Cours 10 : Optimisation non linéaire multidimensionnelle sans contraintes

Christophe Gonzales

LIP6 – Université Paris 6, France

Plan du cours

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 $\mathbb{R}^n$:

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

fonction différentiable

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

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

Gradient (1/3)

Définition du gradient

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

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

$\Longrightarrow$ La variation de f est du 2$^{\text{ème}}$ ordre lorsque $\vec{\nabla} f(x)^T.h = 0$

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

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

Gradient (2/3)

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

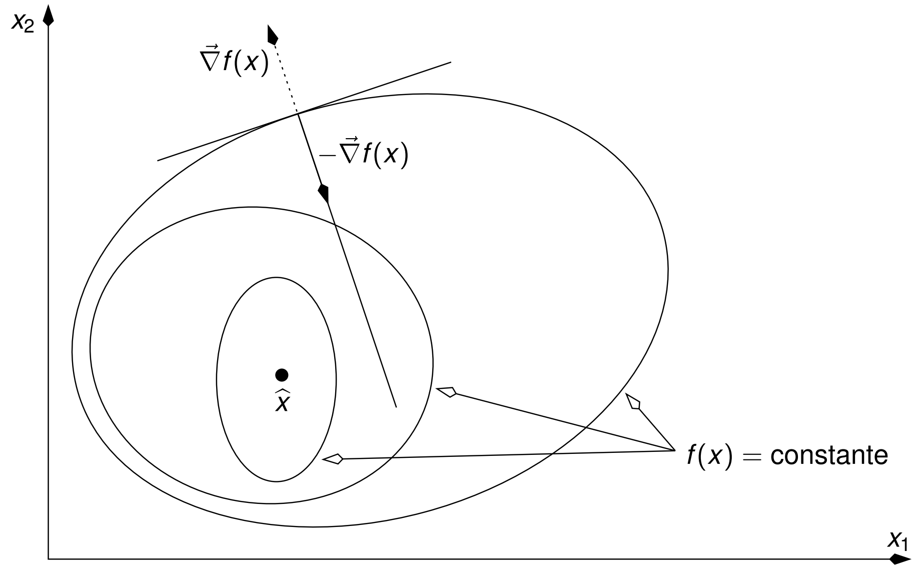
$$\begin{aligned} f(x + \lambda \vec{\nabla} f(x)) &= f(x) + \lambda \vec{\nabla} f(x)^T . \vec{\nabla} f(x) \\ &= f(x) + \lambda \left\| \vec{\nabla} f(x) \right\|^2 \\ &> f(x) \text{ pour } \lambda > 0 \end{aligned}$$

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

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

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

Gradient (3/3)



Minima et points stationnaires

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

Lemme

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

Direction de l'anti-gradient (1/2)

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

Direction de l'anti-gradient (2/2)

Démonstration de la proposition :

$f(x) - f(x+\lambda h) = -\lambda \vec{\nabla} f(x)^T.h + o(\lambda \|h\|) \Longrightarrow$ pour $\|h\| = \left\|\vec{\nabla} f(x)\right\|$:

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

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

$$\Longrightarrow \lim_{\lambda \downarrow 0} \frac{f(x) - f(x+\lambda h)}{\lambda \|h\|} = \frac{-\vec{\nabla} f(x)^T.h}{\left\|\vec{\nabla} f(x)\right\|}$$

où $-\vec{\nabla} f(x).h$ est maximum $\left(\text{sous } \|h\| = \left\|\vec{\nabla} f(x)\right\|\right)$ 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 $\lambda \geq 0$:

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

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

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

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

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

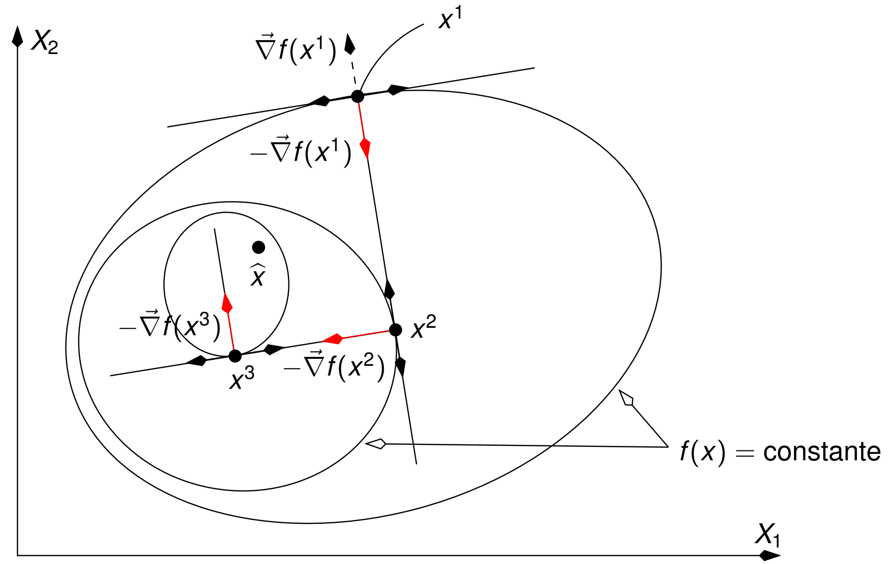
Méthode du gradient (1/2)

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

- partir d'un point initial x^0
- on répète le pas $x \to \widetilde{x}$ précédent : $x^k \to x^{k+1} = \widetilde{(x^k)}$
- 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. $\left|f(x^{k+1}) - f(x^k)\right| < \epsilon$

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

Méthode du gradient (2/2)



Point faible de la méthode du gradient (1/2)

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

$\Longrightarrow$ les gradients en x et $\widetilde{x}$ sont orthogonaux

$\Longrightarrow$ 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)

$\Longrightarrow$ 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 $\widetilde{x}$.

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

- 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»