

Certificat Big Data

Introduction à l'optimisation numérique

Serge Gratton*, Ehouarn Simon*

* Toulouse INP, IRIT
`ehouarn.simon@toulouse-inp.fr`

29 octobre 2019

- 1 Introduction
 - Quelques exemples
 - Rappels et notations
- 2 Premiers théorèmes d'existence et unicité
- 3 Optimisation sans contrainte
 - Caractérisation des extrema
 - Résolution numérique : quelques algorithmes
 - Garanties de convergence ?
- 4 Vers l'optimisation avec contraintes
 - Caractérisation des extrema
 - Résolution numérique

Outline

- 1 Introduction
 - Quelques exemples
 - Rappels et notations
- 2 Premiers théorèmes d'existence et unicité
- 3 Optimisation sans contrainte
 - Caractérisation des extrema
 - Résolution numérique : quelques algorithmes
 - Garanties de convergence ?
- 4 Vers l'optimisation avec contraintes
 - Caractérisation des extrema
 - Résolution numérique

Prévision de la dynamique de l'atmosphère et de l'océan

Modèles numériques

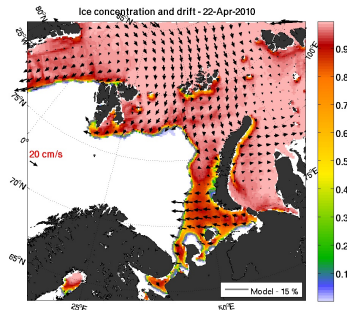
- Nombreuses incertitudes.
 - ▷ Hypothèses simplificatrices, précision numérique, forçages incertains,...
- Très grande dimension.

Observations

- Distribution spatio-temporelle hétérogène.
- Erreurs importantes.

Assimilation de données

- Comment combiner de manière "optimale" l'information incertaine et hétérogène obtenue depuis les modèles numériques et les observations dans le but d'estimer l'état du système.



Glace dans la Mer de Barents

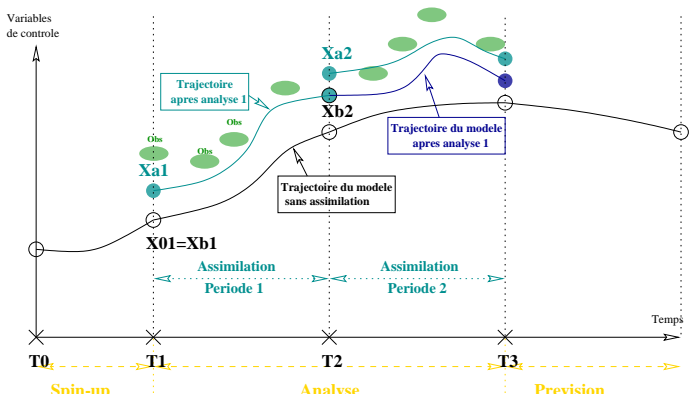
Prévision de la dynamique de l'atmosphère et de l'océan

Minimisation d'une fonctionnelle en très grande dimension

- Algorithme 4D-Var :

$$J(\mathbf{x}_0) = \underbrace{\frac{1}{2}(\mathbf{x}_0 - \mathbf{x}^b)^T \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}^b)}_{\text{écart à l'ébauche}} + \underbrace{\frac{1}{2} \sum_{i=0}^N (H_i[M_{0 \rightarrow i}(\mathbf{x}_0)] - \mathbf{y}_i)^T \mathbf{R}_i^{-1} (H_i[M_{0 \rightarrow i}(\mathbf{x}_0)] - \mathbf{y}_i)}_{\text{écarts aux observations}}$$

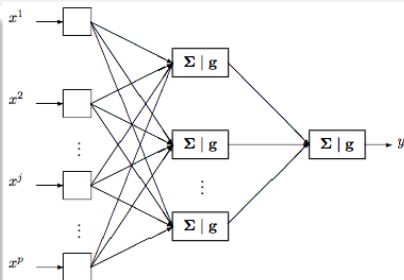
- Fluides géophysiques : $\mathbf{x}_0 \sim 10^9$ variables, $\mathbf{y}_i \sim 10^7$ variables.



Apprentissage par réseaux de neurones

Perceptron

- Signaux d'entrée $(x_i)_{1 \leq p}$.
- Des couches de neurones "cachées".
 - ▷ Connexion en entrée à tous les neurones de la couche précédente.
- Réponse $y = f(x_1, \dots, x_p; \alpha, \beta)$.
 - ▷ Paramètres (α, β) à calibrer.



<http://wikistat.fr/>

Apprentissage du réseau de neurones : estimation (α, β)

- Hypothèse : une seule couche cachée à q neurones et sortie linéaire.
- Données : $\forall 1 \leq i \leq n, (x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$.
- Problème aux moindres carrés :

$$\min_{(\alpha^i)_{1 \leq i \leq q} \in (\mathbb{R}^{p+1})^q, \beta \in \mathbb{R}^{q+1}} h(\alpha^1, \dots, \alpha^q, \beta) = \frac{1}{2} \sum_{i=1}^n (y_i - f(x_i, \alpha^1, \dots, \alpha^q, \beta))^2$$

$$\text{avec } f(x, \alpha^1, \dots, \alpha^q, \beta) = \beta_0 + \sum_{k=1}^q \beta_k g(\alpha_0^k + \sum_{j=1}^p \alpha_j^k x_j).$$

Filtrage collaboratif

Système de recommandations

- Données : utilisateurs émettent un avis sur des produits (films, musiques, etc..).
 - ▷ Ex: triplets (user_id, movie_id, ratings).
- Problème : données manquantes.
 - ▷ Ex: l'utilisateur "user_id" n'a pas noté tous les films de la base de données.
- Peut-on **estimer les évaluations manquantes** pour chaque utilisateur ?
 - ▷ Ex: Futures recommandations de films pour l'utilisateur "user_id".

Une modélisation

- La matrice $R \in \mathcal{M}_{m,n}(\mathbb{R})$ contient les évaluations des utilisateurs.

$R_{i,j}$ = note que l'utilisateur "user_id=i" a donné au film "movie_id=j".

- On cherche $P \in \mathbb{R}^{m,k}$ et $Q \in \mathbb{R}^{n,k}$ telles que $R \approx \hat{R} = PQ^T$.
- Problème d'optimisation :

$$\min_{P,Q} \sum_{(i,j) \text{ t.q. } \exists R_{ij}} \left(r_{ij} - q_j^\top p_i \right)^2 + \lambda (\|p_i\|^2 + \|q_j\|^2)$$

avec (p_i) et (q_j) les lignes de P et Q .

Matrice semi- et définie positive

Définition : Soit $A \in \mathcal{M}(\mathbb{R})$ une matrice symétrique.

- A est semi-définie positive si $\forall x \in \mathbb{R}^n, x^T A x \geq 0$.
- A est définie positive si $\forall x \in \mathbb{R}^n \setminus \{0\}, x^T A x > 0$.

Théorème : Soit $A \in \mathcal{M}(\mathbb{R})$ une matrice symétrique.

- A est semi-définie positive \Leftrightarrow toutes les valeurs propres de A sont positives ou nulles.
- A est définie positive \Leftrightarrow toutes les valeurs propres de A sont strictement positives.

Calcul différentiel

Continue différentiabilité

Définition : Soit $f : \mathcal{O} \subset E \rightarrow F$ avec E, F Banach et \mathcal{O} ouvert de E .

- On dit que f est différentiable en $a \in \mathcal{O}$ si $\exists \ell \in \mathcal{L}_c(E, F), \exists \epsilon : \mathcal{O} \rightarrow F$ telles que

$$\forall x \in \mathcal{O}, f(x) = f(a) + \ell(x - a) + \|x - a\| \epsilon(x)$$

avec $\lim_{x \rightarrow a} \epsilon(x) = 0$. On notera par la suite $\ell = f'(a)$

- On dit que f est continûment différentiable sur \mathcal{O} si f est différentiable en tout point a de \mathcal{O} et si l'application

$$\begin{aligned} f' &: \mathcal{O} \rightarrow \mathcal{L}_c(E, F) \\ a &\mapsto f'(a) \end{aligned}$$

est continue sur \mathcal{O} .

Remarques :

- Si $E = \mathbb{R}^n$ et $F = \mathbb{R}^p$, $\forall x \in \mathcal{O}$, on peut identifier $f'(x)$ avec sa matrice Jacobienne $J_f(x) \in \mathcal{M}_{p,n}(\mathbb{R})$.
- Si $E = \mathbb{R}^n$, muni du produit scalaire canonique, et $F = \mathbb{R}$, on a

$$\forall x \in \mathcal{O}, \forall h \in \mathbb{R}^n, \quad f'(x).h = J_f(x)h = \nabla f(x)^T h,$$

avec $\nabla f(x)$ le gradient de f en x .

Ensemble convexe, applications convexes

Ensemble convexe

Définition : Soit E un espace vectoriel normé. L'ensemble $\mathcal{C} \subset E$ est dit **convexe** si

$$\forall (x, y) \in \mathcal{C}^2, \forall \alpha \in [0, 1], \quad \alpha x + (1 - \alpha)y \in \mathcal{C}$$

Remarque : Autrement dit, si $(x, y) \in \mathcal{C}^2$, alors le segment $[x, y]$ est également contenu dans \mathcal{C} .

Application convexe

Définition : Soit E un espace vectoriel normé.

- Une application $f : \mathcal{C} \subset E \rightarrow \mathbb{R}$ est dite **convexe sur le domaine \mathcal{C} convexe** si

$$\forall (x, y) \in \mathcal{C}^2, \forall \alpha \in [0, 1], \quad f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y).$$

- Une application $f : \mathcal{C} \subset E \rightarrow \mathbb{R}$ est dite **strictement convexe sur le domaine \mathcal{C} convexe** si

$$\forall (x, y) \in \mathcal{C}^2, x \neq y, \forall \alpha \in]0, 1[, \quad f(\alpha x + (1 - \alpha)y) < \alpha f(x) + (1 - \alpha)f(y).$$

Proposition : Soit $f : \mathcal{C} \subset \mathbb{R}^n \rightarrow \mathbb{R}$ une application convexe sur l'ouvert convexe \mathcal{C} . Alors f est continue sur \mathcal{C} .

Ensemble convexe, applications convexes

Convexité et dérivée première

Théorème : Soient $\Omega \subset E$ un ouvert de l'espace vectoriel normé E , et $\mathcal{C} \subset \Omega$ un sous-ensemble convexe de Ω . On suppose que l'application $f : \Omega \rightarrow \mathbb{R}$ est différentiable sur Ω . On a alors:

- f est convexe sur \mathcal{C} convexe si et seulement si

$$\forall (x, y) \in \mathcal{C}^2, f(y) - f(x) \geq f'(x) \cdot (y - x).$$

- f est strictement convexe sur \mathcal{C} convexe si et seulement si

$$\forall (x, y) \in \mathcal{C}^2, x \neq y, f(y) - f(x) > f'(x) \cdot (y - x).$$

Interprétation géométrique

Le graphe de l'application convexe f est toujours **au dessus de son plan tangent** en un point quelconque du domaine \mathcal{C} .



Ensemble convexe, applications convexes

Convexité et dérivée seconde

Théorème : Soient $\Omega \subset E$ un ouvert de l'espace vectoriel normé E , et $\mathcal{C} \subset \Omega$ un sous-ensemble convexe de Ω . On suppose que l'application $f : \Omega \rightarrow \mathbb{R}$ est deux fois différentiable sur Ω . On a alors:

- f est convexe sur \mathcal{C} convexe si et seulement si

$$\forall (x, y) \in \mathcal{C}^2, f''(x)(y - x, y - x) \geq 0.$$

- Si $\forall (x, y) \in \mathcal{C}^2, x \neq y, f''(x)(y - x, y - x) > 0$,

alors f est strictement convexe sur \mathcal{C} convexe.

Corollaire : Supposons $\mathcal{C} = E = \mathbb{R}^n$. Sous les mêmes hypothèses, on a $\forall (x, h) \in (\mathbb{R}^n)^2, f''(x)(h, h) = h^T \nabla^2 f(x) h$, avec $\nabla^2 f(x)$ la matrice hessienne de f en x . On a alors :

- f est convexe sur \mathbb{R}^n si et seulement si $\forall x \in \mathbb{R}^n, \nabla^2 f(x)$ est semi-définie positive.
- Si $\forall x \in \mathbb{R}^n, \nabla^2 f(x)$ est définie positive, alors f est strictement convexe sur \mathbb{R}^n .

Outline

- 1 Introduction
 - Quelques exemples
 - Rappels et notations
- 2 Premiers théorèmes d'existence et unicité
- 3 Optimisation sans contrainte
 - Caractérisation des extrema
 - Résolution numérique : quelques algorithmes
 - Garanties de convergence ?
- 4 Vers l'optimisation avec contraintes
 - Caractérisation des extrema
 - Résolution numérique

Existence de solutions - I

Cadre général

On s'intéresse aux problèmes du type

$$(\mathcal{P}) \quad \min_{x \in \mathcal{C}} f(x) \tag{1}$$

avec f une application de \mathbb{R}^n à valeurs dans \mathbb{R} et $\mathcal{C} \subset \mathbb{R}^n$.

Remarques:

- Si $\mathcal{C} = \emptyset$, (\mathcal{P}) n'admet pas de solution.
- Si \mathcal{C} est fini, (\mathcal{P}) admet au moins une solution.

On s'intéresse dans la suite au cas où \mathcal{C} est non vide et admet un nombre infini d'éléments.

\mathcal{C} compact non vide

Théorème

On suppose que \mathcal{C} est une partie compacte non vide de \mathbb{R}^n . On a

f continue sur $\mathcal{C} \Rightarrow (\mathcal{P})$ admet au moins une solution.

\mathcal{C} fermé non vide

Définition

Soit f une fonction de \mathbb{R}^n à valeurs dans \mathbb{R} .

f est coercive si $f(x) \rightarrow +\infty$ quand $\|x\| \rightarrow +\infty$

Théorème

On suppose que \mathcal{C} est une partie fermée non vide de \mathbb{R}^n . On a

f continue sur \mathcal{C} et coercive $\Rightarrow (\mathcal{P})$ admet au moins une solution.

Unicité de la solution : cas convexe

Convexité de f sur \mathcal{C} convexe

Théorème

On suppose que \mathcal{C} est une partie convexe de \mathbb{R}^n . On a

f convexe sur $\mathcal{C} \Rightarrow$ L'ensemble des solutions de (\mathcal{P}) est vide ou convexe.

Stricte convexité de f sur \mathcal{C} convexe

Théorème

On suppose que \mathcal{C} est une partie convexe de \mathbb{R}^n . On a

f strictement convexe sur $\mathcal{C} \Rightarrow (\mathcal{P})$ admet au plus une solution.

Minimum global VS minimum local

Théorème

On suppose que \mathcal{C} est une partie convexe de \mathbb{R}^n . On a

f convexe sur $\mathcal{C} \Rightarrow$ tout minimum local de f sur \mathcal{C} est solution de (\mathcal{P}) .

Outline

- 1 Introduction
 - Quelques exemples
 - Rappels et notations
- 2 Premiers théorèmes d'existence et unicité
- 3 Optimisation sans contrainte
 - Caractérisation des extrema
 - Résolution numérique : quelques algorithmes
 - Garanties de convergence ?
- 4 Vers l'optimisation avec contraintes
 - Caractérisation des extrema
 - Résolution numérique

Définition du problème

Cadre général

On s'intéresse aux problèmes du type

$$(\mathcal{P}_{sc}) \quad \min_{x \in \mathcal{O}} f(x) \quad (2)$$

avec f une application définie sur un ouvert \mathcal{O} de \mathbb{R}^n et à valeurs dans \mathbb{R} .

Minimisation locale

Définition

Soit f une application définie sur un ouvert \mathcal{O} de \mathbb{R}^n . x^* est un minimum local de f si

$$\exists \epsilon > 0 \text{ tel que } \forall x \in \mathcal{B}(x^*, \epsilon), \quad f(x^*) \leq f(x)$$

avec $\mathcal{B}(x^*, \epsilon)$ la boule ouverte de rayon ϵ centrée en x^* .

Conditions nécessaires d'optimalité

Ordre 1

Théorème

Soit $x^* \in \mathcal{O}$. On suppose que f est différentiable en x^* . On a

$$x^* \text{ est un minimum local de } f \Rightarrow \nabla f(x^*) = 0.$$

Définition

On appelle point critique de f tout $x \in \mathcal{O}$ tel que $\nabla f(x) = 0$.

Ordre 2

Théorème

Soit $x^* \in \mathcal{O}$. On suppose que f est deux fois différentiable en x^* . On a

$$x^* \text{ est un minimum local de } f \Rightarrow \nabla^2 f(x^*) \text{ est semi-définie positive.}$$

Remarques

- Ce sont des **conditions nécessaires** !
Ex: $f(x) = x^3$, $f'(0) = 0$ et $f''(0) \geq 0$, mais 0 n'est pas un minimum local de f .
- Ces conditions ne sont pas vraies si \mathcal{O} n'est pas un ouvert (problèmes avec contraintes).

Conditions suffisantes d'optimalité

Ordre 1

Théorème

Soit $x^* \in \mathcal{O}$. On suppose que \mathcal{O} , ouvert de \mathbb{R}^n , est également convexe. Si f est convexe sur \mathcal{O} et différentiable en x^* , on a

$\nabla f(x^*) = 0 \Rightarrow x^*$ est un minimum local de f , et donc un minimum global de f .

Remarque

Dans ce cas particulier, l'équation $\nabla f(x) = 0$ est une CNS d'optimalité.

Ordre 2

Théorème

Soit $x^* \in \mathcal{O}$ solution de $\nabla f(x^*) = 0$. On suppose que f est deux fois différentiable en x^* . On a

- 1 $\nabla^2 f(x^*)$ est définie positive $\Rightarrow x^*$ est un minimum local de f .
- 2 On suppose, de plus, f deux fois différentiable sur \mathcal{O} .
 $\exists \epsilon > 0$ tel que $\mathcal{B}(x^*, \epsilon) \subset \mathcal{O}$ et $\forall x \in \mathcal{B}(x^*, \epsilon)$, $\nabla^2 f(x)$ est semi-définie positive $\Rightarrow x^*$ est un minimum local de f .

Résolution analytique de (\mathcal{P}_{sc})

Stratégie générale

On suppose f différentiable sur \mathcal{O} autant de fois que nécessaire.

- 1 Démonstration de l'existence et éventuelle unicité des solutions de (\mathcal{P}_{sc}) .
- 2 Recherche des points critiques de f :

$$\text{Résoudre } \nabla f(x) = 0.$$

- 3 Arrêt possible dans certains cas particuliers (non exhaustif) :
 - Si f est convexe sur \mathcal{O} convexe : les points critiques étant exactement les solutions de (\mathcal{P}_{sc}) .
 - Un seul point critique et existence/unicité de la solution de (\mathcal{P}_{sc}) démontrées.

- 4 Recherche des minima locaux parmi les points critiques par une étude au second ordre :

Etude des propriétés spectrales (semi-définie positive, définie positive, etc..) de $\nabla^2 f(x)$, $\forall x$ points critique.

- 5 Recherche des solutions de (\mathcal{P}_{sc}) parmi les minima locaux et points critiques "indéterminés".

Résolution analytique de (\mathcal{P}_{sc})

Exemple : minimisation d'une quadratique strictement convexe

On cherche à résoudre le problème suivant :

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2}x^T A x - b^T x + c$$

avec $A \in \mathcal{M}_n(\mathbb{R})$ symétrique définie positive, $b \in \mathbb{R}^n$ et $c \in \mathbb{R}$.

Ce problème admet une unique solution :

- Existence : f continue sur \mathbb{R}^n , fermé non vide, et coercive (A définie positive).
- Unicité : f strictement convexe ($\forall x \in \mathbb{R}^n, \nabla^2 f(x) = A$, définie positive) sur \mathbb{R}^n convexe.

Cette solution x^* est caractérisée par

$$Ax^* = b.$$

▷ Résolution d'un **système linéaire**, éventuellement de très grande dimension.

A la main ???

- Recherche des points critiques : résolution de systèmes d'équations, éventuellement non-linéaires, en grande dimension.
- Recherche des minima locaux parmi les points critiques : étude des valeurs propres de matrices de grande dimension.

Résolution numérique de (\mathcal{P}_{sc}) ?

Difficultés

- Recherche des points critiques : résolution de systèmes d'équations en grande dimension.
 - ▷ Systèmes linéaires : algorithmes de factorisation (LU, Cholesky) ou méthodes itératives (gradient conjugué, steepest descent,...).
 - ▷ Systèmes non-linéaires : méthodes itératives (Newton, gradient conjugué non-linéaire,...).
 - ▷ Coût et temps de calculs ? Précision des solutions ? Convergence des méthodes itératives ? Obtention de tous les points critiques ? etc..
- Recherche des minima locaux parmi les points critiques : étude des valeurs propres de matrices de grande dimension.
 - ▷ Méthodes itératives (QR, puissance itérée).
 - ▷ Coût et temps de calculs ? Précision des solutions ? Convergence des méthodes ? etc..

Conséquences

Dans beaucoup d'applications, on se contentera de trouver un point critique et/ou de faire décroître la fonctionnelle à minimiser..

Algorithmes de descente

Direction de descente

Définition : Soit $x \in \mathcal{O}$. On suppose f différentiable en x .

$d \in \mathbb{R}^n$ est une direction de descente en x si $\nabla f(x)^T d < 0$.

Remarques :

- Par définition, il ne sera pas question de direction de descente en un point critique.
- Soit $x \in \mathcal{O}$. On suppose f différentiable en x et $\nabla f(x) \neq 0$. Alors $d = -\nabla f(x)$ est une direction de descente en x :

$$\nabla f(x)^T d = -\|\nabla f(x)\|^2 < 0.$$

Intérêt ?

Proposition : On suppose f continument différentiable sur \mathcal{O} . Soient $x \in \mathcal{O}$ et $d \in \mathbb{R}^n$. Si d est une direction de descente de f en x , alors

$$\exists \eta > 0 \text{ tel que } \forall \alpha \in]0, \eta], x + \alpha d \in \mathcal{O} \text{ et } f(x + \alpha d) < f(x).$$

- Il est possible de faire localement décroître f en suivant une direction de descente.

Algorithmes de descente

Algorithme de base

1. Initialisation x_0
2. For $k=0,2, \dots$ Do
3. Calcul d'une direction de descente : d_k t.q. $\nabla f(x_k)^T d_k < 0$.
4. Calcul d'une longueur de pas α_k .
5. Mise à jour : $x_{k+1} = x_k + \alpha_k d_k$.
6. Tests d'arrêt.
7. EndDo

Remarques :

- Algorithme de descente de gradient : $d_k = -\nabla f(x_k)$.
- Quelle stratégie pour la recherche des longueurs de pas α_k ?
- Arrêt si :
 - 1 CN1 ($\nabla f(x) = 0$) : $\|\nabla f(x_k)\| \leq \epsilon_1(\|\nabla f(x_0)\| + \eta)$
 - 2 Stagnation des itérés : $\|x_{k+1} - x_k\| \leq \epsilon_2(\|x_k\| + \eta)$
 - 3 Nombre d'itérations maximum atteint.

Fonction quadratique

Steepest descent

On s'intéresse à la résolution du problème :

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2}x^T A x - b^T x + c$$

avec $A \in \mathcal{M}_n(\mathbb{R})$ symétrique définie positive, $b \in \mathbb{R}^n$ et $c \in \mathbb{R}$ par une méthode de descente de gradients. On choisit :

- Direction de descente : **plus grande pente en x_k** .

$$d_k = -\nabla f(x_k) = b - A x_k$$

- Longueur de pas : **pas optimal** $\min_{\alpha} \phi(\alpha) = f(x_k + \alpha d_k)$.

$$\begin{cases} \phi'(\alpha) = \nabla f(x_k + \alpha d_k)^T d_k = 0 \Leftrightarrow \alpha = \frac{d_k^T d_k}{d_k^T A d_k} \\ \phi''(\alpha) = d_k^T \nabla^2 f(x_k + \alpha d_k) d_k = d_k^T A d_k > 0 \text{ si } d_k \neq 0 \end{cases}$$

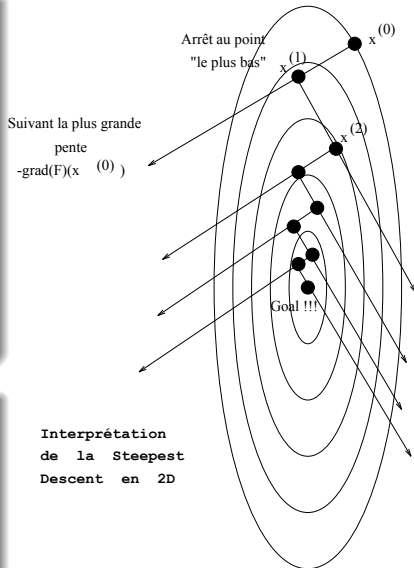
Fonction quadratique

Steepest descent

1. Initialisation x_0
2. For $k=0,2, \dots$ Do
3. Direction de plus grande pente :
 $d_k = b - Ax_k$.
4. Longueur de pas optimale :
$$\alpha_k = \frac{d_k^T d_k}{d_k^T A d_k}.$$
5. $x_{k+1} = x_k + \alpha_k d_k$.
6. Tests d'arrêt.
7. EndDo

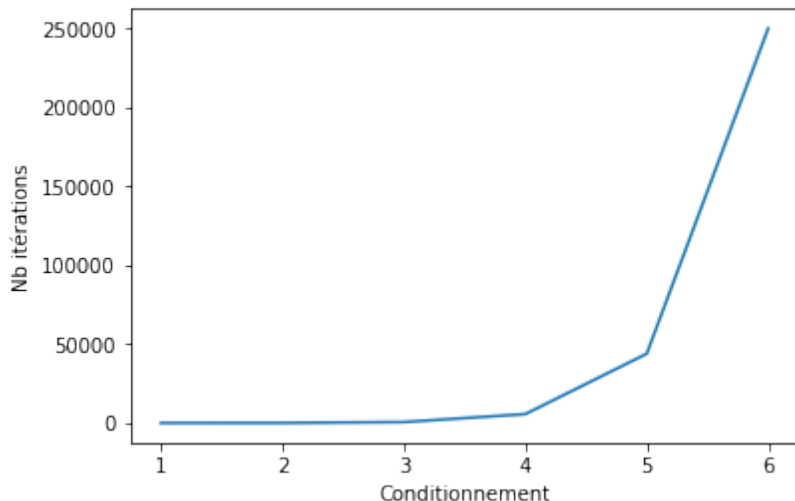
Quelques propriétés

- $\forall k \in \mathbb{N}, d_k \perp d_{k+1}$.
- Si $x^* - x_0 = \beta u$ avec u vecteur propre de A et $\beta \neq 0$, alors convergence en une itération.
- Convergence très lente si $\kappa(A)$ élevé.



Fonction quadratique : steepest descent

Nombre d'itération vs conditionnement de A ($\kappa_2(A) = \frac{|\lambda_{\max}|}{|\lambda_{\min}|}$)



Algorithme

1. Initialisation x_0
2. For $k=0,2, \dots$ Do
3. Si $\nabla^2 f(x_k)$ inversible, calcul de $d_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$.
4. Mise à jour : $x_{k+1} = x_k + d_k$.
4. Tests d'arrêt.
5. EndDo

Remarques :

- La fonction doit être deux fois différentiable et ses dérivées disponibles.
- Mêmes critères d'arrêt que pour les méthodes de descente.
- En pratique, on ne calcule pas l'inverse de $\nabla^2 f(x_k)$:

Résolution du **système linéaire** $\nabla^2 f(x_k) d_k = -\nabla f(x_k)$ à chaque itération.

Méthode de Newton

Interprétations

- ❶ Application de la méthode de Newton pour trouver une racine de l'équation, potentiellement non-linéaire,

$$\nabla f(x) = 0.$$

- ❷ Soit $x_k \in \mathbb{R}^n$. Soit m l'approximation quadratique de f en x_k :

$$\forall x \in \mathbb{R}^n, \quad m(x) = f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2} (x - x_k)^T \nabla^2 f(x_k) (x - x_k).$$

Supposons $\nabla^2 f(x_k)$ définie positive. Alors, x^* le minimum de m est donné par

$$x^* = x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k).$$

▷ Minimisation de l'**approximation quadratique** de f à chaque itération pour laquelle $\nabla^2 f(x_k)$ est définie positive.

- ❸ Si $\nabla^2 f(x_k)$ est définie positive, alors $d_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$ est une direction de descente :

$$d_k^T \nabla f(x_k) = -\nabla f(x_k)^T \nabla^2 f(x_k)^{-1} \nabla f(x_k) < 0$$

▷ **Méthode de descente** à chaque itération pour laquelle $\nabla^2 f(x_k)$ est définie positive.

Méthode de Newton

Convergence

Théorème : Soit $x^* \in \mathcal{O}$, avec \mathcal{O} ouvert convexe. On suppose

h1 f est deux fois continûment différentiable sur \mathcal{O} .

h2 $x \rightarrow \nabla^2 f(x)$ est Lipschitz continue sur \mathcal{O} :

$$\exists \gamma > 0, \forall (x, y) \in \mathcal{O}^2 \quad \|\nabla^2 f(y) - \nabla^2 f(x)\| \leq \gamma \|y - x\|$$

h3 $\nabla f(x^*) = 0$ et $\nabla^2 f(x^*)$ définie positive.

alors $\exists (\delta, K) \in \mathbb{R}_+^2$ tels que

$$\|x^* - x_0\| \leq \delta \Rightarrow \|x^* - x_{k+1}\| \leq K \|x^* - x_k\|^2.$$

Si de plus $K\delta < 1$, alors la suite (x_k) converge vers x^* .

Remarques

- Convergence quadratique : la précision de l'approximation double à chaque itération.
- En pratique, on s'attend à une convergence de l'algorithme au voisinage du minimum local.

Moindres carrés non-linéaires : méthode de Gauss-Newton

Problème

On s'intéresse à la résolution du problème :

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2} \|F(x)\|^2$$

avec $F : \mathbb{R}^n \rightarrow \mathbb{R}^p$ continuement différentiable sur \mathbb{R}^n . On note $J_F(x) \in \mathcal{M}_{p,n}(\mathbb{R})$ la matrice Jacobienne de F évaluée en x .

Algorithme

1. Initialisation x_0
2. For $k=0,2, \dots$ Do
3. Si $J_F(x_k)^T J_F(x_k)$ inversible, calcul de $d_k = -(J_F(x_k)^T J_F(x_k))^{-1} \nabla f(x_k)$.
4. Mise à jour : $x_{k+1} = x_k + d_k$.
4. Tests d'arrêt.
5. EndDo

Remarques :

- Mêmes critères d'arrêt que pour les méthodes de descente.
- De nouveau, on ne calcule pas l'inverse d'une matrice.

Moindres carrés non-linéaires : méthode de Gauss-Newton

Interprétation

- ❶ Linéarisation de F au voisinage de x_k : résolution d'un problème de moindres carrés linéaire.

$$(\mathcal{P}_k) \quad \min_{s \in \mathbb{R}^n} g_k(s) = \frac{1}{2} \|F(x_k) + J_F(x_k)s\|^2$$

- ▷ Minimisation d'une **fonction quadratique convexe** :

$$\forall s \in \mathbb{R}^n, \nabla^2 g_k(s) = J_F(x_k)^T J_F(x_k) \text{ symétrique semi-définie positive.}$$

- ▷ Les minima globaux de g_k sont exactement les solutions du système :

$$J_F(x_k)^T J_F(x_k)s = -\nabla f(x_k).$$

$$\text{avec } \nabla f(x_k) = J_F(x_k)^T F(x_k).$$

- ❷ Si $J_F(x_k)$ est de rang plein, alors $d_k = -(J_F(x_k)^T J_F(x_k))^{-1} \nabla f(x_k)$ est une direction de descente :

$$d_k^T \nabla f(x_k) = -\nabla f(x_k)^T (J_F(x_k)^T J_F(x_k))^{-1} \nabla f(x_k) < 0$$

- ▷ **Méthode de descente** à chaque itération pour laquelle $J_F(x_k)$ est de rang plein.

Moindres carrés non-linéaires : méthode de Gauss-Newton

Convergence

Théorème : Soit $x^* \in \mathcal{O}$, avec \mathcal{O} ouvert convexe. On suppose

h1 f est deux fois continûment différentiable sur \mathcal{O} .

h2 $\nabla f(x^*) = 0$ et $J_F(x^*)$ est de rang plein.

h3 $\rho \left((J_F(x^*)^T J_F(x^*))^{-1} \sum_{i=1}^p F_i(x^*) \nabla^2 F_i(x^*) \right) < 1$.

avec $\rho(A)$ le rayon spectral de la matrice A .

Alors $\exists \delta \in \mathbb{R}_+^*$ tel que $\forall x_0 \in \mathcal{B}(x^*, \delta)$ la suite des itérés (x_k) converge vers x^* .

Remarques

- L'hypothèse (h3) est équivalente à " $\nabla^2 f(x^*)$ est définie positive".
- Convergence linéaire (accélération possible sous certaines conditions supplémentaires).
- En pratique, on s'attend à une convergence de l'algorithme au voisinage du minimum local.

Exemple : convergence de l'algorithme de Newton ?

Moindres carrés non-linéaires

- Estimation des paramètres de l'équation de Michaelis-Menten :

$$V(S) = V_{max} \frac{S}{K_m + S} \quad (3)$$

où V_{max} et K_m sont des **constantes inconnues**, spécifique de l'enzyme utilisée comme catalyseur, S représente la concentration de substrat, et $V(S)$ la vitesse de réaction.

- Observations disponibles des couples $(S_i, V_i)_{1 \leq i \leq p}$.
- Estimation de V_{max} et K_m :

$$\min_{(V_{max}, K_m) \in \mathbb{R}^2} f(V_{max}, K_m) = \frac{1}{2} \sum_{i=1}^p \left(V_i - V_{max} \frac{S_i}{K_m + S_i} \right)^2 \quad (4)$$

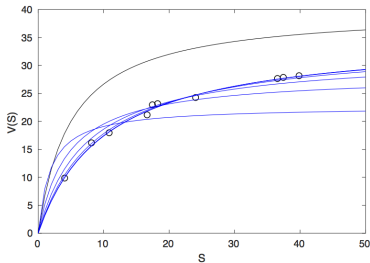
- Application de la méthode de Newton.

Exemple : convergence de l'algorithme de Newton ?

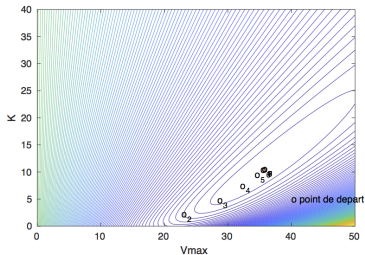
Divergence/convergence selon le point de départ..

Modèle et observations

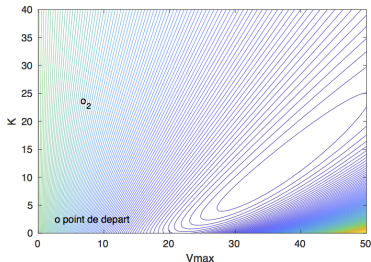
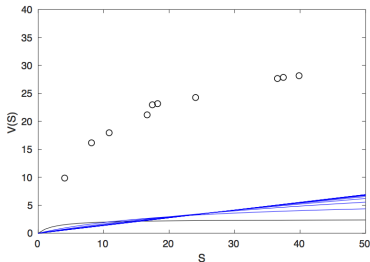
$x_0 = [40, 5]$



Lignes de niveaux de f et itérés



$x_0 = [2.5, 2.5]$



Objectif

Modification des algorithmes de descente afin de garantir une convergence globale :

$\forall x_0 \in \mathcal{O}$, la suite des itérés (x_k) converge vers un point critique de f .

Stratégies classiques

- Recherche de longueurs de pas.
- Algorithme de région de confiance.
- Technique de régularisation.

Recherche de longueurs de pas

Direction de descente

Rappel : On suppose f continument différentiable sur \mathcal{O} . Soient $x \in \mathcal{O}$ et $d \in \mathbb{R}^n$. Si d est une direction de descente de f en x ($\nabla f(x)^T d < 0$), alors

$$\exists \eta > 0 \text{ tel que } \forall \alpha \in]0, \eta], x + \alpha d \in \mathcal{O} \text{ et } f(x + \alpha d) < f(x).$$

Stratégie naive

Calcul de la longueur de pas α tel que $f(x + \alpha d) < f(x)$ avec d direction de descente :

Recherche linéaire basique

1. Initialisation x_0
2. For $k=0, 2, \dots$ Do
3. Calcul d'une direction de descente d_k .
4. Calcul d'une longueur de pas α_k t.q. $f(x_k + \alpha_k d_k) < f(x_k)$.
5. Mise à jour : $x_{k+1} = x_k + \alpha_k d_k$.
7. Tests d'arrêt.
8. EndDo

Recherche de longueurs de pas

Stratégie naive

Calcul de la longueur de pas α tel que $f(x + \alpha d) < f(x)$ avec d direction de descente :

Recherche linéaire basique (et inefficace)

1. Initialisation x_0
2. For $k=0,2, \dots$ Do
3. Calcul d'une direction de descente d_k .
4. Calcul d'une longueur de pas α_k t.q. $f(x_k + \alpha_k d_k) < f(x_k)$.
5. Mise à jour : $x_{k+1} = x_k + \alpha_k d_k$.
7. Tests d'arrêt.
8. EndDo

Exemple : $f(x) = x^2$ avec $x_0 = 2$ et $\forall k \in \mathbb{N}^*, d_k = -1$ et $\alpha_k = 2^{-(k+1)}$

- Par récurrence : $\forall k \in \mathbb{N}^*, x_k = 1 + 2^{-k}$.
- $\forall k \in \mathbb{N}^*, d_k$ est une direction de descente : $f'(x_k)d_k = -2(1 + 2^{-k}) < 0$.
- $\forall k \in \mathbb{N}^*, f(x_k + \alpha_k d_k) - f(x_k) < 0$.
- $\lim_{k \rightarrow \infty} x_k = 1$ et $\lim_{k \rightarrow \infty} f(x_k) = 1$.

▶ Aucune garantie de convergence vers un point critique, voire de convergence !

▶ Conditions supplémentaires sur α_k ..

Recherche de longueurs de pas

Conditions de Wolfe

Définition : Soient $\beta_1 \in]0, 1[$, $\beta_2 \in]\beta_1, 1[$ et d une direction de descente de f en x . Soit $\alpha > 0$. On appelle conditions de Wolfe les deux conditions :

- 1 **Diminution suffisante** : $f(x + \alpha d) \leq f(x) + \beta_1 \alpha \nabla f(x)^T d$,
- 2 **Progrès suffisant** : $\nabla f(x + \alpha d)^T d \geq \beta_2 \nabla f(x)^T d$.

Remarque : dans l'exemple précédent, la condition de progrès suffisant n'est pas vérifiée.

Théorème : Soient $f : \mathbb{R}^n \rightarrow \mathbb{R}$ une fonction différentiable, $x \in \mathbb{R}^n$ et d une direction de descente de f en x . On suppose que f est bornée inférieurement dans la direction d : $\exists c \in \mathbb{R}$ t.q. $\forall \alpha \geq 0, f(x + \alpha d) \geq c$. On a :

- 1 $\forall \beta_1 \in]0, 1[, \exists \eta > 0$ t.q. la première condition de Wolfe soit vérifiée $\forall \alpha \in]0, \eta]$.
- 2 $\forall \beta_1 \in]0, 1[, \forall \beta_2 \in]\beta_1, 1[, \exists \alpha > 0$ vérifiant les deux conditions de Wolfe.

Convergence globale

Théorème : Soit f satisfaisant :

h1 f est continûment différentiable sur \mathbb{R}^n .

h2 f est bornée inférieurement.

h3 $x \rightarrow \nabla f(x)$ est Lipschitz continue sur \mathbb{R}^n :

$$\exists \gamma > 0, \forall (x, y) \in (\mathbb{R}^n)^2 \quad \|\nabla f(y) - \nabla f(x)\| \leq \gamma \|y - x\|$$

On suppose qu'un algorithme de descente soit employé tel que chaque pas vérifie les conditions de Wolfe. Alors soit $\lim_{k \rightarrow \infty} \nabla f(x_k) = 0$, soit $\lim_{k \rightarrow \infty} \frac{\nabla f(x_k)^T d_k}{\|\nabla f(x_k)\| \|d_k\|} = 0$.

Interprétation : Si l'angle entre d_k et $\nabla f(x_k)$ ne converge pas vers $\frac{\pi}{2}$, la limite du gradient de l'itéré est 0, et ce quelque soit x_0 : la suite (x_k) converge vers un point critique de f .

Recherche de longueurs de pas

Longueur de pas optimale

Proposition : Soit d une direction de descente de f en x . Soit ϕ la fonction définie sur \mathbb{R}_+^* par $\phi(\alpha) = f(x + \alpha d)$.

ϕ admet un minimum global α^* sur $\mathbb{R}_+^* \Rightarrow \alpha^*$ vérifie les conditions de Wolfe.

Remarque : la recherche de la longueur de pas optimale requiert la minimisation d'une fonction, ce qui peut s'avérer trop cher à calculer selon les problèmes.

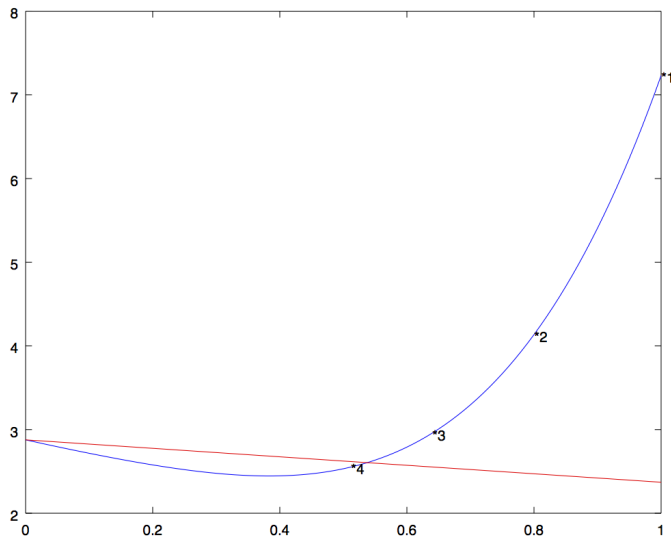
Algorithme de Backtracking

0. Données : x , d direction de descente de f en x , $\beta_1 \in]0, 1[$, $\rho \in]0, 1[$.
1. Initialisation: $\alpha_0 > 0$
2. For $k=0, 2, \dots$ Do
3. Calcul d'une longueur de pas : $\alpha_{k+1} = \rho \alpha_k$.
4. Si α_{k+1} vérifie la première condition de Wolfe : STOP.
8. EndDo

- Cet algorithme fournit un pas satisfaisant la première condition de Wolfe.
 - ▷ Seconde condition de Wolfe non garantie (pas très petits possibles..)
- Intérêts : très simple, uniquement des évaluations de f .

Recherche de longueurs de pas

Algorithme de backtracking : exemple



Recherche de longueurs de pas

Algorithme de recherche linéaire (bisection)

0. Données : x , d direction de descente de f en x , $\beta_1 \in]0, 1[$, $\beta_2 \in]\beta_1, 1[$.
1. Initialisation: $\alpha_0 > 0$, $a = 0, b = \infty$
2. For $k=0, 2, \dots$ Do
3. Si α_k ne satisfait pas la première condition de Wolfe :
$$b = \alpha_k, \alpha_{k+1} = \frac{a+b}{2}.$$
4. Sinon si α_k ne satisfait pas la seconde condition de Wolfe :
$$a = \alpha_k \text{ et } \alpha_{k+1} = \begin{cases} 2a & \text{si } b = \infty \\ \frac{a+b}{2} & \text{sinon} \end{cases}$$
5. Sinon : STOP
6. EndDo

- Cet algorithme fournit un pas satisfaisant les **deux conditions de Wolfe**.
- Interprétation : on commence par faire décroître la longueur de pas jusqu'à satisfaire la première condition de Wolfe. Puis on l'augmente jusqu'à satisfaire la seconde condition de Wolfe.
- Inconvénient : évaluations de ∇f .

Outline

- 1 Introduction
 - Quelques exemples
 - Rappels et notations
- 2 Premiers théorèmes d'existence et unicité
- 3 Optimisation sans contrainte
 - Caractérisation des extrema
 - Résolution numérique : quelques algorithmes
 - Garanties de convergence ?
- 4 Vers l'optimisation avec contraintes
 - Caractérisation des extrema
 - Résolution numérique

Définition du problème

Cadre général

On s'intéresse aux problèmes du type

$$(\mathcal{P}) \quad \min_{x \in \mathcal{C}} f(x) \tag{5}$$

avec f une application de \mathbb{R}^n à valeurs dans \mathbb{R} et $\mathcal{C} \subset \mathbb{R}^n$.

Remarques :

- \mathcal{C} pourra être définie par un système de contraintes d'égalité et/ou d'inégalité :

$$\mathcal{C} = \{x \in \mathbb{R}^n, h(x) = 0 \text{ et } g(x) \leq 0\}$$

où h (resp. g) une application définie de \mathbb{R}^n et à valeurs dans \mathbb{R}^p (resp. \mathbb{R}^m) et

$$g(x) \leq 0 \Leftrightarrow g_i(x) \leq 0 \quad 1 \leq i \leq m.$$

- Si \mathcal{C} est un ouvert de \mathbb{R}^n et f différentiable sur \mathbb{R}^n , alors

$$x^* \text{ solution locale de } (\mathcal{P}) \Rightarrow \nabla f(x^*) = 0.$$

Conditions nécessaires d'optimalité du premier ordre

Cas général : \mathcal{C} quelconque

Définition

Soit x un point d'un ensemble $\mathcal{C} \subset \mathbb{R}^n$. $d \in \mathbb{R}^n$ est une direction tangente à \mathcal{C} en x si il existe une suite de points (x_k) de \mathcal{C} t.q.

$$\forall k \in \mathbb{N}, x_k = x + \alpha_k d_k \in \mathcal{C},$$

avec (d_k) suite de \mathbb{R}^n qui tend vers d et (α_k) suite de réels strictement positifs qui tend vers 0.

Définition

Soit $x \in \mathcal{C}$. On appelle cône tangent à \mathcal{C} en x l'ensemble des directions tangentes à \mathcal{C} en x . On le notera $T(\mathcal{C}, x)$.

Théorème

Soit f différentiable sur \mathbb{R}^n . Si $x^* \in \mathcal{C}$ est une solution locale de (\mathcal{P}) , alors

$$\forall d \in T(\mathcal{C}, x^*), \nabla f(x^*)^T d \geq 0.$$

Conditions nécessaires d'optimalité du premier ordre

Cas particulier : \mathcal{C} convexe

Théorème

Soient f différentiable sur \mathbb{R}^n et \mathcal{C} une partie convexe de \mathbb{R}^n . On a :

- Si $x^* \in \mathcal{C}$ est une solution locale de (\mathcal{P}) , alors

$$\forall x \in \mathcal{C}, \nabla f(x^*)^T (x - x^*) \geq 0.$$

- On suppose de plus f convexe sur \mathcal{C} convexe. On a l'équivalence :

$$x^* \in \mathcal{C} \text{ est une solution locale de } (\mathcal{P}) \Leftrightarrow \forall x \in \mathcal{C}, \nabla f(x^*)^T (x - x^*) \geq 0.$$

Conditions nécessaires d'optimalité du premier ordre

Contraintes d'égalité : $\mathcal{C} = h^{-1}(\{0\})$

On s'intéresse aux problèmes du type

$$(\mathcal{P}_c) \quad \begin{cases} \min f(x) \\ h(x) = 0 \\ x \in \mathbb{R}^n \end{cases} \quad (6)$$

avec h une application définie de \mathbb{R}^n et à valeurs dans \mathbb{R}^p .

Contraintes d'égalité : qualification des contraintes

Définition: Soit $x \in \mathcal{C}$. On suppose h différentiable en x . On appelle cône tangent des contraintes linéarisées en x , noté $T_L(\mathcal{C}, x)$, l'ensemble

$$T_L(\mathcal{C}, x) = \{d \in \mathbb{R}^n, \nabla h(x)^T d = 0\}.$$

Lemme

Soit $x \in \mathcal{C}$. On suppose h différentiable en x . Alors $T(\mathcal{C}, x) \subset T_L(\mathcal{C}, x)$.

Définition: Soit $x \in \mathcal{C}$. On appelle hypothèse de qualification des contraintes en x toute condition suffisante pour avoir $T(\mathcal{C}, x) = T_L(\mathcal{C}, x)$.

► Ex: contraintes linéaires; la famille $(\nabla h_i(x))_{1 \leq i \leq p}$ est libre;..

Conditions nécessaires d'optimalité du premier ordre

Contraintes d'égalité

Définition : On appelle Lagrangien associé au problème (\mathcal{P}) l'application :

$$\begin{aligned} L &: \mathbb{R}^n \times \mathbb{R}^p \rightarrow \mathbb{R} \\ (x, \lambda) &\mapsto f(x) + \lambda^T h(x) \end{aligned}$$

Théorème: Karush-Kuhn-Tucker

On considère le problème (\mathcal{P}_c) . On suppose que

- x^* est une solution locale de (\mathcal{P}_c) vérifiant l'hypothèse de qualification des contraintes,
- f et h sont continûment différentiables au voisinage de x^* .

Alors

$$\exists \lambda^* \in \mathbb{R}^p \text{ t.q. } \begin{cases} \nabla_x L(x^*, \lambda^*) = 0 \\ h(x^*) = 0 \end{cases}$$

Conditions nécessaires d'optimalité du second ordre

Contraintes d'égalité

Théorème: Karush-Kuhn-Tucker

On considère le problème (\mathcal{P}_c) . On suppose que

- x^* est une solution locale de (\mathcal{P}_c) vérifiant l'hypothèse de qualification des contraintes,
- f et h sont deux fois continûment différentiables au voisinage de x^* .

Alors $\exists \lambda^* \in \mathbb{R}^p$ tel que

$$\begin{cases} \nabla_x L(x^*, \lambda^*) = 0 \\ h(x^*) = 0 \end{cases}$$

et

$$\forall d \in T_L(\mathcal{C}, x^*), \quad d^T \nabla_{xx}^2 L(x^*, \lambda^*) d \geq 0.$$

Conditions suffisantes d'optimalité

Ordre 1: contraintes affines et convexité de f

Théorème:

On considère le problème (\mathcal{P}_c) . On suppose que h est affine et f continûment différentiable au voisinage de $x^* \in \mathcal{C}$ et convexe sur \mathcal{C} convexe. Les propositions suivantes sont équivalentes :

- x^* est solution locale de (\mathcal{P}_c) .
- $\exists \lambda^* \in \mathbb{R}^p$ t.q.
$$\begin{cases} \nabla_x L(x^*, \lambda^*) = 0 \\ h(x^*) = 0 \end{cases}$$

Auquel cas x^* est solution globale de (\mathcal{P}_c) .

Ordre 2 : contraintes d'égalité

Théorème:

On considère le problème (\mathcal{P}_c) . On suppose que f et h sont deux fois continûment différentiables sur un ouvert contenant \mathcal{C} . Si $\exists (x^*, \lambda^*) \in \mathbb{R}^n \times \mathbb{R}^p$ tels que

- $$\begin{cases} \nabla_x L(x^*, \lambda^*) = 0 \\ h(x^*) = 0 \end{cases}.$$
- $\forall d \in T_L(\mathcal{C}, x^*), d \neq 0, \quad d^T \nabla_{xx}^2 L(x^*, \lambda^*) d > 0.$

alors x^* est une solution locale de (\mathcal{P}_c) .

Résolution analytique de (\mathcal{P}_c) : contraintes d'égalité

Stratégie générale

On suppose f et h différentiable sur \mathbb{R}^n autant de fois que nécessaire.

❶ Démonstration de l'existence et éventuelle unicité des solutions de (\mathcal{P}_c) .

❷ Recherche des solutions de la CN1 :

$$\text{Résoudre } \begin{cases} \nabla_x L(x, \lambda) = 0 \\ h(x) = 0 \end{cases} \quad \text{et hypothèse de qualification des contraintes en } x.$$

❸ Arrêt possible dans certains cas particuliers.

- Si h est affine et f est convexe sur \mathcal{C} convexe.
- Une seule solution pour la CN1 et existence/unicité de la solution de (\mathcal{P}_c) démontrées.

❹ Recherche des solutions locales parmi les solutions de la CN1 par une étude au second ordre :

$$\text{Etude du signe de } d^T \nabla_{xx}^2 L(x^*, \lambda^*) d, \forall d \in T_L(\mathcal{C}, x^*).$$

❺ Recherche des solutions de (\mathcal{P}_c) parmi les solutions locales et "indéterminés".

Résolution analytique de (\mathcal{P}_c) : contraintes d'égalité

Exemple : minimisation d'une quadratique strictement convexe avec contraintes affines

On cherche à résoudre le problème suivant :

$$\begin{cases} \min f(x) = \frac{1}{2}x^T A x - b^T x + c \\ Cx = d \\ x \in \mathbb{R}^n \end{cases}$$

avec $A \in \mathcal{M}_n(\mathbb{R})$ symétrique définie positive, $C \in \mathcal{M}_{p,n}(\mathbb{R})$ de rang $p \leq n$, $b \in \mathbb{R}^n$, $d \in \mathbb{R}^p$ et $c \in \mathbb{R}$.

Ce problème admet une unique solution :

- Existence : f continue sur \mathbb{R}^n , fermé non vide, et coercive (A définie positive).
- Unicité : f strictement convexe ($\forall x \in \mathbb{R}^n$, $\nabla^2 f(x) = A$, définie positive) sur \mathcal{C} convexe.

Cette solution x^* est caractérisée par

$$\exists \lambda^* \in \mathbb{R}^p \text{ t.q. } \begin{cases} Ax^* + C^T \lambda^* = b. \\ Cx^* = d \end{cases}$$

► Résolution d'un **système linéaire**, éventuellement de très grande dimension.

Contraintes d'égalité

Algorithme du Lagrangien augmenté

0. Données : $\mu_0 > 0$, $\tau > 1$, $\epsilon_0 > 0$, $\eta_0 > 0$, et le point de départ (x_0, λ_0) .
1. For $k=0,1, \dots$ Do
2. Calculer **approximativement** une solution x_k du **problème sans contrainte** :
$$x_{k+1} = \operatorname{argmin}_x L_A(x, \lambda_k, \mu_k) = f(x) + \lambda_k^T h(x) + \frac{\mu_k}{2} \|h(x)\|^2$$

avec x_k comme point de départ, et ϵ_k la tolérance sur le gradient de L_A
3. Test convergence : STOP
4. Si $\|h(x_k)\| \leq \eta_k$, mettre à jour **les multiplicateurs de Lagrange** :
 $\lambda_{k+1} = \lambda_k + \mu_k h(x_k)$
 $\mu_{k+1} = \mu_k$
 Mise à jour des tolérances ϵ_{k+1} et η_{k+1}
5. Sinon, mettre à jour **le paramètre de pénalité** :
 $\lambda_{k+1} = \lambda_k$
 $\mu_{k+1} = \tau \mu_k$
 Mise à jour des tolérances ϵ_{k+1} et η_{k+1}
6. EndDo

Interprétation

Théorème:

On considère le problème (\mathcal{P}_c) . On suppose que f et h sont continûment différentiables sur \mathbb{R}^n . Soit (μ_k) une suite strictement croissante de réels strictement positifs telle que $\lim_{k \rightarrow +\infty} \mu_k = \infty$. Soit (λ_k) une suite bornée de \mathbb{R}^p . Soit (ϵ_k) suite de réels strictement positifs telle que $\lim_{k \rightarrow +\infty} \epsilon_k = 0$. Enfin soit (x_k) une suite de \mathbb{R}^n telle que

$$\|\nabla_x L_A(x_k, \lambda_k, \mu_k)\| \leq \epsilon_k.$$

Soit $(x_{\phi(k)})$ une sous-suite de (x_k) qui converge vers x^* . Si $J_h(x^*)$ est de rang plein, alors

$$\lim_{k \rightarrow +\infty} \lambda_{\phi(k)} + \mu_{\phi(k)} h(x_{\phi(k)}) = \lambda^*$$

avec x^* et λ^* vérifiant

$$\begin{cases} \nabla_x L(x^*, \lambda^*) = 0 \\ h(x^*) = 0 \end{cases}$$

Remarques : Il en résulte le choix de définir $\lambda_{k+1} = \lambda_k + \mu_k h(x_k)$.

- P. Amestoy, P. Berger. Planches du cours d'Algèbre linéaire numérique, ENSEEIHT.
- M. Bierlaire. *Introduction à l'optimisation différentiable*, Presses polytechniques et universitaires romandes, 2006.
- J. Gergaud, S. Gratton, D. Ruiz. *Optimisation numérique : aspects théoriques et algorithmes*, Polycopié du cours d'Optimisation, ENSEEIHT - Sciences du numérique, 2018.
- J. Nocedal, S. Wright. *Numerical Optimization*, Springer Series in Operations Research, 2006.
- Wikistat. *Réseaux de neurones*, <http://wikistat.fr/pdf/st-m-app-rn.pdf>