TC2 - Optimization

CLASS 3 . 19.11.2020

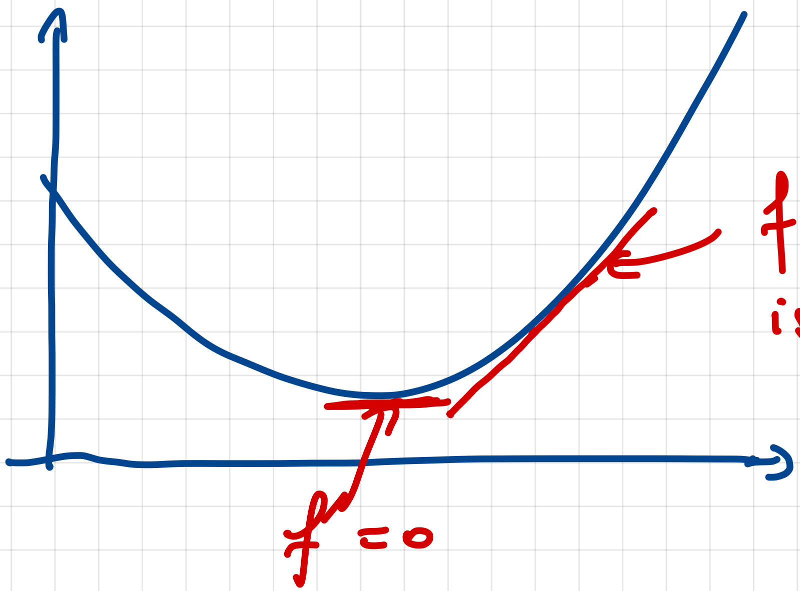
Derivability or differentiability

Let assume $n = 1$, 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 denoted $f'(x)$ and is call derivative of f in x



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

$$\left\{\begin{array}{l}\mathbb{R} \to \mathbb{R} \\ h \mapsto f'(x)\,h\end{array}\right.$$

linear.

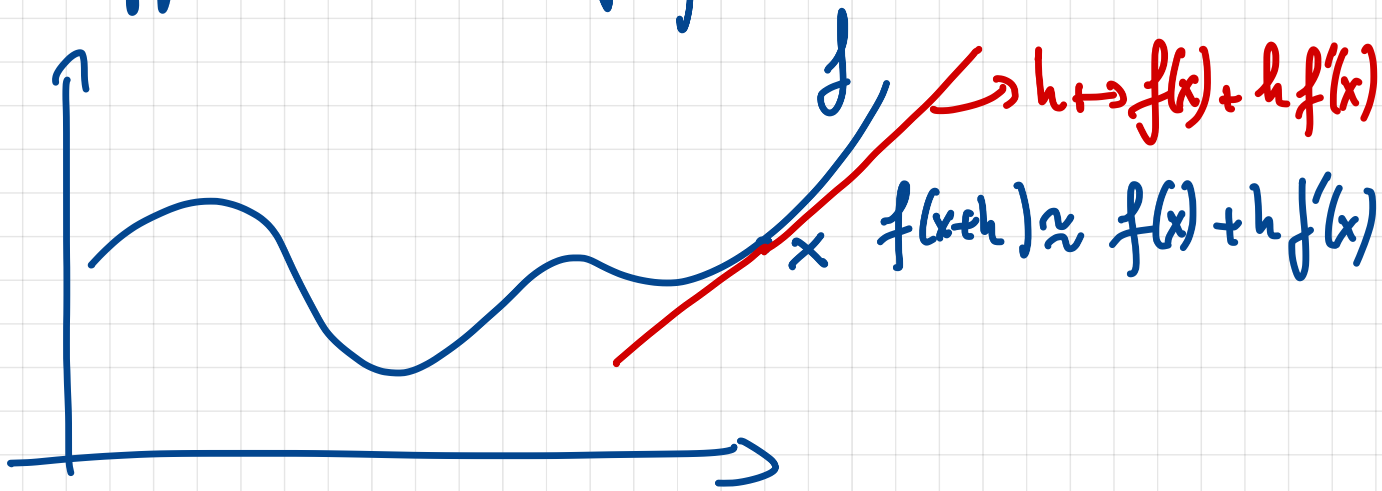
If f is differentiable in x then

$$\underbrace{f(x+h) = f(x) + f'(x)\,h + o(\|h\|)}_{\text{Taylor expansion of } f \text{ in } x \text{ at the first order.}}$$

For h small enough $h \longmapsto f(x+h)$ is approximated by

$$h \longmapsto \underbrace{f(x) + f'(x)\,h}_{\text{first order approximation of } f}$$

Interpret geometrically



. How do we generalize the notion of derivative of a function for $n=1$ 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 $\forall h \in \mathbb{R}^n$ $\quad f(x+h) = f(x) + Df_x(h) + o(\|h\|)$

If $n=1$ $\quad Df_x(h) = f'(x)\,h$

Exercise : $\quad f(x) = Ax \quad A$ is a $n\times n$ matrix | $f(x) = \|x\|^2$

$Df_x = ?$ | $Df_x = ?$

$f(x) = Ax$ $\quad$ A $n \times n$ matrix.

To show that f is differentiable and to find Df_x we need to look at
$$f(x+h) = A(x+h)$$
$$= Ax + Ah$$
$$= f(x) + Ah$$

$h \mapsto Ah$ is linear, so f is differentiable in x

and $\quad Df_x = A \qquad Df_x(h) = Ah$.

If $\quad f(x) = \|x\|^2 = x^T x$
$$f(x+h) = (x+h)^T (x+h) = x^T x + x^T h + h^T x + h^T h$$
$$= x^T x + 2x^T h + \underbrace{h^T h}_{= o(\|h\|)}$$

$\underbrace{h^T x}_{\in \mathbb{R}} = x^T h \qquad h^T x = (h^T x)^T = x^T (h^T)^T = x^T h \qquad (h)^T$

$h \mapsto 2x^T h$ is linear in h

$Df_x = 2x^T$

CHAIN RULE :

$f : \mathbb{R} \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$

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

composition

$$f g' + g f' = (fg)'$$

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

$x \xmapsto{g} x^2$

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

$$f(x)\,g(x) = \sin(x) \quad x^2$$

composition $\neq$ product.

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

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

When $f: \mathbb{R}^n \to \mathbb{R}$ is differentiable in x, there is a specific representation of the differential of f in x. $Df_x: \mathbb{R}^n \to \mathbb{R}$

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

scalar or dot product.

[This comes from Riesz representation theorem]

The vector a has a specific name

$a = \nabla f_x$ [Gradient of f in x]

The gradient can also be defined with partial derivatives.

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

$$f_{x_0}^i : y \in \mathbb{R} \longrightarrow f(x_0^1, \ldots, x_0^{i-1}, y, x_0^{i+1}, \ldots, x_0^n)$$

ith coordinate

$$\frac{\partial f}{\partial x_i} = \left(f_{x_0}^i\right)'$$

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

Exercice : Compute the gradient of

$$\begin{cases} f(x) = x_1 & x \in \mathbb{R}^n \\ f(x) = a^T x & a = \begin{pmatrix} a_1 \\ | \\ a_n \end{pmatrix} \\ f(x) = x^T x \end{cases}$$

- $f(x) = x_1$

$$Df_x = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

$x_1 \xrightarrow{f^1} f(x) = x_1$ $\quad [f^1(x_1)]' = 1$

$x_2 \xrightarrow{f^2} f(x) = x_1$ $\quad f^2$ is constant w.r.t x_2,

$[f^2(x_2)]' = 0$

$$Df_x = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

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

  $= a_1 x_1 + \dots + a_n x_n$

  $$Df_x = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

- $f(x) = \|x\|^2 = x^T x = \sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + \dots + x_n^2$

  $$Df_x = \begin{pmatrix} 2x_1 \\ 2x_2 \\ \vdots \\ 2x_n \end{pmatrix} = 2x$$

This is compliant with what we had before

$Df_x(h) = 2x^T h$

$= \langle Df_x, h \rangle$

$= Df_x^T h$

$\Rightarrow Df_x = 2x.$

GEOMETRICAL INTERPRETATION OF THE GRADIENT

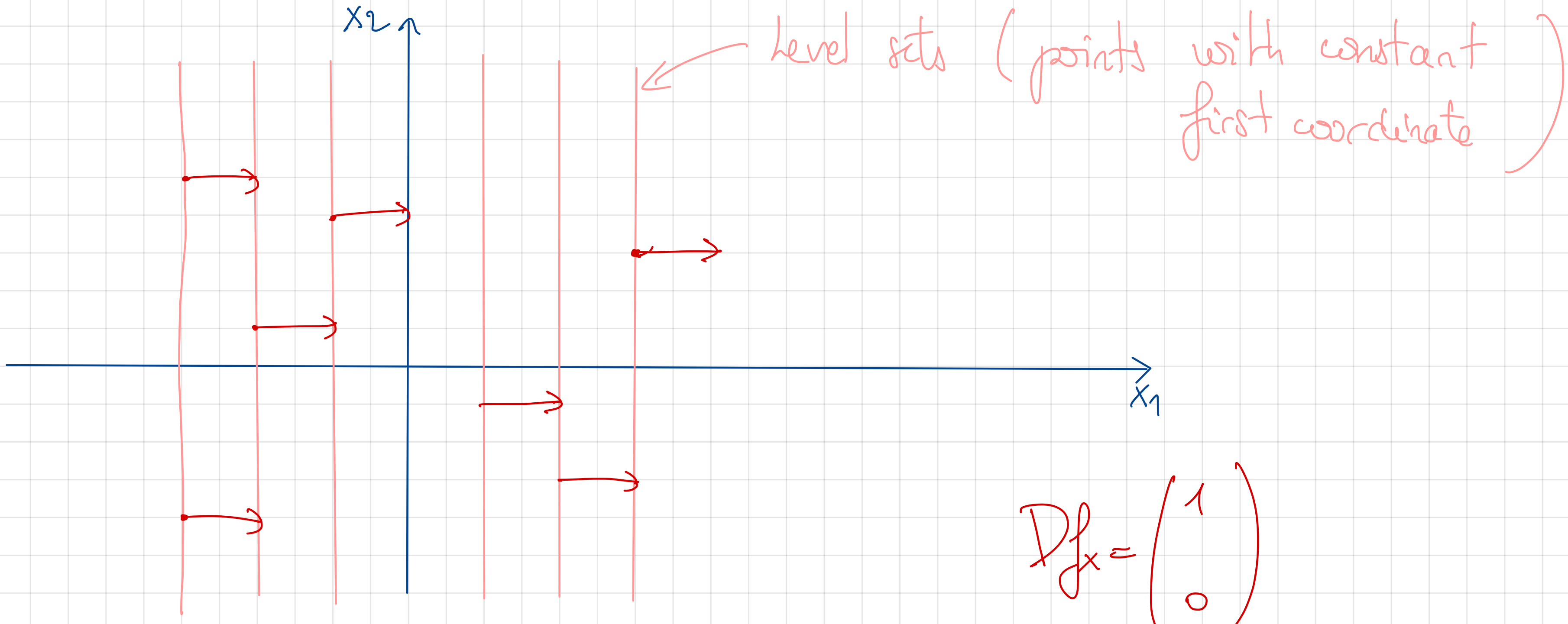
$f_1(x) = x_1 \qquad f_2(x) = \|x\|^2$

Plot on two figures for $n=2$, the level sets of f_1, f_2 and also plot Df_1, Df_2 on the figures.

$$\underset{\substack{\cap \\ \mathbb{R}}}{L_c} = \{ x \in \mathbb{R}^n ; f(x) = c \} \quad \text{level set}$$

$$f_1(x_1, x_2) = x_1$$

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

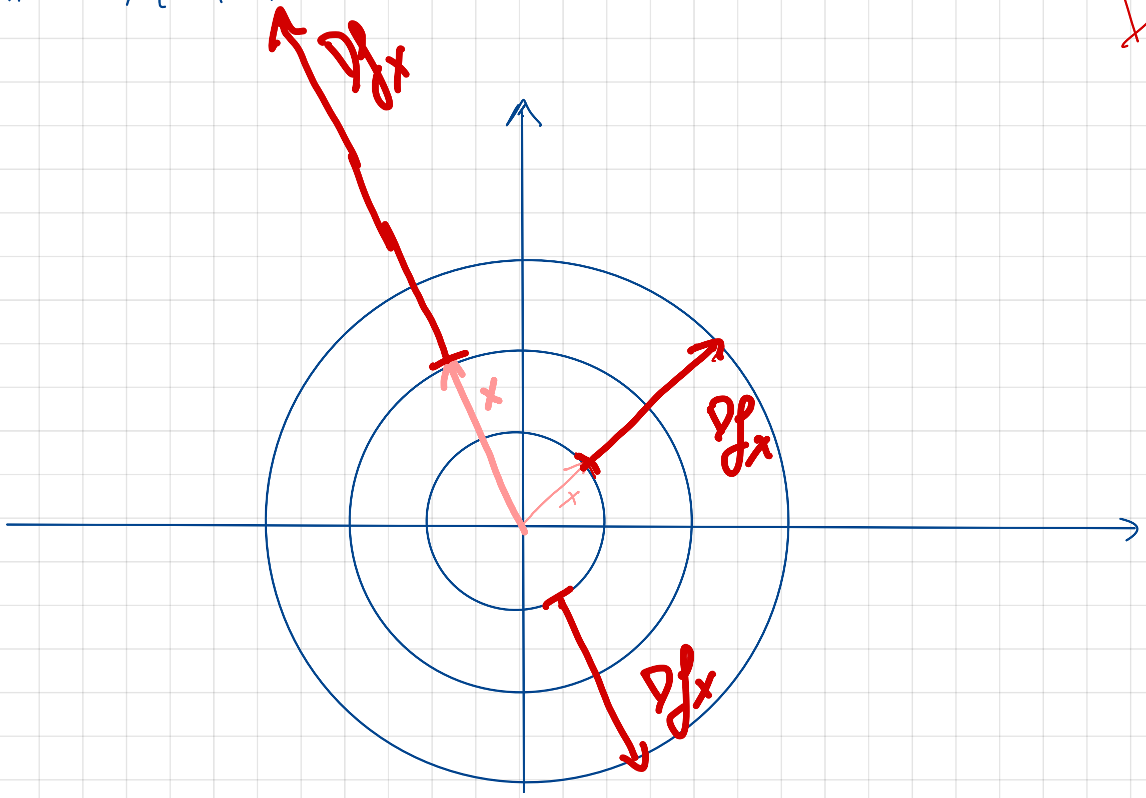


$Df_x = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

On this plot the gradient is orthogonal to the level set.

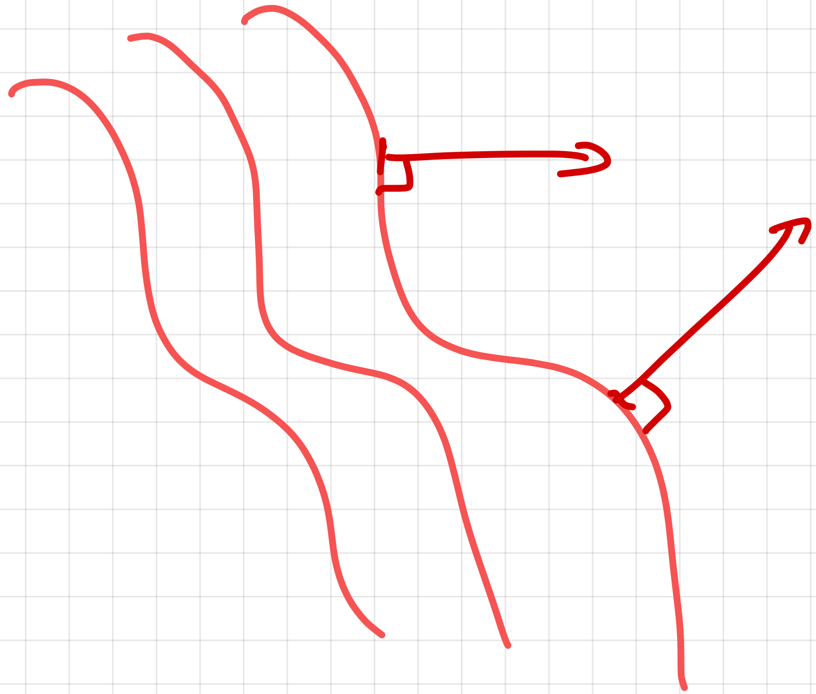
$$f(x_1, x_2) = \|x\|^2 = x_1^2 + x_2^2$$

$$Df_x = 2x$$



The gradient is orthogonal to the level sets.

More generally, the gradient of a differentiable function is orthogonal to its 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.

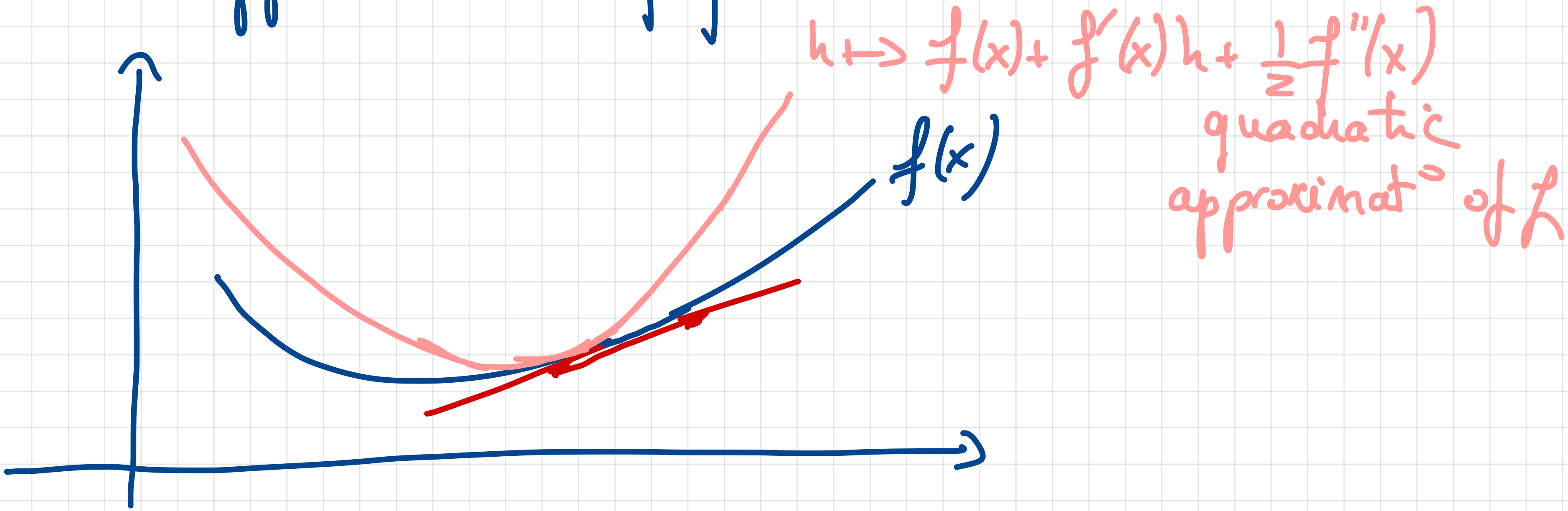
$f''(x)$ is called the second order derivative of f.

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 FORMULA

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



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

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

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

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

"Schwarz theorem"

It is a symmetric matrix.

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

Compute $\nabla^2 f$.

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

$$A = \begin{pmatrix} 9 & 1 \\ 1 & 1 \end{pmatrix} \qquad f(x_1, x_2) = \frac{1}{2} x^T A x \qquad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

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

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

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

$$\frac{\partial^2 f}{\partial x_1 \partial x_2} = \frac{\partial}{\partial x_1}\left(\frac{1}{2}(2x_2 + 2x_1)\right) = \frac{\partial}{\partial x_1}\left(x_2 + x_1\right) = 1 \qquad \nabla^2 f = \begin{pmatrix} 9 & 1 \\ 1 & 1 \end{pmatrix} = A$$

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

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

$$\text{Hessian}(f) = D^2 f = A$$

If A is not symmetric then $D^2 f = \frac{1}{2}(A + A^T)$

Second order Taylor formula:

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

$$f(x+h) = f(x) + Df(x)^T h + \frac{1}{2} h^T D^2 f(x) h + o(\|h\|^2)$$

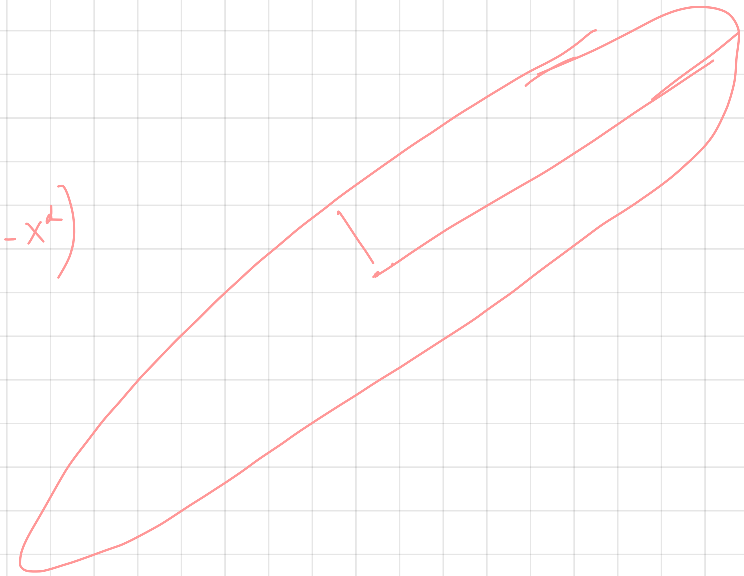
Last time we have seen that a ill-conditionned convex-quadratic problem $f(x)=\frac{1}{2}(x-x^*)^T A\,(x-x^*)$ is a problem where the matrix A is ill conditionned.

where A is symmetric positive definite.

Now we know that A is the Hessian matrix of f.

More generally a function $f:\mathbb{R}^n \to \mathbb{R}$ (not just for convex quadratic functions) where the Hessian matrix exists is ill-conditionned if $D^2 f(x)$ is ill conditionned.

Level sets of $f(x)=\frac{1}{2}(x-x^*)^T A\,(x-x^*)$



GRADIENT DIRECTION VERSUS NEWTON DIRECTION

Gradient direction : $Df(x)$

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

HOME EXERCICE : We go back to the convex quadratic case

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

1) Plot level set of f

2) Plot the gradient direction

3) Compute the Newton direction, plot Newton direction.