

# Optimisation non linéaire et non convexe

Clément Royer

Certificat Chef de Projet IA - Université Paris Dauphine-PSL

11 octobre 2023



- 1 Optimisation lisse
- 2 Descente de gradient

- 1 Optimisation lisse
  - Calcul différentiel et optimisation
  - Solutions et conditions d'optimalité
  - Classes de problèmes remarquables
- 2 Descente de gradient

- 1 Optimisation lisse
  - Calcul différentiel et optimisation
  - Solutions et conditions d'optimalité
  - Classes de problèmes remarquables
- 2 Descente de gradient

## Problème

minimiser  $f(\mathbf{x})$ .  
 $\mathbf{x} \in \mathbb{R}^d$

## Hypothèses

- $f$  minorée par  $f^*$ ;
- $f$  douce/lisse  $\Rightarrow$  les dérivées de  $f$  peuvent être utilisées pour résoudre ce problème.

On considère une fonction **lisse** (ou douce, ou *smooth*)  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ .

On considère une fonction **lisse** (ou douce, ou *smooth*)  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ .

## Dérivée à l'ordre 1

Si  $f$  est continûment dérivable sur  $\mathbb{R}^d$ , on définit pour tout  $\mathbf{x} \in \mathbb{R}^d$  le **gradient de  $f$  en  $\mathbf{x}$**  par

$$\nabla f(\mathbf{x}) := \left[ \frac{\partial f}{\partial x_i}(\mathbf{x}) \right]_{1 \leq i \leq d} \in \mathbb{R}^d.$$

L'ensemble des fonctions continûment dérivables est noté  $\mathcal{C}^1$ . On parle de fonction de classe  $\mathcal{C}^1$ .

On considère une fonction **lisse** (ou douce, ou *smooth*)  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ .



On considère une fonction **lisse** (ou douce, ou *smooth*)  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ .

### Dérivée d'ordre 2

Si  $f$  est *deux fois* continûment dérivable sur  $\mathbb{R}^d$ , on définit pour tout  $\mathbf{x} \in \mathbb{R}^d$  la **matrice hessienne de  $f$  en  $\mathbf{x}$**  par

$$\nabla^2 f(\mathbf{x}) := \left[ \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) \right]_{1 \leq i, j \leq d} \in \mathbb{R}^{d \times d}.$$

Cette matrice est **symétrique**.

L'ensemble des fonctions deux fois continûment dérivables est noté  $\mathcal{C}^2$  (on dira que  $f$  est de classe  $\mathcal{C}^2$ ).

## Développement de Taylor à l'ordre 1

Si  $f \in \mathcal{C}^1$ , pour tous  $\mathbf{x}, \mathbf{h} \in \mathbb{R}^d$ ,

$$f(\mathbf{x} + \mathbf{h}) \approx f(\mathbf{x}) + \nabla f(\mathbf{x})^T \mathbf{h}.$$

pour  $\|\mathbf{h}\|$  suffisamment faible.

## Développement de Taylor à l'ordre 1

Si  $f \in \mathcal{C}^1$ , pour tous  $\mathbf{x}, \mathbf{h} \in \mathbb{R}^d$ ,

$$f(\mathbf{x} + \mathbf{h}) \approx f(\mathbf{x}) + \nabla f(\mathbf{x})^T \mathbf{h}.$$

pour  $\|\mathbf{h}\|$  suffisamment faible.

## Développement de Taylor à l'ordre 2

Si  $f \in \mathcal{C}^2$ , pour tous  $\mathbf{x}, \mathbf{h} \in \mathbb{R}^d$ ,

$$f(\mathbf{x} + \mathbf{h}) \approx f(\mathbf{x}) + \nabla f(\mathbf{x})^T \mathbf{h} + \frac{1}{2} \mathbf{h}^T \nabla^2 f(\mathbf{x}) \mathbf{h}$$

pour  $\|\mathbf{h}\|$  suffisamment faible.

## Définition

Une fonction  $\mathbf{g} : \mathbb{R}^d \rightarrow \mathbb{R}^m$  est dite  $L$ -lipschitzienne si il existe  $L > 0$  telle que

$$\forall (\mathbf{x}, \mathbf{y}) \in (\mathbb{R}^d)^2, \quad \|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|.$$

La valeur  $L$  s'appelle une constante de Lipschitz pour  $\mathbf{g}$ .

- Concept de base, nombreuses variantes.
- $\mathcal{C}_L^{1,1}$  : sous-ensemble de  $\mathcal{C}^1$  des fonctions avec dérivée première  $L$ -lipschitzienne.

## Approximation de Taylor à l'ordre 1

Soit  $f \in \mathcal{C}_L^{1,1}$ . Pour tous  $\mathbf{x}, \mathbf{h} \in \mathbb{R}^d$ ,

$$f(\mathbf{x} + \mathbf{h}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T \mathbf{h} + \frac{L}{2} \|\mathbf{h}\|^2.$$

## Approximation de Taylor à l'ordre 1

Soit  $f \in \mathcal{C}_L^{1,1}$ . Pour tous  $\mathbf{x}, \mathbf{h} \in \mathbb{R}^d$ ,

$$f(\mathbf{x} + \mathbf{h}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T \mathbf{h} + \frac{L}{2} \|\mathbf{h}\|^2.$$

⇒ Une des inégalités majeures en optimisation non linéaire.

- 1 Optimisation lisse
  - Calcul différentiel et optimisation
  - Solutions et conditions d'optimalité
  - Classes de problèmes remarquables
- 2 Descente de gradient

$$\underset{\mathbf{x} \in \mathbb{R}^d}{\text{minimiser}} f(\mathbf{x}) \quad \text{s. c. } \mathbf{x} \in \mathcal{X}$$

## Minimum global

Un point  $\mathbf{x}^*$  est un **minimum global** du problème si  $f(\mathbf{x}^*) \leq f(\mathbf{x}) \forall \mathbf{x} \in \mathbb{R}^d$ .

## Minimum local

Un point  $\mathbf{x}^*$  est un **minimum local** du problème s'il existe  $\epsilon > 0$  tel que

$$f(\mathbf{x}^*) < f(\mathbf{x}) \quad \forall \mathbf{x}, \|\mathbf{x} - \mathbf{x}^*\| \leq \epsilon.$$



- Trouver des minima globaux est difficile en général;
- Trouver et certifier des minima locaux peut aussi être difficile.

- Trouver des minima globaux est difficile en général;
- Trouver et certifier des minima locaux peut aussi être difficile.

## En optimisation lisse/douce

- Les dérivées donnent des informations;
- D'autres hypothèses sur la fonction peuvent aussi aider.

**Problème sans contraintes** minimiser  $_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$ ,  
 $f$  de classe  $\mathcal{C}^1$ .

**Problème sans contraintes** minimiser  $_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$ ,  
 $f$  de classe  $\mathcal{C}^1$ .

Condition nécessaire à l'ordre 1

**Si**  $\mathbf{x}^*$  est un minimum local du problème, **alors**

$$\|\nabla f(\mathbf{x}^*)\| = 0.$$

**Problème sans contraintes** minimiser  $_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$ ,  
 $f$  de classe  $\mathcal{C}^1$ .

Condition nécessaire à l'ordre 1

**Si**  $\mathbf{x}^*$  est un minimum local du problème, **alors**

$$\|\nabla f(\mathbf{x}^*)\| = 0.$$

- Cette condition est seulement nécessaire;
- Un point tel que  $\|\nabla f(\mathbf{x}^*)\| = 0$  peut aussi être un maximum local ou un point selle.

**Problème sans contraintes** minimiser  $_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$ ,  
 $f$  de classe  $\mathcal{C}^2$ .

**Problème sans contraintes** minimiser  $\mathbf{x} \in \mathbb{R}^d$   $f(\mathbf{x})$ ,  
 $f$  de classe  $\mathcal{C}^2$ .

Condition nécessaire à l'ordre 2

Si  $\mathbf{x}^*$  est un minimum local du problème, **alors**

$$\|\nabla f(\mathbf{x}^*)\| = 0 \quad \text{et} \quad \nabla^2 f(\mathbf{x}^*) \succeq 0.$$

**Problème sans contraintes** minimiser  $x \in \mathbb{R}^d$   $f(x)$ ,  
 $f$  de classe  $\mathcal{C}^2$ .

Condition nécessaire à l'ordre 2

Si  $x^*$  est un minimum local du problème, **alors**

$$\|\nabla f(x^*)\| = 0 \quad \text{et} \quad \nabla^2 f(x^*) \succeq 0.$$

Condition suffisante à l'ordre 2

Si  $x^*$  vérifie

$$\|\nabla f(x^*)\| = 0 \quad \text{and} \quad \nabla^2 f(x^*) \succ 0,$$

**alors** c'est un minimum local du problème.



## Minima globaux

- Possibles à trouver pour des problèmes **convexes**.
- Possibles aussi pour certaines classes de problèmes non convexes.

## Minima globaux

- Possibles à trouver pour des problèmes **convexes**.
- Possibles aussi pour certaines classes de problèmes non convexes.

## Minima locaux

- Peuvent être obtenus pour certaines classes de problèmes non convexes.
- En général, peuvent donner des valeurs plus mauvaises que celle des solutions du problème.

## Minima globaux

- Possibles à trouver pour des problèmes **convexes**.
- Possibles aussi pour certaines classes de problèmes non convexes.

## Minima locaux

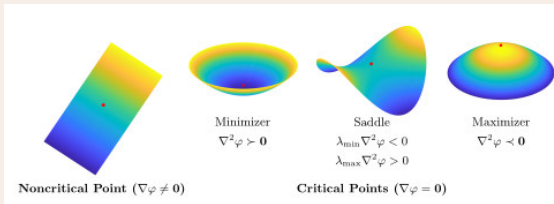
- Peuvent être obtenus pour certaines classes de problèmes non convexes.
- En général, peuvent donner des valeurs plus mauvaises que celle des solutions du problème.

## Points stationnaires

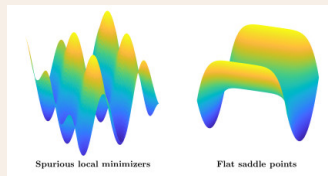
- D'ordre 1 ou 2, vérifient les conditions nécessaires d'optimalité.
- Calculables via des algorithmes.
- Peuvent être des minima/maxima locaux ou des points selles.

# Points remarquables en optimisation

## Cas favorables



## Cas pathologiques



Source : J. Wright et Y. Ma, *High-Dimensional Data Analysis with Low-Dimensional Models*, 2022.

- 1 Optimisation lisse
  - Calcul différentiel et optimisation
  - Solutions et conditions d'optimalité
  - Classes de problèmes remarquables
- 2 Descente de gradient

## Définition

Un ensemble  $\mathcal{C} \in \mathbb{R}^d$  est dit **convexe** si

$$\forall(\mathbf{u}, \mathbf{v}) \in \mathcal{C}^2, \forall t \in [0, 1], \quad t\mathbf{u} + (1 - t)\mathbf{v} \in \mathcal{C}.$$

## Définition

Un ensemble  $\mathcal{C} \in \mathbb{R}^d$  est dit **convexe** si

$$\forall (\mathbf{u}, \mathbf{v}) \in \mathcal{C}^2, \forall t \in [0, 1], \quad t\mathbf{u} + (1 - t)\mathbf{v} \in \mathcal{C}.$$

Exemples :

- $\mathbb{R}^d$ ;
- Droite :  $\{t\mathbf{w} | t \in \mathbb{R}\}$  pour tout  $\mathbf{w} \in \mathbb{R}^d$ ;
- Boule :  $\left\{ \mathbf{w} \in \mathbb{R}^d \mid \|\mathbf{w}\|^2 = \sum_{i=1}^d [\mathbf{w}]_i^2 \leq 1 \right\}$ .

## Définition

Une fonction  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  est dite **convexe** si

$$\forall(\mathbf{u}, \mathbf{v}) \in (\mathbb{R}^d)^2, \forall t \in [0, 1], \quad f(t\mathbf{u} + (1 - t)\mathbf{v}) \leq t f(\mathbf{u}) + (1 - t) f(\mathbf{v}).$$



## Définition

Une fonction  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  est dite **convexe** si

$$\forall (\mathbf{u}, \mathbf{v}) \in (\mathbb{R}^d)^2, \forall t \in [0, 1], \quad f(t\mathbf{u} + (1-t)\mathbf{v}) \leq t f(\mathbf{u}) + (1-t) f(\mathbf{v}).$$

Exemples :

- Fonction linéaire :  $f(\mathbf{w}) = \mathbf{a}^T \mathbf{w} + b$ ;
- Norme au carré :  $f(\mathbf{w}) = \|\mathbf{w}\|^2 = \mathbf{w}^T \mathbf{w}$ .

## Convexité et gradient

Une fonction  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  de classe  $\mathcal{C}^1$  est convexe si et seulement si

$$\forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^d, \quad f(\mathbf{v}) \geq f(\mathbf{u}) + \nabla f(\mathbf{u})^T (\mathbf{v} - \mathbf{u}).$$

## Convexité et gradient

Une fonction  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  de classe  $\mathcal{C}^1$  est convexe si et seulement si

$$\forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^d, \quad f(\mathbf{v}) \geq f(\mathbf{u}) + \nabla f(\mathbf{u})^T (\mathbf{v} - \mathbf{u}).$$

**L'autre inégalité clé en optimisation.**

## Convexité et gradient

Une fonction  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  de classe  $\mathcal{C}^1$  est convexe si et seulement si

$$\forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^d, \quad f(\mathbf{v}) \geq f(\mathbf{u}) + \nabla f(\mathbf{u})^T (\mathbf{v} - \mathbf{u}).$$

**L'autre inégalité clé en optimisation.**

## Convexité et matrice hessienne

Une fonction  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  de classe  $\mathcal{C}^2$  est dite convexe si et seulement si  $\nabla^2 f(\mathbf{x}) \succeq 0$  pour tout vecteur  $\mathbf{x} \in \mathbb{R}^d$ .

minimiser  $f(\mathbf{x})$ ,  $f$  convexe.  
 $\mathbf{x} \in \mathbb{R}^d$

minimiser  $f(\mathbf{x})$ ,  $f$  convexe.  
 $\mathbf{x} \in \mathbb{R}^d$

## Théorème

Tout minimum local de  $f$  est un minimum global.

minimiser  $f(\mathbf{x})$ ,  $f$  convexe.  
 $\mathbf{x} \in \mathbb{R}^d$

## Théorème

Tout minimum local de  $f$  est un minimum global.

## Corollaire

Si  $f$  est de classe  $\mathcal{C}^1$ , tout point  $\mathbf{w}^*$  tel que  $\|\nabla f(\mathbf{x}^*)\| = 0$  est un minimum global de  $f$ .

## Définition

Une fonction  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  est  $\mu$ -fortement convexe si pour tous  $(\mathbf{u}, \mathbf{v}) \in (\mathbb{R}^d)^2$  et  $t \in [0, 1]$ , on a

$$f(t\mathbf{u} + (1-t)\mathbf{v}) \leq t f(\mathbf{u}) + (1-t)f(\mathbf{v}) - \frac{\mu}{2} t(1-t) \|\mathbf{v} - \mathbf{u}\|^2.$$

$\mathbf{x} \mapsto \frac{\mu}{2} \|\mathbf{x}\|^2$  est  $\mu$ -fortement convexe.



## Définition

Une fonction  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  est  $\mu$ -fortement convexe si pour tous  $(\mathbf{u}, \mathbf{v}) \in (\mathbb{R}^d)^2$  et  $t \in [0, 1]$ , on a

$$f(t\mathbf{u} + (1-t)\mathbf{v}) \leq tf(\mathbf{u}) + (1-t)f(\mathbf{v}) - \frac{\mu}{2}t(1-t)\|\mathbf{v} - \mathbf{u}\|^2.$$

$\mathbf{x} \mapsto \frac{\mu}{2}\|\mathbf{x}\|^2$  est  $\mu$ -fortement convexe.

## Théorème

- Une fonction fortement convexe a au plus un minimum global.
- Une fonction continue fortement convexe a un unique minimum global.

## Gradient et convexité forte

Soit  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  de classe  $\mathcal{C}^1$ . Alors,

$$\forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^d, \quad f(\mathbf{v}) \geq f(\mathbf{u}) + \nabla f(\mathbf{u})^T (\mathbf{v} - \mathbf{u}) + \frac{\mu}{2} \|\mathbf{v} - \mathbf{u}\|^2.$$

## Hessienne et convexité forte

Soit  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  de classe  $\mathcal{C}^2$ . Alors

$$f \text{ est } \mu\text{-fortement convexe} \iff \nabla^2 f(\mathbf{x}) \succeq \mu I \quad \forall \mathbf{x} \in \mathbb{R}^d.$$

## Minimisation d'une quadratique convexe

$$\underset{\mathbf{x} \in \mathbb{R}^d}{\text{minimiser}} f(\mathbf{x}) := \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x}, \quad \mathbf{A} \succeq 0.$$

- $\nabla^2 f(\mathbf{x}) = \mathbf{A}$ ;
- Fortement convexe si  $\mathbf{A} \succ 0$  avec  $\mu = \lambda_{\min}(\mathbf{A})$ .

## Minimisation d'une quadratique convexe

$$\underset{\mathbf{x} \in \mathbb{R}^d}{\text{minimiser}} f(\mathbf{x}) := \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x}, \quad \mathbf{A} \succeq 0.$$

- $\nabla^2 f(\mathbf{x}) = \mathbf{A}$ ;
- Fortement convexe si  $\mathbf{A} \succ 0$  avec  $\mu = \lambda_{\min}(\mathbf{A})$ .

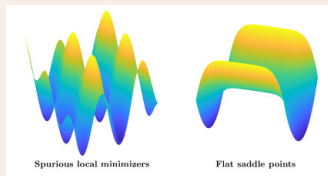
## Projection sur un ensemble fermé convexe

$$\underset{\mathbf{x} \in \mathcal{X}}{\text{minimiser}} \frac{1}{2} \|\mathbf{x} - \mathbf{a}\|^2, \quad \mathcal{X} \text{ fermé convexe.}$$

- Generalise le cas  $\mathcal{X} = \mathbb{R}^d$ ;
- L'objectif est 1-fortement convexe  $\Rightarrow$  il existe une unique solution.

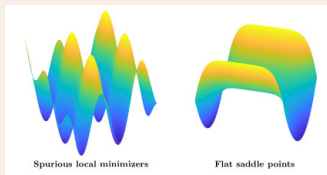
## Problème non convexe pathologique

- Des minima locaux non globaux;
- De “mauvais” points selles.



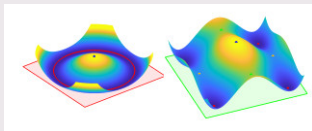
## Problème non convexe pathologique

- Des minima locaux non globaux;
- De “mauvais” points selles.



## Des instances favorables

- Points selles stricts (pas stationnaires à l'ordre 2);
- Équivalence entre minima locaux et globaux.



Source : J. Wright et Y. Ma, *High-Dimensional Data Analysis with Low-Dimensional Models*, 2022.

## Complétion de matrice

$$\underset{X \in \mathbb{R}^{n \times m}, \text{rank}(X) \leq r}{\text{minimiser}} \sum_{(i,j) \in \Omega} (X_{ij} - M_{ij})^2 \quad M \in \mathbb{R}^{n \times m}, \Omega \subset [n] \times [m].$$

- Données : entrées de  $M$  observées.
- Hypothèse :  $M$  est de rang  $r \ll \min(m, n)$ .

## Complétion de matrice

$$\underset{X \in \mathbb{R}^{n \times m}, \text{rank}(X) \leq r}{\text{minimiser}} \sum_{(i,j) \in \Omega} (X_{ij} - M_{ij})^2 \quad M \in \mathbb{R}^{n \times m}, \Omega \subset [n] \times [m].$$

- Données : entrées de  $M$  observées.
- Hypothèse :  $M$  est de rang  $r \ll \min(m, n)$ .

## Formulation factorisée (Burer & Monteiro, '03)

$$\underset{U \in \mathbb{R}^{n \times r}, V \in \mathbb{R}^{m \times r}}{\text{minimiser}} \sum_{(i,j) \in \Omega} \left( [U V^T]_{ij} - M_{ij} \right)^2,$$

- $(n + m)r$  variables ( $\ll nm$ ).
- **Non convexe en  $U$  et  $V$ ...**
- ...mais ne possède que des points selles et des **minima globaux**.



# Exemples de “bons” problèmes non convexes (2)

## Analyse en composantes principales/Calcul de valeurs propres

Partant de données  $\{\mathbf{a}_i\}_{i=1\dots n}$ , trouver la direction de variabilité maximale des  $\mathbf{a}_i$  en résolvant

$$\underset{\mathbf{x} \in \mathbb{R}^d}{\text{minimiser}} -\frac{1}{2} \mathbf{x}^T \mathbf{C} \mathbf{x} \quad \text{s. c.} \quad \|\mathbf{x}\|^2 = 1,$$

avec

$$\mathbf{C} = \frac{1}{n} \sum_{i=1}^n (\mathbf{a}_i - \bar{\mathbf{a}})(\mathbf{a}_i - \bar{\mathbf{a}})^T \quad \bar{\mathbf{a}} = \frac{1}{n} \sum_{i=1}^n \mathbf{a}_i.$$

# Exemples de “bons” problèmes non convexes (2)

## Analyse en composantes principales/Calcul de valeurs propres

Partant de données  $\{\mathbf{a}_i\}_{i=1\dots n}$ , trouver la direction de variabilité maximale des  $\mathbf{a}_i$  en résolvant

$$\underset{\mathbf{x} \in \mathbb{R}^d}{\text{minimiser}} -\frac{1}{2} \mathbf{x}^T \mathbf{C} \mathbf{x} \quad \text{s. c.} \quad \|\mathbf{x}\|^2 = 1,$$

avec

$$\mathbf{C} = \frac{1}{n} \sum_{i=1}^n (\mathbf{a}_i - \bar{\mathbf{a}})(\mathbf{a}_i - \bar{\mathbf{a}})^T \quad \bar{\mathbf{a}} = \frac{1}{n} \sum_{i=1}^n \mathbf{a}_i.$$

## Factorisation orthogonale de tenseur d'ordre 4 (Ge et al. 2015)

$$\underset{\{u_i\} \subset \mathbb{S}^{n-1}}{\text{minimiser}} \sum_{i \neq j} T(u_i, u_i, u_j, u_j).$$

avec  $T \in \mathbb{R}^{n \times n \times n \times n}$ .

- Les solutions **sont** les points nécessaires du second ordre.

## 1 Optimisation lisse

## 2 Descente de gradient

- Algorithmes et descente de gradient
- Descente de gradient et optimisation convexe
- Accélération

## 1 Optimisation lisse

## 2 Descente de gradient

- Algorithmes et descente de gradient
- Descente de gradient et optimisation convexe
- Accélération

minimiser  $f(\mathbf{x})$ .  
 $\mathbf{x} \in \mathbb{R}^d$

minimiser  $f(\mathbf{x})$ .  
 $\mathbf{x} \in \mathbb{R}^d$

## Hypothèses

- $f$  est minorée par  $f_{\text{low}}$ ;
- $f$  est lisse (au moins de classe  $\mathcal{C}^1$ ).

## De manière itérative

- Idée de base : étant donné un point courant, se déplacer vers un point potentiellement meilleur;
- Une itération représente l'ensemble des calculs nécessaires pour ce déplacement.

## Notre but dans le reste du cours

- Proposer des algorithmes;
- Décrire leurs garanties théoriques;
- Vérifier leur intérêt pratique (notebooks).

Pour résoudre minimiser  $\mathbf{x} \in \mathbb{R}^d$   $f(\mathbf{x})$ , l'algorithme devrait satisfaire les propriétés suivantes :

- 1 Les points calculés tendent vers une solution;
- 2 Les valeurs de l'objectif tendent vers la valeur optimale;
- 3 Une condition d'optimalité est satisfaite à la limite.



Pour résoudre minimiser  $\mathbf{x} \in \mathbb{R}^d$   $f(\mathbf{x})$ , l'algorithme devrait satisfaire les propriétés suivantes :

- 1 Les points calculés tendent vers une solution;
- 2 Les valeurs de l'objectif tendent vers la valeur optimale;
- 3 Une condition d'optimalité est satisfaite à la limite.

## Convergence des itérés

L'algorithme génère une suite  $\{\mathbf{x}_k\}_k$  telle que

$$\|\mathbf{x}_k - \mathbf{x}^*\| \rightarrow 0 \quad \text{lorsque } k \rightarrow \infty,$$

où  $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$  est une solution globale du problème.

### Convergence en valeur de fonction

$$f(\mathbf{x}_k) \rightarrow f^* \quad \text{lorsque } k \rightarrow \infty,$$

où  $f^* = \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$ .

## Convergence en valeur de fonction

$$f(\mathbf{x}_k) \rightarrow f^* \quad \text{lorsque } k \rightarrow \infty,$$

où  $f^* = \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$ .

## Convergence vers un point stationnaire d'ordre 1

$$\|\nabla f(\mathbf{x}_k)\| \rightarrow 0 \quad \text{lorsque } k \rightarrow \infty.$$

*Condition plus générale.*

$$\underset{\mathbf{x} \in \mathbb{R}^d}{\text{minimiser}} f(\mathbf{x}), \quad f \in \mathcal{C}^1.$$

Pour tout  $\mathbf{x} \in \mathbb{R}^d$ ,

- ❶ Soit  $\mathbf{x}$  est un minimum local et donc  $\nabla f(\mathbf{x}) = 0$ ;
- ❷ Soit  $f$  décroît **localement** depuis  $\mathbf{x}$  **dans la direction de  $-\nabla f(\mathbf{x})$** .  
*Preuve basée sur Taylor.*

**Entrées :**  $\mathbf{x}_0 \in \mathbb{R}^d$ ,  $\alpha_0 > 0$ ,  $\varepsilon > 0$ ,  $k_{\max} \in \mathbb{N}$ .

Set  $k = 0$ .

- 1 Evaluer  $\nabla f(\mathbf{x}_k)$ ; si  $\|\nabla f(\mathbf{x}_k)\| < \varepsilon$  terminer.
- 2 Poser  $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)$ .
- 3 Incrémenter  $k$  de 1; si  $k = k_{\max}$  terminer, sinon aller à l'étape 1.

**Entrées :**  $\mathbf{x}_0 \in \mathbb{R}^d$ ,  $\alpha_0 > 0$ ,  $\varepsilon > 0$ ,  $k_{\max} \in \mathbb{N}$ .

Set  $k = 0$ .

- 1 Evaluer  $\nabla f(\mathbf{x}_k)$ ; si  $\|\nabla f(\mathbf{x}_k)\| < \varepsilon$  terminer.
- 2 Poser  $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)$ .
- 3 Incrémenter  $k$  de 1; si  $k = k_{\max}$  terminer, sinon aller à l'étape 1.

## Critères d'arrêt

- Convergence :  $\|\nabla f(\mathbf{x}_k)\| < \varepsilon$ ;
- Budget :  $k = k_{\max}$ .

## Pas constant

Si  $f \in \mathcal{C}_L^{1,1}$ , poser  $\alpha_k = \frac{1}{L}$ :

- Garantit une décroissance à chaque itération;
- Mais demande de connaître  $L$ .

## Pas décroissant

Choisir  $\alpha_k$  tel que  $\alpha_k \rightarrow 0$ .

- Garantit une décroissance à partir d'un certain rang;
- Mais force la valeur à décroître.

## En optimisation classique

- Recherche linéaire : À chaque itération,  $\alpha_k$  obtenue par retour arrière (*backtracking*) sur un ensemble de valeurs en ordre décroissants (ex:  $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \dots$ ).
- La valeur renvoyée vérifie une condition type décroissance de la valeur de l'objectif.

## En apprentissage (notamment profond)

$\alpha_k = \text{Learning rate}$

- Utiliser une valeur fixe pendant un certain nombre d'itérations;
- Diminuer progressivement cette valeur selon une règle fixée (*scheduling*).



$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}), \quad f \in \mathcal{C}_L^{1,1}.$$

## Rappels : Descente de gradient

- Itération :  $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)$ , terminer si  $\nabla f(\mathbf{x}_k) = 0$ .
- Choix de base en théorie :  $\alpha_k = \frac{1}{L}$ .

## Résultats théoriques

- Convergence : Montrer que  $\|\nabla f(\mathbf{x}_k)\| \rightarrow 0$ ;
- Vitesse de convergence : Décroissance de  $\|\nabla f(\mathbf{x}_k)\|$ .
- Complexité au pire cas : Effort requis pour obtenir  $\|\nabla f(\mathbf{x}_k)\| \leq \epsilon$  pour  $\epsilon > 0$ .

## Théorème

Si  $f \in \mathcal{C}_L^{1,1}$  et  $\alpha_k = \frac{1}{L}$ , la descente de gradient produit  $\mathbf{w}_k$  tel que  $\|\nabla f(\mathbf{x}_k)\| \leq \epsilon$  en au plus

$$2L(f(\mathbf{x}_0) - f_{\text{low}})\epsilon^{-2} \text{ itérations.}$$

- Même résultat pour d'autres choix pour  $\alpha_k$ , dont la recherche linéaire.
- On dit que la complexité de la descente de gradient est en  $\mathcal{O}(\epsilon^{-2})$ .

## Théorème

Si  $f \in \mathcal{C}_L^{1,1}$  et  $\alpha_k = \frac{1}{L}$ , alors pour tout  $K \geq 1$ , si  $\{\mathbf{x}_k\}$  est la suite des itérés produite par l'algorithme de descente de gradient, on a

$$\min_{0 \leq k \leq K-1} \|\nabla f(\mathbf{x}_k)\| \leq \frac{\sqrt{2L(f(\mathbf{x}_0) - f_{\text{low}})}}{\sqrt{K}}.$$

## Interpretation

- On dit que la vitesse de convergence de la descente de gradient est  $\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$ .
- Il existe une fonction telle que cette vitesse correspond exactement au comportement de la méthode !

## Sur un problème non convexe

- La descente de gradient converge vers un point  $\bar{\mathbf{x}}$  tel que  $\|\bar{\mathbf{x}}\| = 0$ .
- Ce point peut être un point selle, voire un maximum local.

## Théorème (Lee et al, 2015)

Pour **presque tout**  $\mathbf{x}_0 \in \mathbb{R}^d$ , la descente de gradient converge vers un point  $\bar{\mathbf{x}}$  tel que

$$\|\nabla f(\bar{\mathbf{x}})\| = 0 \quad \text{et} \quad \nabla^2 f(\bar{\mathbf{x}}) \succeq 0.$$

## 1 Optimisation lisse

## 2 Descente de gradient

- Algorithmes et descente de gradient
- Descente de gradient et optimisation convexe
- Accélération

$$\underset{\mathbf{x} \in \mathbb{R}^d}{\text{minimiser}} f(\mathbf{x}), \quad f \in \mathcal{C}_L^{1,1}.$$

## Descente de gradient

- Itération:  $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)$ , terminer si  $\nabla f(\mathbf{x}_k) = 0$ .
- Choix typique en théorie :  $\alpha_k = \frac{1}{L}$ .

## Avec la convexité

**Hypothèse :**  $f^* = \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$  est atteint.

- Garanties relativement à un minimum **global**;
- On peut montrer que  $f(\mathbf{x}_k) \rightarrow f^*$ ;
- On peut aussi montrer une convergence vers l'argmin.

## Cas non convexe

- Critère :  $\|\nabla f(\mathbf{x}_k)\|$ ;
- Idée : être proche d'un point stationnaire.

## Cas convexe/fortement convexe

- :  $f(\mathbf{x}_k) - f^* \leq \epsilon$ , avec  $f^* = \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$ ;
- Idée : être proche de la valeur à l'optimum.
- Valeur liée à  $\|\mathbf{x}_k - \mathbf{x}^*\|$  dans le cas fortement convexe.

## Théorème

Si  $f \in \mathcal{C}_L^{1,1}$  est convexe et  $\alpha_k = \frac{1}{L}$ , la descente de gradient calcule  $\mathbf{w}_k$  tel que  $f(\mathbf{x}_k) - f^* \leq \epsilon$  en au plus

- $\mathcal{O}(\epsilon^{-1})$  itérations;
- $\mathcal{O}\left(\frac{L}{\mu} \ln(\epsilon^{-1})\right)$  itérations si  $f$  est  $\mu$ -fortement convexe.

- Cas non convexe :  $\mathcal{O}(\epsilon^{-2})$  pour garantir  $\|\nabla f(\mathbf{x}_k)\| \leq \epsilon$ .
- On dit que la descente de gradient possède une meilleure complexité dans le cas convexe/fortement convexe.



## Théorème

Si  $f \in \mathcal{C}_L^{1,1}$  est convexe et  $\alpha_k = \frac{1}{L}$ , pour tout  $K \in \mathbb{N}$ , on a :

$$f(\mathbf{x}_K) - f^* \leq \frac{L \max_{\mathbf{x} \in \operatorname{argmin}_{\mathbf{v}} f(\mathbf{v})} \|\mathbf{x}_0 - \mathbf{x}\|}{2} \frac{1}{K}$$

pour  $f$  convexe, et

$$f(\mathbf{x}_K) - f^* \leq \left(1 - \frac{\mu}{L}\right)^K (f(\mathbf{x}_0) - f^*).$$

pour  $f$   $\mu$ -fortement convexe.

- Cas non convexe :  $\mathcal{O}(1/\sqrt{K})$  pour  $\min_{0 \leq k \leq K-1} \|\nabla f(\mathbf{x}_k)\|$ .
- On dit que la descente de gradient converge plus rapidement dans le cas fortement convexe que dans le cas convexe.

## 1 Optimisation lisse

## 2 Descente de gradient

- Algorithmes et descente de gradient
- Descente de gradient et optimisation convexe
- Accélération

## Motivation

- En optimisation non convexe,  $\mathcal{O}(1/\sqrt{K})$  est la meilleure vitesse de convergence pour une méthode type gradient;
- Dans le cas convexe, c'est  $\mathcal{O}(1/K^2)$ , mieux que la descente de gradient en  $\mathcal{O}(1/K)$ .

## Comment obtenir cette meilleure vitesse ?

- Stratégies de **gradient accéléré**, basées sur l'idée de **momentum**;
- **Principe** : Réutiliser l'information de l'itération précédente.

## Méthode de la boule lestée (*Heavy ball*, Polyak, 1964)

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k) + \beta(\mathbf{x}_k - \mathbf{x}_{k+1}).$$

- Terme de momentum  $\mathbf{x}_k - \mathbf{x}_{k+1}$ ;
- Optimale sur des quadratiques fortement convexes...
- ... mais ne converge pas toujours pour  $f$  fortement convexe!

Algorithme( $\mathbf{x}_0, \mathbf{x}_{-1} = \mathbf{x}_0$ )

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k + \beta_k(\mathbf{x}_k - \mathbf{x}_{k-1})) + \beta_k(\mathbf{x}_k - \mathbf{x}_{k-1}).$$

- Un appel de gradient par itération;
- Terme de momentum :  $\mathbf{x}_k - \mathbf{x}_{k-1}$  (pas précédent).

Version à deux suites( $\mathbf{x}_0, \mathbf{z}_0 = \mathbf{x}_0$ )

$$\begin{cases} \mathbf{x}_{k+1} &= \mathbf{z}_k - \alpha_k \nabla f(\mathbf{z}_k) \\ \mathbf{z}_{k+1} &= \mathbf{x}_{k+1} + \beta_{k+1}(\mathbf{x}_{k+1} - \mathbf{x}_k). \end{cases}$$

## Longueur de pas $\alpha_k$

- $\alpha_k = \frac{1}{L}$ ;
- Autres : décroissante, recherche linéaire, etc.

## Momentum $\beta_k$

- $f$   $\mu$ -fortement convexe :  $\beta_k = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ ;
- $f$  convexe : Utiliser deux suites

$$t_{k+1} = \frac{1}{2}(1 + \sqrt{1 + 4t_k^2}), t_0 = 0, \quad \beta_k = \frac{t_k - 1}{t_{k+1}}.$$

## Fonctions convexes

- Descente de gradient :  $f(\mathbf{x}_K) - f^* \leq \mathcal{O}\left(\frac{1}{K}\right)$ ;
- Gradient accéléré :  $f(\mathbf{x}_K) - f^* \leq \mathcal{O}\left(\frac{1}{K^2}\right)$ .

## Fonctions $\mu$ -fortement convexes

- Descente de gradient :

$$f(\mathbf{x}_K) - f^* \leq \left(1 - \frac{\mu}{L}\right)^K (f(\mathbf{x}_0) - f^*).$$

- Gradient accéléré

$$f(\mathbf{x}_K) - f^* \leq C \left(1 - \sqrt{\frac{\mu}{L}}\right)^K (f(\mathbf{x}_0) - f^*).$$

## Optimisation non linéaire

- Conditions d'optimalité : Caractérisent des points remarquables au moyen des dérivées.
- Convexité = Contexte favorable pour la minimisation;
- Cas non convexe difficile, mais certaines classes de problèmes ont de bonnes propriétés.

## Descente de gradient

- Algorithme de base pour l'optimisation “douce”;
- Applicable aux problèmes convexes et non convexes !
- Variantes accélérées optimales pour les problèmes convexes.



## Ouvrages :

- J. Wright et Y. Ma, *High-Dimensional Data Analysis with Low-Dimensional Models*, Cambridge University Press, 2022.
- S. J. Wright et B. Recht, *Optimization for Data Analysis*, Cambridge University Press, 2022.
- A. Beck, *First-order methods in optimization*, MOS-SIAM Series on Optimization, 2017.

**Tout à l'heure :** Méthodes stochastiques avec Florentin Goyens.

## Demain

- La gestion des dérivées;
- L'optimisation sans dérivées.

**Tout à l'heure :** Méthodes stochastiques avec Florentin Goyens.

## Demain

- La gestion des dérivées;
- L'optimisation sans dérivées.

**Merci beaucoup !**