Deuxième année SN
Introduction à l'optimisation numérique

Joseph Gergaud, Serge Gratton, Daniel Ruiz, Ehouarn Simon

Toulouse INP, IRIT

15 septembre 2022

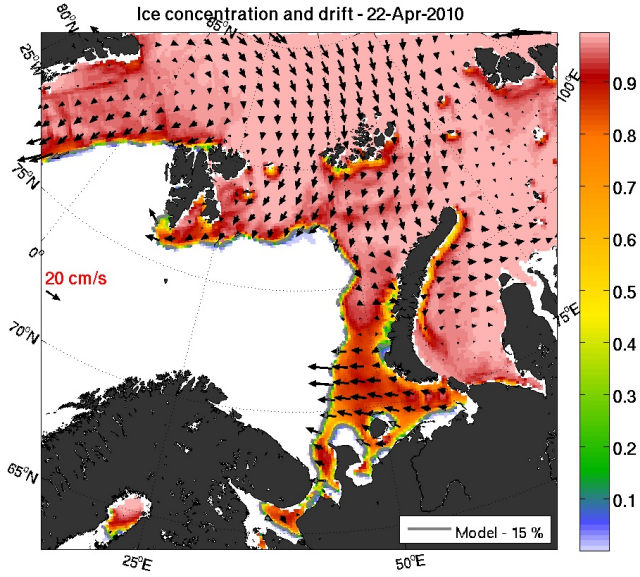
Outline

Outline

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

Modèles numériques

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



Glace dans la Mer de Barents

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.

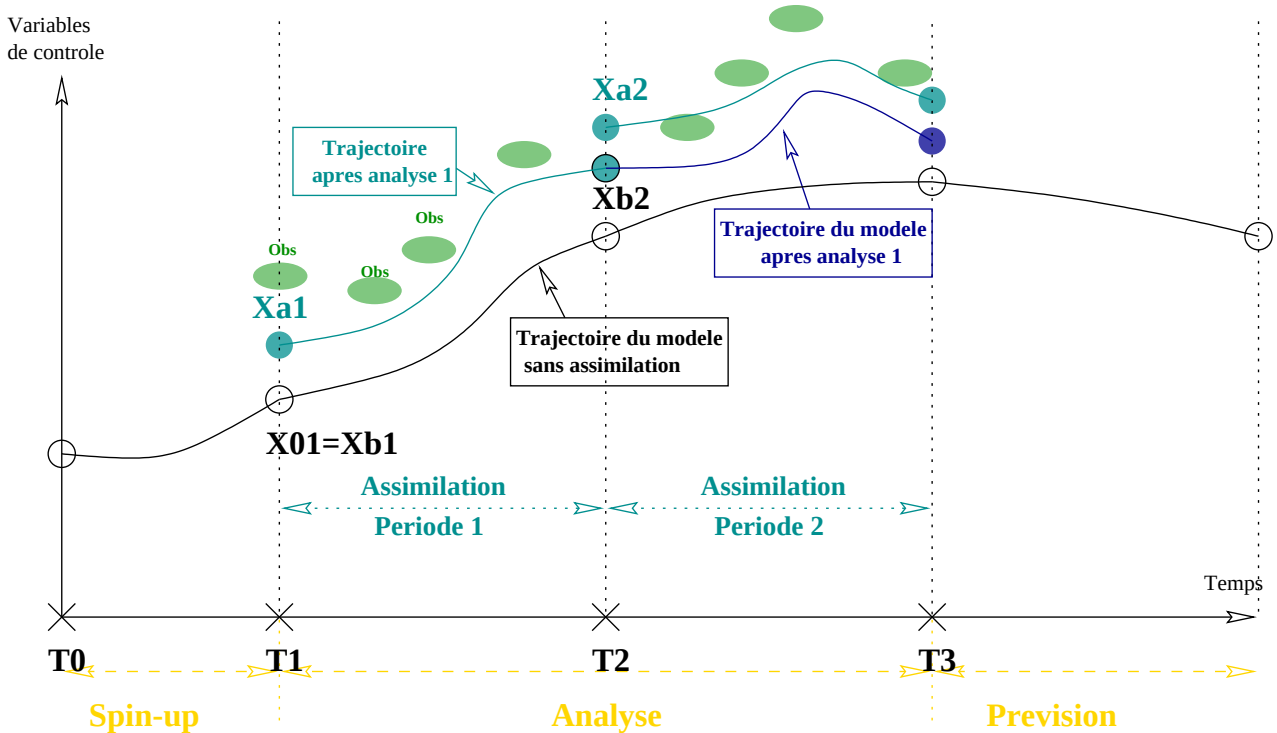
Prévision de la dynamique de l'atmosphère et de l'océan

Minimisation d'une fonctionnelle en très grande dimension

- Algorithme 4D-Var :

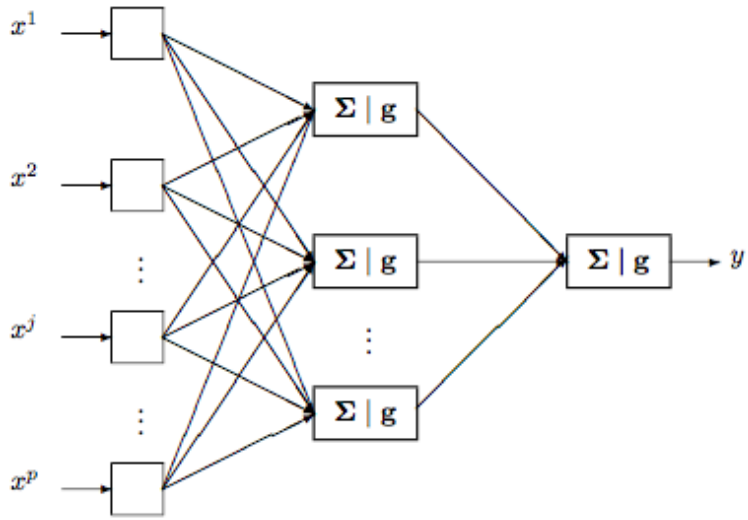
$$J(\mathbf{x_0}) = \underbrace{\frac{1}{2}(\mathbf{x_0} - \mathbf{x}^b)^T \mathbf{B}^{-1}(\mathbf{x_0} - \mathbf{x}^b)}_{\text{écart à l'ébauche}} + \underbrace{\frac{1}{2}\sum_{i=0}^{N}(H_i[M_{0\to i}(\mathbf{x_0})] - \mathbf{y}_i)^T \mathbf{R}_i^{-1}(H_i[M_{0\to i}(\mathbf{x_0})] - \mathbf{y}_i)}_{\text{écarts aux observations}}$$

- Fluides géophysiques : $x_0 \sim 10^9$ variables, $y_i \sim 10^7$ variables.



Apprentissage par réseaux de neurones
Perceptron

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

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

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

Filtrage collaboratif

Système de recommendations

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

$$\min_{P,Q} \sum \left(r_{ij} - q_j^\top p_i\right)^2 + \lambda(||p_i||^2 + ||q_j||^2)$$

Matrice semi- et définie positive

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

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

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 différentiablilité

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

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

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

avec $\exists \lim_{x \to a} \epsilon(x) = 0$. 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

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

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 $f : \mathcal{C} \subset E \to \mathbb{R}$ 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 $f : \mathcal{C} \subset E \to \mathbb{R}$ 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 $f : \mathcal{C} \subset \mathbb{R}^n \to \mathbb{R}$ 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 $f : \Omega \to \mathbb{R}$ est différentiable sur Ω. On a alors:

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

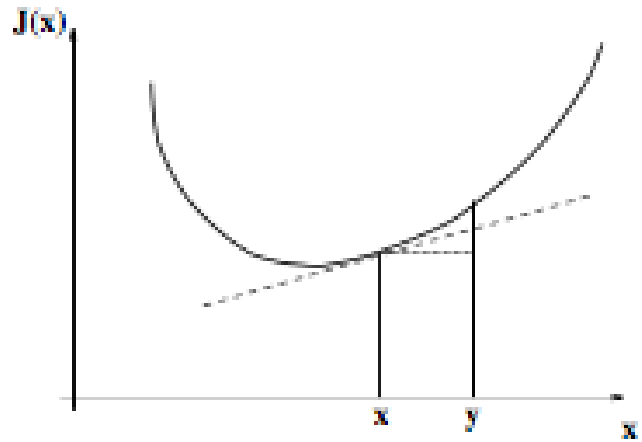
$$\forall (x, y) \in \mathcal{C}^2, f(y) - f(x) \geq f'(x) \cdot (y - x).$$

- 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 géometrique

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 $f : \Omega \to \mathbb{R}$ 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

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'élements.

$\mathcal{C}$ compact non vide

Théorème

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

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

Existence de solutions - II

$\mathcal{C}$ fermé non vide

Définition

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

$$f \text{ est coercive sur } \mathcal{C} \text{ si } f(x) \to +\infty \text{ quand } [\|x\| \to +\infty, x \in \mathcal{C}]$$

Théorème

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

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

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

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

Stricte convexité de f sur $\mathcal{C}$ convexe

Théorème

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

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

Minimum global VS minimum local

Théorème

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

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

Outline

Définition du problème

Cadre général

On s'intéresse aux problèmes du type

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

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. 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 :

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

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

Outline

Définition du problème

Cadre général

On s'intéresse aux problèmes du type

$$(\mathcal{P}) \quad \min_{x \in \mathcal{C}} f(x) \tag{3}$$

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 rééls 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

$$(\mathcal{P}_c) \quad \begin{cases} \min f(x) \\ h(x) = 0 \\ x \in \mathbb{R}^n \end{cases} \tag{4}$$

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

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

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: la famille $(\nabla h_i(x))_{1 \leq i \leq p}$ est libre;..
- Autre : l'application h est affine;

Conditions nécessaires d'optimalité du premier ordre

Contraintes d'égalité

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

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

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

$$\exists \lambda^* \in \mathbb{R}^p \text{ t.q. } \begin{cases} \nabla_x L(x^*, \lambda^*) = 0 \text{ i.e., } \nabla f(x^*) + \sum_{i=1}^{p} \lambda_i^* \nabla h_i(x^*) = 0 \\ h(x^*) = 0 \end{cases}$$

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

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

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 $\exists (x^*, \lambda^*) \in \mathbb{R}^n \times \in \mathbb{R}^p$ tels que

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

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.

1. Démonstration de l'existence et éventuelle unicité des solutions de $(\mathcal{P}_c)$.
2. Recherche des solutions de la CN1 :

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

3. Arrêt possible dans certains cas particuliers.
   - Si h est affine et f est convexe sur $\mathcal{C}$ convexe.
   - Un seule solution pour la CN1 et existence/unicité de la solution de $(\mathcal{P}_c)$ démontrées.
4. Recherche des solutions locales parmi les solutions de la CN1 par une étude au second ordre :

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

5. 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, ... Do
2. Calculer approximativement une solution x_k du problème sans contrainte :

$$x_{k+1} = \text{argmin}_x \quad L_A(x, \lambda_k, \mu_k) = f(x) + \lambda_k^T h(x) + \frac{\mu_k}{2} \|h(x)\|^2$$

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

Contraintes d'égalité

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 $\lim_{k\to+\infty} \mu_k = \infty$. Soit (λ_k) une suite bornée de $\mathbb{R}^p$. Soit (ϵ_k) suite de réels strictement positifs telle que $\lim_{k\to+\infty} \epsilon_k = 0$. 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

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

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 $\lambda_{k+1} = \lambda_k + \mu_k h(x_k)$.

Conditions nécessaires d'optimalité du premier ordre

Contraintes d'égalité det d'inégalité: $\mathcal{C} = h^{-1}(\{0\}) \cap g^{-1}(]-\infty, 0])$

On s'intéresse aux problèmes du type

$$(\mathcal{P}_c) \quad \begin{cases} \min f(x) \\ h(x) = 0 \\ g(x) \leq 0 \\ x \in \mathbb{R}^n \end{cases} \tag{5}$$

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

L'ensemble des contraintes saturées en $\bar{x}$ est

$$J_0(\bar{x}) = \{j \in J = \{1, \ldots, m\}, g_j(\bar{x}) = 0\}. \tag{6}$$

Le cône tangent des contraintes linéarisées en $\bar{x}$ est

$$T_L(C, \bar{x}) = \{d \in \mathbb{R}^n, \nabla h_i(\bar{x})^T d = 0, i = 1 \ldots, m, \\ \nabla g_j(\bar{x})^T d \leq 0, j \in J_0(\bar{x})\}.$$

Conditions nécessaires d'optimalité du premier ordre

Contraintes d'égalité et d'inégalités: qualification des contraintes

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

- Une possibilité: la famille $((\nabla h_i(x))_{1 \leq i \leq p}, (\nabla g_j(x))_{i \in J_{\mathbf{0}}(x)})$ est libre;
- Autre : les applications g et h sont affines;

Conditions nécessaires d'optimalité du premier ordre

Contraintes d'égalité et d'inégalité

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

$$\begin{array}{rccc} L & : & \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m & \to & \mathbb{R} \\ & & (x, \lambda, \mu) & \mapsto & f(x) + \lambda^T h(x) + \mu^T g(x) \end{array}$$

Théorème: Karush-Kuhn-Tucker (KKT1)
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 il existe $(\lambda^*, \mu^*) \in \mathbb{R}^p \times \mathbb{R}^m$ vérifiant

1. $\nabla_x L(x^*, \lambda^*, \mu^*) = 0$;
2. $h(x^*) = 0$;
3. $g(x^*) \leq 0$;
4. $\mu^* \geq 0$;
5. ${\mu^*}^T g(x^*) = 0$, appelée relation de complémentarité.

Conditions nécessaires d'optimalité du second ordre

Contraintes d'égalité et d'inégalité

Théorème: Karush-Kuhn-Tucker (KKT2)
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 il existe des multiplicateurs de Lagrange $\lambda^* = (\lambda_1, \ldots, \lambda_p) \in \mathbb{R}^p$ et $\mu^* \in (\mathbb{R}^+)^m$ tels que

1. KKT1 soit vérifié
2. $(\nabla^2_{xx} L(x^*, \lambda^*) d | d) \geq 0$ pour tout $d \in \mathbb{R}^n$ tel que
   - 2.1 $\nabla h_i(x^*)^T d = 0$ pour tout $i = 1, \ldots, p$;
   - 2.2 $\nabla g_j(x^*)^T d = 0$ pour tout $j \in J_0(x^*)$ et $\mu_j^* > 0$;
   - 2.3 $\nabla g_j(x^*)^T d \leq 0$ pour tout $j \in J_0(x^*)$ et $\mu_j^* = 0$;

Outline

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 décroitre la fonctionnelle à minimiser

Algorithmes de descente

Direction de descente

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

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

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 décroitre f en suivant une direction de descente

Algorithmes de descente

Algorithme de base

1. Initialisation x_0
2. For k=0,2, ... Do
3. Calcul d'une direction de descente : d_k t.q. $\nabla f(x_k)^T dk < 0$.
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.

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

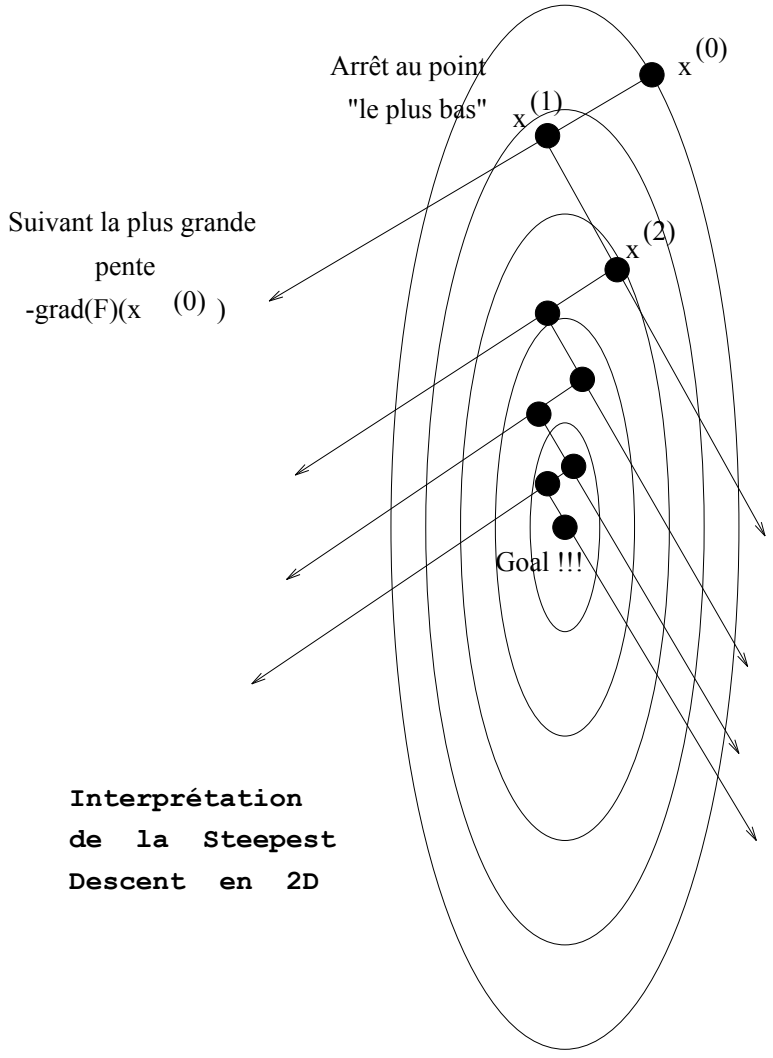
- Longueur de pas : pas optimal $\min_\alpha \phi(\alpha) = f(x_k + \alpha d_k)$.

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

Fonction quadratique

Steepest descent

1. Initialisation x_0
2. For k=0,2, ... 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



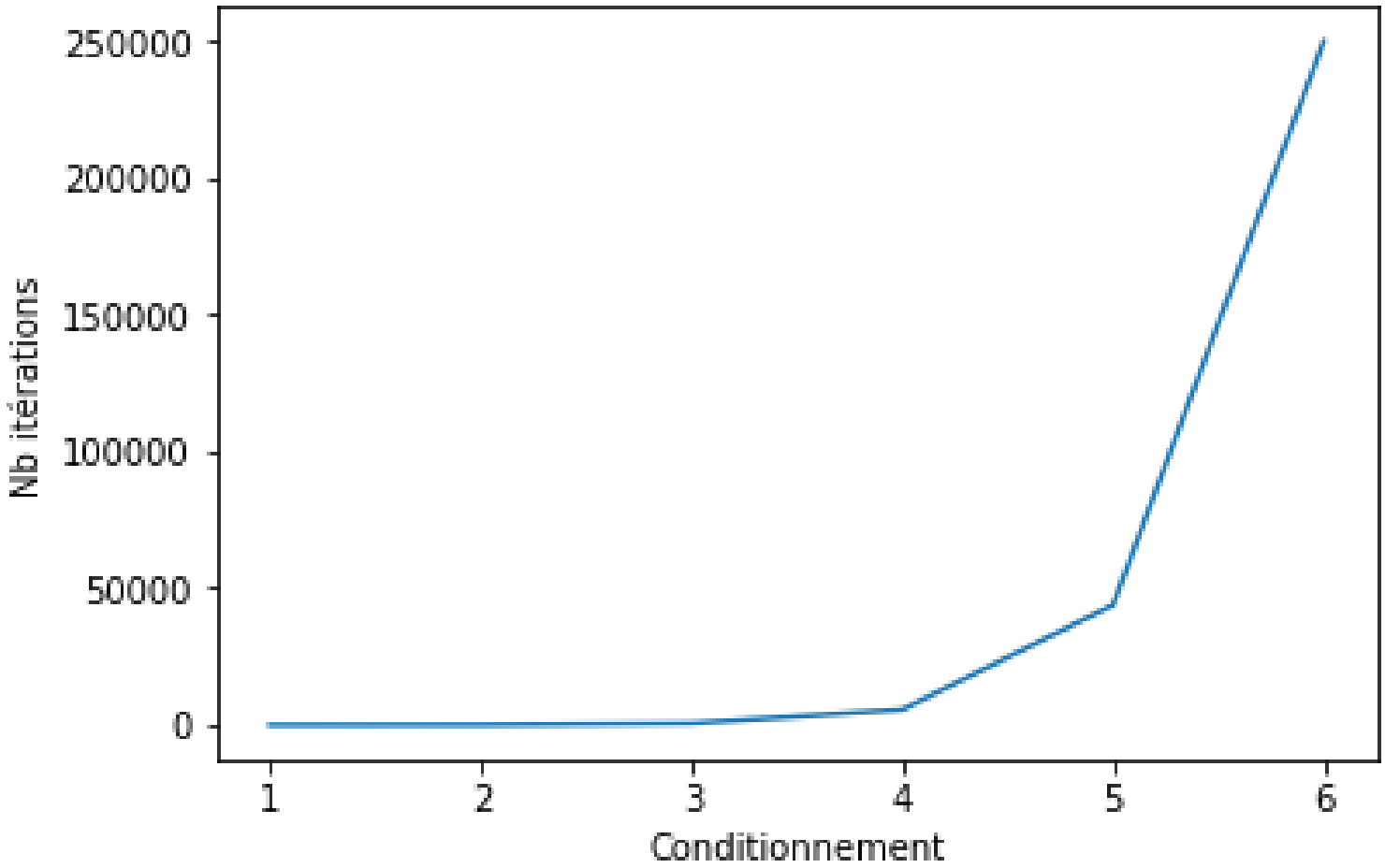
Interprétation de la Steepest Descent en 2D

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.

Fonction quadratique : steepest descent

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



Globalisation des méthodes de descente

Objectif

Modification des algorithmes de descente afin de garantir un 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, ... 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, ... 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 : $f'(x_k)d_k = -2(1+2^{-k}) < 0$.
- $\forall k \in \mathbb{N}^*$, $f(x_k + \alpha_k d_k) - f(x_k) < 0$.
- $\lim_{k\to\infty} x_k = 1$ et $\lim_{k\to\infty} f(x_k) = 1$.

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

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 $f : \mathbb{R}^n \to \mathbb{R}$ 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.

Recherche de longueurs de pas

Convergence globale

Théorème : Soit f satisfaisant :

- h1 f est continûment différentiable sur $\mathbb{R}^n$.
- h2 f est bornée inférieurement.
- h3 $x \to \nabla f(x)$ 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 $\lim_{k\to\infty} \nabla f(x_k) = 0$, soit

$$\lim_{k\to\infty} \frac{\nabla f(x_k)^T d_k}{\|\nabla f(x_k)\| \|d_k\|} = 0.$$

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.

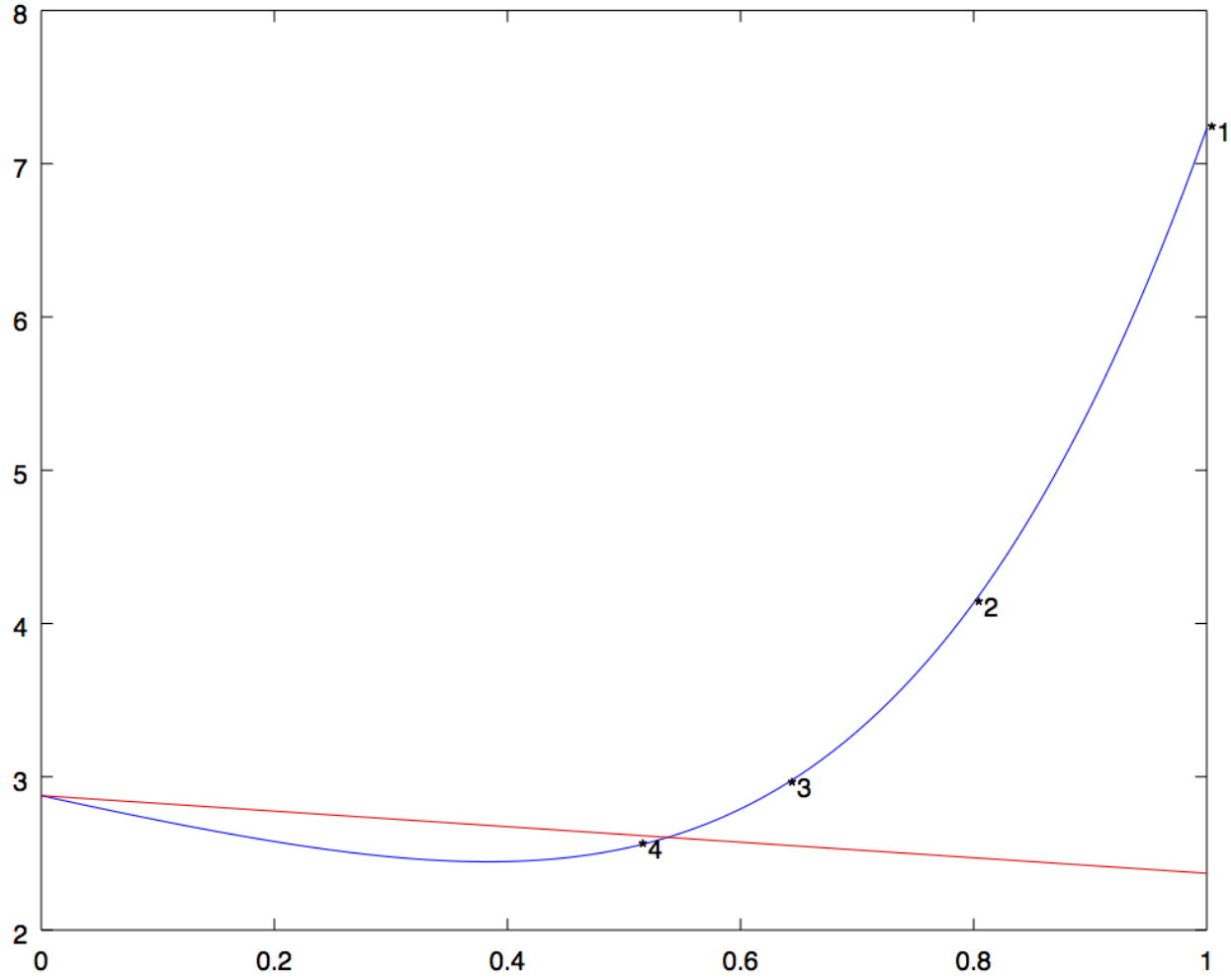
Algorithme de Backtracking

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, ... 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.
- 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 (bissection)

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, ... Do
3. Si α_k ne satisfait pas la première condition de Wolfe :
   $b = \alpha_k$, $\alpha_{k+1} = \dfrac{a+b}{2}$.
4. Sinon si α_k ne satisfait pas le seconde condition de Wolfe :
   $a = \alpha_k$ et $\alpha_{k+1} = \begin{cases} 2a \text{ si } b = \infty \\ \dfrac{a+b}{2} \text{ sinon} \end{cases}$
5. Sinon : STOP
6. EndDo

- Cet algorithme fournit un pas satisfaisant les deux conditions de Wolfe.
- Interprétation : on commence par faire décroitre 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.

Utilisation d'un modèle

Modèle quadratique

Définition : On appelle modèle quadratique de f en x_k une fonction quadratique $m_k(x_k + s)$ telle que $m_k(x_k) = f(x_k)$ et $\nabla m_k(x_k) = \nabla f_k(x_k)$. Il existe alors une matrice $H_k \in \mathbb{R}^{n \times n}$ telle que

$$m_k(x_k + s) = f(x_k) + \nabla f_k(x_k)^T s + \frac{1}{2} s^T H_k s.$$

Définition : Région de confiance. On appelle région de confiance Euclidienne centrée en x_k, de rayon $\Delta_k > 0$ la sphère $\mathcal{B}_k = x_k + \{s, \|s\|_2 \leq \Delta_k\}$.

L'idée de l'algorithme de région de confiance et de résoudre approximativement le problème

$$\min_{x_k + s \in \mathcal{B}_k} m_k(x_k + s).$$

On note $x_{k+1} = x_k + s_k$ le point ainsi obtenu.

Décroissance suffisante

Modèle quadratique

Il est possible de montrer la convergence de l'algorithme si le pas s_k vérifie une condition technique dite de décroissante suffisante.

Définition Condition de décroissante suffisante

$$m_k(x_k) - m_k(x_k + s_k) \geq \kappa_{mdc} \|\nabla m_k(x_k)\|_2 \min\left(\frac{\|\nabla m_k(x_k)\|_2}{\beta_k}, \Delta_k\right), \qquad (7)$$

où $\kappa_{mdc} \in]0, 1[$ et $\beta_k = \|H_k(x)\|_2 + 1$.

Propriété Le point de Cauchy x_k^C qui est, par définition, solution de

$$\begin{cases} \min & m_k(x) \\ t > 0 & \\ x = x_k - t\nabla m(x_k) \in \mathcal{B}_k & \end{cases}$$

vérifie $m_k(x_k) - m_k(x_k^C) \geq \frac{1}{2}\|\nabla m_k(x_k)\|_2 \min\left(\frac{\|\nabla m_k(x_k)\|_2}{\beta_k}, \Delta_k\right)$.

Cout de calcul

Le calcul de x_{k+1} (donc de s_k) est bien moins cher que la résolution du problème initial $\min_x f(x)$ car

1. m_k est une fonction quadratique
2. la décroissance suffisante est obtenue à faible coût, en calculant le point de Cauchy
3. La méthode des régions de confiance a donc un rapport étroit avec la recherche linéaire suivant la direction $-\nabla f_k(x_k)$.
4. Il peut être utile pour réduire le nombre d'itération de continuer à diminuer encore m_k à partir de x_k^C en restant dans la région de confiance. Chaque itération peut devenir coûteuse, et il s'agit de faire un compromis entre le coût et le nombre d'itérations.

Confiance dans le modèle

On introduit le ratio de la réduction observée sur f par rapport à la réduction prédite sur m_k :

$$\rho_k = \frac{f(x_k) - f(x_{k+1})}{m(x_k) - m(x_{k+1})}.$$

Si ρ_k est suffisamment proche de 1, le modèle représente la fonction de manière fiable: la décroissance du modèle est proche de celle de la fonction:

1. on accepte le pas,
2. on augmente éventuellement le rayon de la région de confiance, pour faire des plus grands pas vers la solution.

Si ρ_k est faible, voire négatif, le modèle n'est pas assez fiable:

1. on refuse le pas (dommage, on a calculé pour rien!)
2. on réduit la région de confiance

Notons que pour Δ_k suffisamment petit modèle et fonction sont égaux au premier ordre et donc pour des rayon petits, *rho* va finir par devenir proche de 1.

L'algorithme des régions de confiance

Nous sommes en mesure de présenter à présent l'algorithme des régions de confiance :

Basic trust region algorithm

1. Choose x_0, an initial $\Delta_0 > 0$, and constants $0 < \eta_1 \leq \eta_2 < 1$ and $0 < \gamma_1 \leq \gamma_2 < 1$
2. For k=0,2, ... Do
3. Compute a step s_k that *sufficiently* reduces m_k in $\mathcal{B}_k$.
4. Define $\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m(x_k) - m(x_k + s_k)}$.
5. If $\rho_k \geq \eta_1$ then define $x_{k+1} = x_k + s_k$; otherwise define $x_{k+1} = x_k$
6. Trust region update. Set
   $\Delta_{k+1} \in [\Delta_k, +\infty[$ if $\rho_k \geq \eta_2$ [it. très réussie] or
   $\Delta_{k+1} \in [\gamma_2 \Delta_k, \Delta_k]$ if $\eta_1 \leq \rho_k < \eta_2$ [it. réussie] or
   $\Delta_{k+1} \in [\gamma_1 \Delta_k, \gamma_2 \Delta_k]$ if $\rho_k < \eta_1$ [it. non réussie]
7. If $\|\nabla f(x_k)\|_2 = 0$, exit
8. EndDo

Convergence

Convergence *globale* de l'algorithme

Théorème On suppose que l'algorithme est appliqué à une fonction

- deux fois différentiable,
- bornée inférieurement sur $\mathbb{R}^n$,
- à Hessien borné ($\left\|\nabla^2 f(x)\right\|_2 \leq \kappa_{ufh}$ pour $x \in \mathbb{R}^n$),

et que les modèles m_k sont

- quadratiques,
- ont même valeur et gradient que f en x_k (cohérence au premier ordre)
- ont des Hessien bornés ($\left\|\nabla^2 m(x_k)\right\|_2 \leq \kappa_{umh}$)

alors pour tout x_0, l'algorithme des régions de confiance produit une suite d'itérés telle que $\liminf_{k\to+\infty} \nabla f(x_k) = 0$.

Le déroulement de la preuve

Théorème (Taylor)
Pour tout $s \in \mathbb{R}^n$, $|f(x_k + s) - m(x_k + s)| \leq \frac{1}{2}(\kappa_{ufh} + \kappa_{umh}) \|s\|_2^2$

Théorème (Itérations très réussies si region de confiance assez petite)
Il existe $\kappa_{tr} > 0$ tel que $\Delta_k \leq \kappa_{tr} \|\nabla f(x_k)\|_2 \Longrightarrow \rho_k = \frac{f(x_k) - f(x_k + s_k)}{m(x_k) - m(x_k + s_k)} > \eta_2$.

Théorème (Itérations très réussies si region de confiance assez petite)
Supposons $\exists \epsilon > 0, \forall k, \ \|\nabla f(x_k)\|_2 > \epsilon$. Alors

- Il existe $\kappa_{br} > 0$ tel que $\forall k \ \Delta_k \geq \kappa_{br} \epsilon$
- Le nombre d'itérations réussies est fini. Par conséquent si l'algorithme ne s'arrête pas, le nombre d'itérations non réussies est infini
- Ce dernier point contredit $\forall k, \ \Delta_k \geq \kappa_{br} \epsilon$, car aux itérations non réussies, Δ_k reduit par une constante multiplicative.

En conclusion, soit l'algorithme s'arrête, soit il existe une sous-suite de $(\|\nabla f(x_k)\|_2)$ qui tend vers 0, i.e. $\liminf_{k \to +\infty} \|\nabla f(x_k)\|_2 = 0$.

CG tronqué
Hybridation Cauchy-Newton

Pour obtenir un algorithme plus performant et approcher la convergence locale de l'algorithme de Newton, il convient de choisir un pas s_k qui soit voisin du pas de Newton dans les étapes ultimes de la convergence.

Ceci est réalisé si l'on utilise l'algorithme de gradient conjugué tronqué et si le Hessien du modèle approche celui de la fonction.

Cet algorithme commence par calculer le point de Cauchy puis poursuit la minimisation de la quadratique $m(x_k + s)$ par la méthode des gradients conjugués, en s'arrêtant au premier itéré sortant de la région de confiance $\mathcal{B}_k$.

On a ainsi minimisé davantage $m(x_k + s)$ que $m(x_k^C)$, et donc on a, à la fin de cette procédure de gradient conjugué tronqué, la décroissance suffisante :

$$m(x_k) - m(x_k + s_k) \geq m(x_k) - m(x_k^C) \geq \frac{1}{2} \|\nabla_x m_k(x_k)\|_2 \min\left(\frac{\|\nabla_x m_k(x_k)\|_2}{\beta_k}, \Delta_k \right).$$

CG tronqué

Steihaug Toint Conjugate Gradient algorithm

0. Input parameters: x_k, $\nabla f(x_k)$, $H = \nabla^2 f(x_k)$. Output: s
1. Compute $s_0 = 0$, $g_0 = \nabla f(x_k)$, $p_0 = -g_0$
2. For k=0,2, ... Do
3. $\kappa_k = p_k^T H p_k$
4. If $\kappa_k \leq 0$, then
   compute σ_k as the positive root of $\|s_k + \sigma p_k\|_2 = \Delta_k$
   $s_{k+1} = s_k + \sigma_k p_k$ and stop.
   End If
5. $\alpha_k = r_k^T r_k / \kappa_k$
6. If $\|s_k + \alpha_k p_k\|_2 \geq \Delta_k$, then
   compute σ_k as the positive root of $\|s_k + \sigma p_k\|_2 = \Delta_k$
   $s_{k+1} = s_k + \sigma_k p_k$ and stop.
   End If
4. $s_{k+1} = s_k + \alpha_k p_k$
5. $g_{k+1} = g_k + \alpha_k H p_k$
7. $\beta_k = g_{k+1}^T g_{k+1} / g_k^T g_k$
8. $p_{k+1} = -g_{k+1} + \beta_k p_k$
9. if converged then stop
10. EndDo

en pratique...

Comportement typique

Dans le cas où la convergence a lieu vers un point x^* où le Hessien est défini positif et si $\nabla^2 m_k(x_k) \sim \nabla^2 f_k(x_k)$, le comportement typique de l'algorithme est alors le suivant :

1. les pas deviennent de plus en plus petits (on converge),
2. comme le modèle et la fonction sont cohérents au premier ordre, ρ_k devient proche de 1,
3. la région de confiance a un rayon qui augmente,
4. l'algorithme des gradients conjugués ne rencontre plus le bord de la région de confiance,
5. les gradient conjugués résolvent alors le système $\nabla^2 f(x_k)s_k + \nabla f(x_k) = 0$ ce qui correspond bien à la méthode de Newton, qui a une convergence locale quadratique.

Bibliographie

- 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`