Optimisation non linéaire et non convexe

Yann Chevaleyre, Paul Caillon, Clément Royer

Certificat Chef de Projet IA - Université Paris Dauphine-PSL

9 octobre 2024

Séance de ce matin

1. Optimisation lisse
2. Descente de gradient

Table des matières

Sommaire

Problème

$$\underset{\boldsymbol{x} \in \mathbb{R}^d}{\text{minimiser}} \, f(\boldsymbol{x}).$$

Hypothèses

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

Gradient

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

Gradient

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

Dérivée à l'ordre 1

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

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

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

Calcul différentiel (2)

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

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

Dérivée d'ordre 2

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

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

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

Calcul différentiel (3)

Développement de Taylor à l'ordre 1

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

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

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

Développement de Taylor à l'ordre 1

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

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

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

Développement de Taylor à l'ordre 2

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

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

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

Définition

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

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

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

- Concept de base, nombreuses variantes.
- $\mathcal{C}^{1,1}_L$: 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 $\boldsymbol{x}, \boldsymbol{h} \in \mathbb{R}^d$,

$$f(\boldsymbol{x} + \boldsymbol{h}) \le f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}} \boldsymbol{h} + \frac{L}{2} \|\boldsymbol{h}\|^2.$$

Approximation de Taylor à l'ordre 1

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

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

$\Rightarrow$ **Une des inégalités majeures en optimisation non linéaire.**

Sommaire

Solutions

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimiser}}\, f(\boldsymbol{x})$$

Minimum global

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

Minimum local

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

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

Solutions locales et globales (2)

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

Solutions locales et globales (2)

- 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 $\text{minimiser}_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x})$,
f de classe $\mathcal{C}^1$.

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

Condition nécessaire à l'ordre 1

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

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

Conditions d'optimalité en optimisation douce

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

Condition nécessaire à l'ordre 1

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

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

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

Conditions d'optimalité en optimisation douce (2)

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

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

Condition à l'ordre 2

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

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

Conditions d'optimalité en optimisation douce (2)

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

Condition à l'ordre 2

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

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

Condition à l'ordre 2

Si $\boldsymbol{x}^*$ vérifie

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

alors c'est un minimum local du problème.

Points remarquables en optimisation

Minima globaux

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

Points remarquables en optimisation

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 remarquables en optimisation

Minima globaux

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

Minima locaux

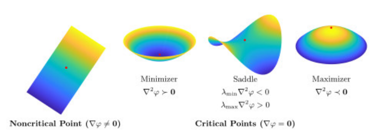
- Peuvent être obtenus pour certaines classes de problèmes non convexes.
- En général, peuvent donner des valeurs plus mauvaises que celle des solutions du problème.

Points stationnaires

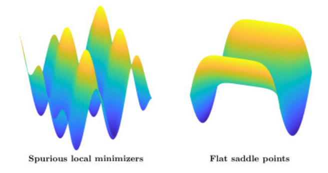
- D'ordre 1 ou 2, vérifient les conditions nécessaires d'optimalité.
- Calculables via des algorithmes.
- Peuvent être des minima/maxima locaux ou des points selles.

Points remarquables en optimisation

Cas favorables



Cas pathologiques



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

Sommaire

Définition

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

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

Définition

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

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

Exemples :

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

Définition

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

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

Définition

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

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

Exemples :

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

Convexité et gradient

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

$$\forall \boldsymbol{x}, \boldsymbol{h} \in \mathbb{R}^d, \quad f(\boldsymbol{x} + \boldsymbol{h}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}} \boldsymbol{h}.$$

Fonctions convexes lisses/douces

Convexité et gradient

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

$$\forall \boldsymbol{x}, \boldsymbol{h} \in \mathbb{R}^d, \quad f(\boldsymbol{x} + \boldsymbol{h}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}} \boldsymbol{h}.$$

L'autre inégalité clé en optimisation.

Fonctions convexes lisses/douces

Convexité et gradient

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

$$\forall \boldsymbol{x}, \boldsymbol{h} \in \mathbb{R}^d, \quad f(\boldsymbol{x}+\boldsymbol{h}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}} \boldsymbol{h}.$$

L'autre inégalité clé en optimisation.

Convexité et matrice hessienne

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

Optimisation et fonction convexe

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

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

Théorème

Tout minimum local de f est un minimum global.

Optimisation et fonction convexe

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

Théorème

Tout minimum local de f est un minimum global.

Corollaire

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

Définition

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

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

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

Fonctions fortement convexes

Définition

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

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

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

Théorème

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

Fonctions fortement convexes

Gradient et convexité forte

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

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

Hessienne et convexité forte

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

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

Exemples de problèmes fortement convexes

Minimisation d'une quadratique convexe

$$\operatorname*{minimiser}_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) := \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}, \quad \boldsymbol{A} \succeq 0.$$

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

Exemples de problèmes fortement convexes

Minimisation d'une quadratique convexe

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

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

Projection sur un ensemble fermé convexe

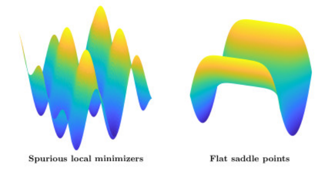
$$\underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimiser}}\, \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{a}\|^2, \quad \mathcal{X} \text{ fermé convexe.}$$

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

Problèmes non convexes remarquables

Problème non convexe pathologique

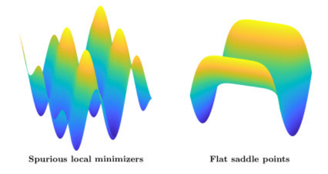
- Des minima locaux non globaux;
- De “mauvais” points selles.



Problèmes non convexes remarquables

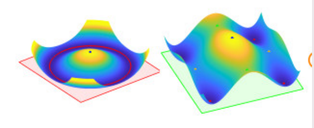
Problème non convexe pathologique

- 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.

Exemples de “bons” problèmes non convexes

Complétion de matrice

$$\underset{X \in \mathbb{R}^{d_1 \times d_2}, \mathrm{rank}(X) \leq r}{\text{minimiser}} \sum_{(i,j) \in \Omega} (X_{ij} - M_{ij})^2 \quad M \in \mathbb{R}^{d_1 \times d_2}, \ \Omega \subset [d_1] \times [d_2].$$

- Données : entrées de M observées.
- Hypothèse : M est de rang $r \ll \min(d_1, d_2)$.

Exemples de “bons” problèmes non convexes

Complétion de matrice

$$\underset{X\in\mathbb{R}^{d_1\times d_2}, \mathrm{rank}(X)\le r}{\text{minimiser}} \sum_{(i,j)\in\Omega} (X_{ij}-M_{ij})^2 \quad M\in\mathbb{R}^{d_1\times d_2},\ \Omega\subset[d_1]\times[d_2].$$

- Données : entrées de M observées.
- Hypothèse : M est de rang $r \ll \min(d_1, d_2)$.

Formulation factorisée (Burer & Monteiro, ’03)

$$\underset{U\in\mathbb{R}^{d_1\times r}, V\in\mathbb{R}^{d_2\times r}}{\text{minimiser}} \sum_{(i,j)\in\Omega} \left([U V^\top]_{ij} - M_{ij}\right)^2,$$

- $(d_1+d_2)r$ variables ($\ll d_1\, d_2$).
- **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 $\{\boldsymbol{a}_i\}_{i=1\dots n}$, trouver la direction de variabilité maximale des $\boldsymbol{a}_i$ en résolvant

$$\operatorname*{minimiser}_{\boldsymbol{x}\in\mathbb{R}^d} -\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{x} \quad \text{s. c.} \quad \|\boldsymbol{x}\|^2 = 1,$$

avec

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

Table des matières

Table des matières

Cadre

$$\underset{\boldsymbol{x} \in \mathbb{R}^d}{\text{minimiser}} \, f(\boldsymbol{x}).$$

Cadre

$$\underset{\boldsymbol{x} \in \mathbb{R}^d}{\text{minimiser}}\, f(\boldsymbol{x}).$$

Hypothèses

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

Comment procéder

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).

Quelles garanties

Pour résoudre $\text{minimiser}_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$, 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 $\text{minimiser}_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x})$, 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 $\{\boldsymbol{x}_k\}_k$ telle que

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

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

Quelles garanties (2)

Convergence en valeur de fonction

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

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

Convergence en valeur de fonction

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

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

Convergence vers un point stationnaire d'ordre 1

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

Condition plus générale.

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

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

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

Algorithme de descente de gradient

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

Set $k = 0$.

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

Algorithme de descente de gradient

Entrées : $\boldsymbol{x}_0 \in \mathbb{R}^d,\ \alpha_0 > 0,\ \varepsilon > 0, k_{\max} \in \mathbb{N}$.
Set $k = 0$.

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

Critères d'arrêt

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

Choix de la longueur de pas α_k

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 α_k tel que $\alpha_k \to 0$.

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

Choix de la longueur de pas α_k (2)

En optimisation classique

- Recherche linéaire : À chaque itération, α_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 value renvoyée vérifie une condition type décroissance de la valeur de l'objectif.

En apprentissage (notamment profond)

$\alpha_k =$*Learning rate*

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

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

Rappels : Descente de gradient

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

Résultats théoriques

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

Théorème

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

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

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

Vitesses de convergence pour la descente de gradient

Théorème

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

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

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 $\bar{\boldsymbol{x}}$ tel que $\|\bar{\boldsymbol{x}}\| = 0$.
- Ce point peut être un point selle, voire un maximum local.

Théorème (Lee et al, 2015)

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

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

Sommaire

Rappels : la descente de gradient

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

Descente de gradient

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

Avec la convexité

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

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

Complexités et vitesses de convergence

Cas non convexe

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

Cas convexe/fortement convexe

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

Résultats de complexité

Théorème

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

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

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

Vitesses de convergence

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:

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

pour f convexe, et

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

pour f μ-fortement convexe.

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

Sommaire

Accélération en optimisation convexe

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)

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

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

Gradient accéléré (Nesterov, 1983)

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

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

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

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

$$\left\{ \begin{array}{lcl} \boldsymbol{x}_{k+1} & = & \boldsymbol{z}_k - \alpha_k \nabla f(\boldsymbol{z}_k) \\ \boldsymbol{z}_{k+1} & = & \boldsymbol{x}_{k+1} + \beta_{k+1}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k). \end{array} \right.$$

Choix des paramètres

Longueur de pas α_k

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

Momentum β_k

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

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

Vitesses de convergence

Fonctions convexes

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

Fonctions μ-fortement convexes

- Descente de gradient :

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

- Gradient accéléré

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

Conclusion

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.

Références

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.

Fin de la seconde partie

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

Demain

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

Fin de la seconde partie

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

Demain

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

Merci beaucoup !