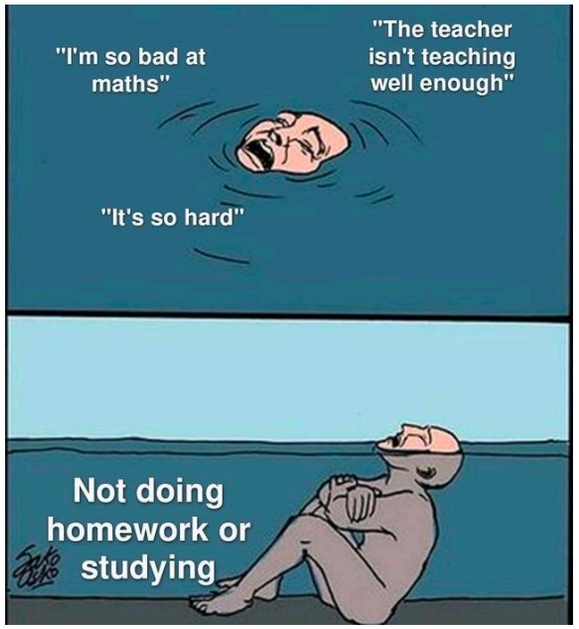
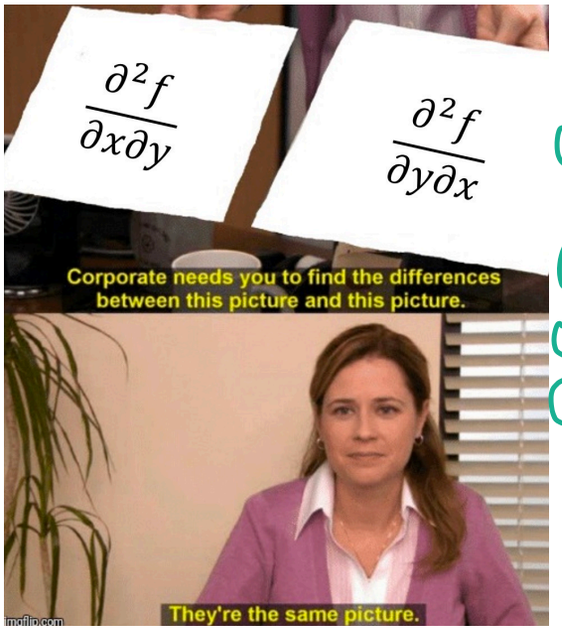
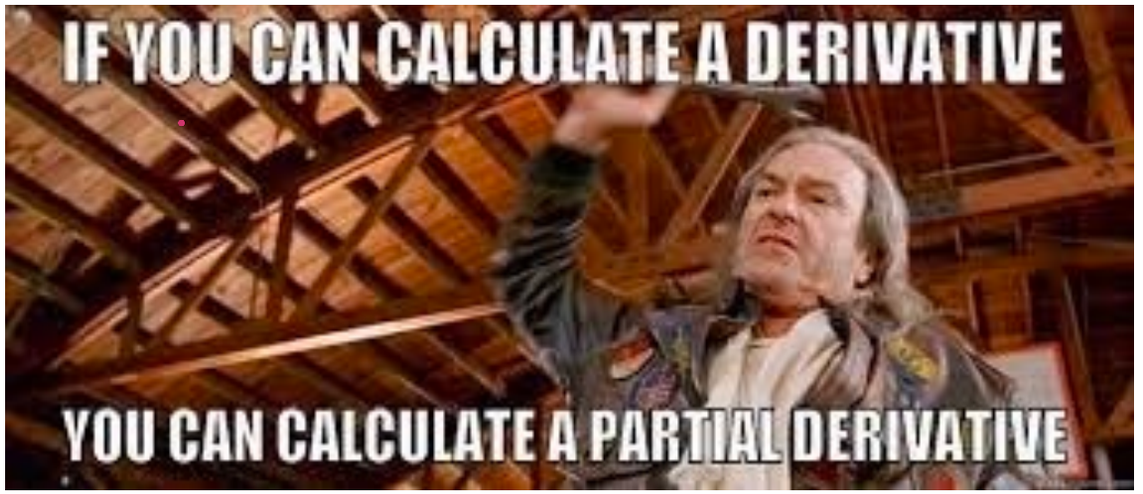
March 4, 2023

DSML : Math for ML.

# Optimization 3 : Gradient Descent in action.




## Recap:

(a) Classification - choosing $\vec{w}$ and $w_0$.
(b) Brute force: very inefficient.
(c) Alternative : Gradient descent
(d) Functions, limits, derivatives.

## Today:

(a) Partial derivatives.
(b) Gradient descent.
(c) Coding gradient descent.

# Detailed Recap.

1] **Derivatives**: $f(x) \rightarrow$ continuous & Differentiable.

$$\frac{d}{dx} f(x) = f'(x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$$

2] **Maxima, Minima**: $f(x) \rightarrow$ continuous & Differentiable.

To find candidate points for maxima & minima:

Find all $x$ such that $f'(x) = 0$.

If: $f''(x) > 0$ : Minima

$f''(x) < 0$ : Maxima.

3] **Rules**:

(a) **Linearity** → If $h(x) = a \cdot f(x) + b \cdot g(x)$ then $h'(x) = a f'(x) + b g'(x)$

(b) **Product** → If $h(x) = f(x) \cdot g(x)$ then $h'(x) = f'(x) g(x) + g'(x) f(x)$

(c) **Quotient** → If $h(x) = \frac{f(x)}{g(x)}$ then $h'(x) = \frac{g(x) \cdot f'(x) - g'(x) f(x)}{(g(x))^2}$

(d) **Chain** → If $h(x) = f(g(x))$ then $h'(x) = f'(g(x)) \cdot g'(x)$.

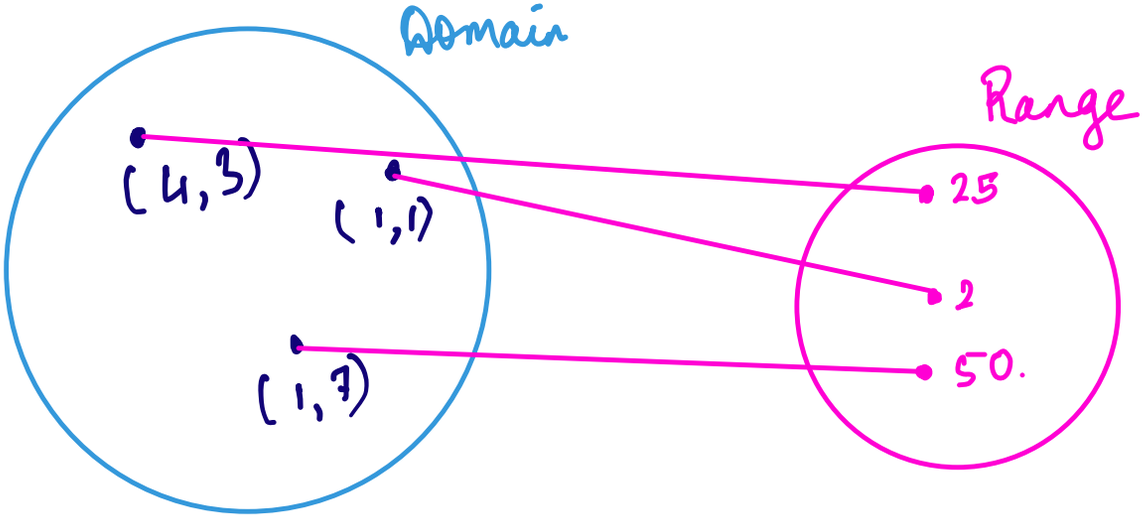
# Multivariable Calculus.

$f(x)$ ← input "variable"

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

two inputs, "Multiple variables".

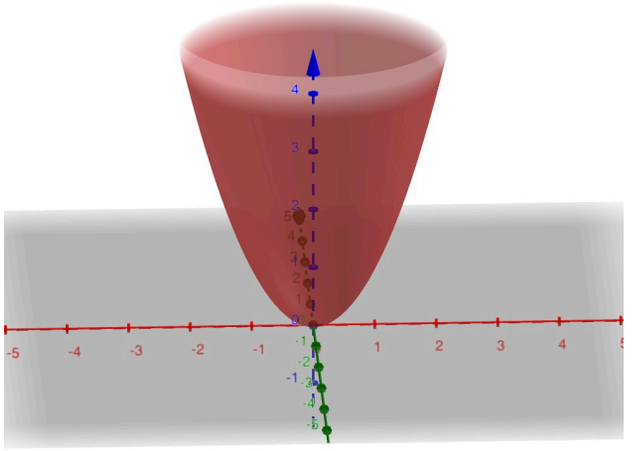
"Functions in multivariable calculus generally take vectors as inputs and give single numbers as outputs."



# Partial derivatives.

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

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



Derivative:

$$\rightarrow \frac{d}{dx}(z) = 2x + 2y \cdot \frac{dy}{dx}$$

$$\rightarrow \frac{d}{dy}(z) = 2x \cdot \frac{dx}{dy} + 2y \cdot$$

## Partial Derivatives:

When we are taking partial derivatives, we treat all other variables as constants.

$$\rightarrow \frac{\partial(z)}{\partial x} = 2x$$

$$\rightarrow \frac{\partial(z)}{\partial y} = 2y$$

Q] $f(w_1, w_2, w_0) = w_1 x_1 + w_2 x_2 + w_0$

treat these as constants.

(a) $\frac{\partial f(w_1, w_2, w_0)}{\partial w_1} = x_1$

(b) $\frac{\partial f(w_1, w_2, w_0)}{\partial w_2} = x_2.$

(c) $\frac{\partial f(w_1, w_2, w_0)}{\partial w_0} = 1$

$$\nabla_{\bar{w}} f(\bar{w}) = \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix}$$

$f(w_1, w_2, w_0) = f(\bar{w})$

where $\bar{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_0 \end{bmatrix}$

Q2] $f(w_1, w_2, w_3, w_0) = w_1 x_1 + w_2 x_2 + w_3 x_3 + w_0$

(a) $\frac{\partial f}{\partial w_1} = x_1$

(b) $\frac{\partial f}{\partial w_2} = x_2$

(c) $\frac{\partial f}{\partial w_3} = x_3$

(d) $\frac{\partial f}{\partial w_0} = 1$

what is $\nabla_{\bar{w}} f(\bar{w}) =$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{bmatrix}$$

$f(w_1, w_2, w_3, w_0) = f(\bar{w})$

where $\bar{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_0 \end{bmatrix}$

Q] $f(x,y) = 3\log(xy) + 4y^2x^3$

(a) $\frac{\partial f}{\partial x} = \frac{3}{x} + 12x^2y^2$

(b) $\frac{\partial f}{\partial y} = \frac{3}{y} + 8x^3y$

What is

$\nabla_{\bar{x}} f(\bar{x})$

$= \begin{bmatrix} 3/x + 12x^2y^2 \\ 3/y + 8x^3y \end{bmatrix}$

$f(x,y) = f(\bar{x})$

where $\bar{x} = \begin{bmatrix} x \\ y \end{bmatrix}$

# Gradient of a vector.

derivative : single variable functions.

gradients : multivariable functions.

$\frac{d}{dx} f(x) = f'(x)$ : $f(\bar{x})$

$f(x, y)$

$f(\bar{x})$ where $\bar{x} = \begin{bmatrix} x \\ y \end{bmatrix}$

$\nabla_{\bar{x}} f(\bar{x})$.

# Computing gradient of a vector.

$$\nabla_{\bar{x}} f(\bar{x}) = \begin{bmatrix} \frac{\partial}{\partial x_1} f(\bar{x}) \\ \frac{\partial}{\partial x_2} f(\bar{x}) \end{bmatrix}$$

where $\bar{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

Q] $f(x_1, x_2) = \log(x_1) + \frac{x_1}{x_2} e^{x_1+x_2}$.

$f(\bar{x})$ where $\bar{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$$\nabla_{\bar{x}} f(\bar{x}) = \begin{bmatrix} \frac{1}{x_1} + \frac{x_1}{x_2} e^{x_1+x_2} + \frac{1}{x_2} e^{x_1+x_2} \\ -\frac{(x_1 e^{x_1+x_2})}{x_2^2} + \frac{x_1}{x_2} e^{x_1+x_2} \end{bmatrix}$$

Q] $f(x_1, x_2, x_3) = a_1 x_1 + a_2 x_2 + a_3 x_3$.

$$f(\bar{x}) = \bar{a}^T \bar{x}$$

where $\bar{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ $\bar{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$

$$\nabla f(\bar{x}) = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \bar{a}$$

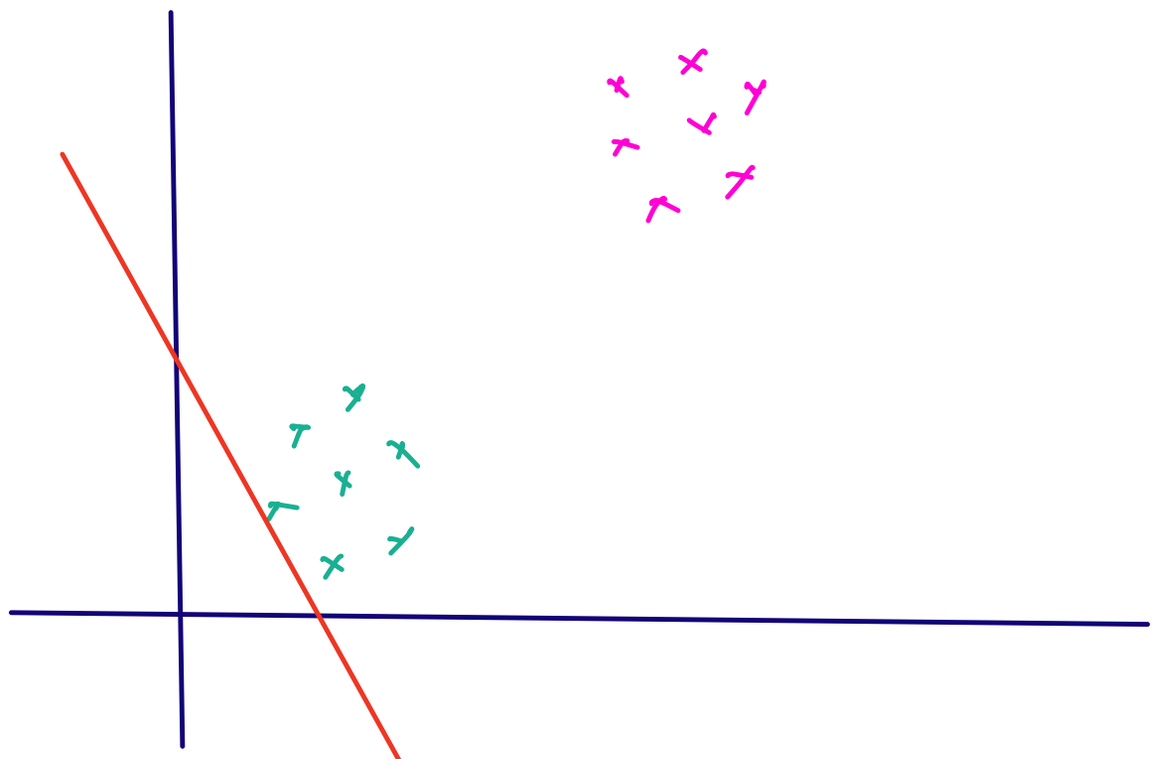
General Rule: $\boxed{\nabla_{\bar{x}} \; \bar{a}^T \bar{x} = \bar{a}}$

$$\underbrace{f(\bar{x})}_{\downarrow} = \underbrace{a_1 x_1 + a_2 x_2 + a_3 x_3 + a_0}_{\bar{a}^T \bar{x} + a_0}$$

$$\bar{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

where $\bar{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$

$$\nabla_{\bar{x}} \; a_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$



Solution: Gradient descent.

$\bar{w}^T \bar{x} + w_0 = 0$.

Initial guess

What we need → An algorithm which can figure out the optimal $\bar{w}, w_0$.

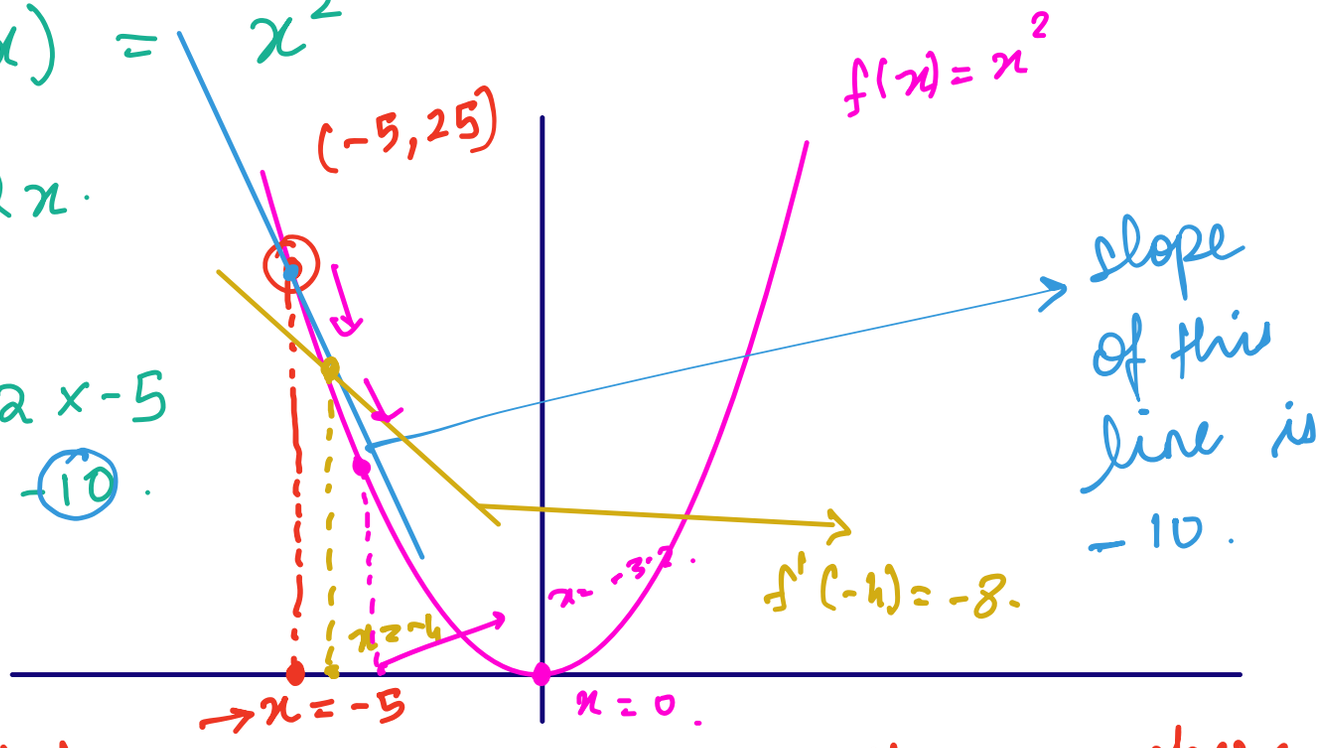
# Intuition for Gradient descent.

$f(x) = x^2$

$\frac{d}{dx} f(x) = 2x$.

$f'(-5) = 2x - 5$

$= -10$.




Step 1: Make an initial guess for the x where our minima lies.

Step 2: Find the slope of the tangent to the function at this point. (-10)

Step 3: Update the guess using the G.D. update rule.

Our guesses are: $x^{(0)}, x^{(1)}, x^{(2)}, x^{(3)}, \ldots\ldots$

$$x^{(t+1)} = x^{(t)} - \eta \cdot \frac{d}{dx} f(x)$$

Gradient update rule in 1-D.

$x^{(0)} \rightarrow$ Random Initial guess.

$x^{(1)} \rightarrow$ we get from G.D. update rule & $x^{(0)}$

$x^{(2)} \rightarrow$ we get from G.D. update rule & $x^{(1)}$

Eta, learning rate.

$$x^{(1)} = x^{(0)} - \eta \cdot \frac{d}{dx} f(x) \qquad \eta = 0.1$$

$$\therefore \quad x^{(1)} = -5 - 0.1 \times (-10)$$

$$= -5 + 1$$

$$= \underline{\underline{-4}}$$

$$x^{(2)} = x^{(1)} - \eta \cdot \frac{d}{dx} f(x)$$

$$x^{(2)} = -4 - (0.1) \times (-8)$$

$$= -4 + 0.8 = -3.2.$$

$$d(h; \bar{w}, w_0) = -\sum_{i=1}^{n} \left( \frac{\bar{w}^T \bar{x}_i + w_0}{\|\bar{w}\|} \right) \cdot y_i$$

$$\underbrace{\nabla_{\bar{w}} d(h; \bar{w}, w_0)}$$

1-D.

$$x^{(t+1)} = x^{(t)} - \eta \cdot \frac{d}{dx} f(x)$$

d-D.

$$\bar{w}^{(t+1)} = \bar{w}^{(t)} - \eta \cdot \nabla_{\bar{w}} d(h; \bar{w}, w_0)$$

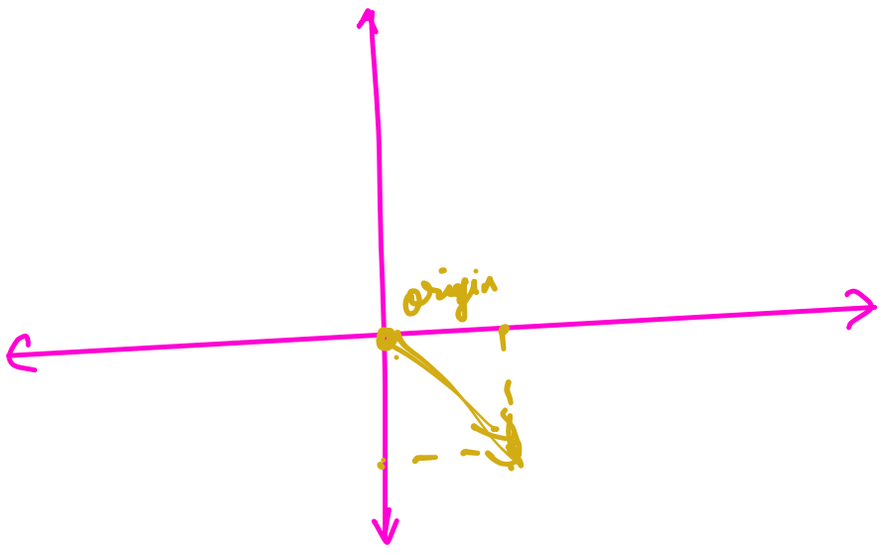
$$\text{Loss} : \sum_{i=1}^{n} \left[ \frac{(w_1 x_{1i} + w_2 x_{2i} + w_0) y_i}{\sqrt{w_1^2 + w_2^2}} \right]$$

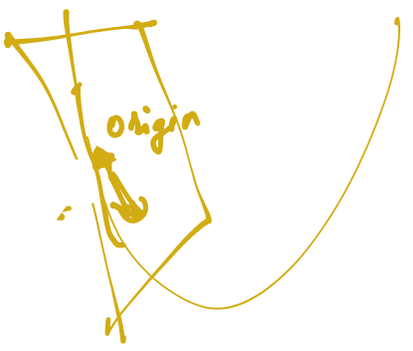
$w_1 = 1$

$w_2 = 1$

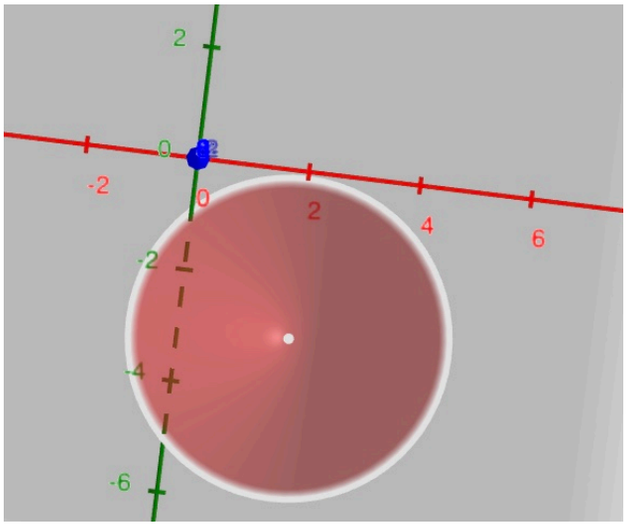
$w_0 = 0$

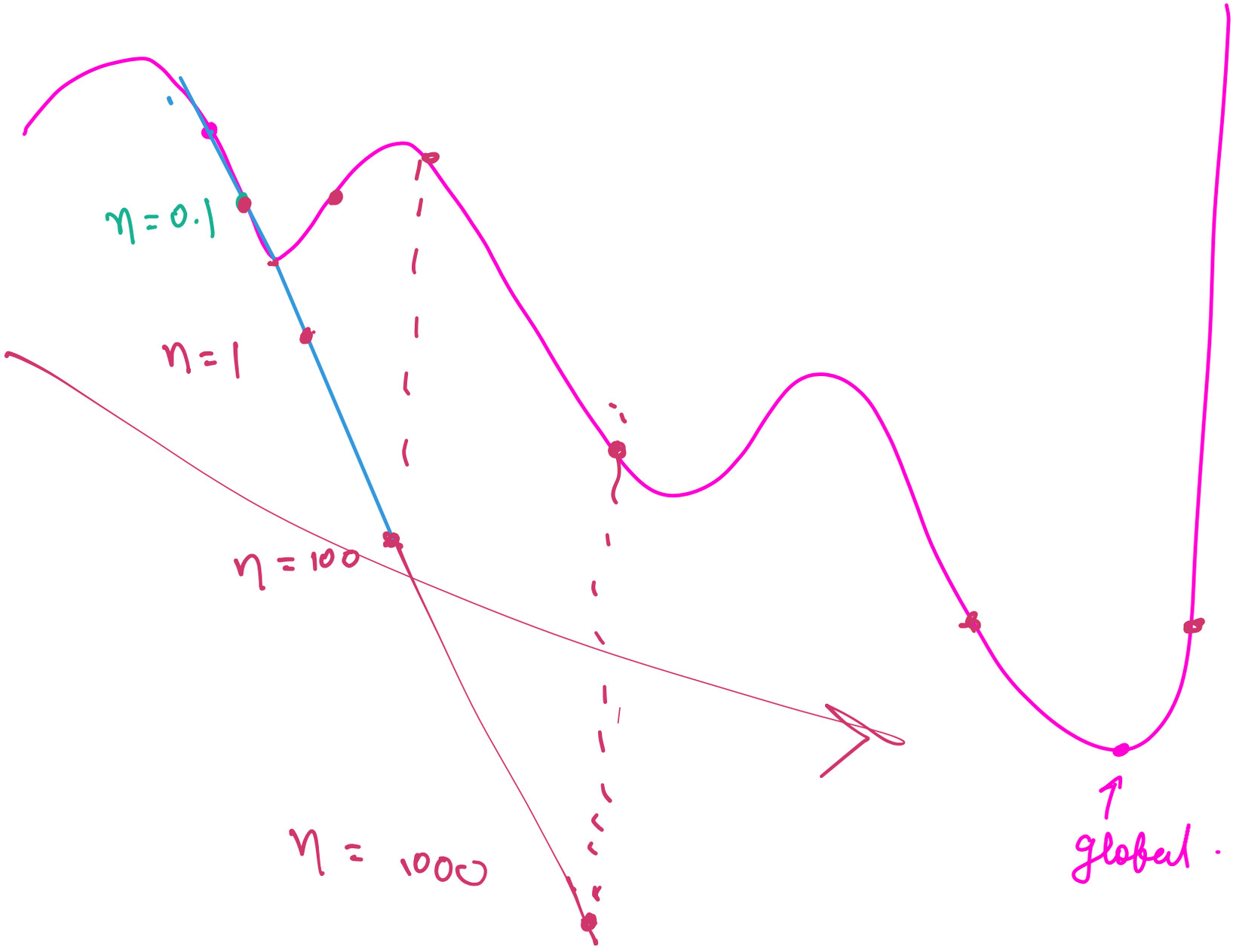
$$\sum \left( \frac{x_{1i} + x_{2i}}{\sqrt{2}} \right) y_i$$



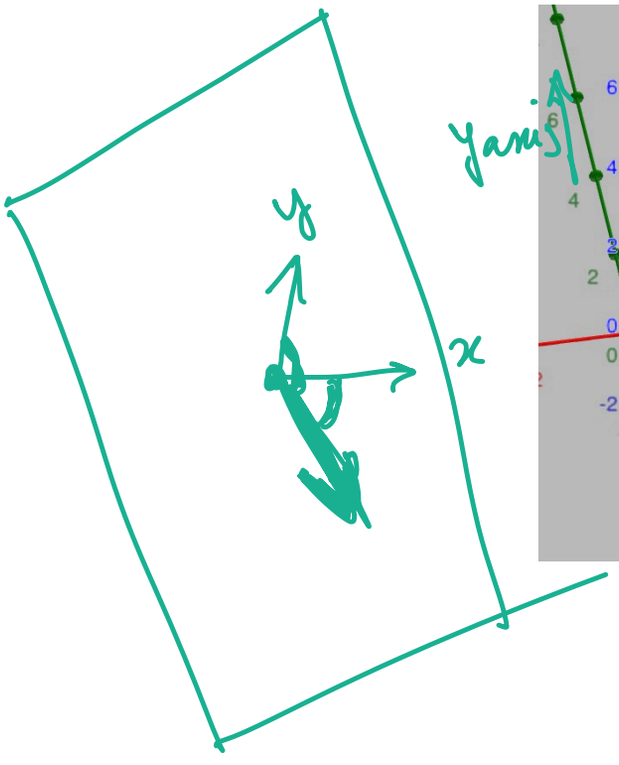



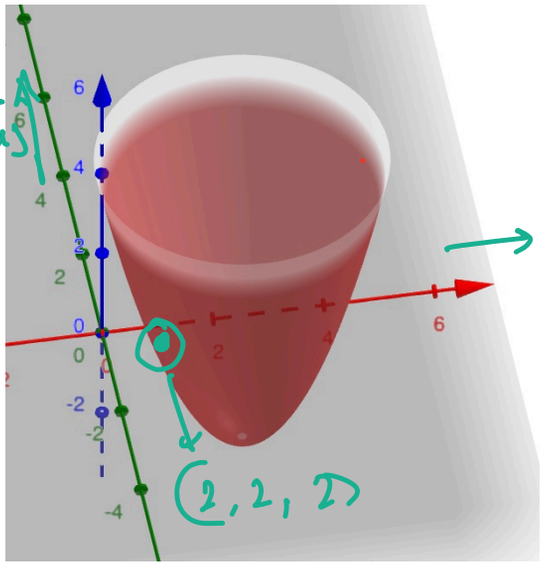

$$\begin{bmatrix} \partial f/\partial x_1 \\ \partial f/\partial y_1 \end{bmatrix}$$

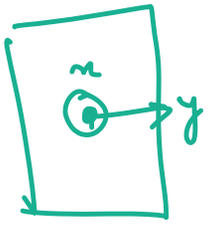

2
0
-2
0
2
4
6
-2
-4
-6

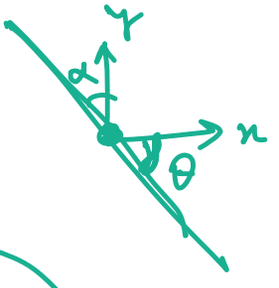

$\eta = 0.1$
$\eta = 1$
$\eta = 100$
$\eta = 1000$
Global.

$$\tan(\alpha) = \frac{\partial f}{\partial y}$$

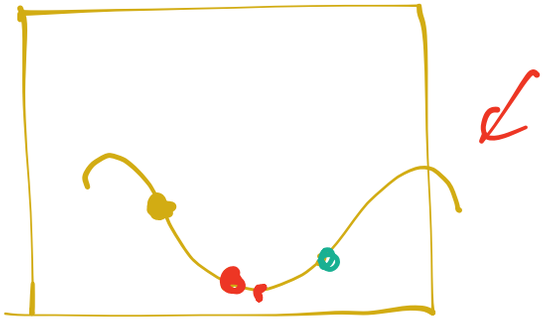


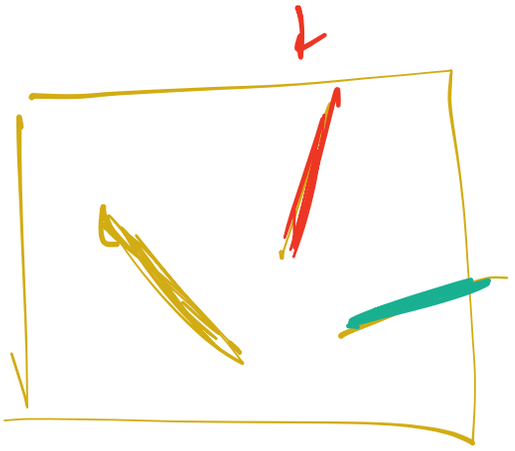




$$\tan(\theta) = \frac{\partial f}{\partial x}$$



Loss plot



Scatter plot