OPTIMIZATION FOR MACHINE LEARNING — CLASS 2

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

REMINDER : Continuous optimization

$$\text{minimize } f(x_1, \dots, x_n) \qquad x_i \in \mathbb{R}$$

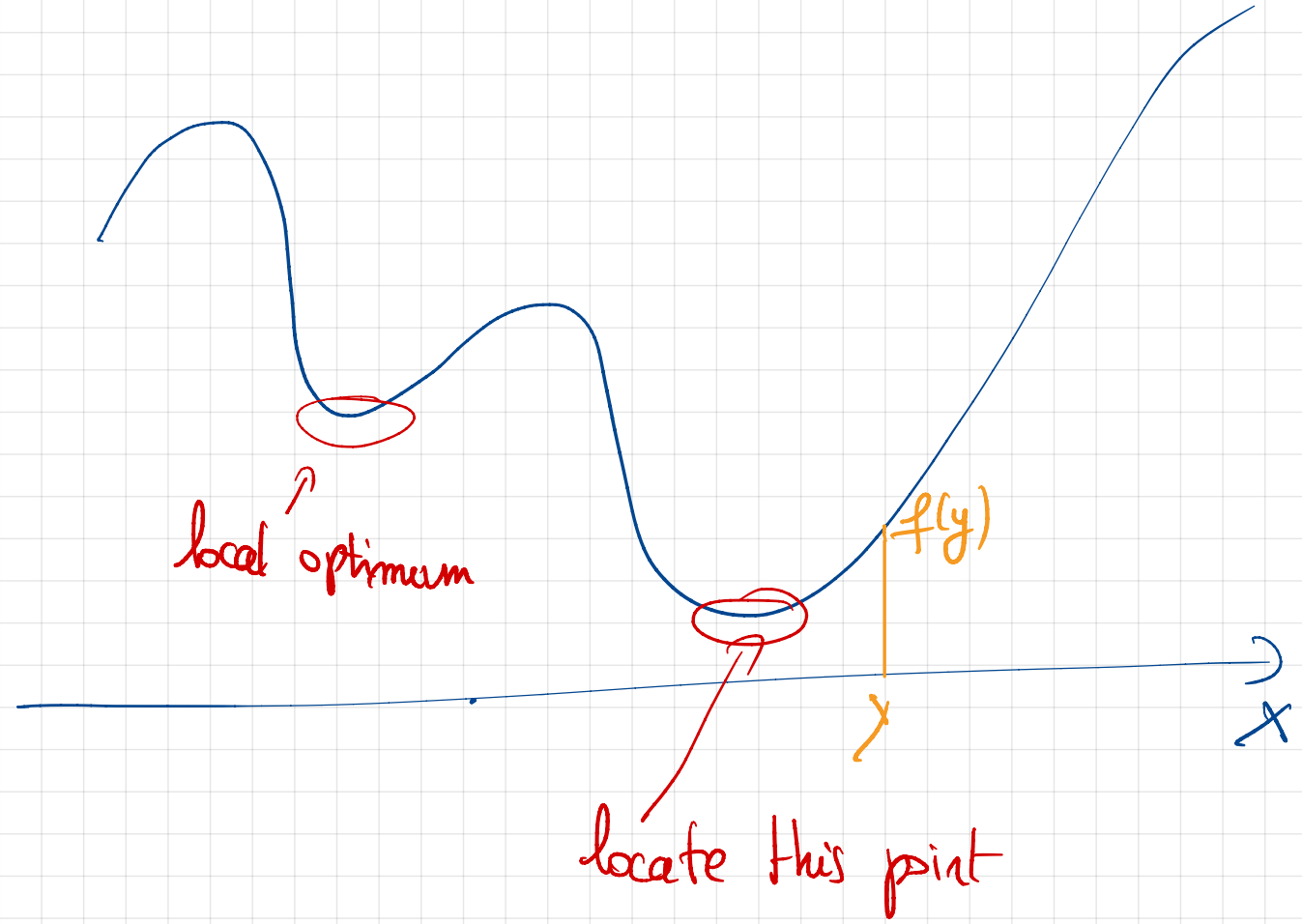
$x = (x_1, \dots, x_n) \in \mathbb{R}^n$ vector space

n: dimension of problem.

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

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

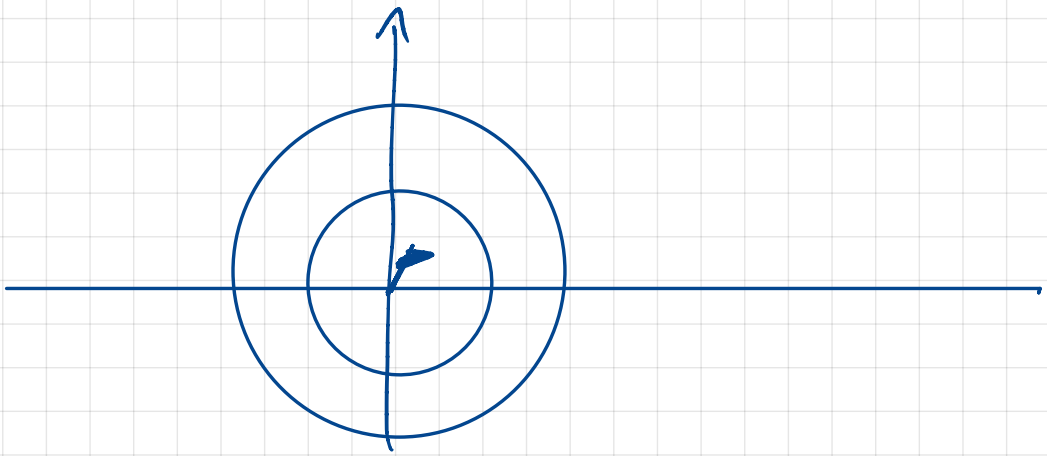
When $n = 1$ $\min_{x \in \mathbb{R}} f(x)$



$n = 2$, 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.

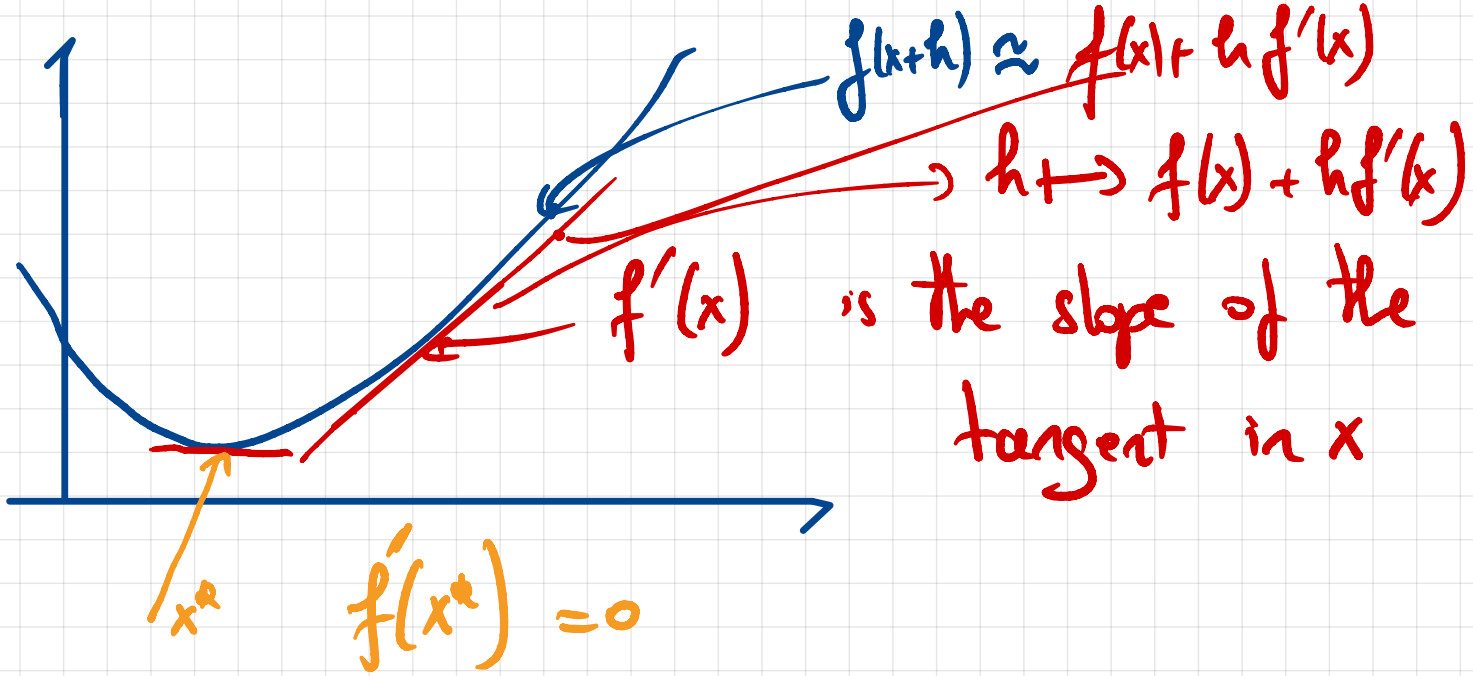


Derivability or differentiability

$n = 1$, 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



If f is differentiable in x then

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

Taylor expansion of f in x, at first order

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

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

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

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

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

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

$f'(x)(h_1 + h_2) = f'(x)h_1 + f'(x)\,h_2$

$f'(x)(\alpha h) = \alpha\,[f'(x).h)$

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

$\mathbb{R}$

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) $f(x) = \|x\|^2$, $x \in \mathbb{R}^n$, $Df_x(h) = 2x^T h$

1) $f(x) = Ax$

$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{n1} & & a_{nn} \end{bmatrix}$ n (rows), n (columns) $\quad x \in \mathbb{R}^n$

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

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

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

$\left.\begin{array}{l} h \mapsto Ah \text{ is Linear} \\ \mathbb{R}^n \longrightarrow \mathbb{R}^n \end{array}\right\}$

so f is differentiable in x and

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

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

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

$$= x^T x + x^T h + h^T x + h^T h \qquad (h^T x = x^T h)$$

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

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

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

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

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

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

$= x^T h$

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

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

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

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

$\left(\quad \right)^T \rightarrow \left(\quad \right)$

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

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

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

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

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

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

CHAIN RULE :

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

$f : \mathbb{R} \longrightarrow \mathbb{R}$ $\quad$ $g : \mathbb{R} \longrightarrow \mathbb{R}$

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

composition

$x \xmapsto{f} \sin(x)$

$x \xmapsto{g} x^2$

$f \circ g(x) = f(g(x)) = \sin(x^2)$

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

[composition & product of functions are different]

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

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

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

$\exists\ a \in \mathbb{R}^n$ 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]

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

LINK BETWEEN DIFFERENTIAL & GRADIENT

The gradient can also be defined with partial derivatives.

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

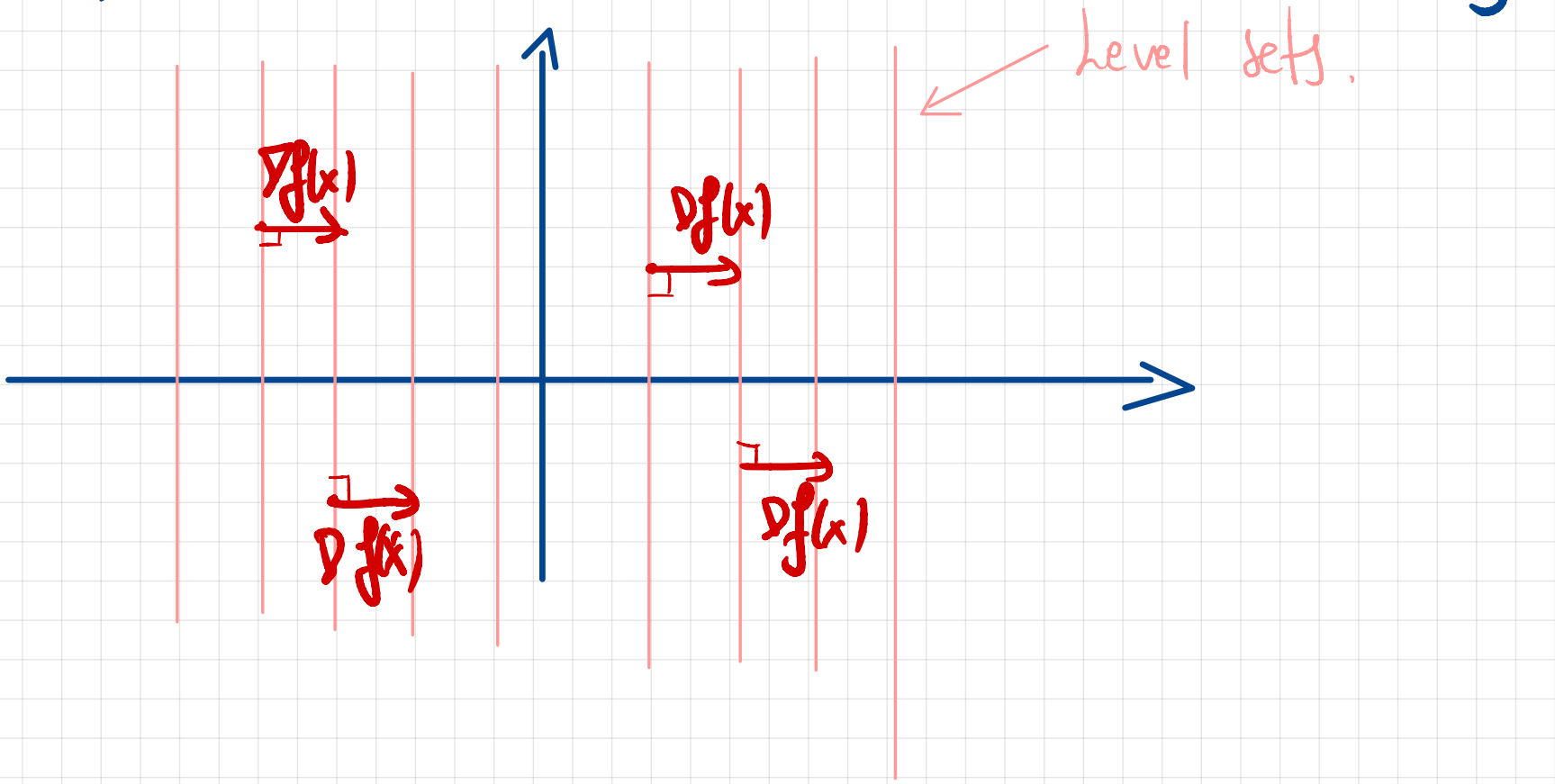
Exercice: Compute the gradient of.

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

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

$f(x) = x^T x$

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



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

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.

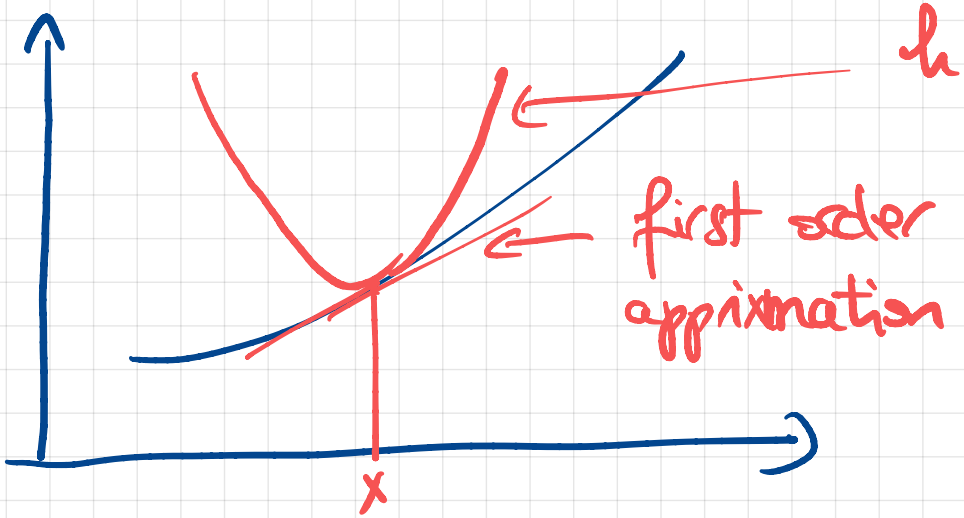
$f''(x)$ is called the second order derivative of f

If f is two times differentiable then

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

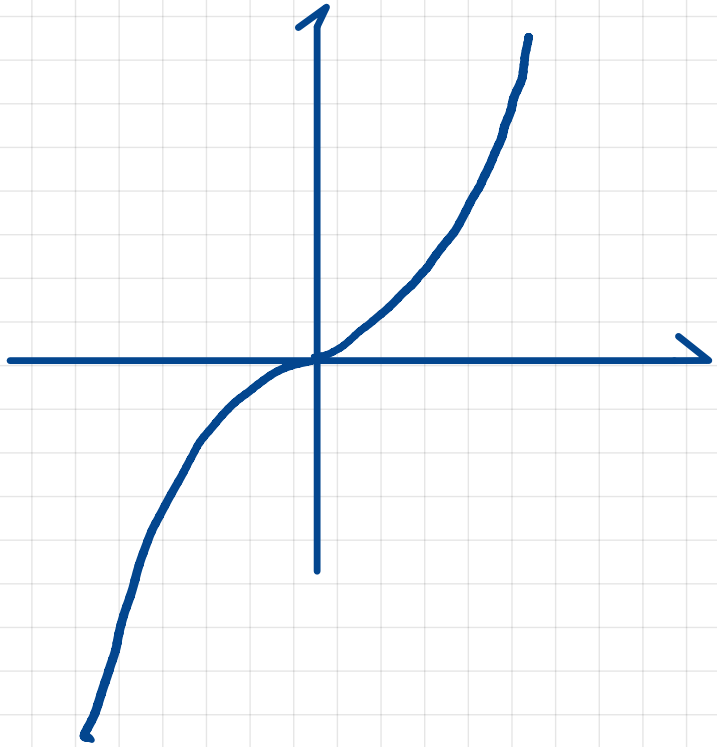
SECOND ORDER TAYLOR EXPANSION FORMULA

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



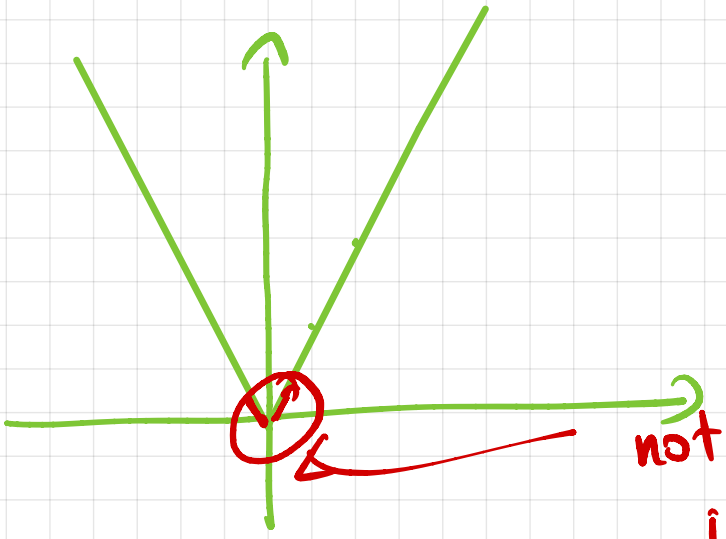
$h \mapsto f(x) + f'(x)h + \frac{1}{2}f''(x)h^2$
quadratic approximation of f in x

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



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

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



not derivable in zero

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

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

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

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

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

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

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

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

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

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

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

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

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

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

DETAIL ABOUT:

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

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

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

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

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

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

SECOND ORDER TAYLOR EXPANSION:

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

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

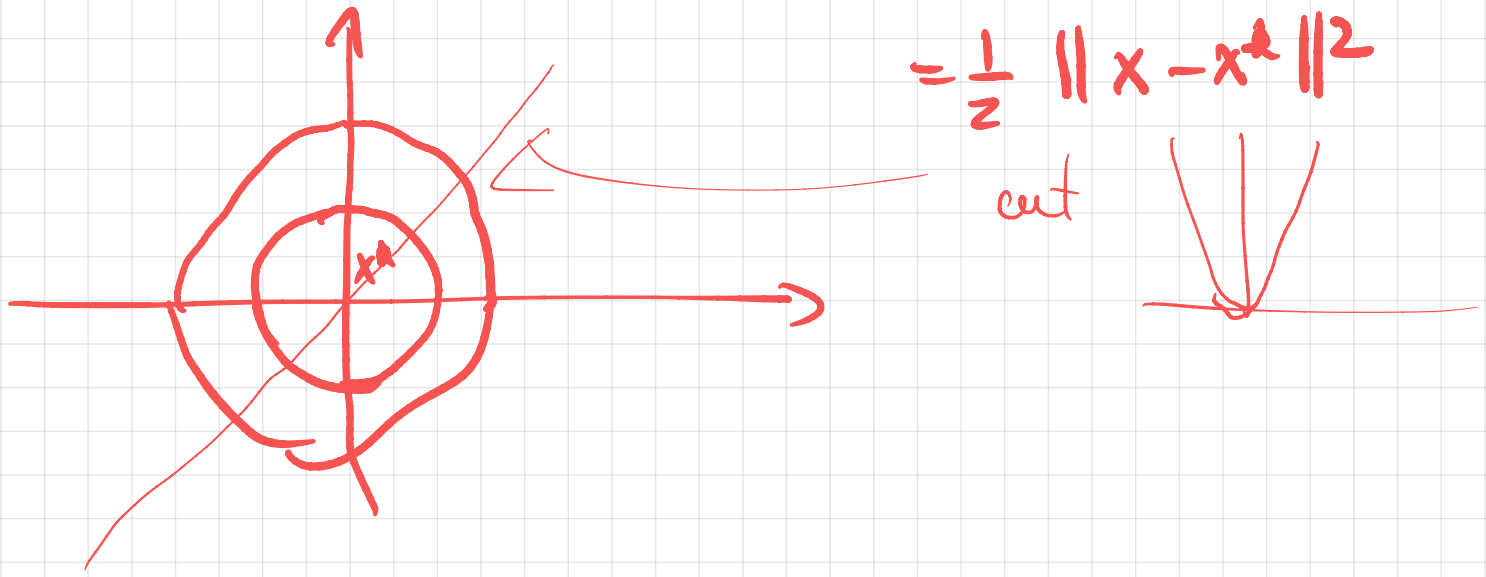
III - conditionning is a difficulty in optimization.

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

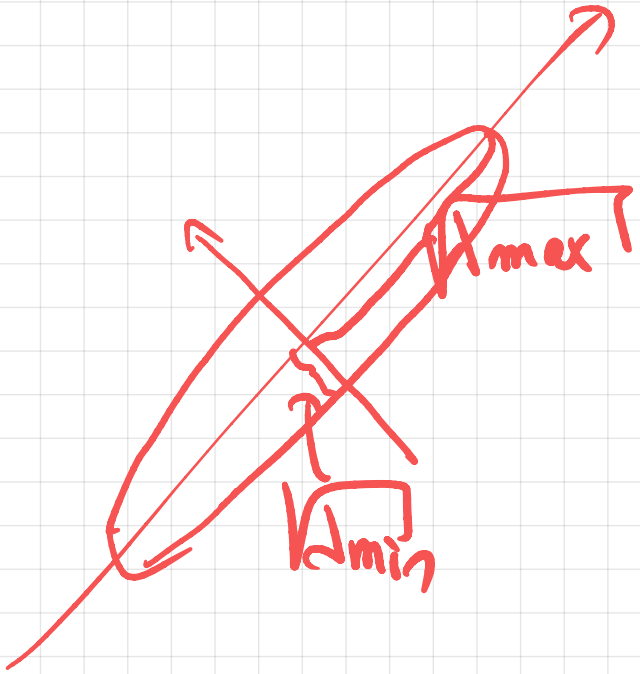
Reminder: If $A = Id = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}$, $f(x) = \frac{1}{2}(x-x^*)^T A (x-x^*)$

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

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



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)

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

A: Symetric matrix

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

More generally (not just for convex quatratic 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 : $-[D^2 f(x)]^{-1} Df(x)$

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