Algorithmes II

(P. Carpentier)

20 mars 2020

Plan du cours

Rappel du problème

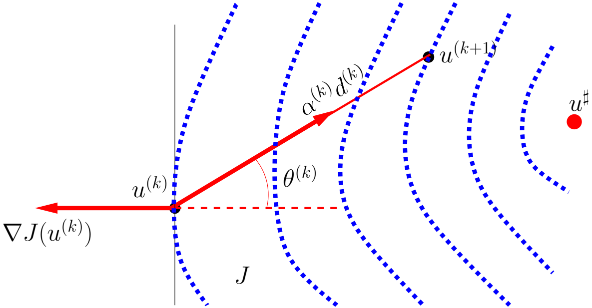
On veut résoudre le problème d'optimisation :

$$\min_{u \in \mathbb{U}} J(u),$$

en utilisant une méthode à direction de descente :

$$u^{(k+1)} = u^{(k)} + \alpha^{(k)} d^{(k)},$$

la direction $d^{(k)}$ étant donc telle que : $\langle \nabla J(u^{(k)}) \, , d^{(k)} \rangle < 0$.



Condition de Zoutendijk (CZ).

On dit que la suite $\{u^{(k)}\}_{k\in\mathbb{N}}$ générée par une méthode à direction de descente satisfait la condition de Zoutendijk si elle vérifie :

$$\exists C > 0,\ \forall k \in \mathbb{N},\ J(u^{(k+1)}) - J(u^{(k)}) \leq -C\left\|\nabla J(u^{(k)})\right\|^2 \cos^2(\theta^{(k)}),$$

$\theta^{(k)}$ étant défini par : $\cos(\theta^{(k)}) = -\dfrac{\left\langle \nabla J(u^{(k)})\,, d^{(k)}\right\rangle}{\left\|\nabla J(u^{(k)})\right\| \left\|d^{(k)}\right\|} > 0.$

Propriété. Si $\{u^{(k)}\}_{k\in\mathbb{N}}$ satisfait (**CZ**) et si J est minorée, alors :

$$\sum_{k\in\mathbb{N}} \left\|\nabla J(u^{(k)})\right\|^2 \cos^2(\theta^{(k)}) < +\infty.$$

Corollaire. Si $\exists \epsilon > 0,\ \forall k \in \mathbb{N},\ \left|\cos(\theta^{(k)})\right| > \epsilon$, on en déduit que :

$$\lim_{k\to+\infty} \left\|\nabla J(u^{(k)})\right\| = 0.$$

Utilisation de la condition de Zoutendijk

Théorème. Soit $J : \mathbb{U} \longrightarrow \mathbb{R}$ continue différentiable à gradient lipschitzien de constante L, et soit $\{u^{(k)}\}_{k\in\mathbb{N}}$ la suite obtenue par une méthode à direction de descente avec recherche linéaire de type Wolfe. Alors, la suite $\{u^{(k)}\}_{k\in\mathbb{N}}$ satisfait la condition (**CZ**).

Preuve. On a par définition : $u^{(k+1)} = u^{(k)} + \alpha^{(k)} d^{(k)}$.

- On utilise la 2-ème condition de Wolfe :

$$
\begin{aligned}
& \left\langle \nabla J(u^{(k+1)}) \,, d^{(k)} \right\rangle \geq \omega_2 \left\langle \nabla J(u^{(k)}) \,, d^{(k)} \right\rangle \\
\Rightarrow \quad & (1-\omega_2) \left| \left\langle \nabla J(u^{(k)}) \,, d^{(k)} \right\rangle \right| \leq \left\langle \nabla J(u^{(k+1)}) - \nabla J(u^{(k)}) \,, d^{(k)} \right\rangle \\
\Rightarrow \quad & (1-\omega_2) \left\| \nabla J(u^{(k)}) \right\| \left\| d^{(k)} \right\| \cos(\theta^{(k)}) \leq L \alpha^{(k)} \left\| d^{(k)} \right\|^2 .
\end{aligned}
$$

- On utilise alors la 1-ère condition de Wolfe :

$$
\begin{aligned}
J(u^{(k+1)}) - J(u^{(k)}) & \leq -\omega_1 \alpha^{(k)} \left\| \nabla J(u^{(k)}) \right\| \left\| d^{(k)} \right\| \cos(\theta^{(k)}) \\
& \leq -\frac{\omega_1 (1-\omega_2)}{L} \left\| \nabla J(u^{(k)}) \right\|^2 \cos^2(\theta^{(k)}).
\end{aligned}
$$

Comme $0 < \omega_1 < \omega_2 < 1$, on en déduit la condition (**CZ**). $\square$

Algorithme du gradient

$$\min_{u\in\mathbb{U}} J(u) \quad \rightsquigarrow \quad \begin{aligned} u^{(k+1)} &= u^{(k)} + \alpha^{(k)} d^{(k)}, \\ d^{(k)} &= -\nabla J(u^{(k)}). \end{aligned}$$

Théorème. On suppose J différentiable, à gradient lipschitzien de constante L et fortement convexe de rapport a. Alors, l'algorithme de gradient à pas fixe converge q-linéairement vers la solution $u^\sharp$ (unique) du problème, pour toute valeur du pas $\alpha \in]0, \frac{2a}{L^2}[$.

Preuve. Elle consiste à majorer la distance à l'optimum $\|u^{(k)} - u^\sharp\|$.

$$\left\|u^{(k+1)} - u^\sharp\right\|^2 = \left\|u^{(k)} - u^\sharp\right\|^2 - 2\alpha\left\langle\nabla J(u^{(k)})\,, u^{(k)} - u^\sharp\right\rangle + \alpha^2\left\|\nabla J(u^{(k)})\right\|^2.$$

On considère les deux derniers termes. D'après les hypothèses, on a :

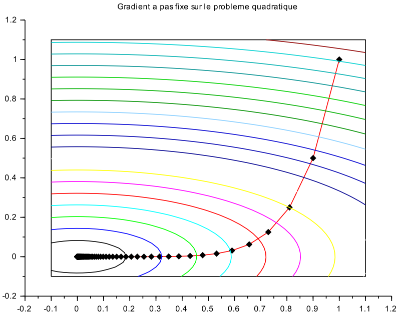
- $\left\langle\nabla J(u^{(k)}) - \nabla J(u^\sharp)\,, u^{(k)} - u^\sharp\right\rangle \geq a\left\|u^{(k)} - u^\sharp\right\|^2$ (forte convexité).
- $\left\|\nabla J(u^{(k)}) - \nabla J(u^\sharp)\right\| \leq L\left\|u^{(k)} - u^\sharp\right\|$ (gradient Lipschitz).

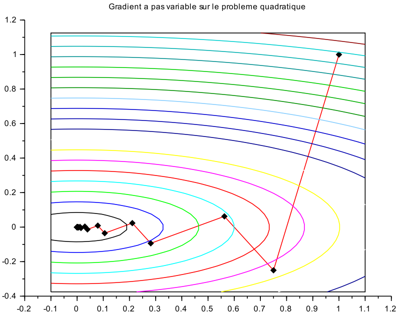
d'où : $\left\|u^{(k+1)} - u^\sharp\right\|^2 \leq \left(1 - 2\alpha a + \alpha^2 L^2\right)\left\|u^{(k)} - u^\sharp\right\|^2$ (car $\nabla J(u^\sharp) = 0$).

Enfin, $1 - 2\alpha a + \alpha^2 L^2 < 1 \;\Rightarrow\; \alpha \in]0, \frac{2a}{L^2}[$ (valeur minimale en $\alpha^\sharp = \frac{a}{L^2}$). □

Exemples pour l’algorithme du gradient

Fonction quadratique : $J(u_1, u_2) = \frac{1}{2}\left(u_1^2 + 5u_2^2\right)$.





Algorithmes de gradient à pas fixe et à pas variable.

$J(u) = \frac{1}{2}u^\top A\, u - b^\top u$, avec $u \in \mathbb{R}^n$ et A symétrique.

Définition. 2 directions d_i et d_j sont conjuguées si $d_i^\top A\, d_j = 0$.

Propriété. Soit A définie positive. On se donne $m\,(\leq n)$ directions $(d_1, \ldots, d_m)$ conjuguées 2 à 2. Alors, elles forment une famille libre.

On se donne un point $u^{(0)} \in \mathbb{U}$ et k directions $(d^{(0)}, \ldots, d^{(k-1)})$.
On note $E^{(k)} = \mathrm{span}(d^{(0)}, \ldots, d^{(k-1)})$.

Théorème. Soit $u^{(k)}$ réalisant le minimum de J sur $u^{(0)} + E^{(k)}$. Soit $d^{(k)} \neq 0$, $\alpha^{(k)} \geq 0$ et soit $u^{(k+1)} = u^{(k)} + \alpha^{(k)} d^{(k)}$. Alors, $u^{(k+1)}$ réalise le minimum de J sur $u_0 + E^{(k+1)}$ si et seulement si :

1. $d^{(k)}$ est conjuguée / $d^{(0)}, \ldots, d^{(k-1)}$ et $\alpha^{(k)} = -\frac{\langle d^{(k)}, \nabla J(^{(k)})\rangle}{\|d^{(k)}\|_A^2}$ dans le cas où $d^{(k)}$ vérifie : $\langle d^{(k)}, \nabla J(u^{(k)})\rangle \neq 0$,
2. $\alpha^{(k)} = 0$ dans le cas contraire.

Algorithme du gradient conjugué.

- Soit $u^{(0)} \in \mathbb{R}^n$ et soit $d^{(0)} = -\nabla J(u^{(0)})$.

  $\rightsquigarrow$ $u^{(1)} = u^{(0)} + \alpha^{(0)} d^{(0)}$, avec $\alpha^{(0)}$ pas optimal de Cauchy.

- Soit $d^{(k)}$ de la forme : $d^{(k)} = -\nabla J(u^{(k)}) + \beta^{(k-1)} d^{(k-1)}$.

  $d^{(k)}$ et $d^{(k-1)}$ conjuguées $\Rightarrow \beta^{(k-1)} = \dfrac{\left\langle d^{(k-1)}, \nabla J(u^{(k)}) \right\rangle_A}{\left\| d^{(k-1)}) \right\|_A^2}$.

  $\rightsquigarrow$ $u^{(k+1)} = u^{(k)} + \alpha^{(k)} d^{(k)}$, avec $\alpha^{(k)}$ pas optimal.

Propriété de l'algorithme. $d^{(k)}$ conjuguée par rapport à $d^{(k-1)}$ implique $d^{(k)}$ conjuguée par rapport à $d^{(0)}, \ldots, d^{(k-1)}$.

Par le théorème précédent, on a engendré en au plus n itérations un point de la forme :

$$u^{(n)} = u^{(0)} + \sum_{k=0}^{n-1} \alpha^{(k)} d^{(k)},$$

qui minimise le critère J dans l'espace $\mathbb{U} = \mathbb{R}^n$ tout entier !

Gradient conjugué dans le cas général

$$\min_{u \in \mathbb{U}} J(u) \quad \rightsquigarrow \quad \begin{aligned} u^{(k+1)} &= u^{(k)} + \alpha^{(k)} d^{(k)}, \\ d^{(k)} &= -\nabla J(u^{(k)}) + \beta^{(k-1)} d^{(k-1)}. \end{aligned}$$

- Variante de Fletcher-Reeves :

$$\beta^{(k-1)} = \frac{\left\| \nabla J(u^{(k)}) \right\|^2}{\left\| \nabla J(u^{(k-1)}) \right\|^2}.$$

- Variante de Polak-Ribière :

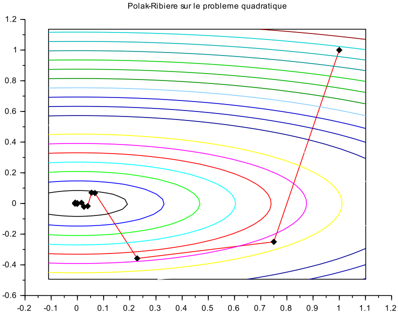
$$\beta^{(k-1)} = \frac{\left\langle \nabla J(u^{(k)}) - \nabla J(u^{(k-1)}) \, , \nabla J(u^{(k)}) \right\rangle}{\left\| \nabla J(u^{(k-1)}) \right\|^2}.$$

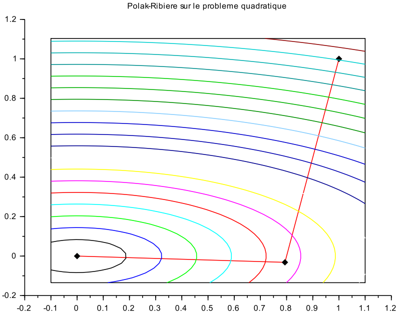
Avantage : $\nabla J(u^{(k)}) \approx \nabla J(u^{(k-1)}) \implies \beta^{(k-1)} \approx 0$
(*réinitialisation automatique en cas de déconjugaison*).

Théorème. On suppose que J est fortement convexe, différentiable à gradient lipschitzien. Alors, l'algorithme de Polak-Ribière avec recherche linéaire exacte converge "*q*-superlinéairement" vers $u^\sharp$.

Exemples pour l'algorithme du gradient conjugué

Fonction quadratique : $J(u_1, u_2) = \frac{1}{2}\left(u_1^2 + 5u_2^2\right)$.





Gradient conjugué (Polak-Ribière) à pas de Wolfe et à pas optimal.

Algorithme de Newton

$$\min_{u \in \mathbb{U}} J(u) \quad \rightsquigarrow \quad u^{(k+1)} = u^{(k)} + \alpha^{(k)} d^{(k)},$$

$$d^{(k)} = -\left[\nabla^2 J(u^{(k)})\right]^{-1} \nabla J(u^{(k)}).$$

L'**approximation au second ordre** de J au point $u^{(k)}$ est :

$$J(u^{(k)}) + \nabla J(u^{(k)})^\top (u - u^{(k)}) + \frac{1}{2}(u - u^{(k)})^\top \left[\nabla^2 J(u^{(k)})\right](u - u^{(k)}),$$

dont la minimisation a pour solution : $u^{(k+1)} = u^{(k)} + 1\, d^{(k)}$.

- Le pas naturel est $\alpha^{(k)} = 1$; on peut effectuer une recherche linéaire de type Wolfe, ou encore employer une technique de globalisation (*exemple* : $\nabla^2 J(u^{(k)}) \rightsquigarrow \nabla^2 J(u^{(k)}) + \mu \mathrm{I}_n,\ \mu > 0$).
- Noter que $d^{(k)}$ n'est pas forcément une direction de descente !

Théorème. On suppose que J est fortement convexe, différentiable et que son hessien est lipschitzien. Alors, l'algorithme de Newton avec pas unitaire converge q-quadratiquement au **voisinage** de $u^\sharp$.

Algorithmes de quasi-Newton

$$\min_{u \in \mathbb{U}} J(u) \quad \rightsquigarrow \quad u^{(k+1)} = u^{(k)} + \alpha^{(k)} d^{(k)},$$

$$d^{(k)} = -W^{(k)} \nabla J(u^{(k)}),$$

avec $W^{(k)}$: approximation de l'inverse du hessien $\left[\nabla^2 J(u^{(k)})\right]^{-1}$.

1) On impose à $W^{(k+1)}$ de vérifier l'équation de la sécante :

$$W^{(k+1)}\left(\nabla J(u^{(k+1)}) - \nabla J(u^{(k)})\right) = u^{(k+1)} - u^{(k)},$$

car au 1-er ordre, $\nabla J(u^{(k+1)}) = \nabla J(u^{(k)}) + \left[\nabla^2 J(u^{(k)})\right]\left(u^{(k+1)} - u^{(k)}\right)$.

2) On impose à $W^{(k+1)}$ d'être symétrique.

D'où n contraintes pour $\frac{n(n+1)}{2}$ coefficients : beaucoup de choix !

Nouvelles notations :

$$\delta_u^{(k)} = u^{(k+1)} - u^{(k)} \quad , \quad \delta_G^{(k)} = \nabla J(u^{(k+1)}) - \nabla J(u^{(k)}).$$

quasi-Newton : algorithme $\mathrm{SR1}$

Première idée : effectuer une correction de rang 1 sur $W^{(k)}$:

$$W^{(k+1)} = W^{(k)} + a^{(k)} v^{(k)} {v^{(k)}}^\top, \quad \text{avec} \quad v^{(k)} \in \mathbb{R}^n,$$

où $a^{(k)} \in \mathbb{R}$ est un simple coefficient de normalisation.

- Sécante :
  $\delta_u^{(k)} = W^{(k+1)} \delta_G^{(k)} = W^{(k)} \delta_G^{(k)} + a^{(k)} v^{(k)} \big({v^{(k)}}^\top \delta_G^{(k)} \big).$
- Choisir $a^{(k)} {v^{(k)}}^\top \delta_G^{(k)} = 1$ implique $v^{(k)} = \delta_u^{(k)} - W^{(k)} \delta_G^{(k)}$.

On obtient la formule $\mathrm{SR1}$ (Symétrique de Rang 1) :

$$W^{(k+1)} = W^{(k)} + \frac{\big(\delta_u^{(k)} - W^{(k)} \delta_G^{(k)}\big) \big(\delta_u^{(k)} - W^{(k)} \delta_G^{(k)}\big)^\top}{\big(\delta_u^{(k)} - W^{(k)} \delta_G^{(k)}\big)^\top \delta_G^{(k)}}.$$

Problèmes pratiques avec SR1 :

- le dénominateur peut s'annuler,
- $W^{(k)}$ peut ne pas être définie positive.

Autre idée : minimiser l'écart entre $W^{(k+1)}$ et $W^{(k)}$ sous :

- $W^{(k+1)}\delta_G^{(k)} = \delta_u^{(k)}$ (équation de la sécante),
- $W^{(k+1)}$ symétrique.

Utilisant une distance matricielle adaptée (Frobenius pondérée), et après des calculs (longs...), on obtient la formule BFGS :

$$W^{(k+1)} = \left(\mathrm{I} - \frac{\delta_u^{(k)}{\delta_G^{(k)}}^\top}{{\delta_G^{(k)}}^\top \delta_u^{(k)}}\right) W^{(k)} \left(\mathrm{I} - \frac{\delta_G^{(k)}{\delta_u^{(k)}}^\top}{{\delta_G^{(k)}}^\top \delta_u^{(k)}}\right) + \frac{\delta_u^{(k)}{\delta_u^{(k)}}^\top}{{\delta_G^{(k)}}^\top \delta_u^{(k)}}.$$

On a de plus : ${\delta_G^{(k)}}^\top \delta_u^{(k)} > 0$ implique $W^{(k+1)}$ définie positive.

Remarque. Appliquant le même principe pour une approximation $M^{(k)}$ du hessien, on obtient après inversion la formule DFP (*moins populaire...*) :

$$W^{(k+1)} = W^{(k)} + \frac{\delta_u^{(k)}{\delta_u^{(k)}}^\top}{{\delta_G^{(k)}}^\top \delta_u^{(k)}} - \frac{W^{(k)}\delta_G^{(k)}{\delta_G^{(k)}}^\top W^{(k)}}{{\delta_G^{(k)}}^\top W^{(k)}\delta_G^{(k)}}.$$

Remarque. Si l'on utilise la règle de Wolfe :

$$\left\langle \nabla J(u^{(k)} + \alpha^{(k)} d^{(k)}) \, , d^{(k)} \right\rangle \geq \omega_2 \left\langle \nabla J(u^{(k)}) \, , d^{(k)} \right\rangle,$$

et retranchant de part et d'autre $\left\langle \nabla J(u^{(k)}) \, , d^{(k)} \right\rangle$ on obtient :

$$\delta_G^{(k)\top} d^{(k)} \geq (\omega_2 - 1) \left\langle \nabla J(u^{(k)}) \, , d^{(k)} \right\rangle > 0.$$

Comme $\delta_u^{(k)} = \alpha^{(k)} d^{(k)}$, on en déduit : $\delta_G^{(k)\top} \delta_u^{(k)} > 0$.

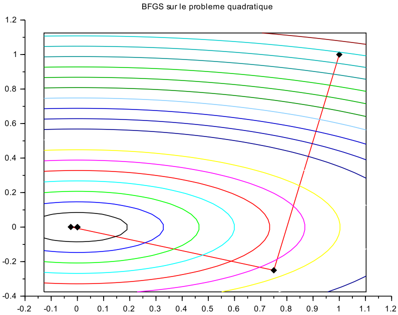
$\rightsquigarrow$ Il faut toujours utiliser la règle de Wolfe avec BFGS !

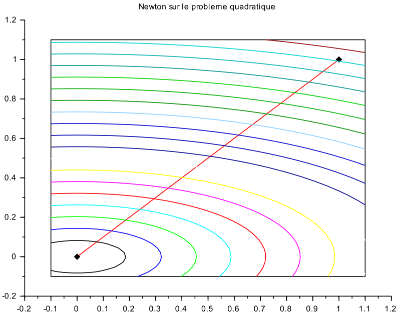
Théorème. On suppose J convexe, différentiable à gradient lipschitzien. On suppose que l'algorithme BFGS est initialisé avec une matrice $W^{(0)}$ symétrique définie positive et que l'on utilise la règle de Wolfe. Alors, $\liminf_{k \to +\infty} \left\| \nabla J(u^{(k)}) \right\| = 0$.

Avec un peu plus d'hypothèses, on prouve la convergence q-superlinéaire.

Un des plus beaux résultats de convergence en optimisation...

Fonction quadratique : $J(u_1, u_2) = \frac{1}{2}\left(u_1^2 + 5u_2^2\right)$.

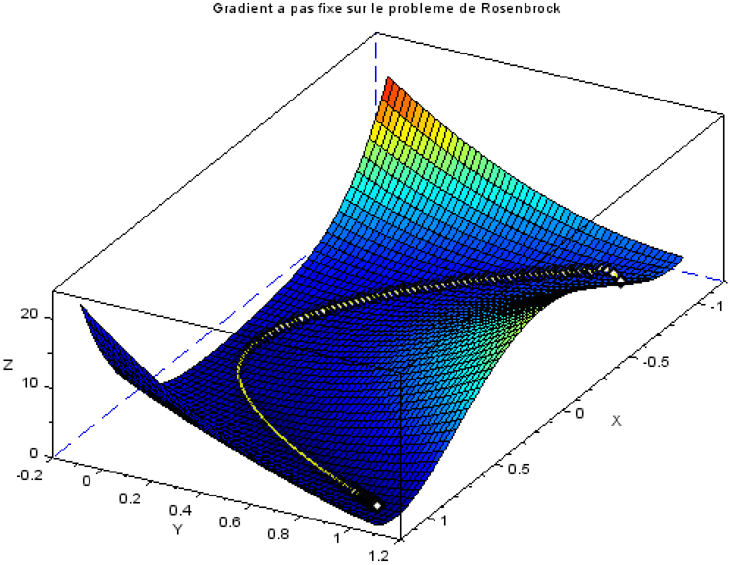




Algorithme BFGS et méthode de Newton.

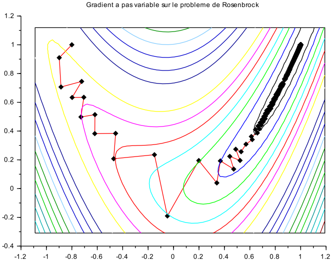
Fonction de Rosenbrock (fonction « banane »)

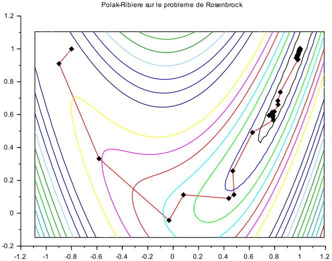
Fonction de Rosenbrock : $J(u_1, u_2) = \left(u_1 - 1\right)^2 + 10\left(u_1^2 - u_2\right)^2$.

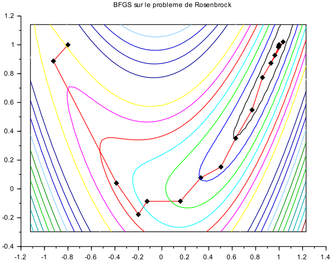


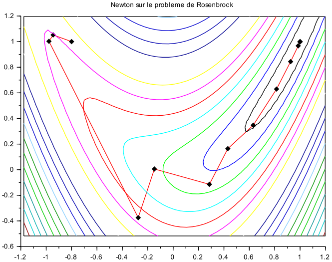
Comparaison sur la fonction de Rosenbrock

Fonction de Rosenbrock : $J(u_1, u_2) = \big(u_1 - 1\big)^2 + 10\big(u_1^2 - u_2\big)^2$.









Quelques références bibliographiques sur les algorithmes

J.-C. Culioli.
Introduction à l'Optimisation.
Ellipses, 1994.

F. Bonnans, J.-C. Gilbert, C. Lemaréchal, C. Sagastizabal.
Numerical Optimization. Theoretical and Practical Aspects.
Springer, 2006.

J. Nocedal, S.J. Wright.
Numerical Optimization.
Springer, 2006.

A. Ruszczynski.
Nonlinear Optimization.
Princeton University Press, 2006.

A.R. Conn, N. Gould, P. Toint.
Trust-Region Methods.
SIAM, 2000.

J.-C. Gilbert.
Éléments d'optimisation différentiable *(notes de cours)*.
https://who.rocq.inria.fr/Jean-Charles.Gilbert/ensta/optim.html