# Optimisation sans contraintes

Joon Kwon

Master 2 — MathSV

jeudi 22 septembre 2022

Dans ce chapitre, $f : \mathbb{R}^d \to \mathbb{R}$ et on considère les problèmes de minimisation sans contrainte.

$$\min_{x \in \mathbb{R}^d} f(x)$$

## Exemple (Régression logistique avec régularisation Ridge)

Soit $d, n \geqslant 1$ entiers, $\lambda > 0$, $a_1, \ldots, a_n \in \mathbb{R}^d$, $b_1, \ldots, b_n \in \mathbb{R}$.

$$\min_{x \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-b_i (a_i^\top x)}\right) + \lambda \|x\|_2^2 \right\}.$$

# Bibliographie


- Stephen Boyd & Lieven Vandenberghe. Convex optimization. Cambridge University Press, 2004.
- Jorge Nocedal & Stephen Wright. Numerical optimization. Springer Science & Business Media, 2006.
- Joseph-Frédéric Bonnans, et al. Numerical optimization : theoretical and practical aspects. Springer Science & Business Media, 2006.
- Wenyu Sun & Ya-Xiang Yuan. Optimization theory and methods : nonlinear programming. Springer Science & Business Media, 2006.
- Kenneth Lange. Optimization. Springer Science & Business Media, 2013.

Conditions d'optimalité

Descente de gradient

Méthode de Newton

# Conditions d'optimalité

# Conditions d'optimalité locale

Dans tout le chapitre, $f$ est une application $f : \mathbb{R}^d \to \mathbb{R}$.

## Théorème (Conditions nécessaires d'optimalité locale)

*Soit $x^*$ un minimiseur local de $f$.*

- *Si $f$ est différentiable en $x^*$, alors $x^*$ est un point critique (i.e. $\nabla f(x^*) = 0$).*
- *Si de plus, $f$ est deux fois différentiable en $x^*$, alors $\nabla^2 f(x^*)$ est semi-définie positive.*

## Remarque

*La première condition n'est pas suffisante : les points d'inflexions ou les points-selles sont des points critiques, mais ne sont pas des minimiseurs.*

## Remarque (Résolution exacte)

*Lorsque $\nabla f$ a une expression suffisamment simple, on peut :*

- *sélectionner les points critiques par résolution exacte de l'équation $\nabla f(x) = 0$,*
- *éventuellement sélectionner les minima locaux,*
- *en déduire les minimiseurs globaux.*

# Conditions suffisantes d'optimalité

## Théorème (Condition suffisante d'optimalité locale)

*Soit $x^* \in \mathbb{R}^d$. On suppose que :*

- *$f$ est deux fois différentiable en $x^*$,*
- *$\nabla f(x^*) = 0$,*
- *$\nabla^2 f(x^*)$ est symétrique définie positive.*

*Alors, $x^*$ est un minimum local de $f$.*

## Remarque

*Cette condition n'est pas nécéssaire.*

## Théorème (Condition suffisante d'optimalité globale)

*Soit $x^*$ un point critique de $f$.*

- *Si $f$ est convexe, alors $x^*$ est un minimiseur global de $f$.*
- *Si $f$ est strictement convexe, alors $x^*$ est l'unique minimiseur global de $f$.*

# Descente de gradient

# Remarques générales sur les algorithmes itératifs

- Lorsqu'une résolution exacte n'est pas possible, on a recours aux algorithmes itératifs pour rechercher des solutions approchées.
- Les algorithmes itératifs construisent une suite d'itérées $(x^{(t)})_{t \geqslant 1}$.
- On espère une convergence vers une solution.

En pratique, différents critères d'arrêt sont possibles.

- Nombre d'itérations $T \geqslant 1$ fixé à l'avance.
- Arrêt lorsque $\left\|\nabla f(x^{(t)})\right\| \leqslant \varepsilon$, pour $\varepsilon > 0$ fixé à l'avance.
- Arrêt lorsque $f(x^{(t)}) \geqslant f(x^{(t-1)}) - \varepsilon$, pour $\varepsilon > 0$ fixé à l'avance.

# Descente de gradient

## Definition (Descente de gradient)

Soit $x^{(1)} \in \mathbb{R}^d$ et $(\gamma^{(t)})_{t \geqslant 1}$ une suite strictement positive. On appelle descente de gradient associée à la fonction objectif $f$, au point initial $x^{(1)}$ et aux pas $(\gamma^{(t)})_{t \geqslant 1}$ la suite $(x^{(t)})_{t \geqslant 1}$ définie par :

$$x^{(t+1)} = x^{(t)} - \gamma^{(t)} \nabla f(x^{(t)}), \qquad t \geqslant 1.$$

## Remarque

- *Méthode de premier ordre i.e. utilise le gradient*
- *Chaque itération correspond à la résolution d'un problème simplifié*

# Garanties quantitatives

## Théorème

*Soit $f : \mathbb{R}^d \to \mathbb{R}$ différentiable et L-régulière, $x^*$ un minimiseur de $f$, et $x^{(1)} \in \mathbb{R}^d$ quelconque. Soit $(x^{(t)})_{t \geqslant 1}$ les itérées de la descente de gradient associée avec un pas $\gamma^{(t)} = 1/L$ $(t \geqslant 1)$, et $T \geqslant 1$.*

- *Alors,*

$$\min_{1 \leqslant t \leqslant T} \left\| \nabla f(x^{(t)}) \right\|_2^2 \leqslant \frac{2L(f(x^{(1)}) - f(x^*))}{T}.$$

- *Si de plus $f$ est convexe,*

$$f(x^{(T+1)}) - f(x_*) \leqslant \frac{2L \left\| x^{(1)} - x^* \right\|_2^2}{T}.$$

- *Soit $K > 0$. Si de plus que $f$ est $K$-fortement convexe,*

$$f(x^{(T+1)} - f(x^*)) \leqslant \frac{\beta}{2} \left(1 - \frac{K}{L}\right)^T \left\| x^{(1)} - x^* \right\|_2^2.$$

# Choix du pas par line-search

**Line-search exact**

$$\gamma^{(t)} = \underset{\gamma>0}{\arg\min} \left\{ f\left( x^{(t)} - \gamma \nabla f(x^{(t)}) \right) \right\}.$$

- Ce calcul de $\gamma^{(t)}$ correspond à une optimisation en dimension 1.
- Relativement coûteux (en calcul) et inutile.
- En pratique, il existe des méthodes moins coûteuses (voir TP) qui donnent d'assez bons pas $\gamma^{(t)}$ : règles d'Armijo, de Wolfe, etc.

# Discussion sur le conditionnement

Soit $A$ une matrice symétrique semi-définie positive de taille $d \times d$.

- Cas extrême. Si $\mathrm{Sp} = \{\lambda\}$, alors $A = \lambda I$. Et la descente de gradient peut minimiser $x \mapsto x^\top A x$ en une itération (avec line-search exact).
- Cas "bien conditionné". S'il y a peu d'écarts entre les valeurs propres de $A$, $x \mapsto x^\top A x$ est facile à minimiser par la descente de gradient.
- Cas "mal conditionné". S'il y a des écarts importants entre les valeurs propres de $A$, $x \mapsto x^\top A x$ est difficle à minimiser par la descente de gradient (i.e. lent).

Au voisinage de $x^*$, $f(x) \simeq \frac{1}{2}(x - x^*)^\top \nabla^2 f(x)(x - x^*)$.

## Conclusion

La performance de la descente de gradient dépend du conditionnement de $\nabla^2 f(x^*)$

# Méthode de Newton

# Méthode de Newton

## Definition (Méthode de Newton)

Soit $x^{(1)} \in \mathbb{R}^d$ et $(\gamma^{(t)})_{t \geqslant 1}$ une suite strictement positive. On appelle méthode de Newton associée à la fonction objectif $f$, au point initial $x^{(1)}$ et aux pas $(\gamma^{(t)})_{t \geqslant 1}$ la suite $(x^{(t)})_{t \geqslant 1}$ définie par :

$$x^{(t+1)} = x^{(t)} - \gamma^{(t)} \left( \nabla^2 f(x^{(t)}) \right)^{-1} \nabla f(x^{(t)}), \qquad t \geqslant 1.$$

- Méthode du second ordre.
- Bien définie si la hessienne est inversible.
- L'itération correspond à minimiser l'approximation d'ordre 2 de $f$ en $x^{(t)}$.
- Solution exacte en une seule itération si $f$ est quadratique.
- Convergence très rapide dans la région où $f$ est bien approximée par son développement d'ordre 2 en $x^*$.
- On peut choisir le pas $\gamma^{(t)}$ par des méthodes de line-search.
- Ne jamais calculer $(\nabla^2 f(x^{(t)}))^{-1}$ (sauf peut-être pour $d$ petit). Il suffit de résoudre $\nabla^2 f(x^{(t)})u = \nabla f(x^{(t)})$. Ce qui est possible dès lors qu'on sait calculer $\nabla^2 f(x^{(t)})u$ pour tout $u \in \mathbb{R}^d$.
- Magré tout, très coûteux pour $d \gg 1$.

# Méthodes quasi-Newton

$$x^{(t+1)} = x^{(t)} - \gamma^{(t)} H^{(t)} \nabla f(x^{(t)}), \qquad t \geqslant 1.$$

- Méthode de Newton où a remplacé $(\nabla^2 f(x^{(t)}))^{-1}$ par une approximation $H^{(t)}$.
- Typiquement, $H^{(t+1)}$ a une expression explicite en $H^{(t)}$, $x^{(t+1)}$, $x^{(t)}$ $\nabla f(x^{(t+1)})$ et $\nabla f(x^{(t)})$.
- Méthode d'ordre 1.
- Nombreuses formules différentes pour $H^{(t)}$ : BFGS, Broyden, DFP, SR1, etc.