$

$x_4 = 0.477$

$x_5 = 0.558$

$x_6 = 0.567$

Second derivative:

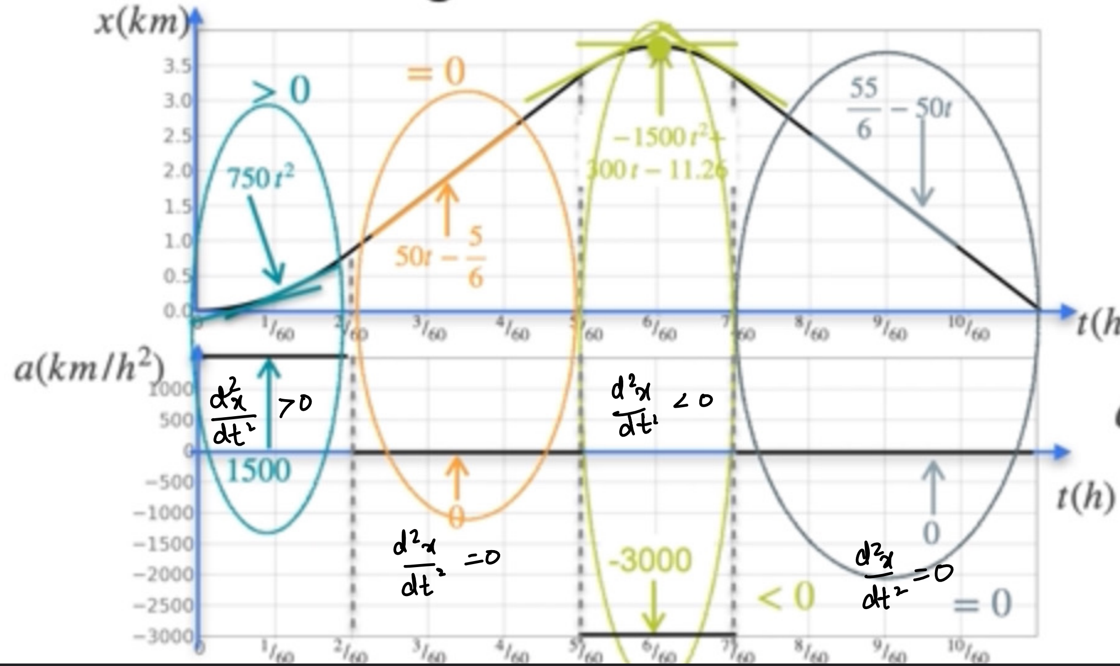
$$\frac{d^2 f(x)}{dx^2} = \frac{d}{dx}\left(\frac{df(x)}{dx}\right) \quad \text{(Leibniz notation)}$$

ex → acceleration

$$\frac{dv}{dt} \text{ or } \frac{d^2x}{dt^2}$$

Lagrange's ⇒ $f''(x)$

Understanding Second Derivative



x Distance

Second derivative tells us about the curvature

a Acceleration $\dfrac{d^2x}{dt^2}$

$\dfrac{d^2x}{dt^2} > 0$ $\cup$ convex or concave up

$\dfrac{d^2x}{dt^2} < 0$ $\cap$ concave down

$\dfrac{d^2x}{dt^2} = 0$ Need more info

Hessian : 2^{nd} Der. of 2 variables

$\dfrac{\partial^2 f}{\partial x^2} \Leftrightarrow f_{xx}(x,y)$ $\qquad$ $\dfrac{\partial^2 f}{\partial y^2} \Leftrightarrow f_{yy}(x,y)$

$\dfrac{\partial^2 f}{\partial y\,\partial x} \Leftrightarrow f_{yx}(x,y) \Longleftrightarrow \dfrac{\partial^2 f}{\partial x\,\partial y} \Leftrightarrow f_{xy}(x,y)$

both are same

Hessian matrix :

$$\begin{bmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{bmatrix}$$

For $f(x,y) = 2x^2 + 3y^2 - xy$

$$H(0,0) = \begin{bmatrix} 4 & -1 \\ -1 & 6 \end{bmatrix}$$

$$\det(H(0,0) - \lambda I) = \det\left(\begin{bmatrix} 4-\lambda & -1 \\ -1 & 6-\lambda \end{bmatrix}\right)$$

$$= (4-\lambda)(6-\lambda) - (-1)(-1)$$

$$= \lambda^2 - 10\lambda + 23$$

$$\Rightarrow \lambda_1 = 6.41 \text{ and } \lambda_2 = 3.5$$

> 0

so $(0,0)$ is min.

For $f(x,y) = -2x^2 + -3y^2 - xy + 15$

$$\nabla f(x,y) = \begin{bmatrix} -4x - y \\ -x - 6y \end{bmatrix}$$

$$H(0,0) = \begin{bmatrix} -4 & -1 \\ -1 & -6 \end{bmatrix}$$

$$\det(H(0,0) - \lambda I) = \lambda^2 + 10\lambda + 23 \rightarrow \lambda_1 = -3.59, \quad \lambda_2 = -6.41 \quad < 0$$

so

$(0,0)$ is max.

At saddle point, all eigenvalues are neither positive nor negative.

Summary

	1 variable $f(x)$	2 variables $f(x, y)$	More variables $f(x_1, x_2, \ldots, x_n)$
(Local) minima	Happy face $f''(x) > 0$	Upper paraboloid $\lambda_1 > 0 \,\&\, \lambda_2 > 0$	All $\lambda_i > 0$
(Local) maxima	Sad face $f''(x) < 0$	Down paraboloid $\lambda_1 < 0 \,\&\, \lambda_2 < 0$	All $\lambda_i < 0$
Need more information	$f''(x) = 0$	Saddle point $\lambda_1 > 0 \,\&\, \lambda_2 < 0$ $\lambda_1 < 0 \,\&\, \lambda_2 > 0$ Or some $\lambda_i = 0$	Some $\lambda_i > 0$ and some $\lambda_j < 0$ OR At least one $\lambda_i = 0$

Newton's method for 2 variables :

1 var $\quad x_{k+1} = x_k - f'(x_k) \cdot f''(x_k)^{-1}$

2 var $\quad \begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - \underbrace{H^{-1}(x_k, y_k)}_{2\times 2} \underbrace{\nabla f(x_k, y_k)}_{2\times 1}$

Example :

$$f(x,y) = x^4 + 0.8y^4 + 4x^2 + 2y^2 - xy - 0.2x^2y$$

$$f(x,y) \xrightarrow{x} 4x^3 + 8x - y - 0.4xy \quad \begin{cases} \xrightarrow{x} 12x^2 + 8 - 0.4y \\ \xrightarrow{y} -1 - 0.4x \end{cases}$$

$$f(x,y) \xrightarrow{y} 3.2y^3 + 4y - x - 0.2x^2 \quad \begin{cases} \xrightarrow{x} -1 - 0.4x \\ \xrightarrow{y} 9.6y^2 + 4 \end{cases}$$

$$\nabla f(x,y) = \begin{bmatrix} 4x^3 + 8x - y - 0.4xy \\ 3.2y^3 + 4y - x - 0.2x^2 \end{bmatrix}$$

$$H(x,y) = \begin{bmatrix} 12x^2 + 8 - 0.4y & -1 - 0.4x \\ -1 - 0.4x & 9.6y^2 + 4 \end{bmatrix}$$

Start at $\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$

$$\nabla f(4,4) = \begin{bmatrix} 277.6 \\ 213.6 \end{bmatrix} \qquad H(4,4) = \begin{bmatrix} 198.4 & -2.6 \\ -2.6 & 157.6 \end{bmatrix}$$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \end{bmatrix} - \begin{bmatrix} 198.4 & -2.6 \\ -2.6 & 157.6 \end{bmatrix}^{-1} \begin{bmatrix} 277.6 \\ 213.6 \end{bmatrix}$$

$$= \begin{bmatrix} 2.58 \\ 2.62 \end{bmatrix}$$

Repeat until you are close to zero

$$\begin{bmatrix} x_8 \\ y_8 \end{bmatrix} = \begin{bmatrix} 4.15 \; 10^{-17} \\ -2.05 \; 10^{-17} \end{bmatrix}$$