

# Cours 10 : Optimisation non linéaire multidimensionnelle sans contraintes

Christophe Gonzales

LIP6 – Université Paris 6, France

## Optimisation d'une fonction différentiable à $n$ variables

- 1 méthode du gradient
- 2 méthode du gradient conjugué (la semaine prochaine)

# Fonction différentiable

$f : C \mapsto \mathbb{R}$  : fonction définie dans un convexe ouvert  $C$  de  $\mathbb{R}^n$  :

$$f : x = \begin{bmatrix} x_1 \\ \dots \\ x_j \\ \dots \\ x_n \end{bmatrix} \mapsto f(x) = f(x_1, \dots, x_j, \dots, x_n)$$

*fonction différentiable*

$f$  est **différentiable** en  $x \in C$  si  $\exists$  dérivées partielles  $\frac{\partial f(x)}{\partial x_j}$ ,  
 $j = 1, \dots, n$  et admet pour approximation du 1<sup>er</sup> ordre la forme  
linéaire qu'elles définissent :

$$f(x+h) = [f(x_1+h_1, \dots, x_n+h_n)] = f(x) + \sum_{j=1}^n \frac{\partial f(x)}{\partial x_j} \cdot h_j + o(\|h\|)$$

## Définition du gradient

$\vec{\nabla}f(x)$  : gradient de  $f$  en  $x$  = le vecteur  $\vec{\nabla}f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \dots \\ \frac{\partial f(x)}{\partial x_j} \\ \dots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$

$$\implies f(x+h) = f(x) + \vec{\nabla}f(x)^T \cdot h + o(\|h\|)$$

$\implies$  La variation de  $f$  est du 2<sup>ème</sup> ordre lorsque  $\vec{\nabla}f(x)^T \cdot h = 0$

$\vec{\nabla}f(x) \neq 0 \implies \left\{ y : f(y) = f(x) + \vec{\nabla}f(x)^T \cdot [y - x] \right\} =$  l'hyperplan tangent en  $x$  à l'hypersurface de niveau  $\{z : f(z) = f(x)\}$

normale de l'hyperplan =  $\vec{\nabla}f(x)$

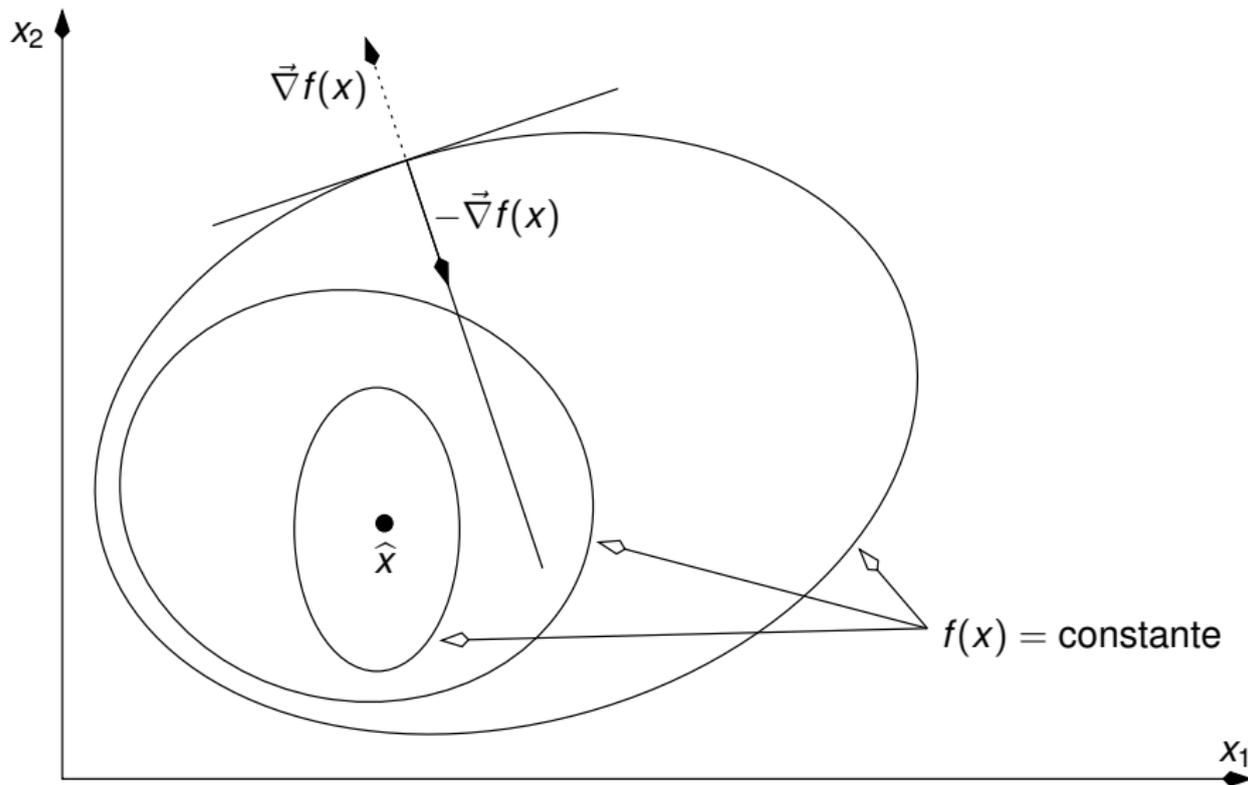
normale de l'hyperplan =  $\vec{\nabla} f(x)$

$$\begin{aligned} f(x + \lambda \vec{\nabla} f(x)) &= f(x) + \lambda \vec{\nabla} f(x)^T \cdot \vec{\nabla} f(x) \\ &= f(x) + \lambda \left\| \vec{\nabla} f(x) \right\|^2 \\ &> f(x) \text{ pour } \lambda > 0 \end{aligned}$$

$\implies$  normale dirigée du côté des points où  $f$  prend des valeurs plus élevées qu'en  $x$

on cherche min  $f \implies \delta = -\vec{\nabla} f(x)$  = direction intéressante

$$\delta = -\vec{\nabla} f(x) = \text{direction de l'anti-gradient}$$



# Minima et points stationnaires

- minimum (global ou absolu) de  $f$  en  $\hat{x} : x \in C \implies f(x) \geq f(\hat{x})$
- minimum local de  $f$  en  $\hat{x} : \exists$  un voisinage  $V$  de  $\hat{x}$  tel que  $x \in V \implies f(x) \geq f(\hat{x})$
- voisinage  $V$  de  $\hat{x}$  dans  $C =$  un sous-ensemble de  $C$  contenant une boule ouverte  $\{y : \|y - x\| < \epsilon\}$
- $\hat{x} =$  point stationnaire de  $f$  si  $\vec{\nabla}f(\hat{x}) = 0$

## Lemme

- $f$  différentiable
- minimum local de  $f$  en  $\hat{x} \implies \hat{x} =$  point stationnaire
- de plus, si  $f$  convexe :  
 $\hat{x} =$  point stationnaire  $\implies \hat{x} =$  minimum local

La méthode du gradient s'appuie sur :

## *Proposition*

La direction de l'anti-gradient est la direction de plus grande pente :

$$\max_{\{h: \|h\| = \|\vec{\nabla} f(x)\|\}} \lim_{\lambda \downarrow 0} \frac{f(x) - f(x + \lambda h)}{\lambda \|h\|}$$

est atteint pour  $h = -\vec{\nabla} f(x)$ .

Démonstration de la proposition :

$$f(x) - f(x + \lambda h) = -\lambda \vec{\nabla} f(x)^T \cdot h + o(\lambda \|h\|) \implies \text{pour } \|h\| = \|\vec{\nabla} f(x)\| :$$

$$\frac{f(x) - f(x + \lambda h)}{\lambda \|h\|} = \frac{-\lambda \vec{\nabla} f(x)^T \cdot h + o(\lambda \|\vec{\nabla} f(x)\|)}{\lambda \|\vec{\nabla} f(x)\|}$$

$$= \frac{-\vec{\nabla} f(x)^T \cdot h}{\|\vec{\nabla} f(x)\|} + \frac{o(\lambda)}{\lambda}$$

$$\implies \lim_{\lambda \downarrow 0} \frac{f(x) - f(x + \lambda h)}{\lambda \|h\|} = \frac{-\vec{\nabla} f(x)^T \cdot h}{\|\vec{\nabla} f(x)\|}$$

où  $-\vec{\nabla} f(x) \cdot h$  est maximum (sous  $\|h\| = \|\vec{\nabla} f(x)\|$ ) pour  $h = -\vec{\nabla} f(x)$

# Pas de la méthode du gradient

variations de  $f$  lorsque l'on part de  $x$  dans la direction de l'anti-gradient = celles de la fonction d'une seule variable  $\lambda \geq 0$  :

$$\varphi : \lambda \mapsto \varphi(\lambda) = f(x - \lambda \vec{\nabla} f(x))$$

$$\text{Or } \varphi'(\lambda) = - \sum_{j=1}^n \frac{\partial f}{\partial x_j}(x - \lambda \vec{\nabla} f(x)) \cdot \vec{\nabla} f_j(x) = - \vec{\nabla} f(x - \lambda \vec{\nabla} f(x)) \cdot \vec{\nabla} f(x)$$

$$\text{en particulier : } \varphi'(0) = - \vec{\nabla} f(x)^T \cdot \vec{\nabla} f(x) = - \left\| \vec{\nabla} f(x) \right\|^2 < 0$$

$\implies \varphi$  strictement décroissante au voisinage de  $\lambda = 0$

tant que  $\varphi'(\lambda) > 0$ , on continue à se déplacer et on s'arrête en :  $\tilde{x} = x - \tilde{\lambda} \vec{\nabla} f(x)$ , où  $\tilde{\lambda}$  est la plus petite valeur de  $\lambda$  solution de  $\varphi'(\lambda) = 0$  (si elle existe)

*Méthode du gradient — Cauchy (1847), Curry (1944)*

- partir d'un point initial  $x^0$
- on répète le pas  $x \rightarrow \tilde{x}$  précédent :  $x^k \rightarrow x^{k+1} = \widetilde{x^k}$
- critères d'arrêts possibles :

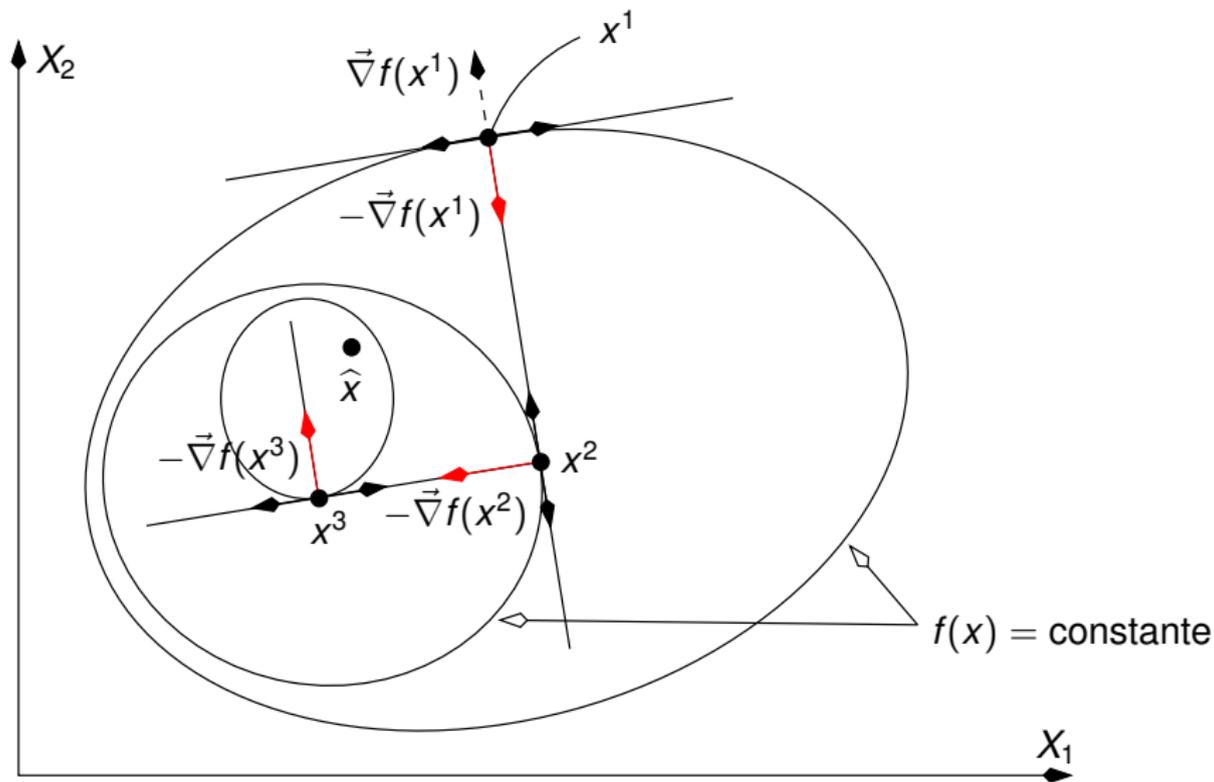
1  $\max_{i=1}^n \left| \frac{\partial f}{\partial x_i}(x^k) \right| < \epsilon$

2  $\left\| \vec{\nabla} f(x^k) \right\| < \epsilon$

3  $|f(x^{k+1}) - f(x^k)| < \epsilon$

Les 3 critères d'arrêt indiquent que  $f$  est proche d'être stationnaire

# Méthode du gradient (2/2)



$$\varphi'(\tilde{\lambda}) = -\vec{\nabla}f(\tilde{x})^T \cdot \vec{\nabla}f(x) = 0$$

⇒ les gradients en  $x$  et  $\tilde{x}$  sont orthogonaux

⇒ à chaque pas on prend une direction orthogonale à la direction précédente

cheminement de la méthode du gradient : «en zigzag»

## Point faible de la méthode du gradient (2/2)

⇒ pour éviter les zigzags et accélérer la convergence, on peut avoir recours à l'un des procédés suivants :

- **diminuer le pas** : ne pas aller jusqu'en  $\tilde{x}$ .

*Polyak (66)* : effectuer des pas prédéterminés en imposant une suite  $(\lambda^k)$  telle que  $\lambda^k \downarrow 0$  et  $\sum_k \lambda^k = +\infty$   
⇒  $(x^k)$  tend vers  $\hat{x}$

- **utiliser d'autres directions que l'anti-gradient** :

*Forsythe (1968), Luenberger (1973)* :  
toutes les  $m$  itérations, au lieu de partir dans la direction de l'anti-gradient en  $x^k$ , «couper» en partant dans la direction  $\delta = x^k - x^{k-m}$

- **utiliser des directions «conjuguées»**