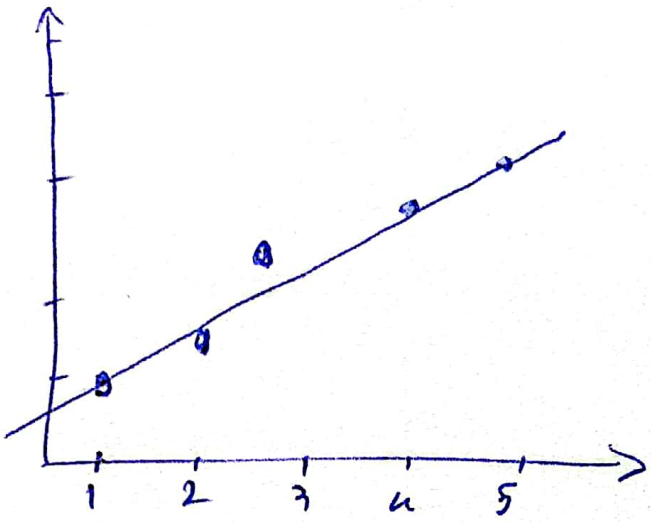
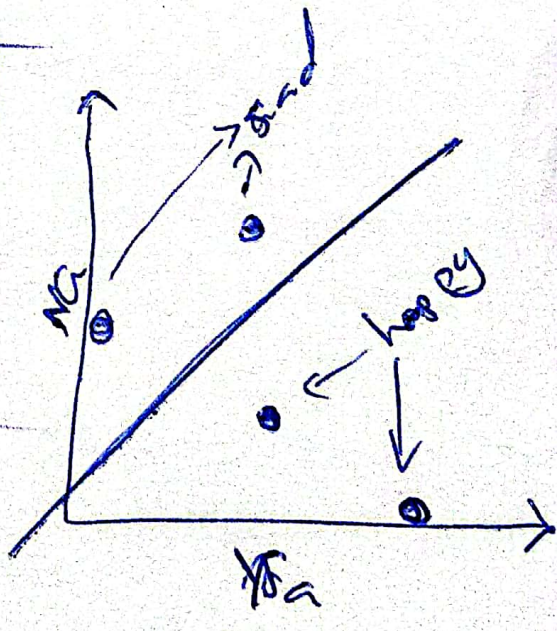
- A linear regression problem:



- A classification problem:

a sentiment analysis model

Sentence	Ya	Na	mood
Ya Ya Ya	3	0	happy
Na Na	0	2	Sad
Ya Na Ya	2	1	happy
Ya Na Na Na	1	3	Sad



maths concepts used in training a model

- Gradients
- Derivatives
- optimization
- Loss and cost functions
- Gradient Descent

models

Linear regression

Classification

~~Linear~~

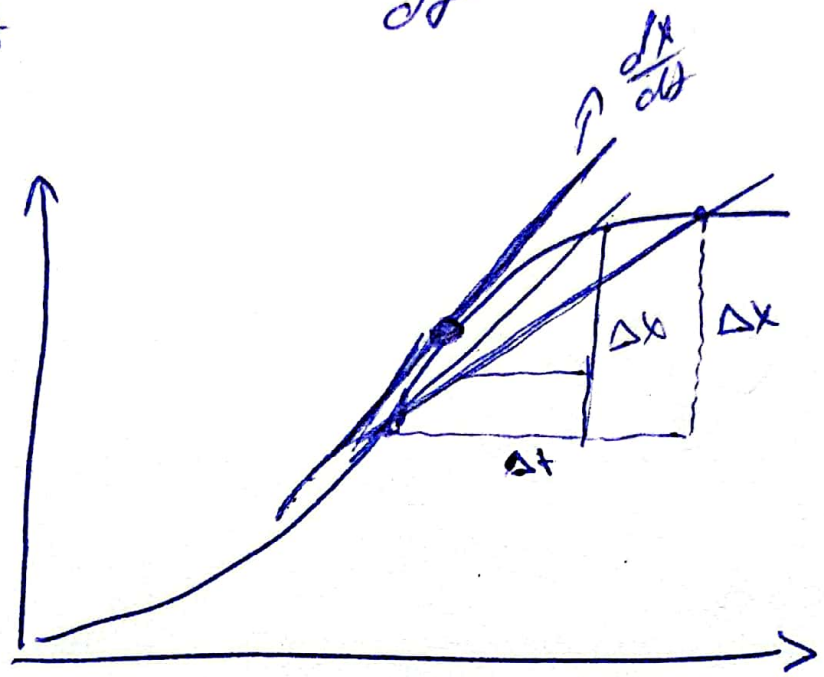
"Derivative"

Derivative: the instantaneous velocity.

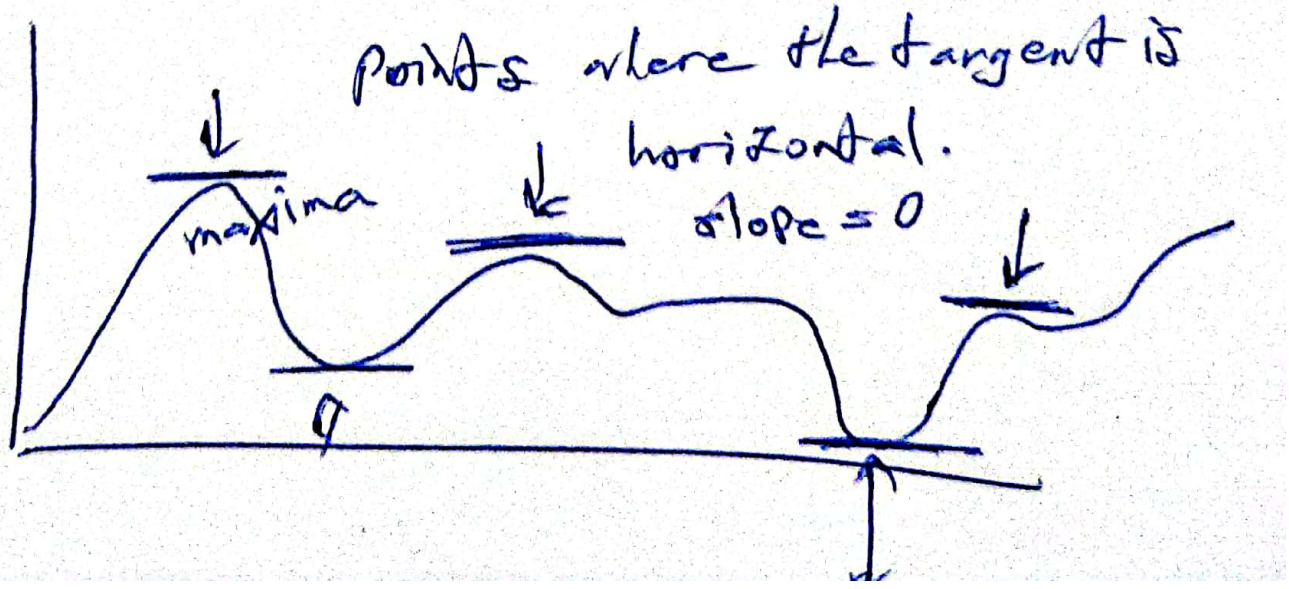
$$\text{slope} = \frac{\text{rise}}{\text{run}} \Rightarrow \text{slope} = \frac{\text{change in distance } (\Delta x)}{\text{change in time } (\Delta t)}$$

CamScanner

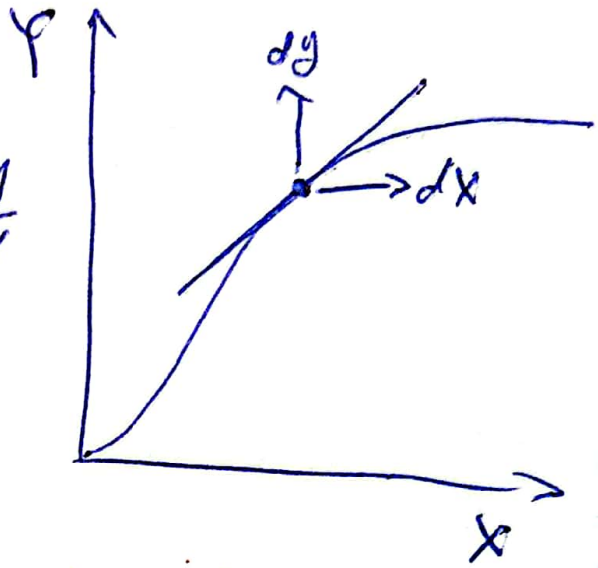
$\frac{\Delta x}{\Delta t} \rightarrow \rightarrow \rightarrow \frac{dx}{dt}$



Derivative: the derivative of a function at a point is precisely the slope of the tangent at that particular point.



slope at a point $= \frac{dy}{dx}$



if: $y = f(x)$

slope $= f'(x)$

↳ Lagrange's notation.

$$\frac{dy}{dx} = \frac{d}{dx} f(x)$$

↳ Leibniz's notation.

$y = f(x) = c$

$f'(x) = 0$

$f(x) = ax + b$

$f'(x) = a$

Quadratics

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

$f'(x) = 2x$

Cubics

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

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

$$y = f(x) = \frac{1}{x} = x^{-1}$$

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

---

so:

$$f(x) = x^n$$

$$f'(x) = \frac{d}{dx} f(x) = n x^{n-1}$$

Inverse function

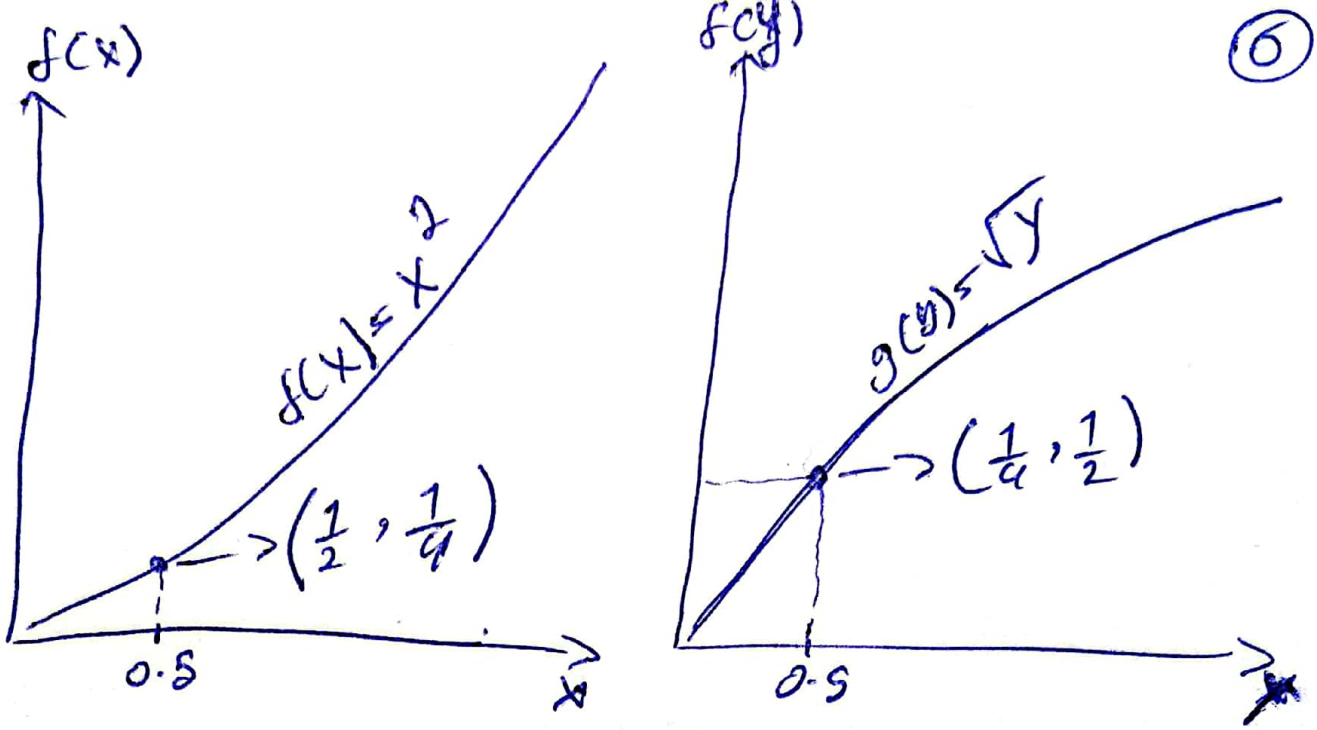
$$x \xrightarrow[\downarrow x^2]{f} x^2 \xrightarrow[\sqrt{x}]{g} x \qquad \text{for } x > 0$$

if $g(x)$ and $f(x)$ are inverses

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

$$g(f(x)) = x$$

CamScanner



if $g(y) = \sqrt{y}$ $\quad g'(y) = \frac{1}{2\sqrt{y}}$

- if f and g are inverse functions, then the derivative of g is 1 over the derivative of x.

---

if $f(x) = \sin(x)$

$\Rightarrow f'(x) = \cos(x)$

if $f(x) = \cos(x)$

$\Rightarrow f'(x) = -\sin(x)$

"Euler's number"

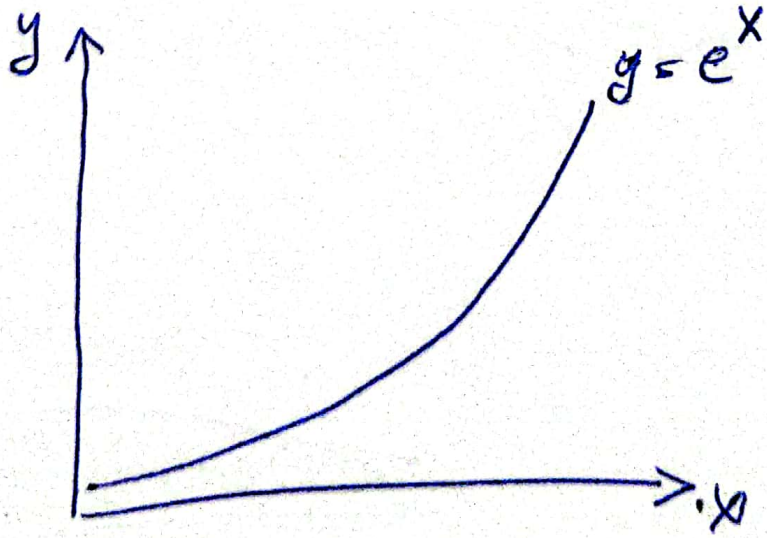
$$2.718281828\ldots$$

$$\frac{n}{\left(1+\frac{1}{n}\right)^n} = \frac{1}{2} = \frac{10}{2.594} = \frac{100}{2.717} \cdots \frac{\infty}{e}$$

$\Leftrightarrow$ if $f(x) = e^x$

then: $f'(x) = e^x$

"exponential function"



Logarithm

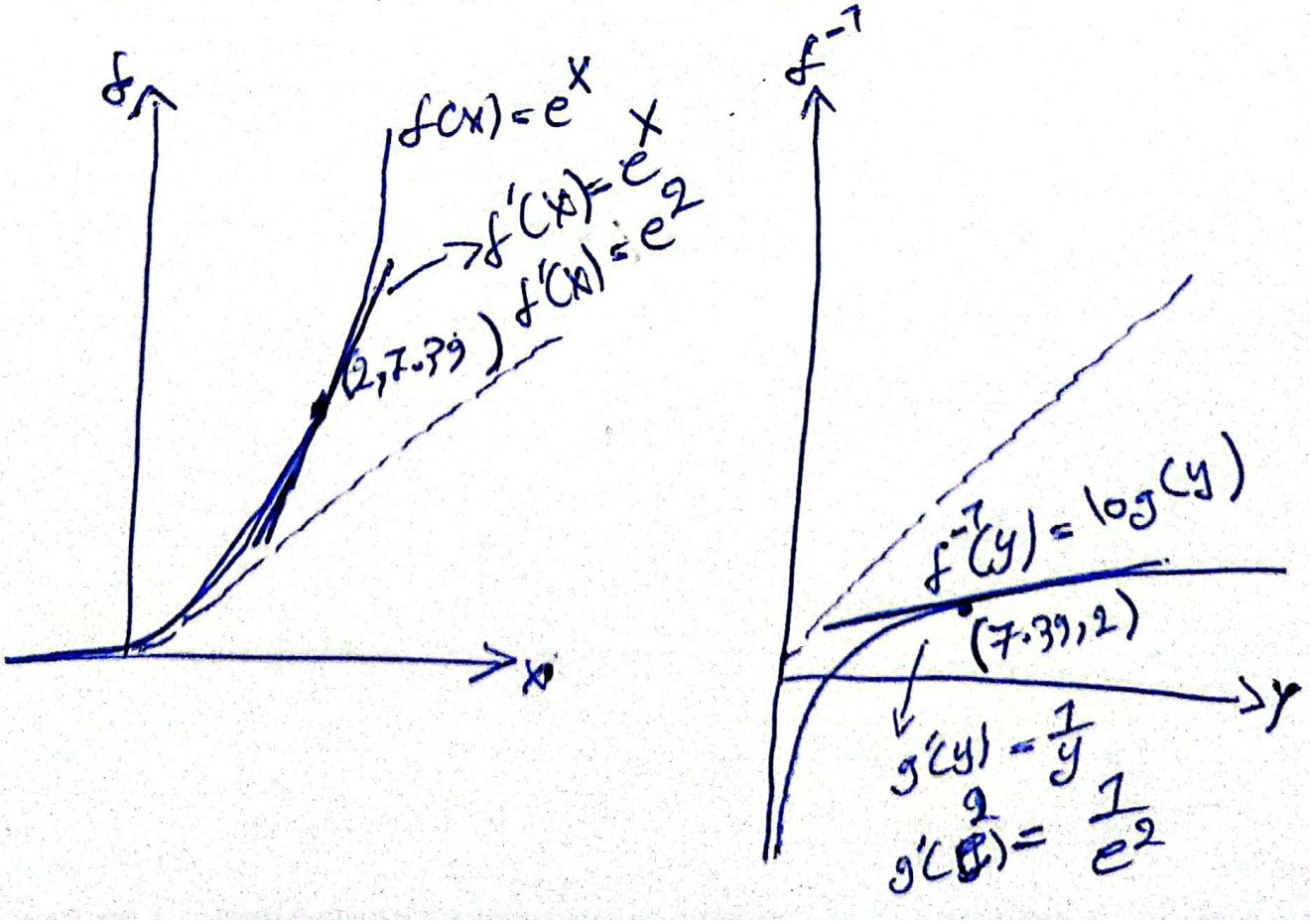
$e^{?} = 3 \Rightarrow \log_e(3)$

$e^{?} = X \Rightarrow \log(X)$ natural logarithm

$\hookrightarrow \ln(X)$

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

$e^{\log(x)} = x$ $\quad \log(e^y) = y$



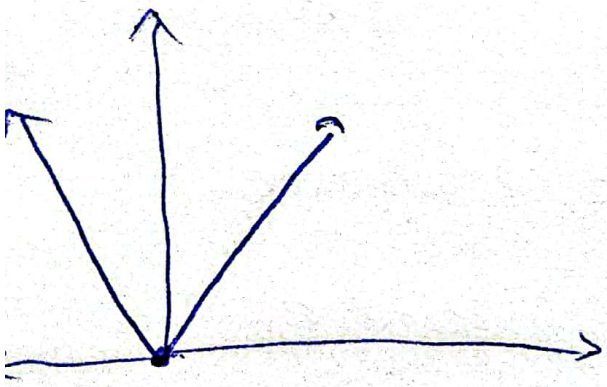
$$\frac{d}{dy} f^{-1}(y) = \frac{1}{f'(f^{-1}(y))}$$

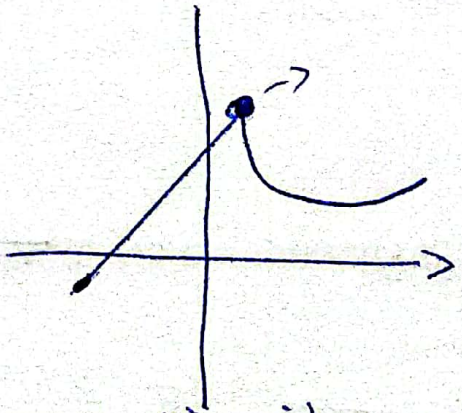
$$\frac{d}{dy} \log(y) = \frac{1}{e^{\log(y)}} = \frac{1}{y}$$

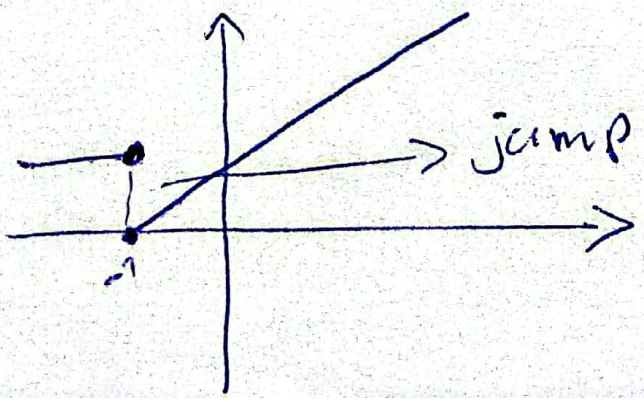
$$\frac{d}{dy} \log(y) = \frac{1}{y}$$

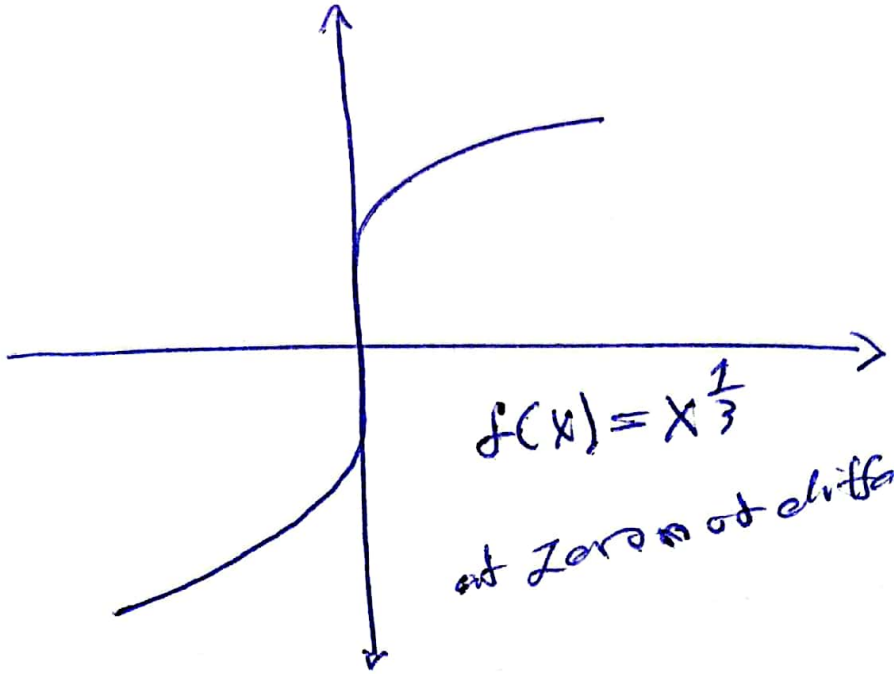
"non-differentiable functions"

$f(x) = |x|$ non-differentiable









non-diffantiable functions

1) any function with a cusp or a corner

2) any function with a jump discontinuity

3) any function with a virtical tangent

properties of derivative

1) multiplication ~~of~~ by a scalar

~~if~~ if $f = cg$

then $f' = cg'$

2) sum rules

if $f(x)=2x$ and $g(x)=x^2$

$(f+g)'(x) = ?$

if $f = g + h$

then $f' = g' + h'$ $\Rightarrow f'(x) + g'(x) = 2 + 2x$

---

3) the product rule

if $f(x) = xe^x$

$f'(x) = ?$

if $f = gh$

then $f' = g'h + gh'$

$1 \cdot e^x + x \cdot e^x$

~~$x \cdot e^x + x \cdot e^x$~~

$e^x + xe^x$

---

4) the chain rule

$$\frac{d}{dt} g(h(t)) = \frac{dg}{dh} \cdot \frac{dh}{dt}$$

$$= g'(h(t)) \cdot h'(t)$$

$$\frac{d}{dt} f(g(h(t))) = \frac{df}{dg} \cdot \frac{dg}{dh} \cdot \frac{dh}{dt}$$

$$= f'(g(h(t))) \cdot g'(h(t)) \cdot h'(t)$$

CamScanner

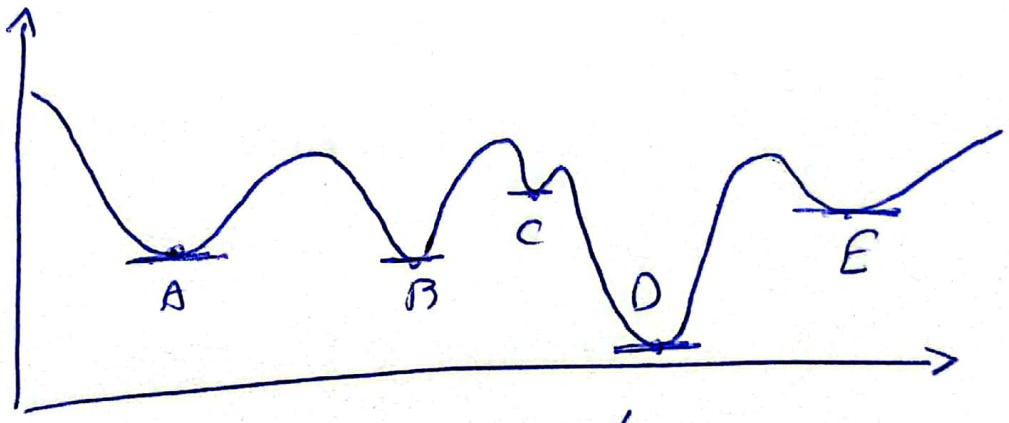
if $f(x) = e^{2x}$

$f'(x) = ?$

$e^{2x} \cdot 2$

---

optimization



slope = 0 in A, B, C and D

D is global minimum.

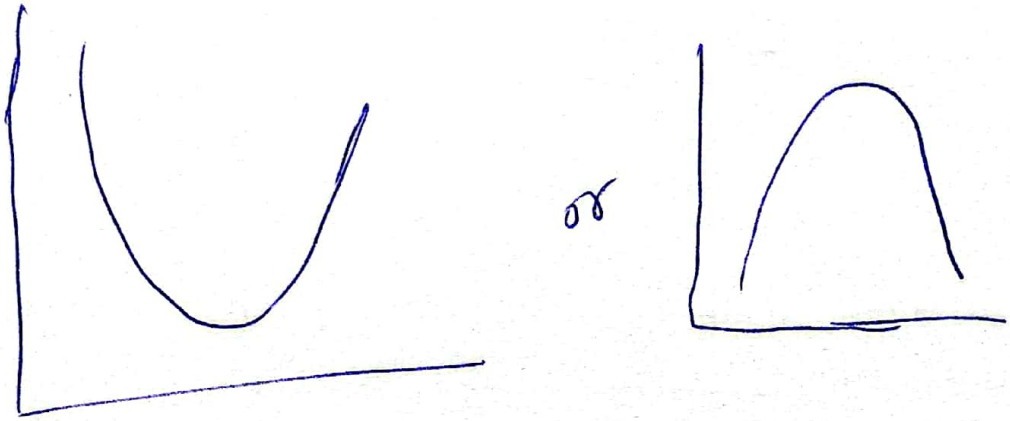
A, B, C and E are local minima.

"Two power line problem"

$(x-a)^2 + (x-b)^2$ = total cost

↓ (we should minimize the total cost.)

aggravation function



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

minimum or maximum point

$$2(x-a) + 2(x-b) = 0$$

$$2x - 2a + 2x - 2b = 0$$

$$4x - 2a - 2b = 0$$

$$x = \frac{a+b}{2}$$

CamScanner

"three point line problem"

$$\frac{d}{dx}\left[(x-a)^2+(x-b)^2+(x-c)^2\right]=0$$ minimum point

$$2(x-a)+2(x-b)+2(x-c)^2=0$$

$$2x-a+2x-2b+2x-2c=0$$

$$3x-a-b-c=0$$

$$\boxed{x=\frac{a+b+c}{3}}$$

"In general The square loss"

$$\text{minimize}\ (x-a_1)^2+(x-a_2)^2+\cdots+(x-a_n)^2$$

$$\text{solution: } x=\frac{a_1+a_2+\cdots+a_n}{n}$$

"optimization of log loss"

coin toss
T = 0.7 , H = 0.3

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

$$\frac{d}{dp}\left[p^7 (1-p)^3\right] = 0$$

$$7p^6 (1-p)^3 + p^7 \cdot 3(1-p)(-1) = 0$$

$$= p^6 (1-p)^2 (7 - 10p) = 0$$

$p = 0$ ✗ $\quad p = 1$ ✗ $\quad p = 0.7$ ✓

maximizing of $g(p)$ is the same as

maximizing of $\log(g(p))$ so:

$$\log\left(p^7 \cdot (1-p)^3\right) = 0$$

$$\Rightarrow \log p^7 + \log\left((1-p)^3\right) = 0$$

CamScanner

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

$$\frac{dG(p)}{dp} = \frac{d}{dp}\left[7\log(p) + 3\log(1-p)\right]$$

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

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

$$7(1-p) - 3p \leq 0$$

$$7 - 7p - 3p \leq 0$$

$$-10p \leq -7$$

$$p \leq \frac{7}{10} \leq 0.7$$

Derivative of product is hard, but the derivative of sums is much easier so we use logarithm to make a product problem to be a sum problem.

"Tangent Planes"

"Partial derivative"

$f(x,y) = x^2 + y^2$ → plot in 3-Dimensions

fix $y = 4 \Rightarrow f(x,4) = x^2 + 4^2$

$\frac{d}{dx}(f(x,4)) = 2x$ → Partial derivative of f with respect to x

fix $x = 2 \Rightarrow f(2,y) = 2^2 + y^2$

$\frac{d}{dy}(f(2,y)) = 2y$ ↔ Partial derivative of f with respect to y

if we fix one variable, so the function will be a function with only one variable. and the slope of this function will be "Partial derivative"

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

$\frac{df}{dx} = 6y^3x$ → Partial derivative of f with respect to x.

$\frac{df}{dy} = 9x^2y^2$ → Partial derivative of f with respect to y.

Gradient

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

$\frac{df}{dx} = 2x$

$\frac{df}{dy} = 2y$

$\text{Gradient} = \begin{bmatrix} 2x \\ 2y \end{bmatrix}$

↓

a vector contains both Partial derivatives.

in general Gradient → $\nabla f = \begin{bmatrix} \frac{df}{dx} \\ \frac{df}{dy} \end{bmatrix}$

this vector defines the slope of the two lines of the plane.

↑ tangent

if $f(x,y) = x^2 + y^2$, calculate the gradient f, ∇f, at $(2,3)$

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

$\frac{df}{dx} = 2x$ $\qquad \nabla f = \begin{bmatrix} 2x \\ 2y \end{bmatrix}$

$\frac{df}{dy} = 2y$

so $\nabla f = \begin{bmatrix} 4 \\ 6 \end{bmatrix}$

Gradient and minima and maxima

The minimum of the this function $f(x,y) = x^2 + y^2$ is where when both of the slops of tangent lines are zero in other words the tangent plane in this case (minimas) is Parallel to the floor.

So minimas in this case is:

$\frac{df}{dx} = 0$ and $\frac{df}{dy} = 0$

so

$2x = 0$ and $2y = 0$

$(x, y) = (0, 0)$ ← minimum point.

---

if $f(x,y) = 85 + \frac{1}{90}x^2(x-5)y^2(y-5)$

$\frac{df}{dx} = ?$

$\frac{df}{dx} = \cancel{\frac{1}{45}x} - \frac{1}{90}x(3x-12)y^2(y-5)$

$\frac{df}{dy} = -\frac{1}{90}x^2(x-5)y(3y-12)$

find the minimum

$x=0$ → $x=4$ → $y=0$ → $y=5$

$-\frac{1}{90}[x][(3x-12)][y^2][(y-5)] = 0$?

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

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

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

$$\frac{dE}{dm} = 28m + 12b - 42$$

$$\frac{dE}{db} = 6b + 12m - 20$$

$$\begin{cases} 28m + 12b - 42 = 0 \\ (6b + 12m - 20 = 0) \times 2 \end{cases}$$

$$12b + 24m - 40 = 0$$

$$-\ 12b + 28m - 42 = 0$$

$$-4m + 2 = 0$$

$$\boxed{m = \frac{1}{2}}$$

$$\boxed{b = \frac{7}{3}}$$

$E(m=\frac{1}{2}, b=\frac{7}{3}) \approx 4.167$ the minimum value of the cost.

"Gradient descent"

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

$$f'(x) = e^x - \frac{1}{x}$$

$$e^x - \frac{1}{x} = 0$$

$$e^x = \frac{1}{x}$$

$x = 0.5671\ldots$ → omega constant.

in gradient descent:

$$\text{new point} = \text{old point} - (\text{slope} \times \underbrace{\text{learning-rate}}_{\alpha})$$

$$X_1 = X_0 - \underbrace{[f'(X_0) \times \text{learning-rate}]}_{\alpha}$$

Gradient Descent Algorithm

function: $f(X)$ goal: find minimum of $f(X)$

step 1:

Define a learning rate (α)

choose a starting point (X_0)

step 2:

update: $X_k = X_{k-1} - \alpha f'(X_{k-1})$

step 3:

repeat step 2 until you are close enough to the true minimum X^*

exp: $f(X) = e^X - \log(X)$ $f'(X) = e^X - \frac{1}{X}$

start: $X = 0.05$ rate: $\alpha = 0.005$

find: $f'(0.05) = -18.9$

move by $-0.005 f'(0.05)$ $X \longrightarrow 0.1447$

CamScanner

find:

$f'(0.1447) = -5.7552$

more by $-0.005 f'(0.1447)$ $x_1 \longrightarrow 0.1735$

Idea for gradient descent

Initial position: (x_0, y_0)

direction of ~~gradient ascent~~ greatest ascent: ∇f

direction of ~~gradient descent~~ greatest descent: $-\nabla f$

updated position: $(x_0, y_0) - a\nabla f$

learning rate

better point

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

start: $x = 0.5$, $y = 0.6$

$$\nabla f = \begin{bmatrix} \frac{df}{dx} \\ \frac{df}{dy} \end{bmatrix}$$

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

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

CamScanner

$$\nabla f = \begin{bmatrix} -\frac{1}{90} x(3x-72) y^2 (y-6) \\ -\frac{1}{90} x^2 (x-6) y (5y-72) \end{bmatrix}$$

$$\nabla f(0.5, 0.6) = \begin{bmatrix} -0.7734 \\ -0.0935 \end{bmatrix}$$

move by: $-0.05 \cdot \nabla f(05, 0.6)$

↙ learning-rate

$x \longmapsto 0.5057$

$y \longrightarrow 0.6047$

⋮

repeat the process many times.

Linear regression : Gradient Descent

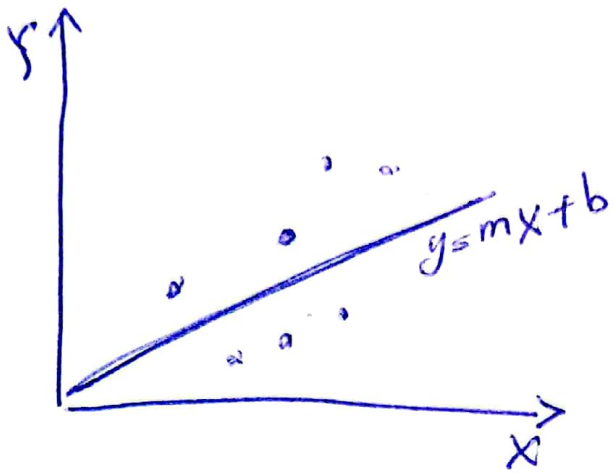
$$\nabla E = [28m + 72b - 42, 6b + 72m - 20]$$

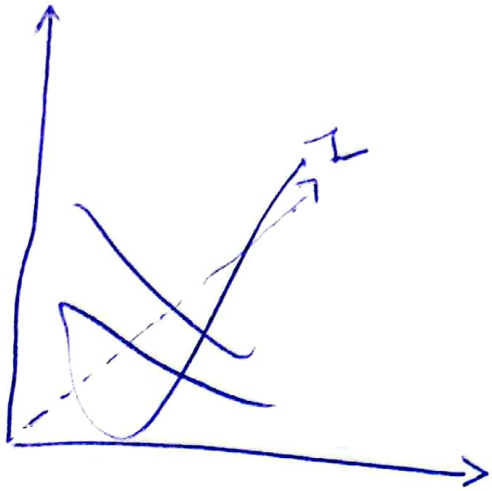
$m = ?$
$b = ?$

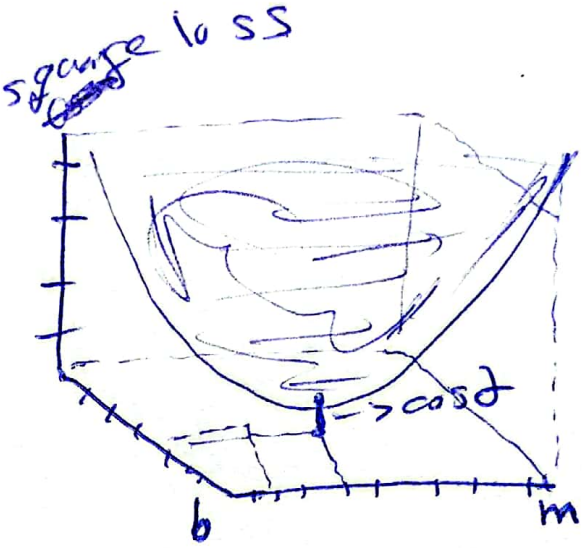
start with (m_0, b_0) → some random points (initial)

iterate:

$$(m_{k+1}, b_{k+1}) = (m_k, b_k) - \alpha \nabla E(m_k, b_k)$$







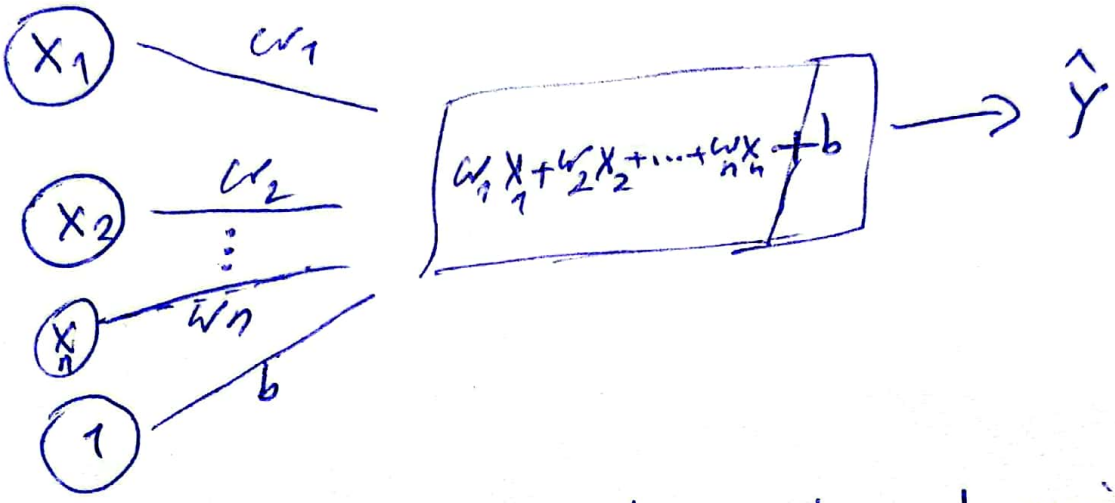
$$\mathcal{L}(m, b) = \frac{1}{2n}\left[(mx_1 + b - y_1)^2 + \cdots + (mx_n + b - y_n)^2\right]$$

cost function

$$\begin{bmatrix} m_n \\ b_n \end{bmatrix} = \begin{bmatrix} m_{n-1} \\ b_{n-1} \end{bmatrix} - \alpha \nabla \mathcal{L}_1 (m_{n-1}, b_{n-1})$$

"regression with a perceptron"

"Single Layer neural network Perceptron"



the goal is to find the optimal weights and bias. $(w_1, w_2, \dots, b)$

how: minimize the error in the prediction with using the loss function

Regression with a Perceptron

Prediction function:

$$\hat{y} = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b$$

Loss function

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

main goal:

find ~~the~~ $w_1, w_2, \ldots, w_n, b$ that give $\hat{y}$ with the least error.

~~To find optimal values for: $w_1, w_2, \ldots, w_n, b$~~

we use gradient descent

$$w_1 \longrightarrow w_1 - \alpha \frac{dL}{dw_1}$$

$$w_2 \longrightarrow w_2 - \alpha \frac{dL}{dw_2}$$

$$\vdots$$

$$w_n \longrightarrow w_n - \alpha \frac{dL}{dw_n}$$

$$b \longrightarrow b - \alpha \frac{dL}{db}$$

Regression with a Perceptron

Prediction function:

$$\hat{y} = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b$$

Loss function

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

main goal:

find ~~the~~ $w_1, w_2, \ldots, w_n, b$ that give $\hat{y}$ with the least error.

~~To find optimal values for: $w_1, w_2, \ldots, w_n, b$~~

we use gradient descent

$$w_1 \longrightarrow w_1 - \alpha \frac{dL}{dw_1}$$

$$w_2 \longrightarrow w_2 - \alpha \frac{dL}{dw_2}$$

$$\vdots$$

$$w_n \longrightarrow w_n - \alpha \frac{dL}{dw_n}$$

$$b \longrightarrow b - \alpha \frac{dL}{db}$$

CamScanner

~~$\hat{y} = w_1 x_1 + w_2 x_2 + b$~~

~~$\frac{dL}{db} =$~~

prediction function: $\hat{y} = w_1 x_1 + w_2 x_2 + b$

loss function: $L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$

$$\frac{dL}{db} = \frac{dL}{d\hat{y}} \times \frac{d\hat{y}}{db} = -(y - \hat{y}) \times 1$$

$$\frac{dL}{dw_1} = \frac{dL}{d\hat{y}} \times \frac{d\hat{y}}{dw_1} = -(y - \hat{y}) \times x_1$$

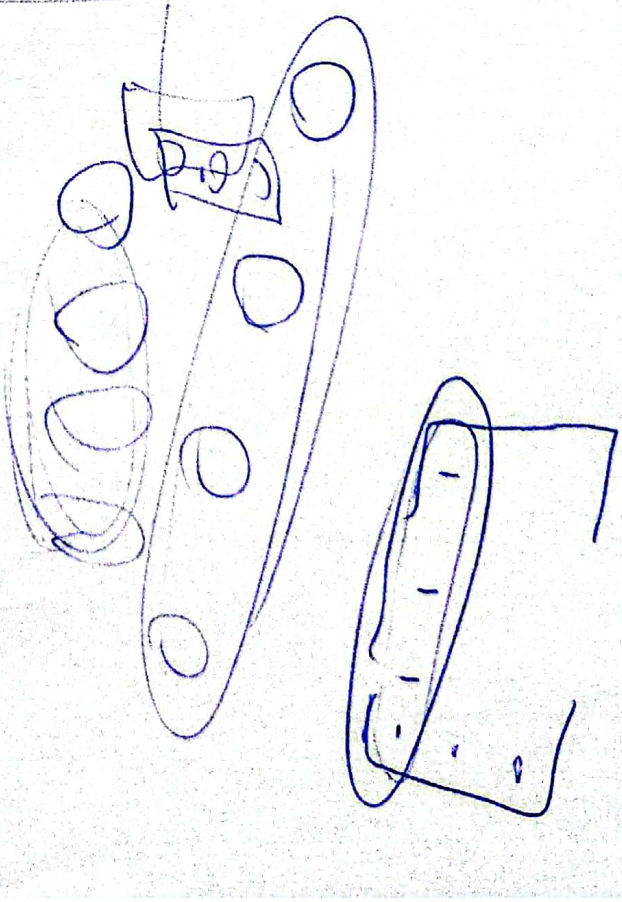
$$\frac{dL}{dw_2} = \frac{dL}{d\hat{y}} \times \frac{d\hat{y}}{dw_2} = -(y - \hat{y}) \times x_2$$

$$\frac{dL}{d\hat{y}} = -(y - \hat{y})$$

$$\frac{d\hat{y}}{db} = 1$$

$$\frac{d\hat{y}}{dw_1} = x_1$$

$$\frac{d\hat{y}}{dw_2} = x_2$$

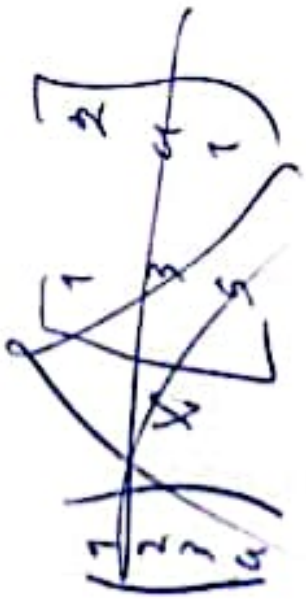


So: update the parameters

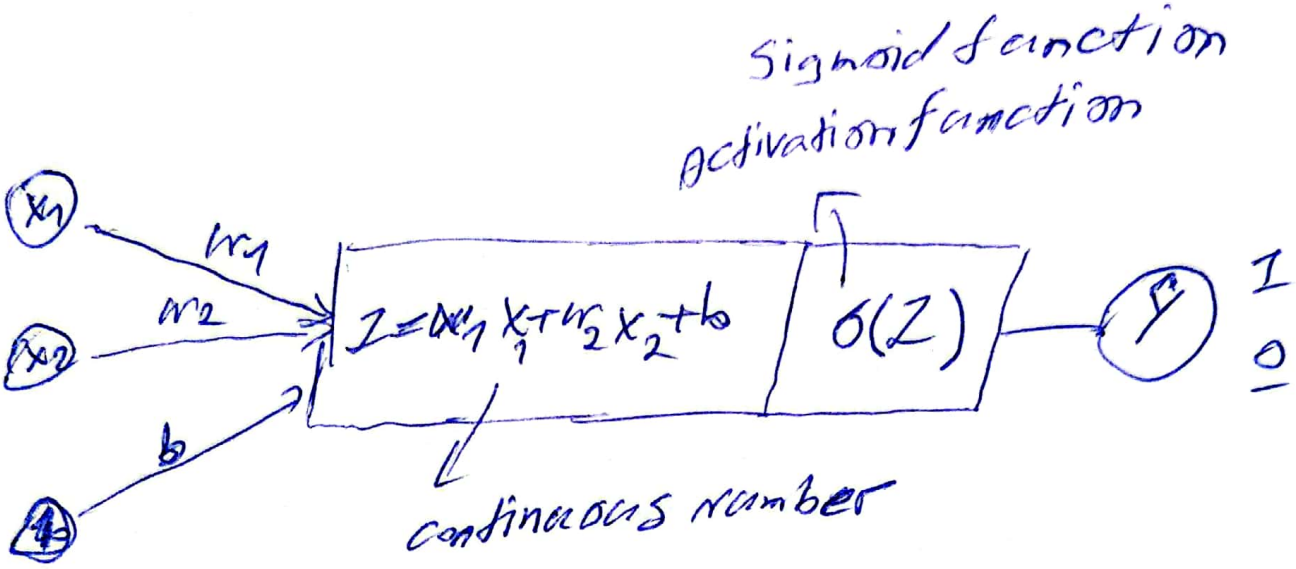
$$w_1 \longrightarrow w_1 - \alpha\left(-x_1(y-\hat{y})\right)$$

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

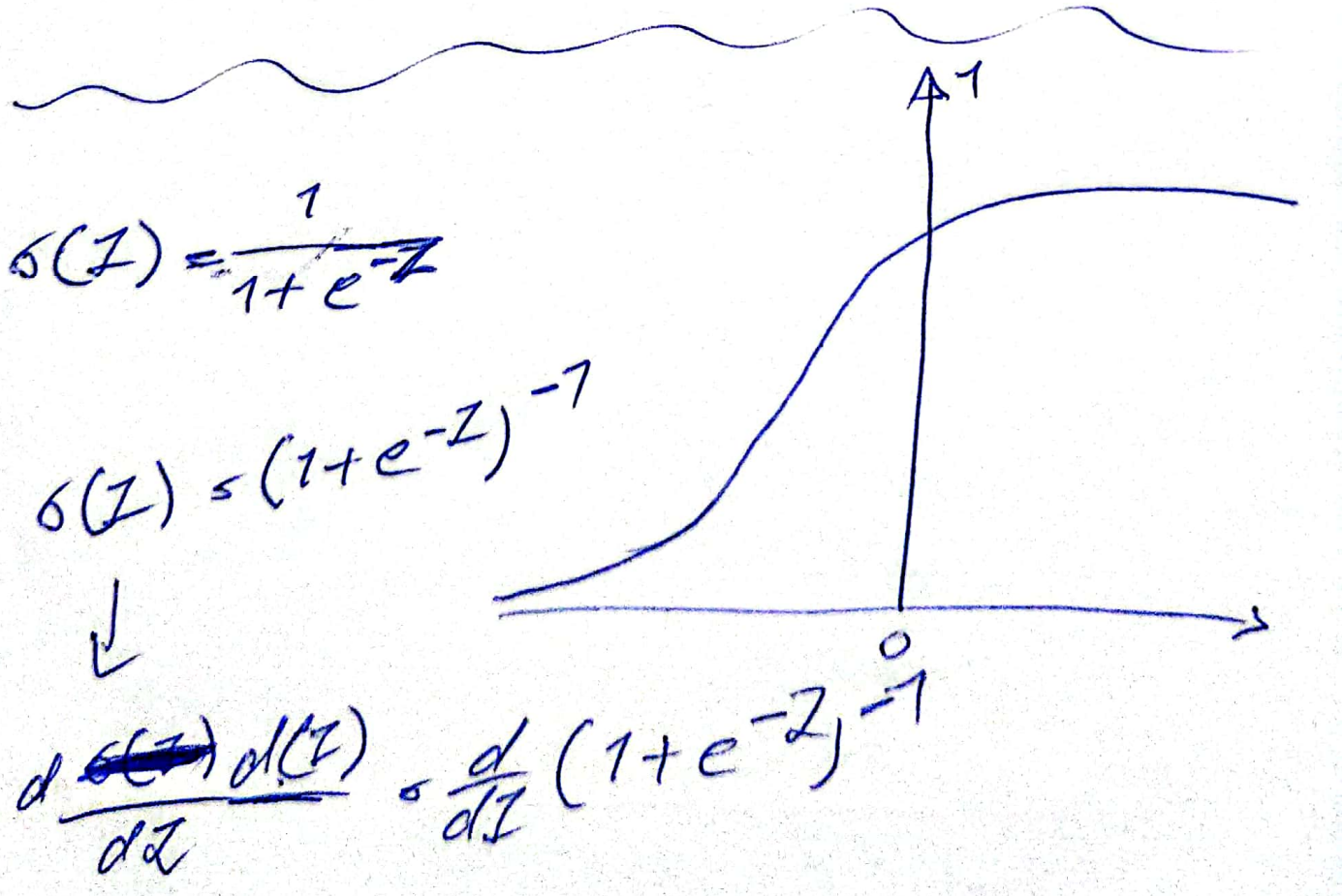
$$b \longrightarrow b - \alpha\left(-(y-\hat{y})\right)$$



classification



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



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

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

$$\frac{d\sigma(Z)}{dZ} = \frac{d}{dZ}(1+e^{-Z})^{-1}$$

$$\frac{d}{dI}\sigma(I) = -1(1+e^{-I})^{-2-1}\cdot\left(\frac{d}{dI}(1+e^{-I})\right)$$

$$= -1(1+e^{-I})^{-2}\cdot\left(\frac{d}{dI}(1)+\frac{d}{dI}(e^{-I})\right)$$

$$= -1(1+e^{-I})^{-2}\cdot(0+e^{-I}(\frac{d}{dI}(-I)))$$

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

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

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

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

$$\frac{d}{dI}\sigma(I) = \frac{e^{-I}+1-1}{(1+e^{-I})^2}$$

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

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

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

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

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

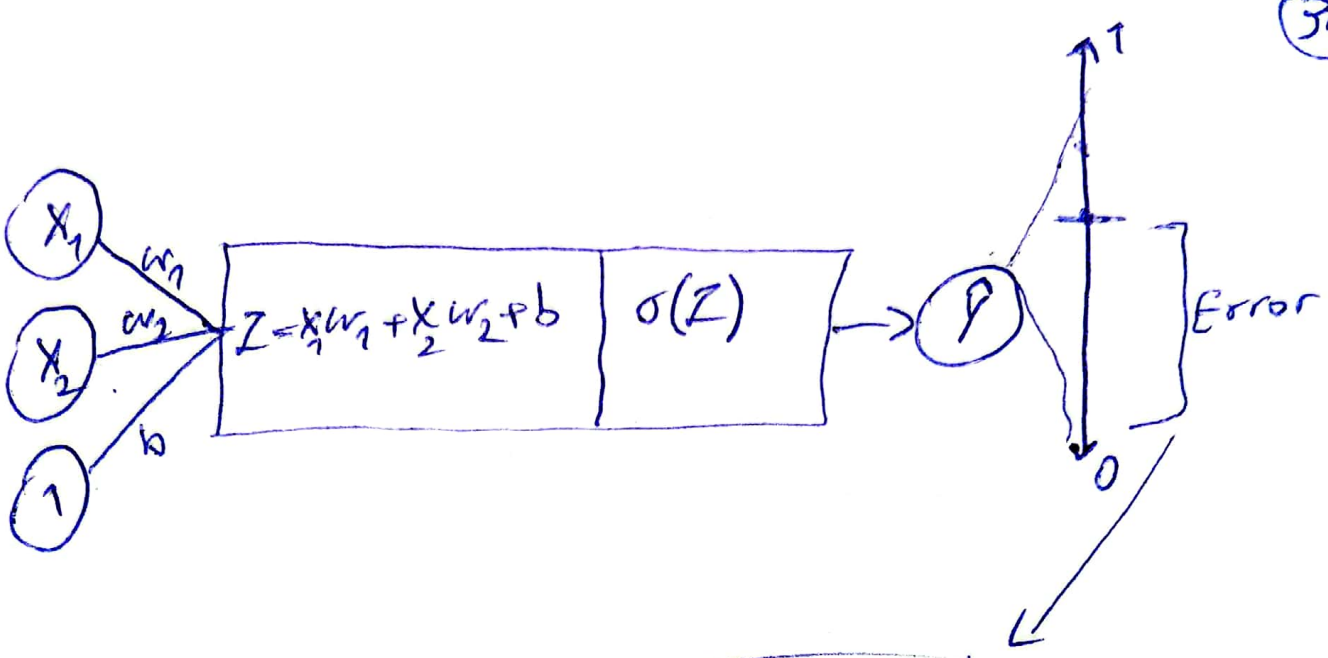
Recall that: $\sigma(z) = \frac{1}{1+e^{-z}}$

So:

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

$\hat{y}$ $\hat{y}$

$$\frac{d\hat{y}}{dy} \quad \hat{y}(1-\hat{y})$$



Log loss

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

note:

$L(y, \hat{y})$ is a large number if y and $\hat{y}$ are far from each other and it is a small number if y and $\hat{y}$ are close to each other.

To find optimal values for:

w_1, w_2, b

we need gradient descent:

$$w_1 = w_1 - \alpha \frac{dL}{dw_1}$$

$$w_2 = w_2 - \alpha \frac{dL}{dw_2}$$

$$b = b - \alpha \frac{dL}{db}$$

(?)

$$\frac{dL}{dw_1} = \frac{dL}{d\hat{y}} \cdot \frac{d\hat{y}}{dw_1} \quad ?$$

$$\frac{dL}{dw_2} = \frac{dL}{d\hat{y}} \cdot \frac{d\hat{y}}{dw_2} \quad ?$$

$$\frac{dL}{db} = \frac{dL}{d\hat{y}} \cdot \frac{d\hat{y}}{db} \quad ?$$

$$\hat{y} = \sigma(w_1 x_1 + w_2 x_2 + b)$$

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

$$\frac{dL}{d\hat{y}} = \frac{-y}{\hat{y}} + \frac{1-y}{1-\hat{y}}$$

$$= \frac{-y + y\hat{y} + \hat{y} - y\hat{y}}{\hat{y}(1-\hat{y})}$$

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

$$\hat{y} = \sigma(w_1 x_1 + w_2 x_2 + b)$$

$$\frac{d\hat{y}}{dw_1} = \hat{y}(1-\hat{y}) x_1$$

$$\frac{d\hat{y}}{dw_2} = \hat{y}(1-\hat{y}) x_2$$

$$\frac{d\hat{y}}{db} = \hat{y}(1-\hat{y})$$

So:

$$\frac{dL}{db} = \frac{dL}{d\hat{y}} \cdot \frac{d\hat{y}}{db} = \frac{-(y-\hat{y})}{\hat{y}(1-\hat{y})} \cdot \hat{y}(1-\hat{y})$$

$$\frac{dL}{dw_1} = \frac{dL}{d\hat{y}} \cdot \frac{d\hat{y}}{dw_1} = \frac{-(y-\hat{y})}{\hat{y}(1-\hat{y})} \cdot \hat{y}(1-\hat{y}) x_1$$

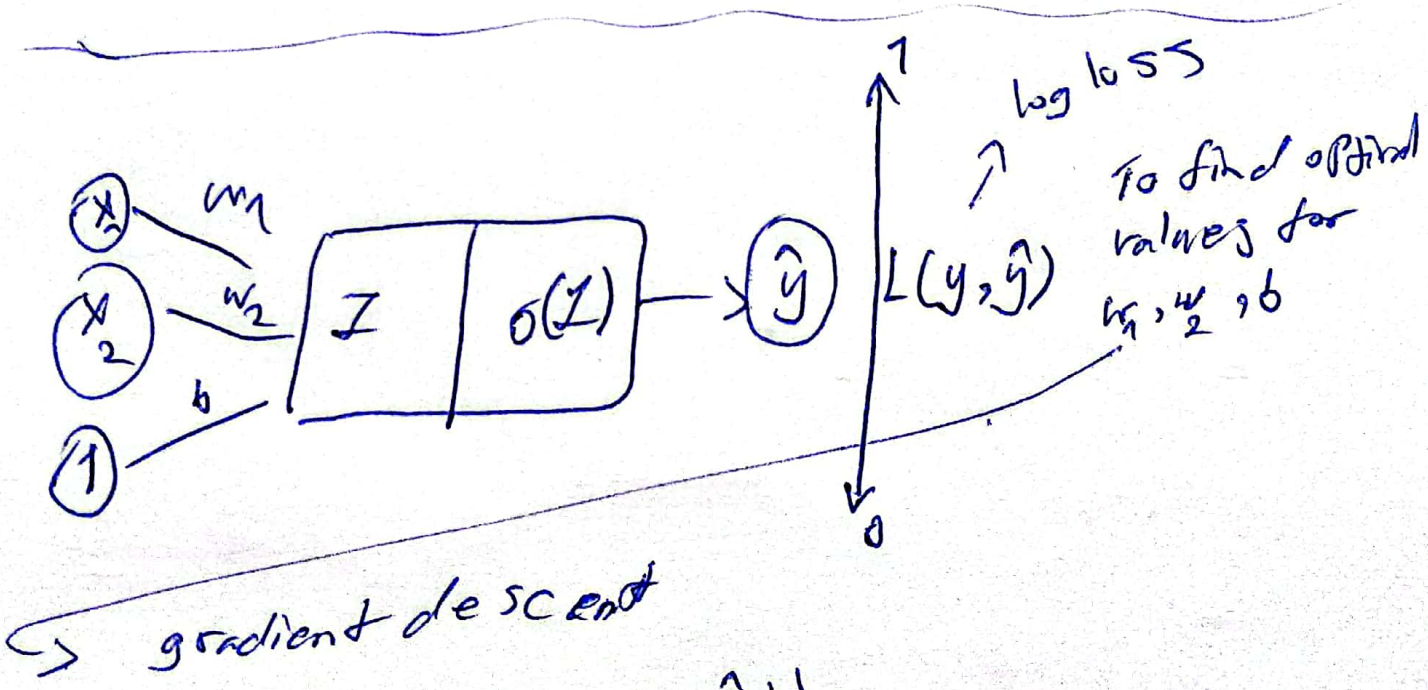
$$\frac{dL}{dw_2} = \frac{dL}{d\hat{y}} \cdot \frac{d\hat{y}}{dw_2} = \frac{-(y-\hat{y})}{\hat{y}(1-\hat{y})} \cdot \hat{y}(1-\hat{y}) x_2$$

CamScanner

$$\frac{dL}{db} = -(y - \hat{y})$$

$$\frac{dL}{dw_1} = -(y - \hat{y}) x_1$$

$$\frac{dL}{dw_2} = -(y - \hat{y}) x_2$$



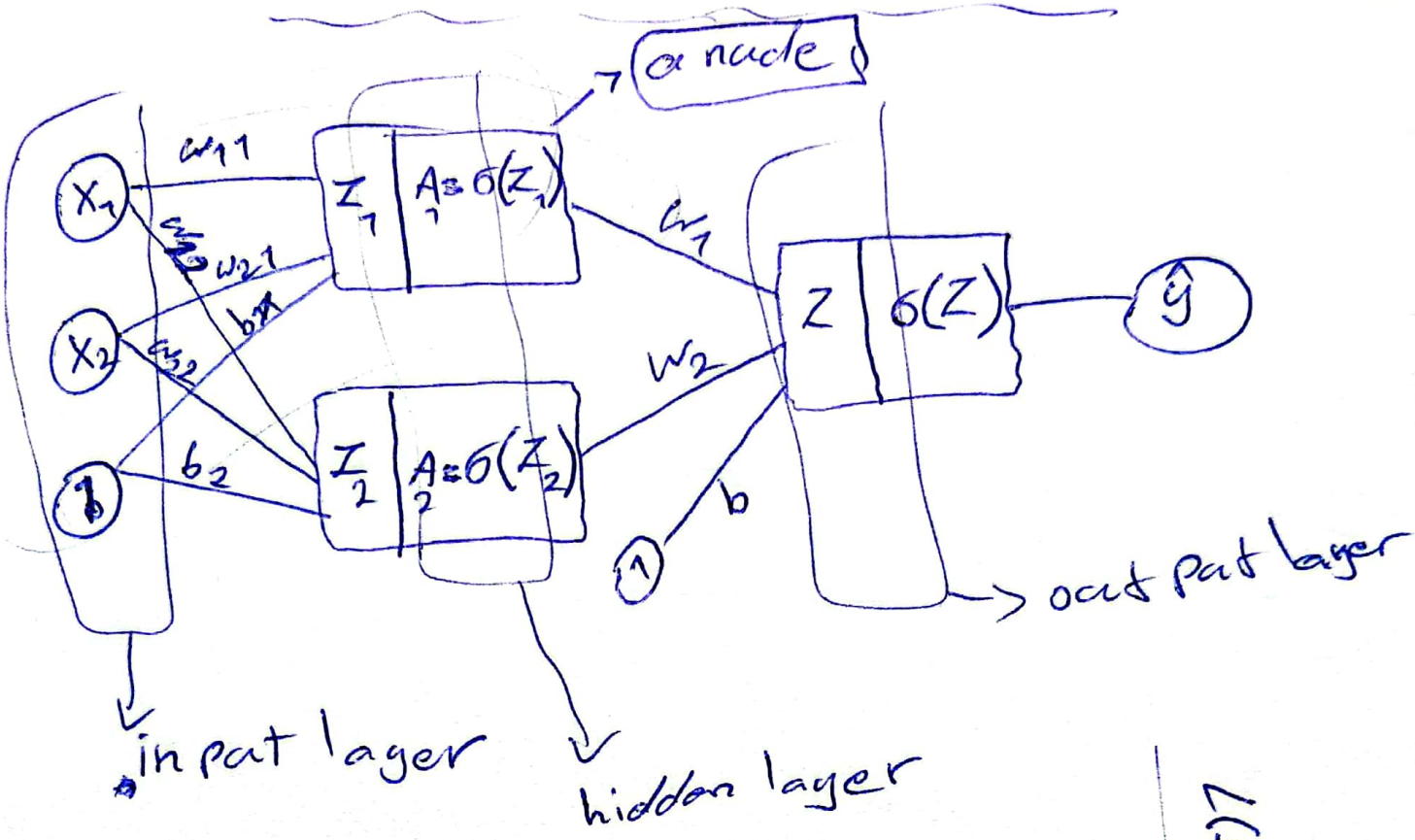
$$w_1 \longrightarrow w_1 - \alpha(-x_1(y - \hat{y}))$$

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

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

A neural network of depth 2

classifications



$$\begin{cases} z_1 = x_1 w_{11} + x_2 w_{21} + b_1 \\ A_1 = \sigma(z_1) \end{cases}$$

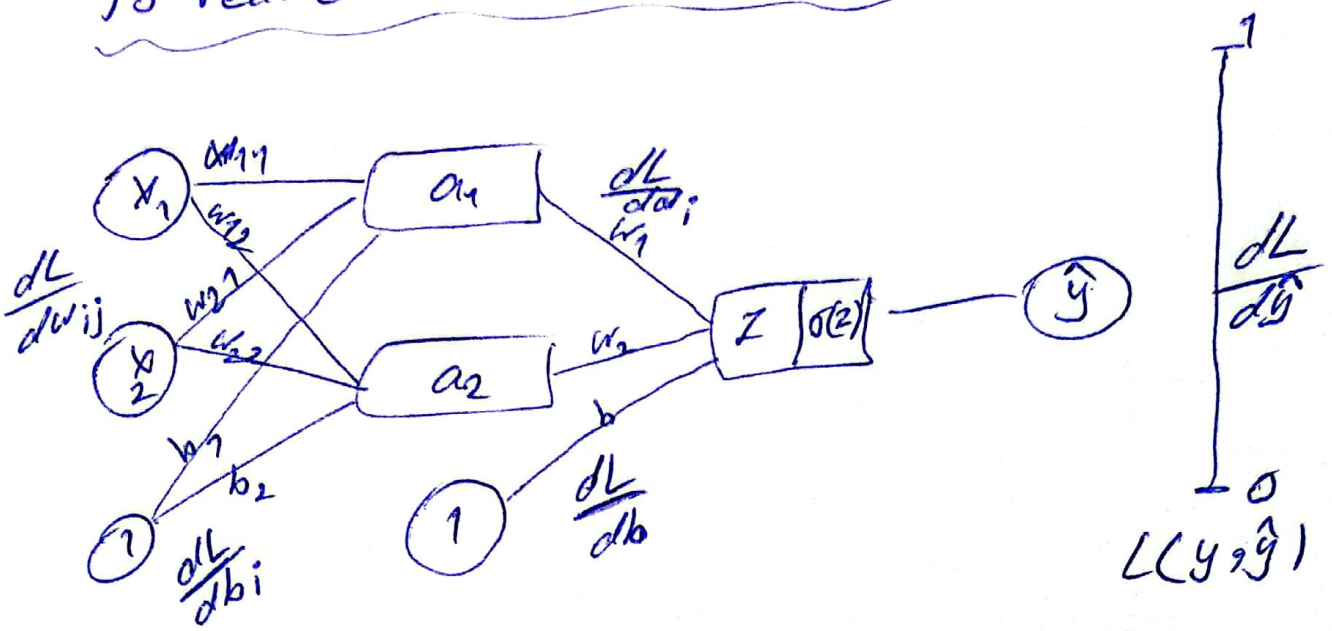
$$\begin{cases} z_2 = x_1 w_{12} + x_2 w_{22} + b_2 \\ A_2 = \sigma(z_2) \end{cases}$$

$$\begin{cases} z = A_1 w_1 + A_2 w_2 + b \\ \hat{y} = \sigma(z) \end{cases}$$

log loss

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

To reduce the loss function: $L(y, \hat{y})$



$$\frac{dL}{dw_{11}} = \frac{dz_1}{dw_{11}} \cdot \frac{da_1}{dz_1} \cdot \frac{dz}{da_1} \cdot \frac{d\hat{y}}{dz} \cdot \frac{dL}{d\hat{y}}$$

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

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

perform gradient descent

$$w_{11} \longrightarrow w_{11} - a\frac{dL}{dw_{11}}$$

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

$$\frac{dL}{db_1} = \frac{dz_1}{db_1} \cdot \frac{da_1}{dz_1} \cdot \frac{dz}{da_1} \cdot \frac{d\hat{y}}{dz} \cdot \frac{dL}{d\hat{y}}$$

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

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

perform gradient descent

$$b_1 \longrightarrow b_1 - \alpha(-w_1 a_1(1-a_1)(y-\hat{y}))$$

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

$$w_{12} \longrightarrow w_{11} + \alpha x_2 w_1 a_1 (1-a_1)(y-\hat{y})$$

$$b_1 \longrightarrow b_1 + \alpha w_1 a_1 (1-a_1)(y-\hat{y})$$

$$w_{21} \longrightarrow w_{21} + \alpha x_1 w_2 a_2 (1-a_2)(y-\hat{y})$$

$$w_{22} \longrightarrow w_{22} + \alpha x_2 w_2 a_2 (1-a_2)(y-\hat{y})$$

$$b_2 \longrightarrow b_2 + \alpha w_2 a_2 (1-a_2)(y-\hat{y})$$

update the first layer of network

$$\frac{dL}{dw_1} = \frac{dz}{dw_1} \cdot \frac{d\hat{y}}{dz} \cdot \frac{dL}{d\hat{y}}$$

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

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

GD

$$w_1 \longrightarrow w_1 - a(-a_1(y-\hat{y}))$$

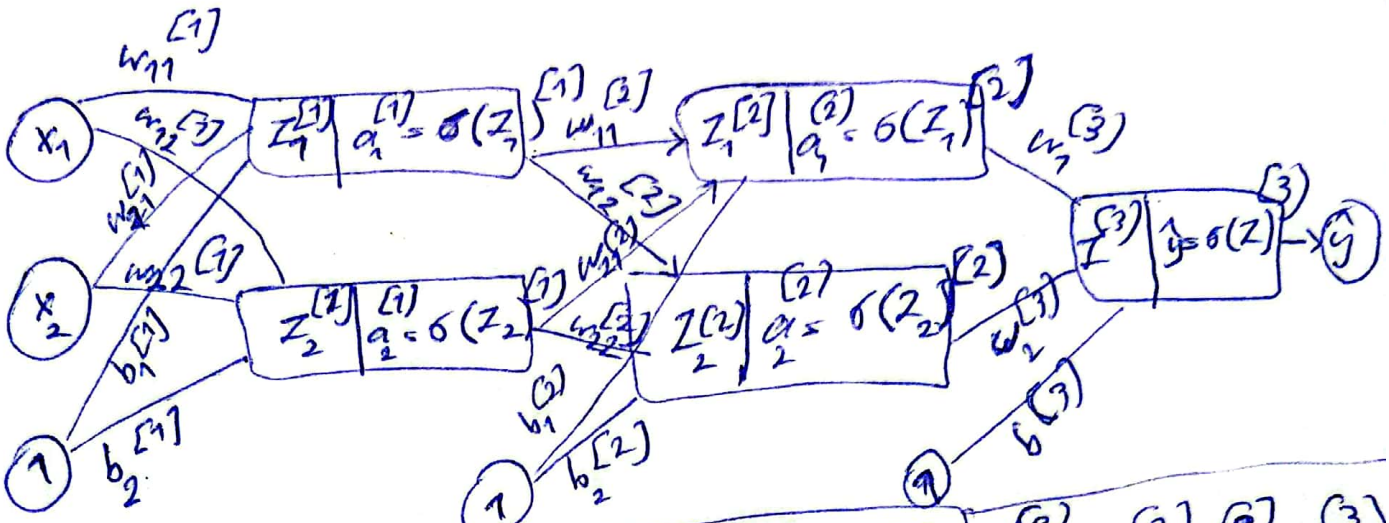
$$\longrightarrow w_1 + aa_1(y-\hat{y})$$

$$w_2 \longrightarrow w_2 + aa_2(y-\hat{y})$$

$$b \longrightarrow b + a(y-\hat{y})$$

update the second layer of NN

Back propagation intro



$$Z^{[1]} = w^{[1]} x + b^{[1]} \qquad Z^{[2]} = w^{[2]} a^{[1]} + b^{[2]} \qquad Z^{[3]} = w^{[3]} \cdot a^{[2]} + b^{[3]}$$

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

$$\frac{dL}{dw^{[3]}} = \frac{dZ^{[3]}}{dw^{[3]}} \cdot \frac{da^{[3]}}{dZ^{[3]}} \cdot \frac{dL}{da^{[3]}}$$

$$\frac{dL}{db^{[3]}} = \frac{dZ^{[3]}}{db^{[3]}} \cdot \frac{da^{[3]}}{dZ^{[3]}} \cdot \frac{dL}{da^{[3]}}$$

$$\frac{dL}{dw^{[2]}} = \frac{dL}{da^{[3]}} \cdot \frac{da^{[3]}}{dZ^{[3]}} \cdot \frac{dZ^{[3]}}{da^{[2]}} \cdot \frac{da^{[2]}}{dZ^{[2]}} \cdot \frac{dZ^{[2]}}{dw^{[2]}}$$

$$\frac{dL}{db^{[2]}} = \frac{dL}{da^{[3]}} \cdot \frac{da^{[3]}}{dZ^{[3]}} \cdot \frac{dZ^{[3]}}{da^{[2]}} \cdot \frac{da^{[2]}}{dZ^{[2]}} \cdot \frac{dZ^{[2]}}{dw^{[2]}}$$

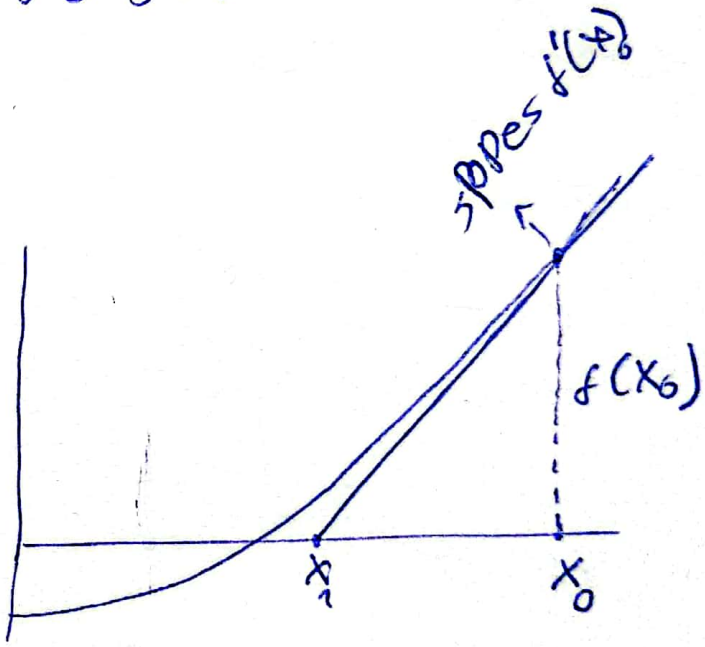
CamScanner

$$\frac{dL}{dw^{[1]}} = \frac{dL}{da^{[3]}} \cdot \frac{da^{[3]}}{dZ^{[3]}} \cdot \frac{dZ^{[3]}}{da^{[2]}} \cdot \frac{da^{[2]}}{dZ^{[2]}} \cdot \frac{dZ^{[2]}}{da^{[1]}} \cdot \frac{da^{[1]}}{dZ^{[1]}} \cdot \frac{dZ^{[1]}}{dw^{[1]}}$$

$$\frac{dL}{db^{[1]}} = \frac{dL}{da^{[3]}} \cdot \frac{da^{[3]}}{dZ^{[3]}} \cdot \frac{dZ^{[3]}}{da^{[2]}} \cdot \frac{da^{[2]}}{dZ^{[2]}} \cdot \frac{dZ^{[2]}}{da^{[1]}} \cdot \frac{da^{[1]}}{dZ^{[1]}} \cdot \frac{dZ^{[1]}}{db^{[1]}}$$

"newton's method"

- In principle it is used to find the zeros of a function.



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

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

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

newton's method:
Goal: find a zero of $f(x)$

NM for optimization:
Goal: minimize $g(x)$
→ find zeros of $g'(x)$

if: $f(x) \longrightarrow g'(x)$

then: $f'(x) \longrightarrow (g'(x))'$

NM	NMO
1) start with some x_0	1) start with some x_0
2) update: $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$	2) update: $x_{k+1} = x_k - \frac{g'(x_k)}{(g'(x_k))'}$
3) repeat 2) until you find the root	3) repeat 2) until you find the candidate for minimum.

exp:

$$g(x) = e^x - \log(x) \qquad g'(x) = e^x - \frac{1}{x}$$

($f(x) = g'(x)$)

find the zero of derivative:

$$g'(x) = e^x - \frac{1}{x} \xrightarrow{\text{der}} e^x + \frac{1}{x^2} \quad (f'(x))$$

$$x_{k-1} = x_k - \frac{e^x - \frac{1}{x}}{e^x + \frac{1}{x^2}}$$

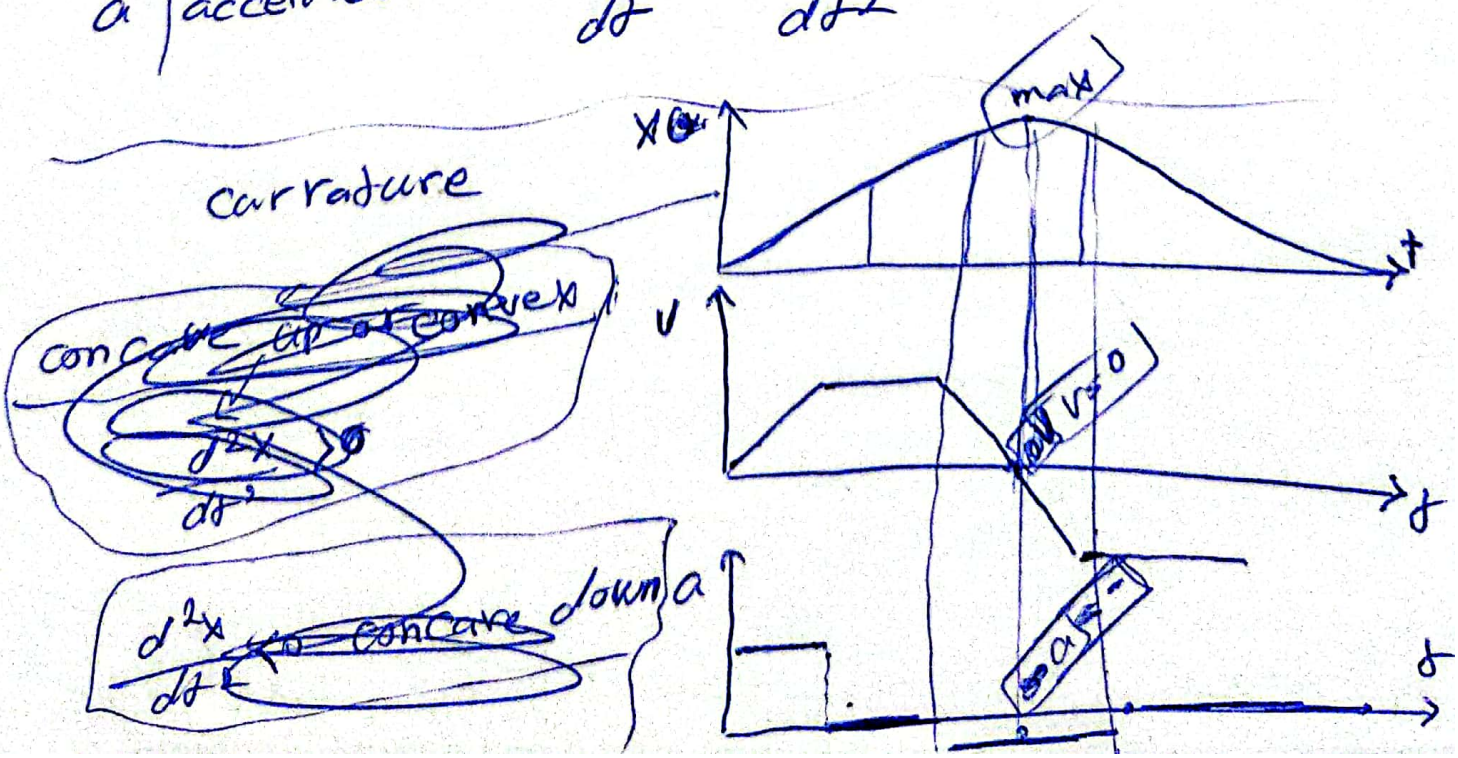
"The second derivative"

$$\frac{d^2 f(x)}{dx^2} = \frac{d}{dx}\left(\frac{df(x)}{dx}\right)$$ → leibniz notation

or

$f''(x)$ → lagrange notation

x	distance	
v	velocity	$\frac{dx}{dt}$
a	acceleration	$\frac{dv}{dt} = \frac{d^2 x}{dt^2}$



$\frac{d^2x}{dt^2} > 0$ concave up or convex.

$\frac{d^2x}{dt^2} < 0$ concave down.

$\frac{d^2x}{dt^2} = 0$ need more information.

The Hessian

	1 variable	2 variable
function	$f(x)$	$f(x,y)$
first derivative	$f'(x)$ rate of change of $f(x)$	$f_x(x,y)$ rate of change w.r.t x $f_y(x,y)$ " " " w.r.t y $\nabla f = \begin{bmatrix} f_x(x,y) \\ f_y(x,y) \end{bmatrix}$
second derivative	$f''(x)$ rate of change of the rate of change of $f(x)$	$Hm = \begin{bmatrix} f_{xx}(x,y) & f_{xy}(x,y) \\ f_{yx}(x,y) & f_{yy}(x,y) \end{bmatrix}$

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

$$f_x(x,y) = 4x - y$$

$$f_y(x,y) = 6y - x$$

$x \rightarrow f_x''(x,y) = 4$

$y \rightarrow f_y''(x,y) = -1$

$x \rightarrow f_x''(x,y) = -1$

$y \rightarrow f_y''(x,y) = 6$

$$\frac{d^2f}{dx^2} = f_{xx}(x,y) = 4$$

$$\frac{d^2f}{dy^2} = f_{yy}(x,y) = 6$$

$$\text{Hessian} = \begin{bmatrix} 4 & -1 \\ -1 & 6 \end{bmatrix}$$

$$\frac{d^2f}{dx\,dy} = f_{xy}(x,y) = -1$$

$$\frac{d^2f}{dy\,dx} = f_{yx}(x,y) = -1$$

CamScanner

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

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

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

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

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

$$\begin{cases} \lambda_1 = 6.41 \\ \lambda_2 = 3.59 \end{cases} > 0$$

if means the matrix (H) is positive definite

So the function is concave up and the point (0,0) is a minimum!

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

$$f_x'(x,y) = -4x - y$$

$$f_{xx}''(x,y) = -4$$

$$f_{xy}''(x,y) = -1$$

$$f_y'(x,y) = -6y - x$$

$$f_{yx}''(x,y) = -1$$

$$f_{yy}''(x,y) = -6$$

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

Hessian at (0,0)

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

gradient

n eigen values of Hessian "M"

$$\det(H(0,0) - \lambda I) = (-4-\lambda)(-6-\lambda) - (-1)(-1)$$

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

$$\lambda_1 = -3.59 < 0$$

$$\lambda_2 = -6.41$$

(0,0) is a maximum

Saddle Point

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

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

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

$$\det(H(0,0) - \lambda I) = (4-\lambda)(4-\lambda) - 0$$

$$\begin{bmatrix} \lambda_1 = -4 \\ \lambda_2 = +4 \end{bmatrix}$$ are can not conclude anything

---

NM

1 variable: $$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

or

$$x_k - f''(x_k)^{-1} \cdot f'(x_k)$$

2 variable: $$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - H^{-1}(x_k, y_k) \cdot \nabla f(x_k, y_k)$$

CamScanner

exp:

$$f(x,y) = x^4 + 0.9y^4 + 4x^2 + 2y^2 - xy - 0.2x^2y$$

$f(x,y)$ → x → $4x^3 + 8x - y - 0.4xy$ $[f'_x(x,y)]$

$f(x,y)$ → y → $3.2y^3 + 4y - x - 0.2x^2$ $[f'_y(x,y)]$

$f'_x(x,y)$ → x → $12x^2 + 8 - 0.4y$

$f'_x(x,y)$ → y → $-1 - 0.4x$

$f'_y(x,y)$ → x → $-1 - 0.4x$

$f'_y(x,y)$ → y → $10.8y^2 + 4$

$$H(0,0) = \begin{bmatrix} 12x^2 + 8 - 0.04y & -1 - 0.4x \\ -1 - 0.4x & 9.6y^2 + 4 \end{bmatrix}$$

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

start at some point $\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$

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

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

$$= \begin{bmatrix} 2.56 \\ 2.62 \end{bmatrix}$$

$\nabla f(2.58, 2.61) = \begin{bmatrix} 84.25 \\ 63.4 \end{bmatrix}$ $H(2.58, 2.61) = \begin{bmatrix} 80.83 & -2.032 \\ -2.032 & 69.39 \end{bmatrix}$

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} 80.83 & -2.032 \\ -2.032 & 69.39 \end{bmatrix}^{-1} \cdot \begin{bmatrix} 84.25 \\ 63.4 \end{bmatrix}$$

$$= \begin{bmatrix} 1.59 \\ 1.67 \end{bmatrix}$$

...

needed 8 steps

$$\begin{bmatrix} x_8 \\ y_8 \end{bmatrix} = \begin{bmatrix} 4.75 \times 10^{-17} \\ -2.05 \times 10^{-17} \end{bmatrix}$$ → last zero