OPTIMIZATION FOR MACHINE LEARNING 2022 CLASS 2

- Google doc shared document.
- Be active in chat
- Have a pen & paper

REMINDER: Continuous optimization

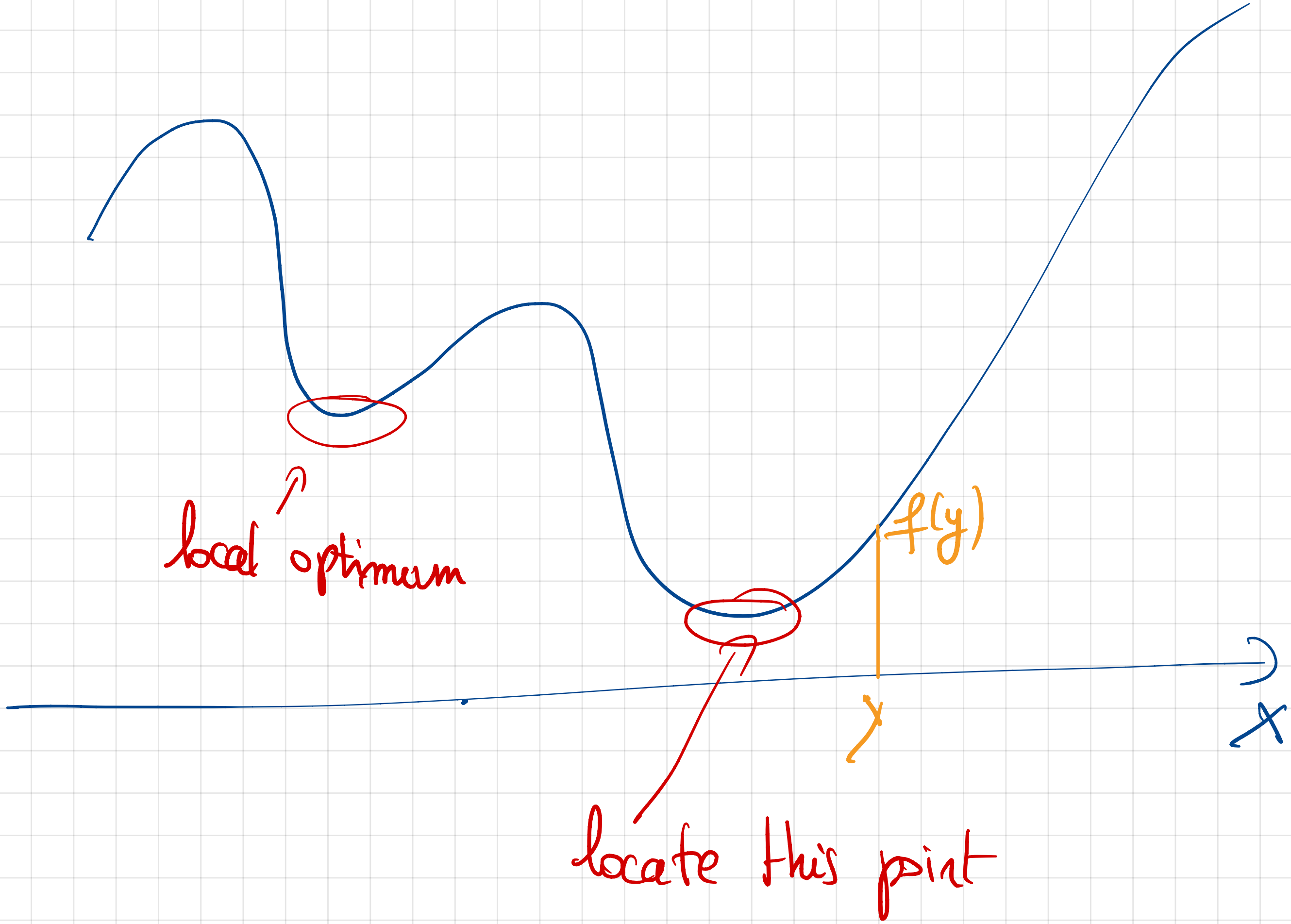
$$\text{minimize } f(\underset{\mathbb{R}}{\underset{\Uparrow}{x_1}}, \dots, \underset{\mathbb{R}}{\underset{\Uparrow}{x_n}})$$

$x = (x_1, \dots, x_n) \in \mathbb{R}^n$
vector space
n: dimension of problem.

Look for $\underset{\mathbb{R}^n}{\underset{\Uparrow}{x^*}}$ such that

$$f(x^*) \leq f(x) \quad \forall x \ (\in \mathbb{R}^n)$$

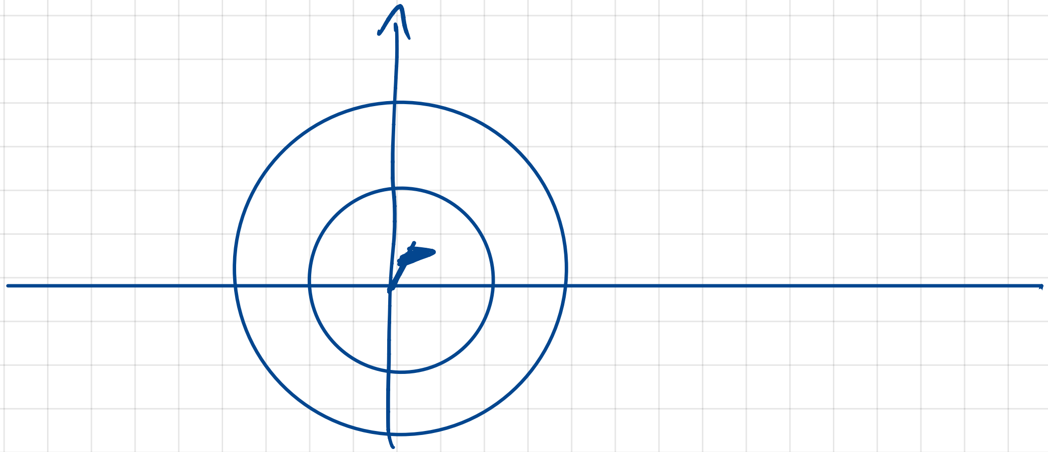
When $n = 1$ $\quad \min_{x \in \mathbb{R}} f(x)$



$n = 2$, we can represent functions via level sets.

$$L_c = \{ x \in \mathbb{R}^n \mid f(x) = c \}$$

$f(x) = x_1^2 + x_2^2$, what is the geometric shape of its level sets.



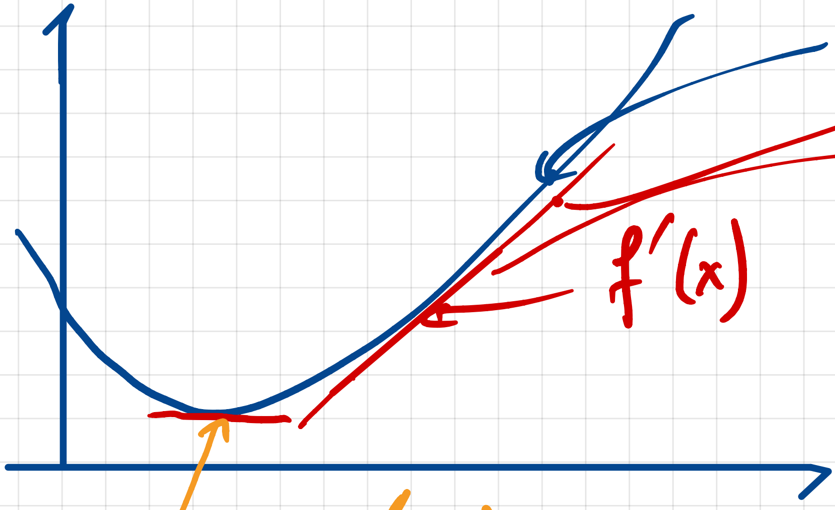
Derivability or differentiability

$n = 1$, let $f : \mathbb{R} \longrightarrow \mathbb{R}$

we say that f is derivable / differentiable in x if

$\lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$ exists , the limit is denotes $f'(x)$

and it is called the derivative of f in x



$f(x+h) \approx f(x) + h f'(x)$

$h \mapsto f(x) + h f'(x)$

$f'(x)$ is the slope of the tangent in x

x^* $\quad f'(x^*) = 0$

If f is differentiable in x then

$$\underbrace{f(x+h) = f(x) + f'(x)h + o(\|h\|)}$$

Taylor expansion of f in x, at first order

For h small enough $h \mapsto f(x+h)$ is approximately equal to $h \longmapsto f(x) + f'(x)h$

$g(h) \in o(\|h\|)$ $\qquad \dfrac{g(h)}{\|h\|} \xrightarrow[h \to 0]{} 0$

$g(h)$ is a small o of h if it goes faster to 0 than $\|h\|$.

<u>example</u> $\quad g(h) = \|h\|^2 \; (= |h|^2) \in o(\|h\|)$

$$\frac{g(h)}{\|h\|} = \frac{\|h\|^2}{\|h\|} = \|h\| \xrightarrow[h \to 0]{} 0$$

. How do we generalize derivative from $n=1$ to $n>1$?

Differential of $f:\mathbb{R}^n \longrightarrow \mathbb{R}^m$

Let $f:\mathbb{R}^n \longrightarrow \mathbb{R}^m$, we say that f is differentiable in x if there exists a linear transformation $Df_x:\mathbb{R}^n \longrightarrow \mathbb{R}^m$ such that $\forall h \in \mathbb{R}^n$ $f(x+h) = f(x) + Df_x(h) + o(\|h\|)$

If $n=1$, $Df_x(h) \overset{?}{=} \underbrace{f'(x)\,h}_{\text{Linear in } h\,?}$

$$\left\{\begin{array}{l} f'(x)(h_1+h_2) = f'(x)h_1 + f'(x)\,h_2 \\ f'(x)(\underset{\mathbb{R}}{\alpha} h) = \alpha\left[f'(x)\cdot h\right] \end{array}\right)$$

$h \mapsto f'(x)\,h$
Linear in h

Exercise : 1) $f(x) = Ax$ where A is a $n \times n$ matrix

$x \in \mathbb{R}^n$ $(\Rightarrow Ax \in \mathbb{R}^n)$

$Df_x = A$

2) $f(x) = \|x\|^2$, $Df_x(h) = 2x^T h$

$x \in \mathbb{R}^n$

1) $f(x) = Ax$

$$A = \begin{array}{cc} \begin{bmatrix} a_{11} & \cdots & \cdots \\ & \ddots & \\ a_{1n} & & a_{nn} \end{bmatrix} & n \\ n & \end{array}$$

$x \in \mathbb{R}^n$

$$f(\underset{\mathbb{R}^n}{\underset{\Uparrow}{x}} + \underset{\mathbb{R}^n}{\underset{\Uparrow}{h}}) =$$

(we try to find a linear mapping L st $f(x+h) = f(x) + L(h) + o(\|h\|)$

$$f(x+h) = A(x+h) = Ax + Ah = f(x) + \underbrace{Ah}_{\text{Linear in } h} + \underset{o(\|h\|)}{\overset{\|}{0}}$$

$$\begin{cases} h \mapsto Ah & \text{is Linear} \\ \mathbb{R}^n \longrightarrow \mathbb{R}^n \end{cases}$$

so f is differentiable in x and

$$Df_x = A \qquad Df_x(h) = Ah$$

If $f(x) = \|x\|^2 = x^T x$ $\quad f : \mathbb{R}^n \to \mathbb{R}$

$$f(x+h) = (x+h)^T (x+h)$$

$$= x^T x + x^T h + \overbrace{h^T x}^{= x^T h} + h^T h$$

$$= x^T x + \underbrace{2x^T h}_{\text{Linear in } h} + \underbrace{h^T h}_{= \|h\|^2 = o(\|h\|)}$$

$$Df_x : h \mapsto 2x^T h$$

$$h^T x \stackrel{?}{=} x^T h$$

$$\underbrace{h^T x}_{\in \mathbb{R}}$$

$$\underbrace{(h^T x)^T}_{} = h^T x$$

$$= x^T (h^T)^T$$

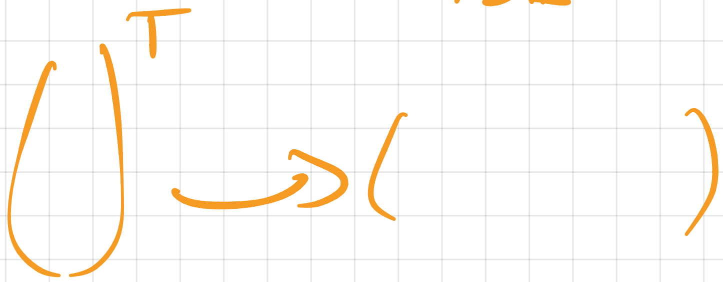
$$= x^T h$$

We have $h^T x = x^T h$

$$\|x\|^2 = x^T x$$

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \qquad \underbrace{(x_1, \dots, x_n)}_{x^T} \begin{pmatrix} x \end{pmatrix}$$

$$= \sum_{i=1}^{n} x_i^2$$



$$(ab)^T = b^T a^T$$

Why: $h \stackrel{L}{\longmapsto} 2x^T h$ linear.

$$L(h_1 + h_2) = L(h_1) + L(h_2)$$

$$L(\overset{\in \mathbb{R}}{\lambda} h_1) = \overset{\in \mathbb{R}}{\lambda} L(h_1)$$

$$\to L(h_1 + h_2) = 2x^T (h_1 + h_2)$$

$$= 2x^T h_1 + 2x^T h_2$$

$$= L(h_1) + L(h_2)$$

CHAIN RULE :

$$\left[\left(f(x) g(x) \right)' = f(x) g'(x) + g(x) f'(x) \right]$$

$f : \mathbb{R} \longrightarrow \mathbb{R}$ $\qquad$ $g : \mathbb{R} \longmapsto \mathbb{R}$

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

composition

$x \xmapsto{f} \sin(x)$ $\qquad$ $f \circ g(x) = f(g(x)) = \sin(x^2)$

$x \xmapsto{g} x^2$ $\qquad$ $f(x) g(x) \overset{?}{=} \sin(x) \cdot x^2$

[composition & product of functions are different]

$$\boxed{D(f \circ g)_x (h) = Df_{g(x)} \left(Dg_x (h) \right)}$$

We go back to $f: \mathbb{R}^n \longrightarrow \mathbb{R}$ $[m = 1]$

when $f: \mathbb{R}^n \longrightarrow \mathbb{R}$ is differentiable in x, there is a specific representation of the differential of f in x

$Df_x : \mathbb{R}^n \longrightarrow \mathbb{R}$

$$\exists \; a \in \mathbb{R}^n \text{ such that } Df_x(h) = \langle a, h \rangle$$
$$= a^T h$$

[This comes from the Riesz representation theorem]

The vector a has a specific name $a = \nabla f_x$

[Gradient of f in x]

$$\boxed{Df_x(h) = \langle \nabla f_x , h \rangle}$$

LINK BETWEEN DIFFERENTIAL & GRADIENT

The gradient can also be defined with partial derivatives.

$$\nabla f_x = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix}$$

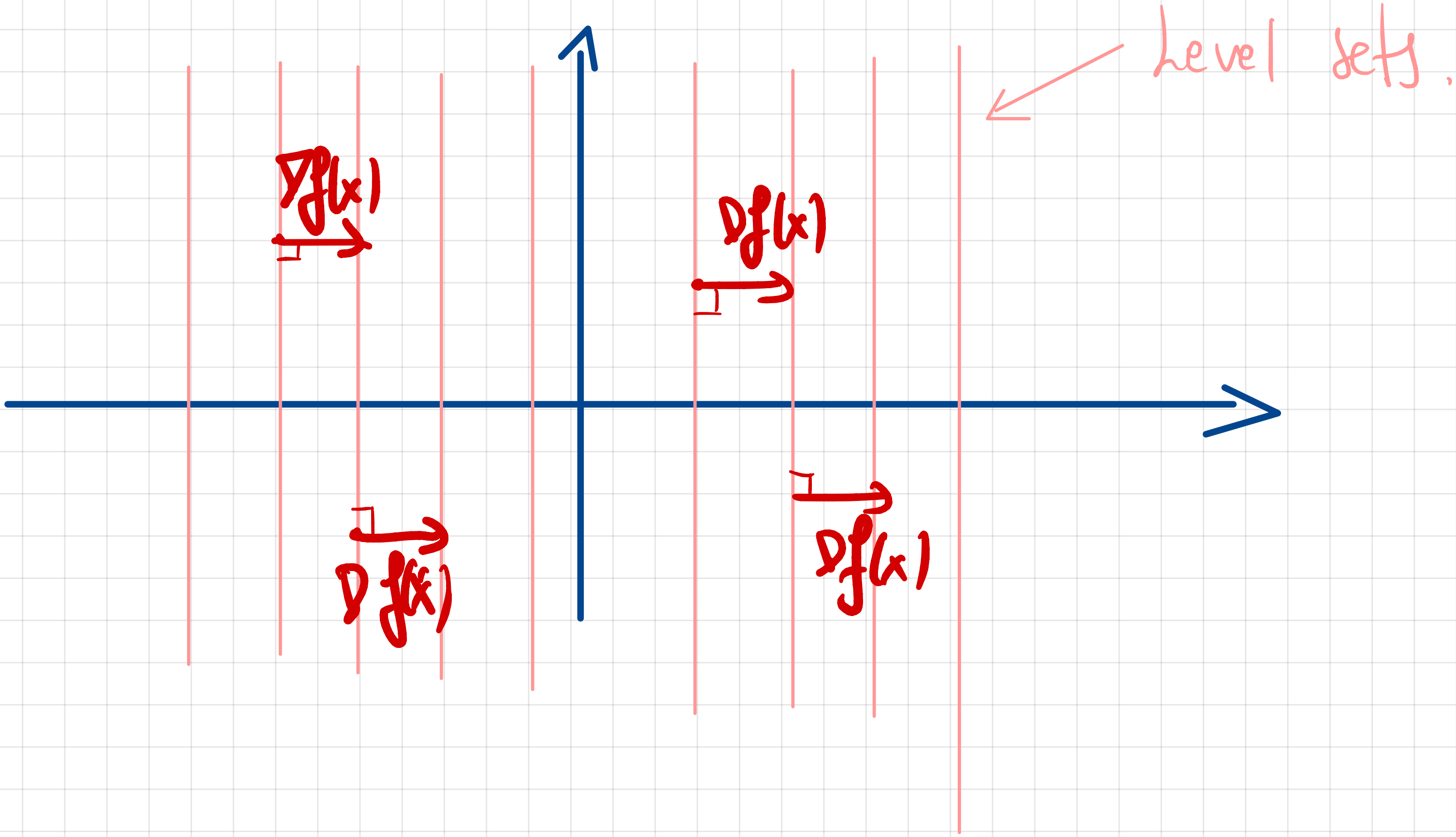
<u>Exercise:</u> Compute the gradient of:

$f(x) = x_1 \qquad x \in \mathbb{R}^n$

$f(x) = a^T x \qquad a = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$

$f(x) = x^T x$

$$f(x_1, x_2) = x_1 \qquad L_c = \{ x = (x_1, x_2) \in \mathbb{R}^2 \mid x_1 = c \}$$



$$\nabla f_x = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

The gradient vector is orthogonal to the level sets.

Second order derivability / differentiability

$n = 1$ (1D-case)

Let $f : \mathbb{R} \to \mathbb{R}$ be differentiable on $\mathbb{R}$ and let

$f' : x \to f'(x)$ be its derivative function

If f' is derivable / differentiable, then we denote $f''(x)$ its derivative.

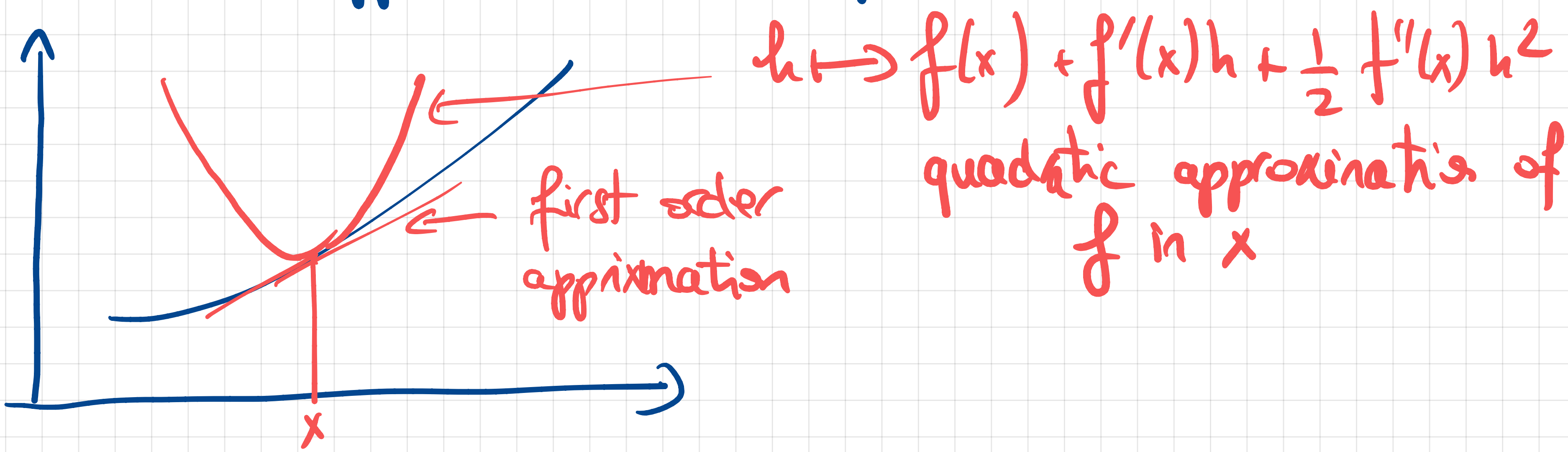
$f''(x)$ is called the second order derivative of f

If f is two times differentiable then

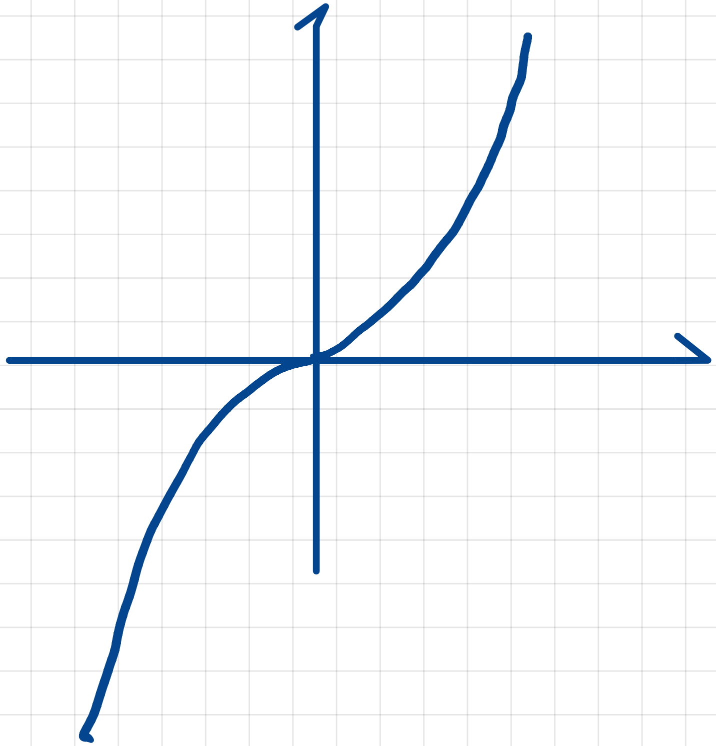
$$f(x+h) = f(x) + f'(x)\,h + \frac{1}{2} f''(x)\,h^2 + o(\|h\|^2)$$

SECOND ORDER TAYLOR | EXPANSION
FORMULA

for h small enough $h \mapsto f(x) + f'(x)h + \frac{1}{2} f''(x) h^2$ (which is quadratic in h) approximates f. This is called a second order approximation of f

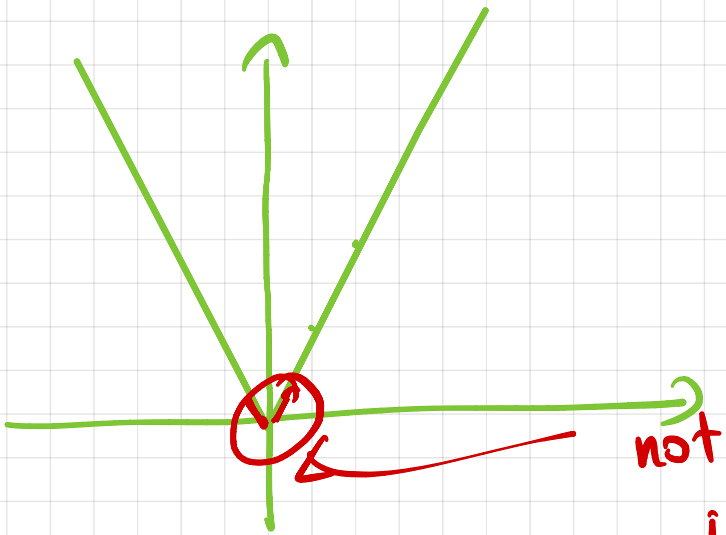


$$f(x) = \begin{cases} x^2 & \text{if } x \geq 0 \\ -x^2 & \text{if } x \leq 0 \end{cases} \qquad x \in \mathbb{R}$$



$$f'(x) = \begin{cases} 2x & x \geq 0 \\ -2x & x \leq 0 \end{cases}$$

$$f'(x) = 2|x|$$



not derivable
in zero

We want to generalize second order derivative to functions $f: \mathbb{R}^n \longrightarrow \mathbb{R}$

The Hessian matrix generalizes $f''(x)$

$$\text{Hessian}(x) = \nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \ddots & \\ \vdots & & \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix}$$

The Hessian matrix is symmetric $\quad \frac{\partial^2 f}{\partial x_1 \partial x_n} = \frac{\partial^2 f}{\partial x_n \partial x_1}$

Schwarz Theorem

Example: Compute the Hessian matrix for $f(x) = \frac{1}{2} x^T A x$

A symmetric $n \times n$ matrix.

Start with $A = \begin{pmatrix} 9 & 1 \\ 1 & 1 \end{pmatrix}$

$$\frac{\partial^2 f}{\partial x_1 \partial x_1} \overset{?}{=} 9$$

$$f(x) = \frac{1}{2} x^T \begin{pmatrix} 9 & 1 \\ 1 & 1 \end{pmatrix} x = \frac{1}{2}\left(9x_1^2 + x_2^2 + 2x_1x_2\right)$$

$$\frac{\partial f}{\partial x_1} = \frac{1}{2}\left(2 \cdot 9 x_1 + 2 x_2\right)$$
$$= 9x_1 + x_2$$

$$\frac{\partial^2 f}{\partial x_1 \partial x_1} = \frac{\partial}{\partial x_1}\left[9x_1 + x_2\right] = 9$$

$$\frac{\partial^2 f}{\partial x_2 \partial x_1} = \frac{\partial}{\partial x_2}\left[9x_1 + x_2\right) = 1$$

$$\frac{\partial f}{\partial x_2} = \frac{1}{2}\left(2x_2 + 2x_1\right) = x_2 + x_1$$

$$\frac{\partial^2 f}{\partial x_2 \partial x_2} = \frac{\partial}{\partial x_2}\left[x_2 + x_1\right] = 1$$

$$\nabla^2 f = \begin{pmatrix} 9 & 1 \\ 1 & 1 \end{pmatrix} = A$$

If $f(x) = \frac{1}{2} x^T A x$ with A symmetric . A: $n \times n$

$$\nabla^2 f(x) = A$$

If A is not symmetric : $\nabla^2 f(x) = \frac{1}{2}\left(A + A^T\right)$

DETAIL ABOUT:

$$f(x) = \frac{1}{2} x^T \underbrace{\begin{pmatrix} 9 & 1 \\ 1 & 1 \end{pmatrix}} x = \frac{1}{2}\left(9x_1^2 + x_2^2 + 2x_1x_2\right)$$

$$\begin{pmatrix} 9 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 9x_1 + x_2 \\ x_1 + x_2 \end{pmatrix}$$

$$\frac{1}{2} x^T \begin{pmatrix} 9x_1 + x_2 \\ x_1 + x_2 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} x_1 & x_2 \end{pmatrix}\begin{pmatrix} 9x_1 + x_2 \\ x_1 + x_2 \end{pmatrix}$$

$$= \frac{1}{2} x_1 (9x_1 + x_2) + x_2 (x_1 + x_2)$$

$$= \frac{1}{2}\left(9x_1^2 + x_1x_2 + x_1x_2 + x_2^2\right)$$

$$= \frac{1}{2}\left(9x_1^2 + 2x_1x_2 + x_2^2\right)$$

SECOND ORDER TAYLOR EXPANSION:

If $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable, then

$$f(\underset{\substack{\uparrow \\ \mathbb{R}^n}}{x} + \underset{\substack{\uparrow \\ \mathbb{R}^n}}{h}) = f(x) + \nabla f(x)^T h + \frac{1}{2} h^T D^2 f(x) h + o(\|h\|^2)$$

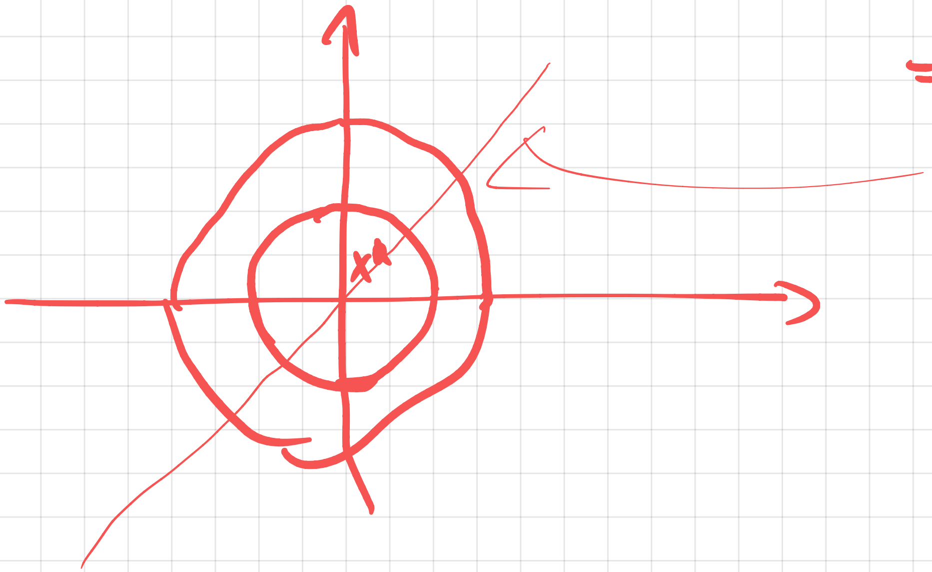
Ill-conditionning is a difficulty in optimization.

For a convex-quadratic problem $f(x) = \frac{1}{2}(x-x^*)^T A (x-x^*)$ where A is symmetric positive definite.

Reminder: If $A = Id = \begin{pmatrix} 1 & \\ & \ddots & \\ & & 1 \end{pmatrix}$, $f(x) = \frac{1}{2}(x-x^*)^T A (x-x^*)$

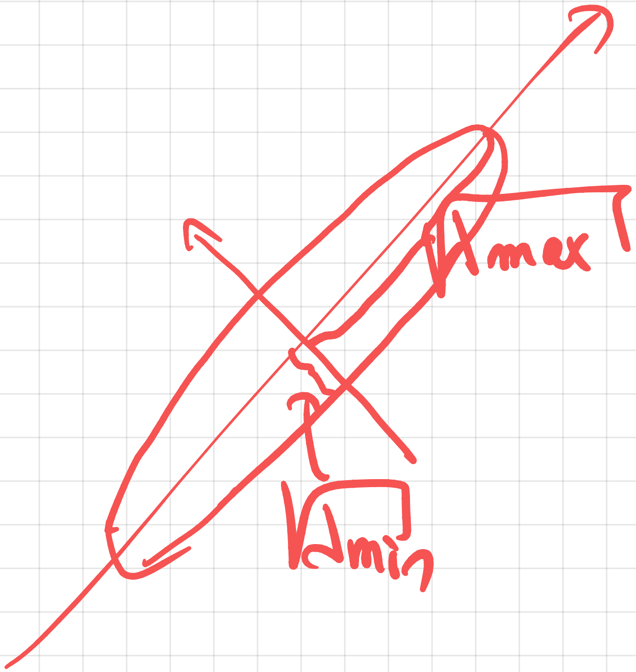
$= \frac{1}{2}(x-x^*)^T (x-x^*)$

$= \frac{1}{2} \| x - x^* \|^2$



cut

If $A \neq Id$, the level sets are ellipsoid.



λ_{max} : largest square root of A

λ_{min} : smallest square root of A

For a ill-conditioned problem we have a large ratio between the largest axis of ellipsoid and smallest axis, equivalently we have a large ratio between the largest eigenvalue of A and the smallest eigenvalue of A.

for a ill-conditionned problem, the condition number of the matrix A is large (of the order of 10^6 or higher)

$$\text{cond}(A) = \frac{\lambda_{max}(A)}{\lambda_{min}(A)}$$

Symmetric matrix

A ill-conditionned convex-quadratic problem is a problem with a ill-conditionned Hessian matrix.

More generally (not just for convex quadratic functions), a function $f : \mathbb{R}^n \to \mathbb{R}$ where the Hessian matrix is ill-conditionned is said to be ill-conditionned.

GRADIENT DIRECTION VERSUS NEWTON DIRECTION

Gradient direction: $Df(x)$

Newton direction: $-\left[D^2 f(x)\right]^{-1} Df(x)$

Exercice: $f(x) = \frac{1}{2} x^T H x$, $x \in \mathbb{R}^2$ $H = \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix}$

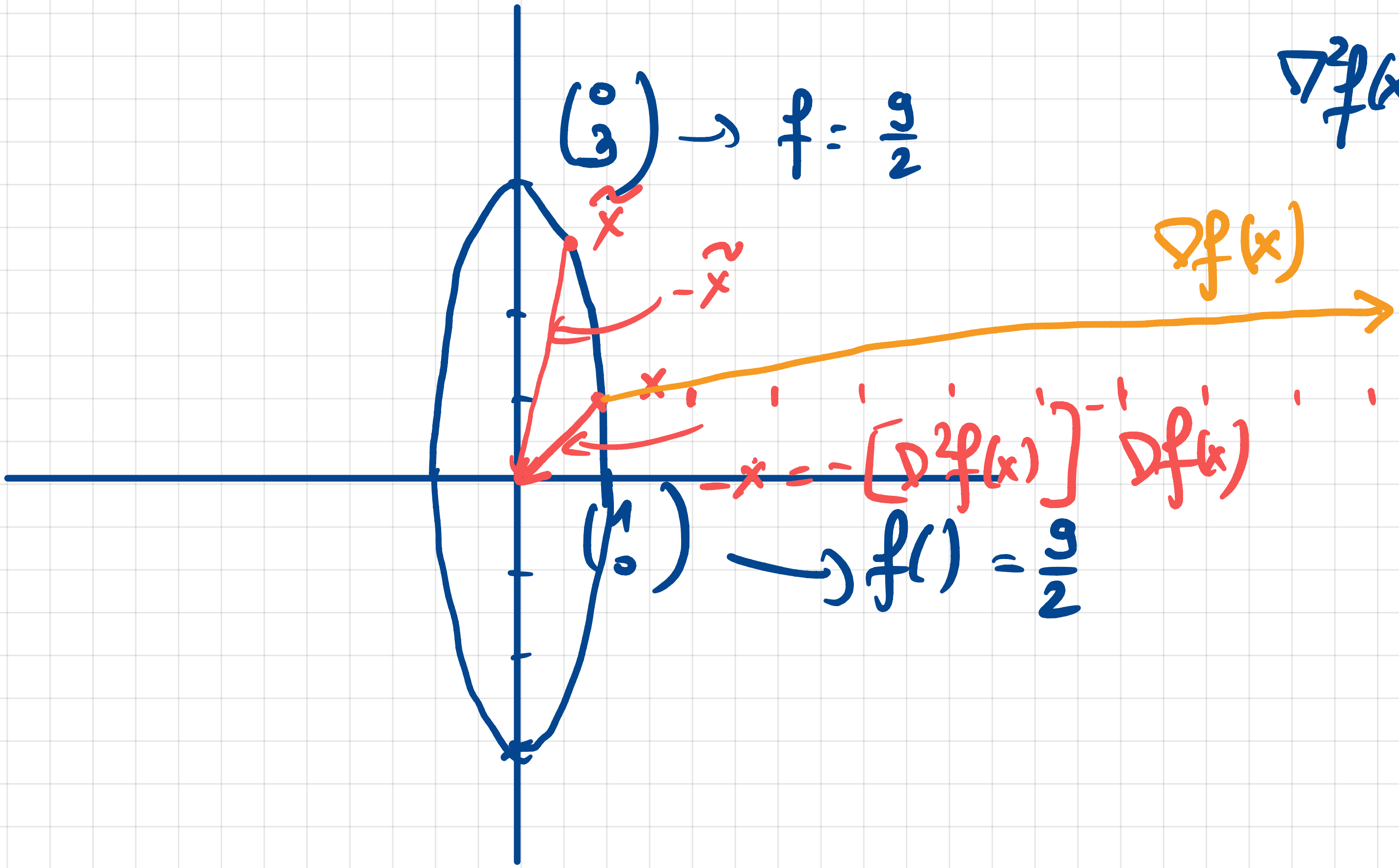
1) Plot level sets of f
2) Plot the gradient direction at different x
2) Compute & plot the Newton direction

Correction of previous exercise.

$$f(x) = \frac{1}{2}\left(9x_1^2 + x_2^2\right)$$

$$\nabla f(x) = \begin{pmatrix} 9x_1 \\ x_2 \end{pmatrix}$$

$$\nabla^2 f(x) = \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix}$$



$$\nabla^2 f(x) = \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix} \qquad \left[\nabla^2 f(x)\right]^{-1} = \begin{pmatrix} \frac{1}{9} & 0 \\ 0 & 1 \end{pmatrix}$$

If $D = \begin{pmatrix} \lambda_1 & & (0) \\ & \ddots & \\ (0) & & \lambda_n \end{pmatrix}$ is diagonal $D^{-1} = \begin{pmatrix} \frac{1}{\lambda_1} & & (0) \\ & \ddots & \\ (0) & & \frac{1}{\lambda_n} \end{pmatrix}$

<u>why</u> : Indeed $D D^{-1} = Id = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$

<u>Newton direction</u>: $-\left[\nabla^2 f(x)\right]^{-1} \nabla f(x) = -\begin{pmatrix} \frac{1}{9} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 9x_1 \\ x_2 \end{pmatrix}$

$$= -\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = -x$$

Iterative algorithm

for $t = 0, 1, \dots$

$$x_{t+1} = x_t + \underbrace{0{,}01}_{\eta \text{ (learning rate}} \left(-Df(x_t)\right)$$

Same with Newton direction

$$x_{t+1} = x_t + \eta \left(-[D^2 f(x_t)]^{-1} Df(x_t)\right)$$

We observe that the Newton direction points towards the optimum on convex quadratic problems independently of the condition number of the Hessian matrix.

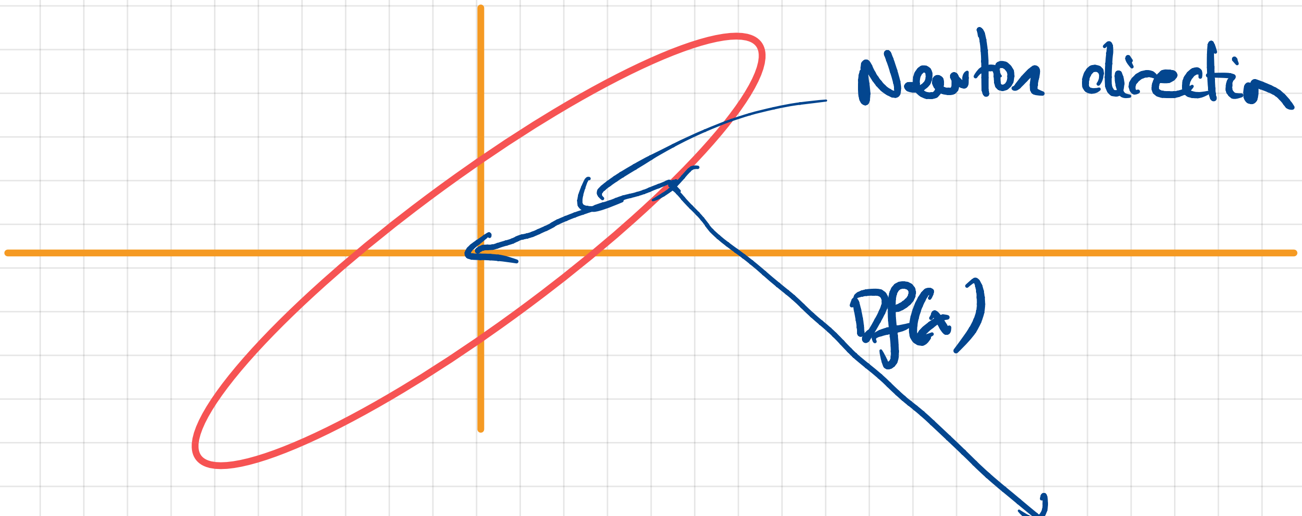
Whereas $-Df(x)$ points towards the optimum at $x = \begin{pmatrix}1\\1\end{pmatrix}$ if and only if $D^2 f(x) = Id$ (and thus the condition number equal to 1).

If the Hessian matrix is not diagonal anymore

$f(x) = \frac{1}{2} x^T A x$

A sym. pos. def

A not diagonal



$$Df(x) = Ax$$

$$D^2 f(x) = A$$

Newton direction: $-[A]^{-1} Ax = -\mathrm{Id}\, x = -x$

Optimality conditions

Assume

Optimality conditions

Assume $f: \mathbb{R} \longrightarrow \mathbb{R}$ is differentiable ($f'(x)$ exists for all x)

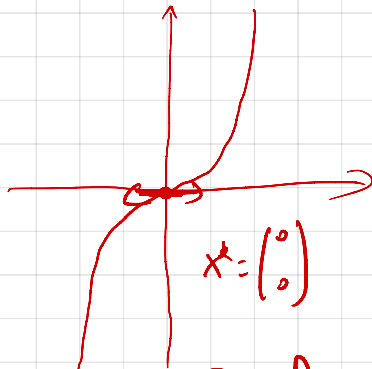
Which one of the following statements are correct:

① $f'(x^*) = 0 \implies x^*$ is a local optimum of f WRONG

$f(x) = x^3$

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

② x^* is a local optimum $\implies f'(x^*) = 0$ CORRECT



③ $f'(x^*) = 0 \implies x^*$ is a global optimum WRONG (same as ① for counter-example)

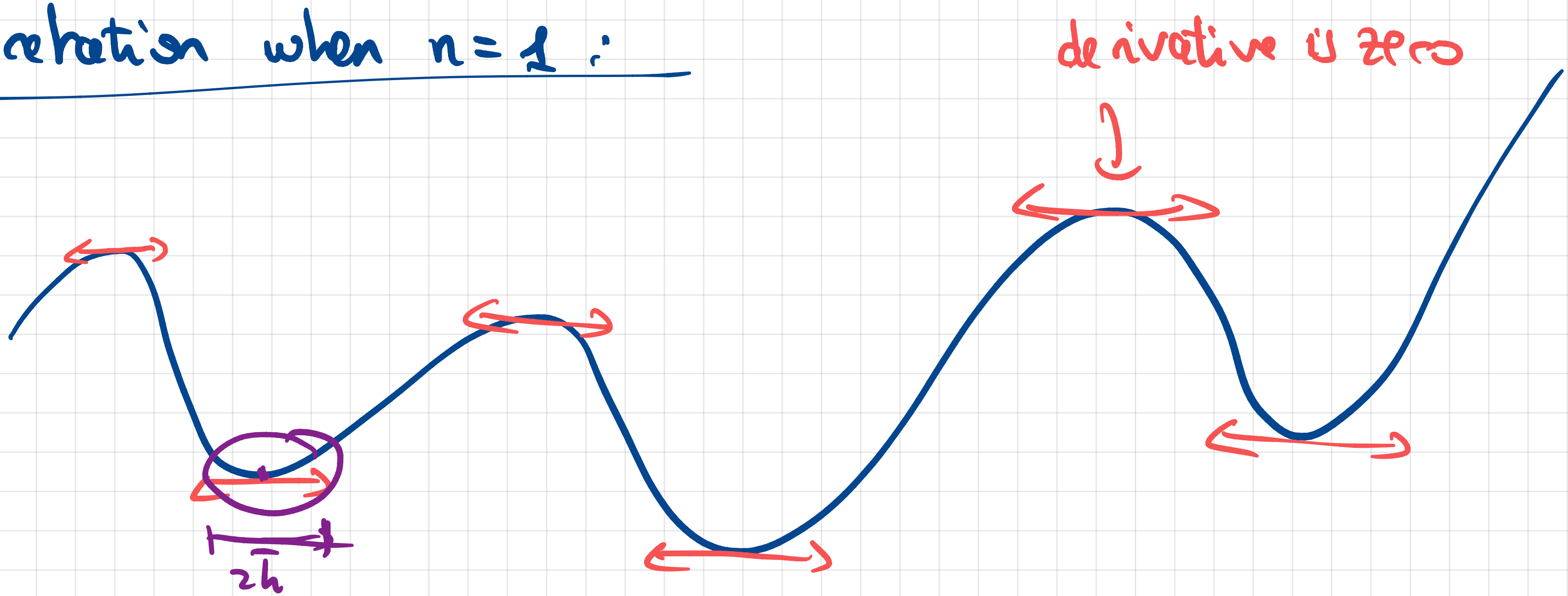
④ x^* is a global optimum $\implies f'(x^*) = 0$ CORRECT

THEOREM (first order necessary condition)

Let $f: \mathbb{R}^n \longrightarrow \mathbb{R}$ be a differentiable function. If x^* is a local optimum of f (minimum or maximum) then $\nabla f(x^*) = 0$

Remark: we talk about first order condition because it involves only first order derivative.

Interpretation when $n=1$:



PROOF for $n=1$:

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

assume that x^* is a local minimum : $f(x^*) \leq f(x^*+h)$ $\forall h$ small enough

$\left[\exists\, \bar{h} \text{ such } \forall\, h \le \bar{h} \quad f(x^*) \le f(x^*+h)\right]$

$$A(h) = \frac{f(x^*+h) - f(x^*)}{h} \qquad \ge 0$$

if $h \ge 0$ $\quad A(h) \ge 0$

if $h \le 0$ $\quad A(h) \le 0$

$$\underbrace{\lim_{\substack{h \to 0 \\ h \ge 0}} \underbrace{A(h)}_{\ge 0} = f'(x^*) \ge 0 \;, \quad \lim_{\substack{h \to 0 \\ h \le 0}} A(h) = f'(x^*) \le 0}$$

$$\Rightarrow f'(x) = 0$$

SECOND ORDER NECESSARY AND SUFFICIENT CONDITIONS:

Let's assume that f is twice continuously differentiable.

NECESSARY CONDITION: If x^* is a local minimum, then

$$\nabla f(x^*) = 0 \quad \text{and} \quad \nabla^2 f(x) \text{ is positive semi-definite.}$$

(if $n=1$ x^* local minimum $\Rightarrow f'(x^*) = 0$, $f''(x) \geq 0$)

[A sym. matrix is positive if $\forall y$ $y^T A y \geq 0$
definite $y^T A y = 0 \Rightarrow y = 0$]

positive definite $y^T A y > 0 \quad \forall y \neq 0$

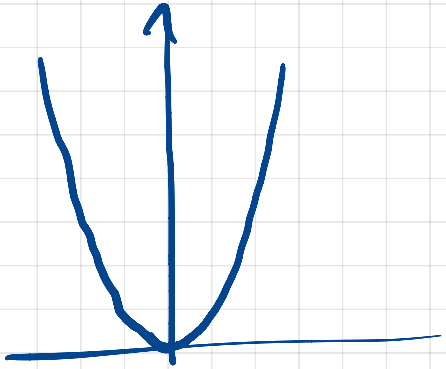
positive semi-definite $y^T A y \geq 0 \quad \forall y$

Not sufficient: $f(x) = x^3$, $f'(x) = 0$ $f''(x) = 0 \geq 0$, yet it not a local minimum.

SUFFICIENT CONDITION: If x^* such that $Df(x^*) = 0$ and $D^2 f(x)$ is positive definite, then x^* is a strict local minimum.

(if $n=1$, x^* such that $f'(x) = 0$ $f''(x) > 0 \Rightarrow x^*$ is a strict local optimum.

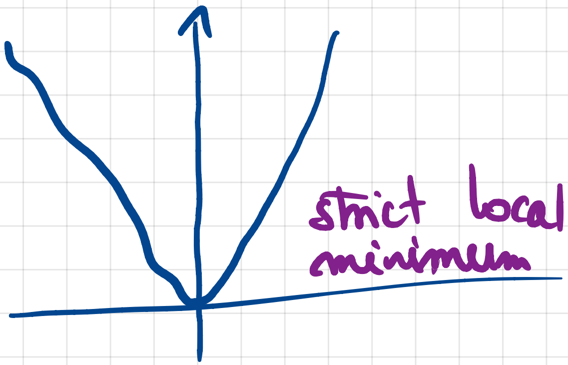
Example: $f(x) = x^2$, $f'(x) = 2x$ $f''(x) = 2$

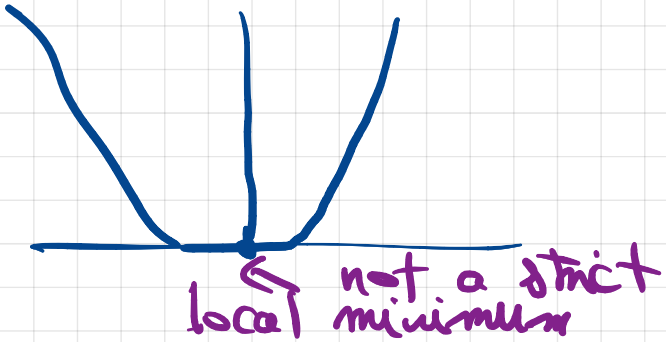


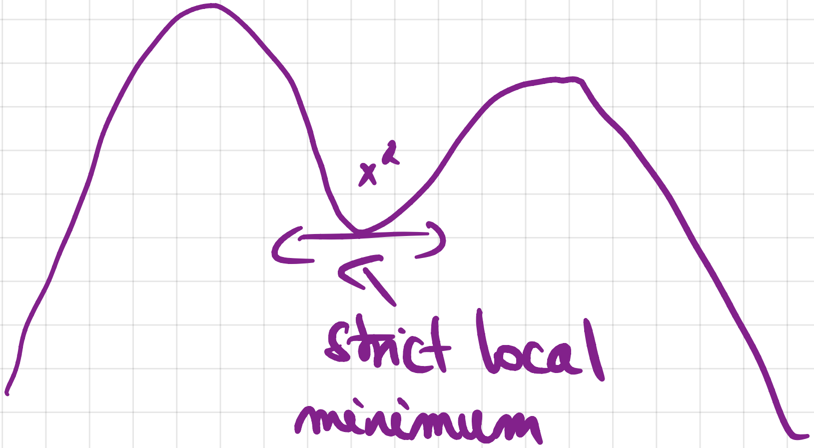
0 satisfies $f'(0) = 0$ $f''(0) = 2 > 0$

then 0 is a stric local minimum of the function

strict local minimum:







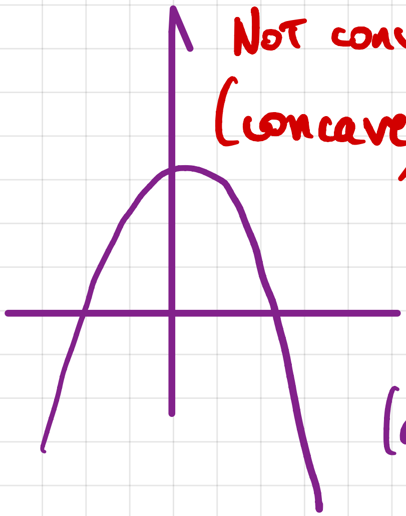
strict local minimum

$f'(x) = 0$

$f''(x) = 0$

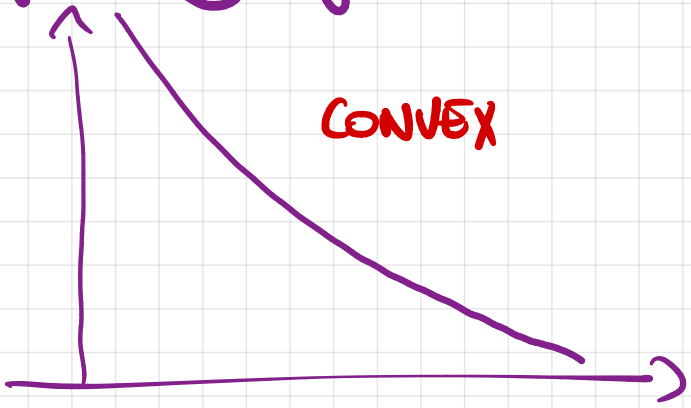
CONVEXITY:

Which of the following functions are convex?



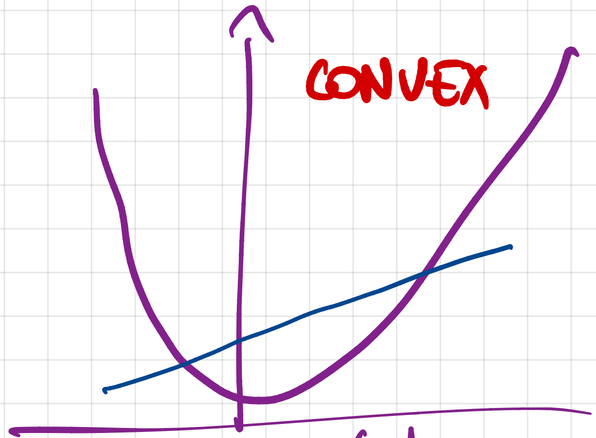
Not convex (concave)

(a)



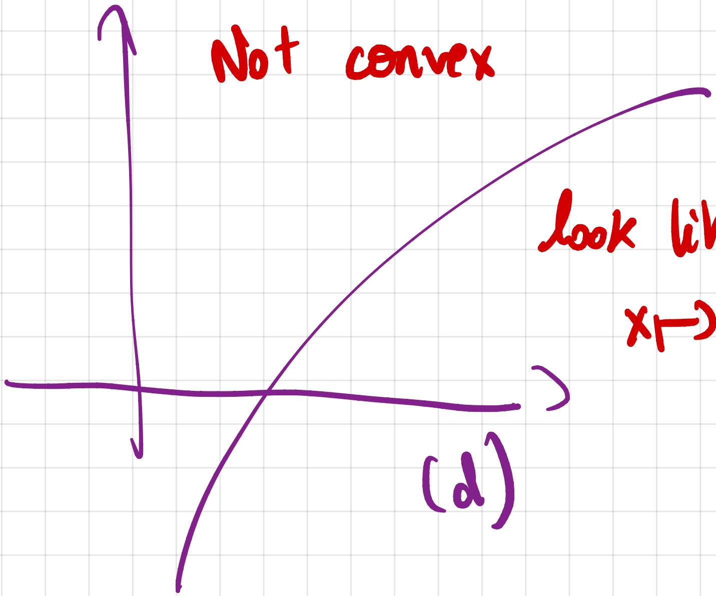
CONVEX

(b)



CONVEX

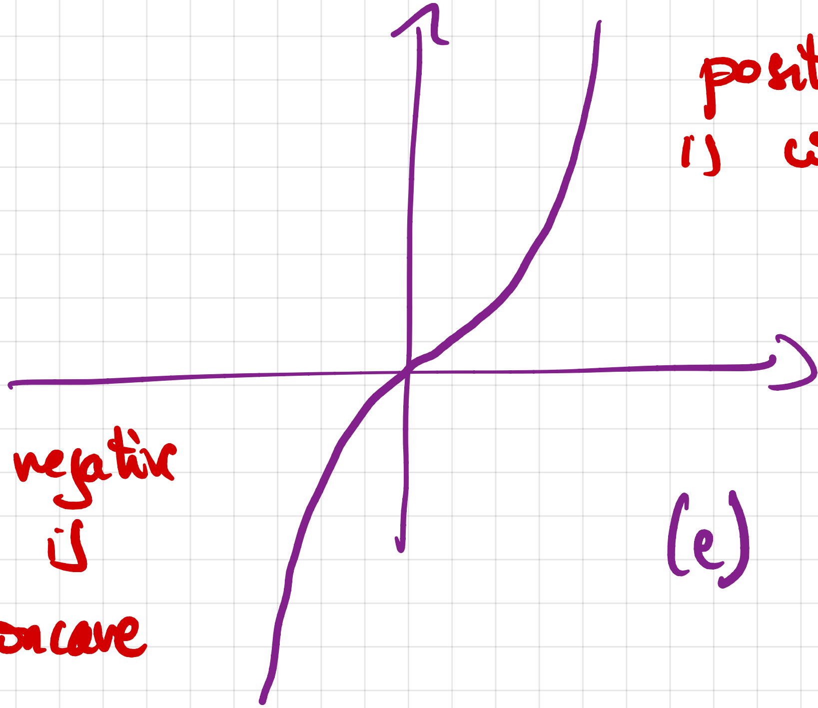
(c)



Not convex

look like $x \mapsto \ln(x)$

(d)



positive part is convex

negative part is concave

(e)

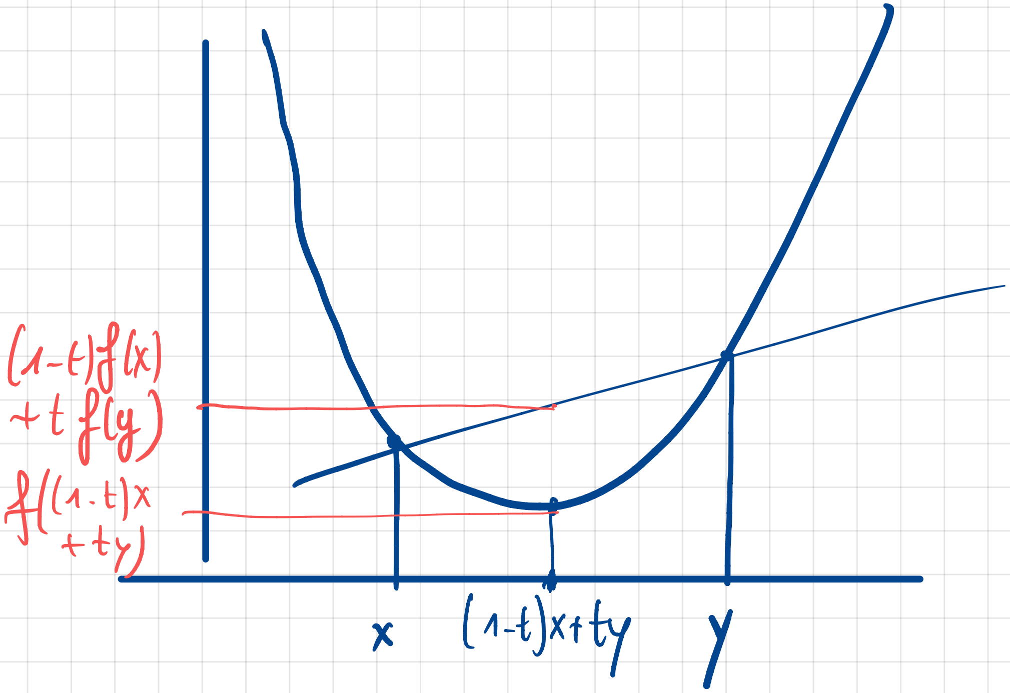
CONVEX FUNCTIONS

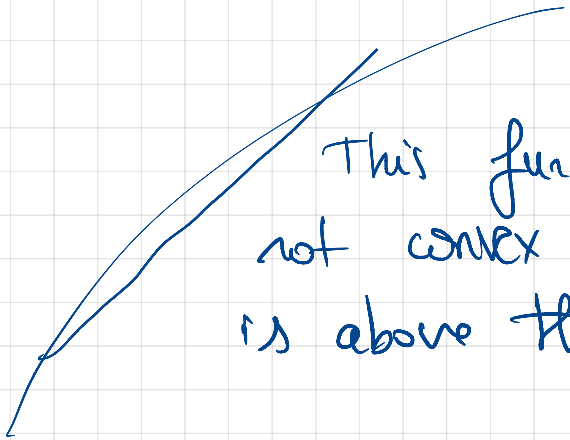
Let $f : U \subset \mathbb{R}^n \longrightarrow \mathbb{R}$. We say that f is convex, if for all $x, y \in U$

open convex set

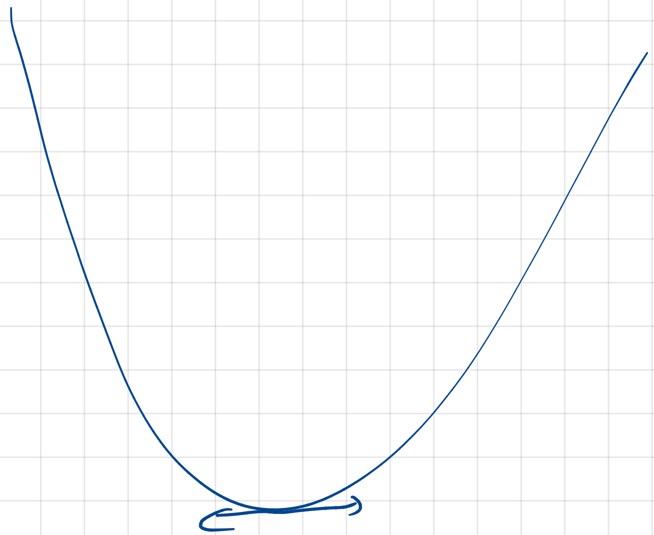
$\forall t \in [0,1]$

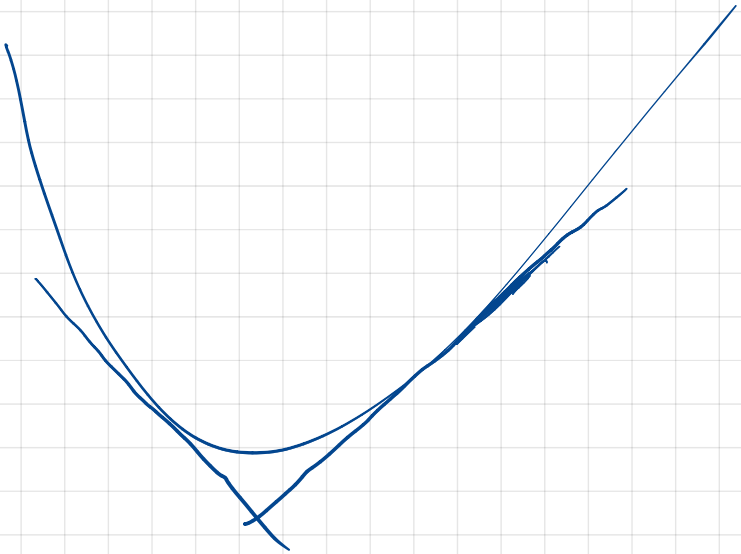
$$f\left((1-t)x + ty\right) \leq (1-t)f(x) + t f(y)$$





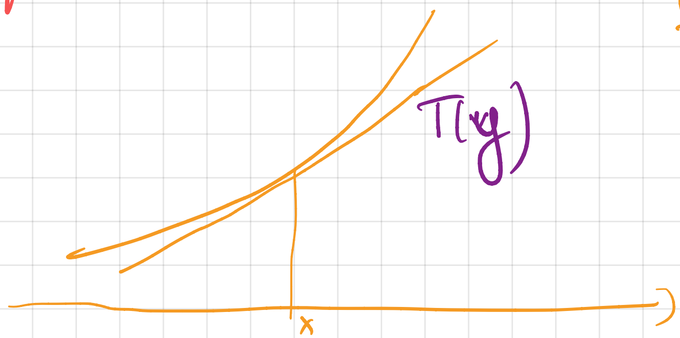
This function is not convex because f is above the line.





Intuition: for a convex function that is differentiable the tangent is below the curve.

Exercise: translate this property into an equation (you can assume $n = 1$)



Equation of the tangent in x

$$\boxed{y \overset{T}{\longmapsto} f'(x)(y-x) + f(x)}$$

$$T(x) = f'(x)(x-x) + f(x) = f(x)$$

↳ $T(x)$ goes though $(x, f(x))$

The slope of $T(x)$ is the derivative of f in x.

If $n=1$, f is differentiable, then f is convex if and only if for all x and y, $f(y) \geq f'(x)(y-x) + f(x)$

↳ This property translates that for a convex function the curve is above the tangent.

THEOREM: If f is differentiable, then f is convex if and only if for all x, y

$$f(y) - f(x) \geq Df(x)^T (y-x)$$

If $n=1$, f is twice continuously differentiable, then f is convex iff $f''(x) \geq 0$.

THEOREM: If f is twice continuously differentiable, then f is convex if and only if $D^2 f(x)$ is positive semi-definite for all x.

Definition: A function is concave if and only if $-f$ is convex.

Examples:

$f(x) = x^2$ is convex because $f''(x) = 2 \geq 0$

$f(x) = -x^2$ is concave because x^2 is convex.

$f(x) = \log(x)$ $\quad f'(x) = \frac{1}{x}$, $f''(x) = -\frac{1}{x^2} \leq 0$

$\hookrightarrow$ f is concave.

$f(x) = x$ $\quad f$ is convex (and concave)

$f''(x) = 0$

Other examples of convex functions:

- $f(x) = \frac{1}{2} x^T A x$ $\quad A$ sym pos semi definite, then f is convex
- $f(x) = a^T x + b$ $\quad a \in \mathbb{R}^n$, $b \in \mathbb{R}^n$ [linear slope]
- the negative of the entropy : $f(x) = -\sum_{i=1}^{n} x_i \ln(x_i)$

EXERCICE : Let $f : U \subset \mathbb{R}^n \to \mathbb{R}$ be a convex and differentiable function. convex open subset of $\mathbb{R}^n$

Prove that if $Df(x^*) = 0$, then x^* is a global minimum of the function.

If f is convex and differentiable,

$$\forall y, x \qquad f(y) - f(x) \geq Df(x)^T (y - x)$$

If x^* is such that $Df(x^*) = 0$, then $Df(x^*)^T(y - x^*) = 0$ and the previous equation gives

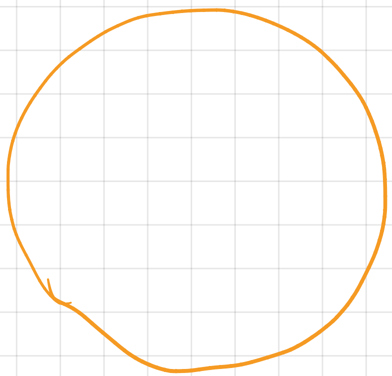
$$f(y) \geq f(x^*) \quad \forall y$$

which means that x^* is a global minimum of f.

The important consequence is that for convex and differentiable functions, critical points, ie points where $\nabla f(x^*) = 0$ are global minima of the function.

We assumed that $f : U \subset \mathbb{R}^n \to \mathbb{R}$ where U is an open convex set.

"open : intuition, ball with boundary":

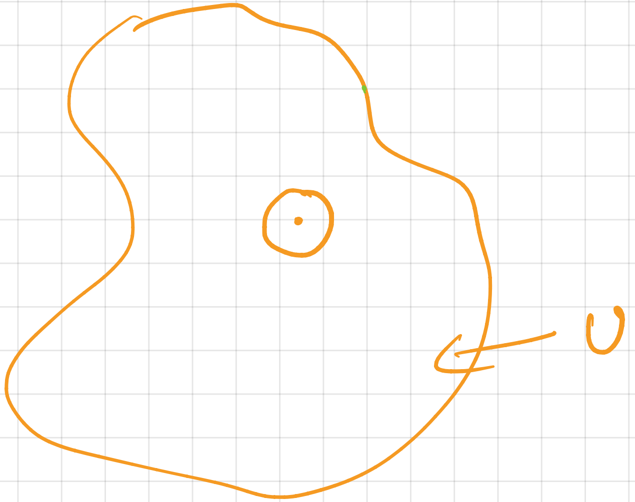


$[0,1]$ closed $\quad]0,1[$

$]0,1($, $(0,1)$, $]0,1[$

Same notation for open intervall (excluding 0 and 1 from $[0,1]$

$]0,1[\cup]2,3[$ is also an open .



U is open, if $\forall x \in U$, I can put a small open ball in U which is fully in U

$]0,1[\cup]1,2[$