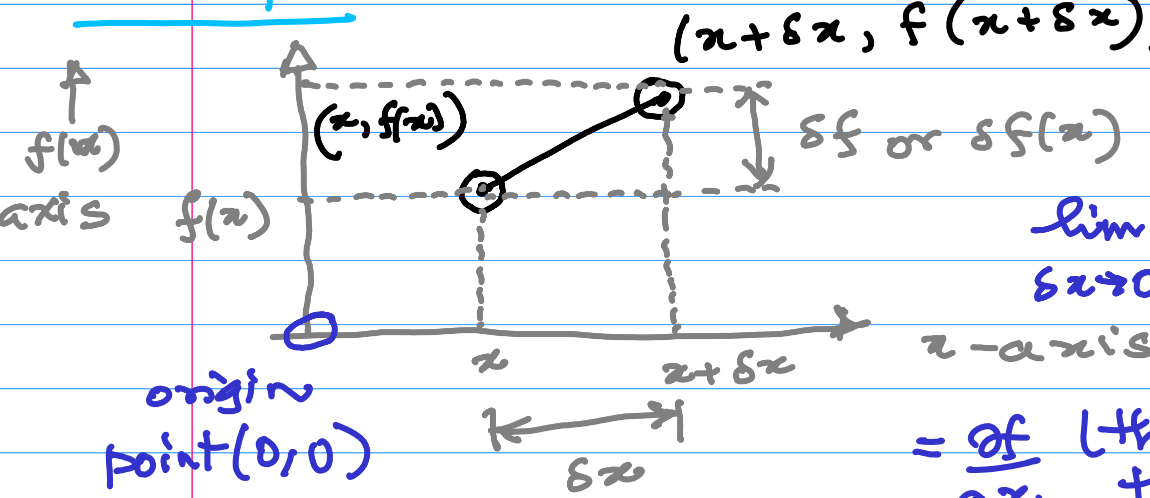
DERIVATIVES

The most general derivative is NOT the total derivative, but the partial derivatives.

$\boxed{\text{1-D}}$ one independent variable x → scalar
one dependent variable $f(x)$ → scalar function

Example: e.g., audio signal $f(t)$



$$\lim_{\delta x \to 0} \frac{f(x+\delta x) - f(x)}{(x+\delta x)-(x)}$$

$$= \frac{\partial f}{\partial x}$$ (this is also the total derivative df/dx)

Why? 1 scalar variable.

$$\lim_{\delta x \to 0}: \underbrace{f(x+\delta x) - f(x)}_{\delta f \text{ or } \delta f(x)} = \left(\frac{\partial f}{\partial x}\right) \underrightarrow{\delta x}$$

small change in the dependent variable, or the function

small change in the independent variable.

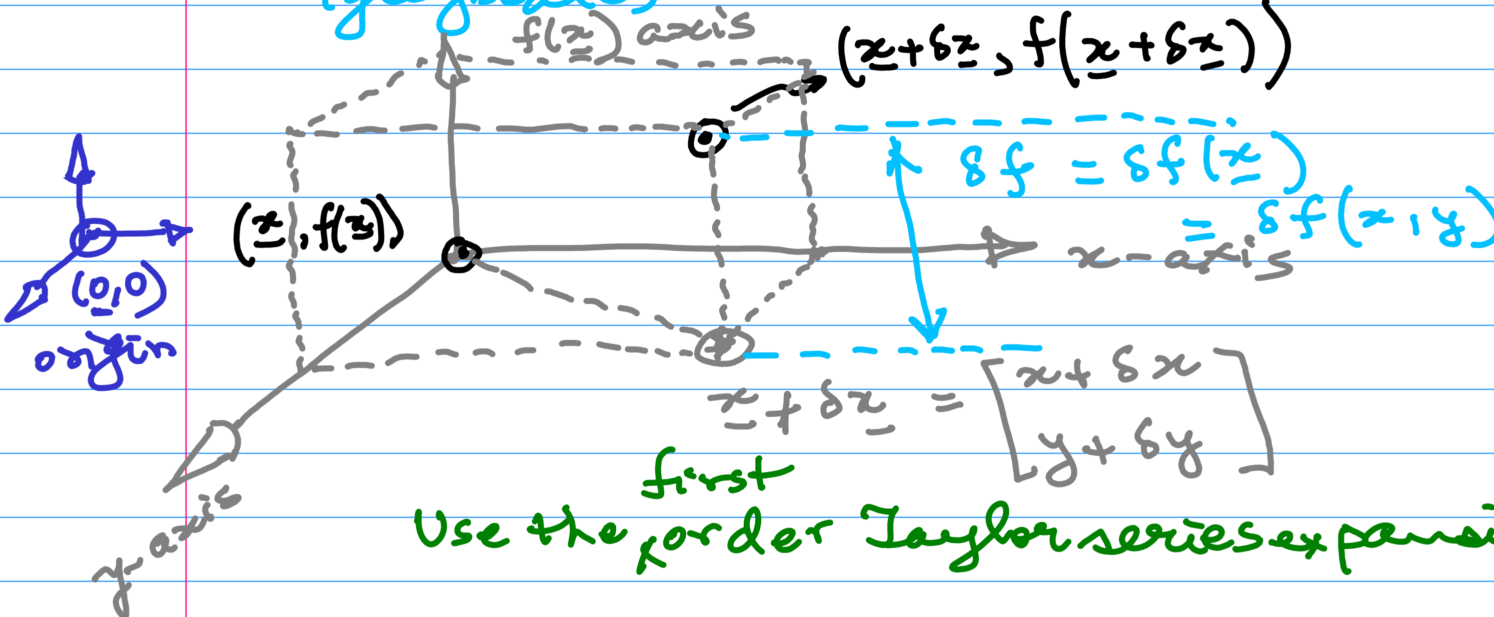
$\boxed{\text{2-D}}$ two independent variables $\underset{\sim}{x} = \begin{bmatrix} x \\ y \end{bmatrix}$

2 → vector

one dependent variable $\underbrace{f(\underset{\sim}{x})}$ → scalar.

Example $I(x,y)$ or $I(\underline{x})$

the intensity of a pixel at position $\underline{x} = \begin{bmatrix} x \\ y \end{bmatrix}$ (greyscale)



Use the first order Taylor series expansion

$$f(\underline{x}+\delta\underline{x}) = f(\underline{x}) + \nabla_f \cdot \delta\underline{x}$$

$$\underbrace{f(\underline{x}+\delta\underline{x}) - f(\underline{x})}_{\delta f \text{ or } \delta f(\underline{x}) \text{ or } \delta f(x,y)} = \nabla f \cdot \delta \underline{x}$$

$$\underbrace{\begin{bmatrix} \partial f/\partial x \\ \partial f/\partial y \end{bmatrix} \cdot \begin{bmatrix} \delta x \\ \delta y \end{bmatrix}}$$

$$\underbrace{\frac{\partial f}{\partial x}\delta x + \frac{\partial f}{\partial y}\delta y}$$

$$= \sum_{var:\, x,y} \left(\frac{\partial f}{\partial\, var}\right)(\delta\, var)$$

$\boxed{\text{3-D}}$ Three independent variables $\underline{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$ → vector

one dependent variable $f(\underline{x})$ → scalar

Example: Video (greyscale) $I(x,y,t)$: Intensity at pixel (x,y) in frame # t.

Impossible to visualise $I(x, y, t) \rightarrow$ 4-D entity

Use the first order Taylor series approximation:

$$f(\underline{x} + \delta\underline{x}) = f(\underline{x}) + \vec{\nabla} f \cdot \delta\underline{x}$$

$$\underbrace{f(\underline{x}+\delta\underline{x}) - f(\underline{x})}_{\delta f \text{ or } \delta f(\underline{x}) \text{ or, } \delta f(x, y, z)} = \underbrace{\begin{bmatrix} \partial f/\partial x \\ \partial f/\partial y \\ \partial f/\partial z \end{bmatrix} \cdot \begin{bmatrix} \delta x \\ \delta y \\ \delta z \end{bmatrix}}$$

$$= \frac{\partial f}{\partial x}\delta x + \frac{\partial f}{\partial y}\delta y + \frac{\partial f}{\partial z}\delta z$$

$$= \sum_{var = x, y, z} \left(\frac{\partial f}{\partial var}\right)(\delta var)$$

Generalise to a function of D variables

$f(\underline{x})$, $\underline{x} = \begin{bmatrix} x_D \\ \vdots \\ x_1 \end{bmatrix}$ 1st order Taylor series expansion

$$\left.\begin{array}{l} \delta f \text{ or } \delta f(\underline{x}) \\ \text{or, } \delta f(x_1 \ldots x_D) \end{array}\right\} = \vec{\nabla} f \cdot \delta\underline{x}$$

$$= \sum_{i=1}^{D} \frac{\partial f}{\partial x_i} \delta x_i$$

Take-home point: The total change is always a summation

$$\delta f = \delta f(\underline{x}) = \sum_{i=1}^{D} \frac{\partial f}{\partial x_i} \delta x_i$$

Consider another variable t
(all the x_i's are functions of this variable t)

$$\frac{\delta f}{\delta t} = \sum_{i=1}^{D} \frac{\partial f}{\partial x_i} \frac{\delta x_i}{\delta t}$$

We can take the limit as $\delta t \to 0$

$$\frac{\partial f}{\partial t} = \sum_{i=1}^{D} \frac{\partial f}{\partial x_i} \frac{\partial x_i}{\partial t}$$

Here, we had one variable t, so the partial derivative is also the total derivative.

$$\frac{df}{dt} = \sum_{i=1}^{D} \frac{\partial f}{\partial x_i} \frac{dx_i}{dt}$$

Now, consider a set of variables t_j : $j \in \{1, D\}$
(all the x_i's are functions of t_j

$i \in \{1, D\}$ $j \in \{1, D\}$

$$\boxed{\forall j : \frac{\partial f}{\partial t_j} = \sum_{i=1}^{D} \frac{\partial f}{\partial x_i} \frac{\partial x_i}{\partial t_j}}$$

Chain Rule

Compact Moral of the story: if f depends on many x_i, then

$$\frac{\partial f}{\partial t} = \sum_{\forall i} \frac{\partial f}{\partial x_i} \frac{\partial x_i}{\partial t}$$

Perceptron & MLP: some closing notes

(*) Key difference: MLP and the perceptron: MLP uses continuous sigmoidal non-linearities in the hidden units, whereas the perceptron uses a step function non-linearities

(*) Variants: Skip layers: either direct connection, or with a small first layer weight (so that over its operating range, the hidden unit is effectively linear), compensating with a large weight value from the hidden unit to the output

(*) Sparse network (CNN)

Math overview/recap: → 1st order Taylor series approximation

$$\underbrace{E(\underline{w}+\delta\underline{w})}_{\text{minimise}} = E(\underline{w}) + (\nabla E)\cdot(\delta\underline{w})$$

error function

weights (All)

Overall aim: to find a weight vector, which minimises an error function $E(\underline{w})$

scalar

vector

At the extremum, $\nabla E = \underline{0}$

max/min/saddle point

vector

$$\begin{bmatrix} \partial E/\partial w_2 \\ \partial E/\partial w_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

(2-D)

higher order terms

Local Quadratic Approximation

$$E(\underline{w}+\delta\underline{w}) = E(\underline{w}) + \underbrace{\nabla E \cdot (\delta\underline{w})}_{(\delta\underline{w})^T \nabla E} + \frac{1}{2}(\delta\underline{w})^T H (\delta\underline{w}) + \ldots$$

H → Hessian

Example: first order

$$I(x+p, y+q, t+\tau) = I(x,y,t) + pI_x + qI_y + \tau I_t$$

Second order term

$$H_{ij} = \left.\frac{\partial^2 E}{\partial w_i \partial w_j}\right|_{\underline{w}}$$

$$E(\underline{w}+\delta\underline{w}) = E(\underline{w}) + (\delta\underline{w})^T \underbrace{\nabla E}_{=\underline{0}} + \frac{1}{2}(\delta\underline{w})^T H (\delta\underline{w})$$

$= \underline{0}$, at the extremum

Extremum:

$$E(\underline{w}+\delta\underline{w}) = E(\underline{w}) + \frac{1}{2}(\delta\underline{w})^T H (\delta\underline{w})$$

→ a geometric interpretation