UFR
DE MATHÉMATIQUES
ET INFORMATIQUE

UNIVERSITÉ
PARIS
DESCARTES

Optimisation (MML1E31)
Notes de cours

Master 1 Mathématiques et Modélisation (MM)

2017-2018

Bruno GALERNE
Bureau 812-F
bruno.galerne@parisdescartes.fr

Table des matières

Introduction

Ce cours est une introduction aux problèmes d'optimisation. Le cours se focalise essentiellement sur des problèmes d'optimisation sans contrainte en dimension finie. Après une introduction des différentes notions mathématiques nécessaires (rappels de calcul différentiel, conditions d'optimalité, convexité, etc.), une part importante est donnée à l'exposition des différents algorithmes classiques d'optimisation, l'étude théorique de leur convergence, ainsi que la mise en œuvre pratique de ces algorithmes. Le logiciel libre de calcul scientiffique **Octave** sera utilisé en séance de Travaux Pratiques (TP).

Octave est téléchargeable gratuitement ici :
`https://www.gnu.org/software/octave/`

Les principaux ouvrages de référence pour ce cours sont :

[CIARLET] Philippe G. CIARLET, *Introduction à l'analyse numérique matricielle et à l'optimisation*, cinquième édition, Dunod, 1998

[BOYD & VANDENBERGHE] Stephen BOYD and Lieven VANDENBERGHE *Convex Optimization*, Cambridge University Press, 2004.
Ouvrage téléchargeable gratuitement ici :
`http://stanford.edu/~boyd/cvxbook/`

[ALLAIRE & KABER] Grégoire ALLAIRE et Sidi Mahmoud KABER, *Algèbre linéaire numérique*, Ellipses, 2002

La page web dédiée à ce cours est ici :
`http://w3.mi.parisdescartes.fr/~bgalerne/m1_optimisation/`

Chapitre 1

Rappels et compléments de calculs différentiels

Les références principales pour ce chapitre sont le chapitre A.4 de [BOYD & VANDENBERGHE] et le chapitre 7 de [CIARLET]. Dans ce cours on se placera toujours sur des espaces vectoriels normés de dimensions finis que l'on identifie à $\mathbb{R}^n$, $n \geqslant 1$.

1.1 Cadre et notation

n et m sont des entiers supérieurs ou égaux à 1. Par convention les vecteurs de $\mathbb{R}^n$ sont des vecteurs colonnes. On note $\langle \cdot, \cdot \rangle$ le produit scalaire canonique et $\| \cdot \|$ la norme euclidienne associée. On note $\mathcal{M}_{m,n}(\mathbb{R})$ l'ensemble des matrices de taille $m \times n$ à coefficients réelles et $\mathcal{M}_n(\mathbb{R}) = \mathcal{M}_{n,n}(\mathbb{R})$ l'ensemble des matrices carrées de taille $n \times n$. La transposée d'une matrice A est notée A^T. On a donc pour tous $x, y \in \mathbb{R}^n$, $\langle x, y \rangle = x^T y$ et par conséquent, pour tout $A \in \mathcal{M}_{m,n}(\mathbb{R})$, $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$,

$$\langle y, Ax \rangle = \langle A^T y, x \rangle.$$

Remarque (Notation de la transposée)**.** La notation A^T correspond plutôt à une convention anglo-saxonne. Elle a été choisie pour ce polycopié car elle est plus simple à taper en LaTeXet est plus proche de l'opération transposée en octave, notée `A'` . Toutefois les étudiants sont libres d'utiliser la notation classique ${}^tA$ pour leurs prises notes et leurs copies d'examen.

1.2 Différentielle et gradient

1.2.1 Applications linéaires et matrices associées

On désigne par $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ l'ensemble des applications linéaires de $\mathbb{R}^n$ dans $\mathbb{R}^m$. On identifie un élément de $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ à une matrice rectangulaire de taille $m \times n$ correspondant à la matrice de l'application dans les bases canoniques de $\mathbb{R}^n$ et $\mathbb{R}^m$: si $\varphi \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ alors pour tout $x \in \mathbb{R}^n$, $\varphi(x) = Ax$ avec A la matrice dont les colonnes sont les images par φ des vecteurs de la base canonique $(e_1, \ldots, e_n)$ de la base canonique de $\mathbb{R}^n$,

$$\varphi(x) = \varphi\left(\sum_{k=1}^{n} x_k e_k\right) = \sum_{k=1}^{n} x_k \varphi(e_k) = \begin{pmatrix} \varphi(e_1) & \cdots & \varphi(e_n) \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = Ax.$$

1.2.2 Différentielle

Dans tout ce chapitre, Ω désigne un *ensemble ouvert* de $\mathbb{R}^n$.

Définition 1.1. Soit $f : \Omega \to \mathbb{R}^m$. La fonction f est *différentiable au point* $x \in \Omega$ si il existe une matrice $\mathrm{d}f(x) \in \mathcal{M}_{m,n}(\mathbb{R})$ telle que au voisinage de x on ait

$$f(y) = f(x) + \mathrm{d}f(x)(y-x) + \|y-x\|\varepsilon(y-x)$$

avec $\lim_{y\to x} \varepsilon(y-x) = 0$, *i.e.*, $\|y-x\|\varepsilon(\|y-x\|) = \underset{y\to x}{\mathrm{o}}(\|y-x\|)$. On appelle $\mathrm{d}f(x)$ la différentielle de f au point x, ou encore la matrice jacobienne de f au point x. On dit que f est *différentiable* si f est différentiable en tout point de Ω. On dit que f est *continûment différentiable* si f est différentiable et l'application $x \mapsto \mathrm{d}f(x)$ est continue.

La fonction affine $f(y) = f(x) + \mathrm{d}f(x)(y-x)$ est l'approximation à l'ordre 1 de f au point x. La différentielle peut être calculée à partir des dérivées partielles des composantes de f : Si on note $f = (f_1, \dots, f_m)^T$ les composantes de f, alors pour tout $(i,j) \in \{1,\dots,m\} \times \{1,\dots,n\}$,

$$(\mathrm{d}f(x))_{i,j} = \frac{\partial f_i}{\partial x_j}(x).$$

Exercice 1.2. Soit A une matrice de $\mathcal{M}_{m,n}(\mathbb{R})$ et $b \in \mathbb{R}^m$. Montrer que la fonction $f : \mathbb{R}^n \to \mathbb{R}^m$ définie par $f(x) = Ax + b$ est différentiable sur $\mathbb{R}^n$ et calculer sa différentielle en tout point $x \in \mathbb{R}^n$. En déduire que f est $\mathcal{C}^1$ (et même $\mathcal{C}^\infty$) sur $\mathbb{R}^n$.

Solution de l'exercice 1.2. Pour tous $x, h \in \mathbb{R}^n$,

$$f(x+h) = A(x+h) + b = f(x) + Ah$$

donc f est différentiable en x et $\mathrm{d}f(x) = A$. Comme $x \mapsto \mathrm{d}f(x)$ est constante, c'est une application $\mathcal{C}^\infty$ donc f est aussi $\mathcal{C}^\infty$.

1.2.3 Gradient

Dans ce cours on s'intéressera plus particulièrement à des fonctions à valeurs réelles, ce qui correspond au cas $m = 1$. La matrice jacobienne de $f : \Omega \to \mathbb{R}$ est alors une matrice ligne de taille $1 \times n$. La transposée de cette matrice est un vecteur de $\mathbb{R}^n$ appelé *gradient* de f au point x et noté $\nabla f(x)$. Pour tout $h \in \mathbb{R}^n$,

$$\mathrm{d}f(x)h = \nabla f(x)^T h = \langle \nabla f(x), h \rangle.$$

Ainsi, pour $f : \Omega \to \mathbb{R}$, f est différentiable si et seulement si il existe un vecteur $\nabla f(x)$ tel que

$$f(y) = f(x) + \langle \nabla f(x), y-x \rangle + \|y-x\|\varepsilon(y-x) \quad \text{avec } \lim_{y\to x} \varepsilon(y-x) = 0.$$

$\nabla f(x)$ s'interprète comme le vecteur de plus forte augmentation de f au voisinage de x. En particulier, $\nabla f(x)$ est orthogonal au ligne de niveaux de la fonction f.

Exercice 1.3. Soit A une matrice de $\mathcal{M}_n(\mathbb{R})$. Soit $f : \mathbb{R}^n \to \mathbb{R}$ l'application $f : x \mapsto \frac{1}{2}\langle Ax, x \rangle$.

1. Montrer que f est différentiable sur $\mathbb{R}^n$ et déterminer $\nabla f(x)$ pour tout x.
2. Quelle est l'expression de ∇f si A est symétrique ?

3. Quel est le gradient de l'application $x \mapsto \frac{1}{2}\|x\|^2$?

Solution de l'exercice 1.3.

1. Pour tous $x, h \in \mathbb{R}^n$,

$$\begin{aligned} f(x+h) &= \frac{1}{2}\langle A(x+h), x+h\rangle \\ &= \frac{1}{2}\langle Ax, x\rangle + \frac{1}{2}\langle Ax, h\rangle + \frac{1}{2}\langle Ah, x\rangle + \frac{1}{2}\langle Ah, h\rangle \quad = f(x) + \frac{1}{2}\langle (A+A^T)x, h\rangle + \frac{1}{2}\langle Ah, h\rangle. \end{aligned}$$

   Or par Cauchy-Shwarz, $|\frac{1}{2}\langle Ah, h\rangle| \leqslant \frac{1}{2}\|Ah\|\|h\| \leqslant \frac{1}{2}\|A\|_{\mathcal{M}_n(\mathbb{R})}\|h\|^2$, donc c'est bien en o de $\|h\|$. D'où f est différentiable en tout point $x \in \mathbb{R}^n$ et

$$\nabla f(x) = \frac{1}{2}(A + A^T)x.$$

2. Si A est symétrique, alors $\nabla f(x) = \frac{1}{2}(A + A^T)x = Ax$.
3. C'est le cas particulier où $A = I_n$ qui est symétrique, donc $\nabla f(x) = x$.

1.2.4 Dérivation des fonctions composées

Théorème 1.4 (Dérivation des fonctions composées)**.** Soient $f : \mathbb{R}^n \to \mathbb{R}^m$ et $g : \mathbb{R}^m \to \mathbb{R}^p$ deux fonctions différentiables. Soit $h = g \circ f : \mathbb{R}^n \to \mathbb{R}^p$ la fonction composée définie par $h(x) = g(f(x))$. Alors h est différentiable sur $\mathbb{R}^n$ et pour tout $x \in \mathbb{R}^n$,

$$\mathrm{d}h(x) = \mathrm{d}g(f(x))\mathrm{d}f(x).$$

Remarque. On peut énoncer une version locale du résultat précédent car, comme le suggère la formule $\mathrm{d}h(x) = \mathrm{d}g(f(x))\mathrm{d}f(x)$, pour que h soit différentiable en x, il suffit que f soit différentiable en x et que g soit différentiable en $f(x)$.

Exemple 1.5. Déterminons le gradient de l'application $g : \mathbb{R}^n \to \mathbb{R}$ définie par

$$g(x) = f(Ax + b)$$

où A est une matrice de $\mathcal{M}_{m,n}(\mathbb{R})$, $b \in \mathbb{R}^m$ et $f : \mathbb{R}^m \to \mathbb{R}$ est une application différentiable. On a $g(x) = f \circ h(x)$ avec $h(x) = Ax + b$. Comme h est affine on a $\mathrm{d}h(x) = A$ en tout point x. On a donc d'après la règle de dérivation des fonctions composées

$$\mathrm{d}g(x) = \mathrm{d}f(h(x))\mathrm{d}h(x) = \mathrm{d}f(Ax + b)A.$$

Donc $\nabla g(x) = \mathrm{d}g(x)^T = A^T\mathrm{d}f(Ax + b)^T = A^T\nabla f(Ax + b)$.

1.3 Différentielle d'ordre deux et matrice hessienne

Dans le cadre général où $f : \Omega \to \mathbb{R}^m$, si f est différentiable alors l'application différentielle $\mathrm{d}f : x \mapsto \mathrm{d}f(x)$ est une application de l'ouvert Ω vers l'espace vectoriel $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$. Si cette application est elle-même différentiable en x, alors on obtient une différentielle $\mathrm{d}(\mathrm{d}f)(x)(\cdot)$ qui appartient à $\mathcal{L}(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m))$ que l'on identifie à une application bilinéaire $\mathrm{d}^2f(x) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$ qui est symétrique d'après le théorème de Schwarz. $\mathrm{d}^2f(x)$ est appelée la

différentielle d'ordre deux de l'application f au point x. f est *deux fois différentiable* si elle est différentiable sur tout Ω. f est *deux fois continûment différentiable* sur Ω, si f est deux fois différentiable et si l'application $x \mapsto \mathrm{d}^2 f(x)$ est continue. On note $\mathcal{C}^2(\Omega)$ l'ensemble des fonctions deux fois continûment différentiables.

Dans le cas où $m = 1$, c'est-à-dire où $f : \Omega \to \mathbb{R}$ est à valeurs réelles, $\mathrm{d}^2 f(x)$ est une forme bilinéaire symétrique dont la matrice s'écrit

$$\nabla^2 f(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x)\right)_{1 \leqslant i,j \leqslant n}.$$

Cette matrice est appelée *matrice hessienne* de f au point x. On a alors pour tous vecteurs $h, k \in \mathbb{R}^n$,

$$\mathrm{d}^2 f(x)(h,k) = \langle \nabla^2 f(x)h, k\rangle = k^T \nabla^2 f(x) h = h^T \nabla^2 f(x) k.$$

Pour ce cours on aura constamment besoin de calculer le gradient et la matrice hessienne de fonctionnelles $f : \Omega \to \mathbb{R}$ deux fois différentiables. En pratique on utilise la proposition suivante.

Proposition 1.6 (La matrice hessienne est la différentielle du gradient)**.** Soit $f : \Omega \to \mathbb{R}$ une fonction différentiable sur Ω et deux fois différentiable au point $x \in \Omega$. Alors, la matrice hessienne $\nabla^2 f(x)$ de f au point x est la différentielle de l'application gradient $x \mapsto \nabla f(x)$ au point x.

Exercice 1.7. Démontrer la Proposition 1.6 ci-dessus.

Solution de l'exercice 1.7. En explicitant avec les dérivées partielles, la différentielle est la matrice

$$\left(\frac{\partial}{\partial x_j}\left(\nabla f(x)\right)_i\right)_{1 \leqslant i,j \leqslant n} = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x)\right)_{1 \leqslant i,j \leqslant n} = \nabla^2 f(x).$$

Il est difficile de donner une règle de dérivation des fonctions composées pour l'ordre deux. Voici toutefois deux règles à connaître pour ce cours.

Composition avec une fonction scalaire : On considère $f : \Omega \to \mathbb{R}$ une fonction deux fois différentiable et $g : \mathbb{R} \to \mathbb{R}$ une fonction deux fois dérivable. Alors, $h = g \circ f$ est deux fois différentiable et

$$\nabla^2 h(x) = g'(f(x))\nabla^2 f(x) + g''(f(x))\nabla f(x)\nabla f(x)^T.$$

Composition avec une fonction affine : Soit $g : \mathbb{R}^n \to \mathbb{R}$ définie par

$$g(x) = f(Ax + b)$$

où A est une matrice de $\mathcal{M}_{m,n}(\mathbb{R})$, $b \in \mathbb{R}^m$ et $f : \mathbb{R}^m \to \mathbb{R}$ est une application deux fois différentiable. Alors g est deux fois différentiable et

$$\nabla^2 g(x) = A^T \nabla^2 f(Ax + b) A,$$

formule qui s'obtient facilement en dérivant l'expression $\nabla g(x) = A^T \nabla f(Ax + b)$ montrée précédemment.

1.4 Formules de Taylor

Les formules de Taylor se généralisent aux fonctions de plusieurs variables. On se limite aux fonctions à valeurs réelles.

Théorème 1.8 (Formules de Taylor pour les fonctions une fois dérivable)**.** Soit $f : \Omega \to \mathbb{R}$ une fonction.

(a) Définition de la différentielle = Formule de Taylor-Young à l'ordre 1 : Si f est différentiable en $x \in \Omega$, alors

$$f(x+h) = f(x) + \langle \nabla f(x), h\rangle + \|h\|\varepsilon(h) \quad \text{avec } \lim_{h\to 0} \varepsilon(h) = 0.$$

On considère maintenant un point h fixé tel que le segment $[x, x+h]$ soit inclus dans Ω.

(b) Formule des accroissements finis : Si f est continue sur Ω et différentiable sur $]x, x+h[$, alors

$$|f(x+h) - f(x)| \leqslant \sup_{y\in]x,x+h[} \|\nabla f(y)\|\|h\|.$$

(c) Formule de Taylor-Maclaurin : Si f est continue sur Ω et différentiable sur $]x, x+h[$, alors il existe $\theta \in]0,1[$ tel que

$$f(x+h) = f(x) + \langle \nabla f(x+\theta h), h\rangle.$$

(d) Formule de Taylor avec reste intégral : Si $f \in \mathcal{C}^1(\Omega)$ alors

$$f(x+h) = f(x) + \int_0^1 \langle \nabla f(x+th), h\rangle dt.$$

Preuve. On applique les formules de Taylor à la fonction $\varphi(t) = f(x+th)$, $t \in [0,1]$. □

Théorème 1.9 (Formules de Taylor pour les fonctions deux fois dérivable)**.** Soit $f : \Omega \to \mathbb{R}$ une fonction.

(a) Formule de Taylor-Young à l'ordre 2 : Si f est différentiable dans Ω et deux fois différentiable en $x \in \Omega$, alors

$$f(x+h) = f(x) + \langle \nabla f(x), h\rangle + \frac{1}{2}\langle \nabla^2 f(x)h, h\rangle + \|h\|^2\varepsilon(h) \quad \text{avec } \lim_{h\to 0} \varepsilon(h) = 0.$$

On considère maintenant un point h fixé tel que le segment $[x, x+h]$ soit inclus dans Ω.

(b) Formule des accroissements finis généralisée : Si $f \in \mathcal{C}^1(\Omega)$ et f est deux fois différentiable sur $]x, x+h[$, alors

$$|f(x+h) - f(x) - \langle \nabla f(x), h\rangle| \leqslant \frac{1}{2} \sup_{y\in]x,x+h[} \|\nabla^2 f(y)\|_{\mathcal{M}_n(\mathbb{R})}\|h\|^2.$$

où $\|\cdot\|_{\mathcal{M}_n(\mathbb{R})}$ désigne la norme subordonnée des matrices pour la norme euclidienne.

(c) Formule de Taylor-Maclaurin : Si $f \in \mathcal{C}^1(\Omega)$ et f est deux fois différentiable sur $]x, x+h[$, alors il existe $\theta \in]0,1[$ tel que

$$f(x+h) = f(x) + \langle \nabla f(x), h\rangle + \frac{1}{2}\langle \nabla^2 f(x+\theta h)h, h\rangle.$$

(d) Formule de Taylor avec reste intégral : Si $f \in \mathcal{C}^2(\Omega)$ alors

$$f(x+h) = f(x) + \langle \nabla f(x), h\rangle + \int_0^1 (1-t)\langle \nabla^2 f(x+th)h, h\rangle dt.$$

Chapitre 2

Problèmes d'optimisation : Existence et unicité des solutions

La référence principale pour ce chapitre est le chapitre 8 de [CIARLET].

2.1 Cadre et vocabulaire

On appelle *problème d'optimisation* tout problème de la forme

$$\text{Trouver } x^\star \text{ tel que } x^\star \in U \text{ et } \quad f(x^\star) = \min_{x \in U} f(x),$$

où U est une partie donnée de $\mathbb{R}^n$ et $f : \mathbb{R}^n \to \mathbb{R}$ est une fonction donnée que l'on appelle *fonctionnelle* du problème d'optimisation. Le but de l'optimisation est de proposer des algorithmes permettant d'approcher les solutions $x^\star$ au sens où, partant d'un vecteur initial $x^{(0)}$ quelconque, on construit explicitement une suite de vecteurs $(x^{(k)})_{k \geqslant 0}$ convergeant vers une solution $x^\star$.

Le problème d'optimisation est dit *sans contraintes* si $U = \mathbb{R}^n$ et *sous contraintes* sinon.

On dit que le problème est *convexe* si f et U sont convexes.

Dans ce cours on s'intéressera à résoudre des problèmes d'optimisation convexes, sans contraintes, et de dimension finie.

On établira dans ce chapitre des conditions d'existence et d'unicité des solutions de problèmes d'optimisation. Dans les chapitres suivants, on s'intéressera à l'élaboration d'algorithmes itératifs pour la résolution effectives de tels problèmes d'optimisation convexes, sans contraintes et de dimension finie.

Bien sûr, les méthodes développées dans ce cours permettent également de trouver les valeurs maximales de fonctions f. Pour cela il suffit de remplacer f par $-f$ puisque

$$\max_{x \in U} f(x) = \min_{x \in U} -f(x).$$

Extremums des fonctions réelles Soit $f : U \to \mathbb{R}$, où $U \subset \mathbb{R}^n$. On dit que la fonction f admet en un point $x \in U$ un *minimum local* (respectivement un *maximum local*) s'il existe un $\varepsilon > 0$ tel que pour tout $y \in U \cap B(x, \varepsilon)$, $f(y) \geqslant f(x)$ (resp. $f(y) \leqslant f(x)$). On dit que la fonction admet un *extremum local* en x si elle admet soit un minimum soit un maximum local en x.

Par abus de langage, on dira que x est un minimum local pour dire que la fonction f admet un minimum local en x.

On dit qu'un minimum local x est strict s'il existe un $\varepsilon > 0$ tel que pour tout $y \in U \cap B(x, \varepsilon)$, $y \neq x$, $f(y) > f(x)$. On définit de même la notion de maximum strict.

Enfin, on dit qu'un minimum x est *global* si pour tout $y \in U$, $f(y) \geqslant f(x)$. Si $W \subset U$, on dira qu'un minimum $x \in W$ est global sur W si pour tout $y \in W$, $f(y) \geqslant f(x)$. On défini de même la notion de maximum global.

2.2 Existence de solutions pour les fonctions coercives et continues

La première question concernant un problème d'optimisation est celle de l'existence d'une solution. Si on cherche à minimiser une fonction $f : U \subset \mathbb{R}^n \to \mathbb{R}$ continue sur U, alors il est bien connue que si U est compact (*i.e.* fermé et borné) la fonction f est bornée et atteint ses bornes sur U. Elle admet donc au moins un minimum global $x^\star \in U$. La notion de fonction coercive permet d'étendre ce type de raisonnement pour des fonctions définies sur des domaines non bornés.

Définition 2.1 (Fonctions coercives)**.** Une fonction $f : \mathbb{R}^n \to \mathbb{R}$ est dite *coercive* si

$$\lim_{\|x\| \to +\infty} f(x) = +\infty.$$

Théorème 2.2. Soient U une partie non vide fermée de $\mathbb{R}^n$ et $f : \mathbb{R}^n \to \mathbb{R}$ une fonction continue, coercive si l'ensemble U est non borné. Alors il existe au moins un élément $x^\star \in U$ tel que

$$f(x^\star) = \inf_{x \in U} f(x).$$

Preuve. Soit x_0 un point quelconque de U. La coercivité de f entraîne qu'il existe un réel $r > 0$ tel que

$$\|x\| \geqslant r \Rightarrow f(x) > f(x_0).$$

Donc,

$$\inf_{x \in U} f(x) = \inf_{x \in U \cap \overline{B(0,r)}} f(x).$$

Comme l'ensemble $U \cap \overline{B(0,r)}$ est fermé et borné et que f est continue, f est bornée et atteint ses bornes sur le compact $U \cap \overline{B(0,r)}$, ce qui assure l'existence d'un minimum (global) dans U (qui est inclus dans $U \cap \overline{B(0,r)}$). □

2.3 Extremums locaux et dérivabilité

On va maintenant chercher à caractériser les minimums locaux des fonctions différentiables. Dans toute la suite du chapitre Ω désigne un sous-ensemble ouvert de $\mathbb{R}^n$.

Théorème 2.3 (Condition nécessaire d'extremum local)**.** Soit $f : \Omega \to \mathbb{R}$ une fonction à valeurs réelles. Si la fonction f admet un extremum local en un point $x \in \Omega$ et si elle est différentiable en ce point, alors

$$\nabla f(x) = 0 \quad \text{(ou encore } \mathrm{d}f(x) = 0\text{)}.$$

On dit qu'un point x est un *point critique* de la fonctionnelle f si $\nabla f(x) = 0$.

Exercice 2.4.

1. Prouver le Théorème 2.3 en considérant l'application $\varphi : t \mapsto f(x+th)$ pour un vecteur $h \in \mathbb{R}^n$ quelconque.
2. Montrer que la conclusion du théorème est fausse si Ω n'est pas un ouvert.
3. Montrer que la réciproque du théorème est fausse : $\nabla f(x) = 0$ n'implique pas que x soit un extremum local.

Solution de l'exercice 2.4.

1. Soit $h \in \mathbb{R}^n$. Comme Ω est un ouvert, il existe $\varepsilon > 0$, tel que pour tout $t \in]-\varepsilon, \varepsilon[$, $x+th \in \Omega$. La fonction $\varphi : t \mapsto f(x+th)$ est donc définie sur $]-\varepsilon, \varepsilon[$. Elle est dérivable en $t = 0$, et d'après la formule de dérivation des fonctions composées, $\varphi'(0) = \mathrm{d}f(x)h = \langle \nabla f(x), h \rangle$. φ ayant un extremum en $t = 0$, on sait que $\varphi'(0) = 0$ (en utilisant le cas bien connu des fonctions réelles de la variable réelle). On en déduit que $\langle \nabla f(x), h \rangle = 0$. Ceci est vrai pour tout $h \in \mathbb{R}^n$, donc on a bien $\nabla f(x) = 0$.
2. On considère par exemple pour $n = 1$, $f(x) = x$ sur le domaine fermé $\Omega = [0, 1]$ qui admet un minimum local en $x = 0$ (au bord du domaine...).
3. Par exemple, toujours sur $\mathbb{R}$, $f(x) = x^3$ a une dérivée nulle en 0 mais 0 n'est pas un extremum.

On s'intéresse maintenant aux conditions nécessaires et suffisantes faisant intervenir la dérivée seconde.

Théorème 2.5 (Condition nécessaire de minimum local pour la dérivée seconde)**.** Soit $f : \Omega \to \mathbb{R}$ une fonction différentiable dans Ω. Si la fonction f admet un minimum local en un point $x \in \Omega$ et si f est deux fois différentiable en x, alors pour tout $h \in \mathbb{R}^n$,

$$\langle \nabla^2 f(x)h, h \rangle \geqslant 0 \quad \text{(ou encore } \mathrm{d}^2 f(x)(h,h) \geqslant 0),$$

autrement dit la matrice hessienne $\nabla^2 f(x)$ est positive.

Preuve. Soit $h \in \mathbb{R}^n$. Il existe un intervalle ouvert $I \subset \mathbb{R}$ contenant l'origine tel que

$$t \in I \Rightarrow (x+th) \in \Omega \quad \text{et} \quad f(x+th) \geqslant f(x).$$

La formule de Taylor-Young donne

$$f(x+th) = f(x) + t\langle \nabla f(x), h \rangle + \frac{t^2}{2}\langle \nabla^2 f(x)h, h \rangle + t^2 \|h\|^2 \varepsilon(th)$$

avec $\lim_{y \to 0} \varepsilon(y) = 0$. Comme x est un minimum dans l'ouvert Ω, d'après le Théorème 2.3 on a $\nabla f(x) = 0$. Ainsi,

$$0 \leqslant f(x+th) - f(x) = \frac{t^2}{2}\langle \nabla^2 f(x)h, h \rangle + t^2 \|h\|^2 \varepsilon(th)$$

En divisant par $\frac{t^2}{2}$ on en déduit que pour tout $t \neq 0$,

$$\nabla^2 f(x)h, h \rangle + 2\|h\|^2 \varepsilon(th) \geqslant 0.$$

En faisant tendre t vers 0 on obtient bien que $\langle \nabla^2 f(x)h, h \rangle \geqslant 0$. □

Comme le montre le résultat suivant, la dérivée seconde permet souvent de déterminer la nature d'un point critique, c'est-à-dire de déterminer si un point critique est bien un minimum local, un maximum local, ou ni l'un ni l'autre.

Théorème 2.6 (Condition suffisante de minimum local pour la dérivée seconde)**.** Soit $f : \Omega \to \mathbb{R}$ une fonction différentiable dans Ω et x un point critique de f (*i.e.* tel que $\nabla f(x) = 0$).

(a) Si la fonction f est deux fois différentiable en x et si

$$\forall h \in \mathbb{R}^n \setminus \{0\}, \quad \langle \nabla^2 f(x)h, h\rangle > 0$$

(*i.e.* la matrice hessienne $\nabla^2 f(x)$ est définie positive), alors la fonction f admet un minimum local strict en x.

(b) Si la fonction f est deux fois différentiable dans Ω, et s'il existe une boule $B \subset \Omega$ centrée en x telle que

$$\forall y \in B,\ \forall h \in \mathbb{R}^n, \quad \langle \nabla^2 f(y)h, h\rangle \geqslant 0$$

alors la fonction f admet en minimum local en x.

Preuve. (a) Comme $\nabla^2 f(x)$ est définie positive, il existe un nombre $\alpha > 0$ tel que

$$\forall h \in \mathbb{R}^n, \quad \langle \nabla^2 f(x)h, h\rangle \geqslant \alpha \|h\|^2$$

(en prenant par exemple $\alpha = \lambda_{\min}(\nabla^2 f(x))$ la plus petite valeur propre de $\nabla^2 f(x)$). D'après la formule de Taylor-Young,

$$f(x+h) = f(x) + \langle \nabla f(x), h\rangle + \frac{1}{2}\langle \nabla^2 f(x)h, h\rangle + \|h\|^2 \varepsilon(h) \geqslant f(x) + (\frac{1}{2}\alpha - |\varepsilon(h)|)\|h\|^2$$

avec $\lim_{h\to 0} \varepsilon(h) = 0$. Soit $r > 0$ tel que pour tout $h \in B(0, r)$, $|\varepsilon(h)| < \frac{1}{2}\alpha$. Alors, pour tout $h \in B(0, r)$, $f(x + h) > f(h)$, donc x est bien un minimum strict.

(b) Soit h tel que $x + h \in B$. Alors, comme f est deux fois différentiable, d'après la formule de Taylor-Maclaurin il existe $y \in]x, x + h[$ tel que

$$f(x+h) = f(x) + \langle \nabla f(x), h\rangle + \frac{1}{2}\langle \nabla^2 f(y)h, h\rangle = f(x) + \frac{1}{2}\langle \nabla^2 f(y)h, h\rangle,$$

donc $f(x + h) \geqslant f(x)$ pour tout h tel que $x + h \in B$, x est bien un minimum local de f. □

2.4 Ensembles convexes

On rappelle qu'étant donnés deux vecteurs x et $y \in \mathbb{R}^n$, $[x, y]$ désigne le segment entre x et y, à savoir

$$[x, y] = \{\theta x + (1 - \theta)y,\ \theta \in [0, 1]\}.$$

Définition 2.7 (Ensembles convexes)**.** On dit qu'un ensemble $U \subset \mathbb{R}^n$ est *convexe* si

$$\forall x, y \in U, \quad [x, y] \subset U,$$

soit encore si

$$\forall x, y \in U, \forall \theta \in [0, 1], \quad \theta x + (1 - \theta)y \in U.$$

(autrement dit U contient tout segment rejoignant n'importe quel couple de ses points).

Voici quelques exemples d'ensembles convexes :
— Un sous-espace vectoriel est convexe
— Un hyperplan est convexe.
— La boule unité d'une norme est convexe.
— Toute intersection d'ensembles convexes est convexe.
— Un hyper-rectangle $[a_1, b_1] \times \cdots \times [a_n, b_n]$ est convexe. Plus généralement le produit cartésien $C = C_1 \times \cdots \times C_k$ d'ensembles convexes $C_1 \subset \mathbb{R}^{n_1}, \ldots, C_n \subset \mathbb{R}^{n_k}$ est un ensemble convexe de l'espace produit $\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}$.
— L'image d'un ensemble convexe par une application linéaire est convexe (voir exercice ci-dessous). En particulier, les translations, rotations, dilatations, projections d'ensembles convexes sont convexes.

Exercice 2.8. Soit $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ une application linéaire.

1. Montrer que si $U \subset \mathbb{R}^n$ est convexe alors l'image directe $W = \varphi(U) = \{\varphi(x),\ x \in U\}$ de U par φ est un ensemble convexe de $\mathbb{R}^m$.
2. Montrer que si $W \subset \mathbb{R}^m$ est un ensemble convexe alors l'image réciproque $U = \varphi^{-1}(W) = \{x \in \mathbb{R}^n,\ \varphi(x) \in W\}$ est un ensemble convexe de $\mathbb{R}^n$.

Solution de l'exercice 2.8.

1. Soient y_1 et $y_1 \in W$. Alors il existe x_1 et x_2 dans U tel que $y_1 = \varphi(x_1)$ et $y_2 = \varphi(x_2)$. Soit $\theta \in [0, 1]$. Alors, par linéarité,

$$\theta w_1 + (1-\theta)w_2 = \theta\varphi(x_1) + (1-\theta)\varphi(x_2) = \varphi(\theta x_1 + (1-\theta)x_2).$$

   Or comme U est convexe, $\theta x_1 + (1-\theta)x_2 \in U$, et donc, $\theta y_1 + (1-\theta)y_2 \in W = \varphi(U)$. $W = \varphi(U)$ est bien un ensemble convexe.
2. Soient x_1 et $x_2 \in U = \varphi^{-1}(W)$ et $\theta \in [0, 1]$. Comme W est convexe

$$\varphi(\theta x_1 + (1-\theta)x_2) = \theta\varphi(x_1) + (1-\theta)\varphi(x_2) \in W.$$

   Donc $\theta x_1 + (1-\theta)x_2 \in U = \varphi^{-1}(W)$, c'est bien un ensemble convexe.

Théorème 2.9 (Condition nécessaire de minimum local sur un ensemble convexe)**.** Soit $f : \Omega \to \mathbb{R}$ et U une partie convexe de Ω. Si la fonction f est différentiable en un point $x \in U$ et si elle admet en x un minimum local par rapport à l'ensemble U, alors

$$\forall y \in U, \quad \langle \nabla f(x), y - x\rangle \geqslant 0 \quad \text{(ou encore } \mathrm{d}f(x)(y-x) \geqslant 0.$$

En particulier si U est un sous-espace affine de $\mathbb{R}^n$ (c'est-à-dire $U = x + F$ avec F un sous-espace vectoriel de V), alors

$$\forall y \in U, \quad \langle \nabla f(x), y - x\rangle = 0 \quad \text{(ou encore } \mathrm{d}f(x)(y-x) = 0.$$

Preuve. Soit $y = x + h$ un point quelconque de l'ensemble U. U étant convexe, les points $x + \theta h$, $\theta \in [0, 1]$, sont tous dans U. La dérivabilité de f en x permet d'écrire

$$f(x + \theta h) - f(x) = \theta\langle \nabla f(x), h\rangle + \theta\|h\|\varepsilon(\theta h),$$

avec $\lim_{\theta\to 0} \varepsilon(\theta h) = 0$. Comme le membre de gauche est positif, on a nécessairement $\langle \nabla f(x), h\rangle = \langle \nabla f(x), y - x\rangle \geqslant 0$ (dans le cas contraire pour θ assez petit le membre de droite serait < 0). Les cas des sous-espaces affines $U = u + F$, on remarque que si $x + h \in U$ alors $x - h$ appartient également à U et donc on a la double inégalité $\langle \nabla f(x), h\rangle \geqslant 0$ et $\langle \nabla f(x), -h\rangle \geqslant 0$ et donc $\langle \nabla f(x), h\rangle = 0$. □

Remarque. L'interprétation géométrique du théorème précédent est très importante. Si u est un minimum local par rapport au convexe U tel que $\nabla f(u) \neq 0$, alors $\nabla f(u)$ est orienté vers l'intérieur du convexe. En effet, la condition $\langle \nabla f(u), v-u \rangle \geqslant 0$ signifie que l'angle formé par les vecteurs $\nabla f(u)$ et $v-u$ est un angle aigu. Dans le cas d'un espace affine $U = u + F$, cela revient à une condition d'orthogonalité $\nabla f(u) \in F^{\perp}$. Dans ce cours on ne considérera pas de problème de minimisation sous contrainte convexe. En revanche, on sera amené à constamment minimiser des fonctionnelles sur des sous-espaces affines, et en premier lieu des droites.

2.5 Fonctions convexes

2.5.1 Définition et exemples

Définition 2.10 (Fonctions convexes)**.** Soit $U \subset \mathbb{R}^n$ un ensemble convexe. Soit $f : U \to \mathbb{R}$ une fonction à valeurs réelles.

— f est *convexe* si

$$\forall x, y \in U,\ \forall \theta \in [0,1], \quad f(\theta x + (1-\theta)y) \leqslant \theta f(x) + (1-\theta) f(y).$$

— f est *strictement convexe* si

$$\forall x, y \in U,\ x \neq y,\ \forall \theta \in]0,1[, \quad f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta) f(y).$$

Une fonction f est (strictement) *concave* si son opposée $x \mapsto -f(x)$ est (strictement) convexe.

Remarque. On peut également restreindre θ à $]0,1[$ pour la définition de la convexité.

Voici quelques exemples de fonctions convexes :

— Sur $\mathbb{R}$, la fonction $x \mapsto x^2$ est strictement convexe.
— Sur $\mathbb{R}$, la fonction $x \mapsto |x|$ est convexe mais pas strictement convexe.
— De même, sur $\mathbb{R}^n$, la fonction $x \mapsto \|x\|^2$ est strictement convexe la fonction $x \mapsto \|x\|$ est convexe mais pas strictement convexe.
— Le sup d'une famille quelconque de fonctions convexes est convexe.
— La composée d'une fonction affine et d'une fonction convexe est convexe (voir ci-dessous).
— Sur $\mathbb{R}^n$, les fonctions affines $f(x) = \langle a, x \rangle + b$ (avec $a \in \mathbb{R}^n$ et $b \in \mathbb{R}$) sont les seules fonctions à la fois convexes et concaves (voir Exercice 2.20).

Exercice 2.11. Soient $A \in M_{m,n}(\mathbb{R})$ et $b \in \mathbb{R}^m$. Montrer que si $f : \mathbb{R}^m \to \mathbb{R}$ est convexe alors la fonction $g : \mathbb{R}^n \to \mathbb{R}$ définie par $g(x) = f(Ax+b)$ est également convexe.

Solution de l'exercice 2.11. Soient $x, y \in \mathbb{R}^n$ et $\theta \in [0,1]$. On a, par linéarité puis convexité,

$$g(\theta x + (1-\theta)y) = f(\theta(Ax+b) + (1-\theta)(Ay+b)) \leqslant \theta f(Ax+b) + (1-\theta) f(Ay+b) = \theta g(x) + (1-\theta) g(y).$$

Donc g est bien convexe.

Théorème 2.12 (Continuité des fonctions convexes)**.** Soit $U \subset \mathbb{R}^n$ un ensemble convexe d'intérieur non vide et $f : U \to \mathbb{R}$ une fonction convexe sur U. Alors f est continue sur l'intérieur de U.

On admet la preuve de ce résultat qui se démontre en revenant au cas des fonctions de la variable réelle en restreignant f dans chaque direction. On notera qu'une fonction convexe peut être discontinue au bord de son domaine (par valeur supérieure).

Etant donnée une fonction $f : U \subset \mathbb{R}^n \to \mathbb{R}$, on appelle sous-ensemble de niveau α de f l'ensemble

$$C_\alpha = \{x \in U,\ f(x) \leqslant \alpha\}.$$

Proposition 2.13. Soit $U \subset \mathbb{R}^n$ un ensemble convexe et $f : U \to \mathbb{R}$ une fonction convexe sur U. Alors pour tout $\alpha \in \mathbb{R}$, l'ensemble C_α est convexe. En particulier, l'ensemble des minimums globaux de f est un ensemble convexe (qui peut être vide).

Preuve. Soit $\alpha \in \mathbb{R}$. Soient x_1 et x_2 dans C_α et $\theta \in [0, 1]$. Alors, comme f est convexe,

$$f(\theta x_1 + (1-\theta)x_2) \leqslant \theta f(x_1) + (1-\theta)f(x_2) \leqslant \theta\alpha + (1-\theta)\alpha = \alpha$$

donc $\theta x_1 + (1-\theta)x_2 \in C_\alpha$ ce qui prouve bien que C_α est convexe. L'ensemble des minimums globaux de f n'est autre que le sous-ensemble de niveau $p^* = \inf_{x\in U} f(x)$ de f, et il est donc bien convexe. □

Remarque. La réciproque de la proposition précédente est fausse. Il existe des fonctions non convexes dont tous les sous-ensembles de niveaux sont convexes, comme par exemple $x \mapsto -e^{-x}$ sur $\mathbb{R}$. Un autre exemple : Sur $\mathbb{R}$, tous les ensembles de niveaux de la fonction $\mathbb{1}_{[0,+\infty[}$ sont convexes mais cette fonction n'est pas convexe. En effet,

$$C_\alpha = \{x \in \mathbb{R},\ \mathbb{1}_{[0,+\infty[}(x) \leqslant \alpha\} = \begin{cases} \emptyset & \text{si } \alpha < 0, \\ ]-\infty, 0[& \text{si } \alpha \in [0, 1[, \\ \mathbb{R} & \text{si } \alpha > 1, \end{cases}$$

qui sont bien tous des ensembles convexes. En revanche la fonction n'est pas convexe car par exemple

$$1 = f(0) > \frac{1}{2}f(1) + \frac{1}{2}f(-1) = \frac{1}{2}.$$

2.5.2 Caractérisation des fonctions convexes différentiables et deux fois différentiables

Avant de considérer l'influence de la convexité sur l'existence et l'unicité de minimums, nous donnons des caractérisations de la notion de convexité pour les fonctions différentiables et deux fois différentiables.

Le théorème ci-dessous exprime le fait qu'une fonction différentiable est convexe si et seulement si son graphe est au-dessus de chacun des ses plans tangents.

Théorème 2.14 (Convexité et dérivabilité première)**.** Soit $f : \Omega \to \mathbb{R}$ une fonction différentiable dans l'ouvert Ω et soit $U \subset \Omega$ un sous-ensemble convexe.

(a) La fonction f est convexe sur U si et seulement si pour tout $x, y \in U$,

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x\rangle \quad \text{(ou encore } f(y) \geqslant f(x) + \mathrm{d}f(x)(y-x)\text{)}.$$

(b) La fonction f est strictement convexe sur U si et seulement si pour tout $x, y \in U$, $x \neq y$,

$$f(y) > f(x) + \langle \nabla f(x), y - x\rangle \quad \text{(ou encore } f(y) > f(x) + \mathrm{d}f(x)(y-x)\text{)}.$$

Preuve. (a) : $\Rightarrow$: Soient x, y deux points distincts de U et $\theta \in]0,1[$. Comme f est convexe,

$$f((1-\theta)x + \theta y) \leqslant (1-\theta)f(x) + \theta f(y)$$

et on a donc

$$\frac{f(x+\theta(y-x)) - f(x)}{\theta} \leqslant f(y) - f(x).$$

En passant à la limite $\theta \to 0$ on a

$$\langle \nabla f(x), y - x \rangle = \lim_{\theta \to 0} \frac{f(x+\theta(y-x)) - f(x)}{\theta} \leqslant f(y) - f(x).$$

$\Leftarrow$: Réciproquement supposons que pour tout $x, y \in U$,

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle.$$

Soient x, y deux points distincts de U et $\theta \in]0,1[$. En appliquant l'inégalité aux deux couples $(\theta x + (1-\theta)y, y)$ et $(\theta x + (1-\theta)y, x)$ on a

$$f(y) \geqslant f(\theta x + (1-\theta)y) + \langle \nabla f(\theta x + (1-\theta)y), \theta(y-x) \rangle,$$

et

$$f(x) \geqslant f(\theta x + (1-\theta)y) + \langle \nabla f(\theta x + (1-\theta)y), (1-\theta)(x-y) \rangle.$$

En multipliant par $(1-\theta)$ et θ ces deux inégalités, on obtient en les sommant

$$\theta f(x) + (1-\theta) f(y) \geqslant f(\theta x + (1-\theta)y),$$

donc f est bien convexe.

(b) : La preuve de l'implication indirecte est identique en remplaçant les inégalités larges par des inégalités strictes. En revanche pour l'implication directe, le passage à la limite change les inégalités strictes en inégalités larges, donc on ne peut pas conclure aussi rapidement. Pour cela on se donne cette fois-ci deux poids $0 < \theta < \omega < 1$. Alors, comme f est strictement convexe et $(1-\theta)x + \theta y \in [x, (1-\omega)x + \omega y]$ avec

$$(1-\theta)x + \theta y = \frac{\omega - \theta}{\omega} x + \frac{\theta}{\omega}((1-\omega)x + \omega y)$$

$$f((1-\theta)x + \theta y) < \frac{\omega - \theta}{\omega} f(x) + \frac{\theta}{\omega} f((1-\omega)x + \omega y).$$

D'où,

$$\frac{f(x+\theta(y-x)) - f(x)}{\theta} < \frac{f(x+\omega(y-x)) - f(x)}{\omega} < f(y) - f(x).$$

En passant à la limite $\theta \to 0$ on a alors

$$\langle \nabla f(x), y - x \rangle \leqslant \frac{f(x+\omega(y-x)) - f(x)}{\omega} < f(y) - f(x).$$

□

Théorème 2.15 (Convexité et dérivabilité seconde)**.** Soit $f : \Omega \to \mathbb{R}$ une fonction deux fois différentiable et soit $U \subset \Omega$ un sous-ensemble convexe.

(a) La fonction f est convexe sur U si et seulement si pour tout $x, y \in U$,

$$\langle \nabla^2 f(x)(y-x), y-x \rangle \geqslant 0$$

(b) Si pour tout $x, y \in U$, $x \neq y$,

$$\langle \nabla^2 f(x)(y-x), y-x \rangle > 0$$

alors f est strictement convexe sur U.

En particulier, si $\Omega = U$ est un ouvert convexe, alors

(a) f est convexe sur Ω si et seulement si pour tout $x \in \Omega$ la matrice hessienne $\nabla^2 f(x)$ est positive.

(b) Si pour tout $x \in \Omega$ la matrice hessienne $\nabla^2 f(x)$ est définie positive, alors f est strictement convexe sur Ω.

Preuve. Soient x et $y = x + h$ deux points distincts de U. Alors, comme f est deux fois différentiable, d'après la formule de Taylor-Maclaurin il existe $z \in]x, x+h[$ tel que

$$f(x+h) = f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} \langle \nabla^2 f(z)h, h \rangle.$$

Mais $z \in]x, x+h[$, donc il existe $\theta \in]0, 1[$ tel que $z = \theta x + (1-\theta)y$, soit $z - x = (1-\theta)(y-x) = (1-\theta)h$. Ainsi

$$f(x+h) = f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} \frac{1}{(1-\theta)^2} \langle \nabla^2 f(z)(z-x), (z-x) \rangle.$$

Si par hypothèse $\langle \nabla^2 f(z)(z-x), (z-x) \rangle$ est positif (resp. strictement positif) on déduit du théorème de caractérisation des fonctions convexes différentiables que f est convexe (resp. strictement convexe).

Il reste à montrer que si f est convexe alors pour tout $x, y \in U$, $\langle \nabla^2 f(x)(y-x), y-x \rangle \geqslant 0$. Soient $x, y = x + h \in U$. En appliquant la formule de Taylor-Young en x pour l'accroissement th (avec $t \in [0, 1]$),

$$f(x+th) = f(x) + t\langle \nabla f(x), h \rangle + \frac{t^2}{2} \langle \nabla^2 f(x)h, h \rangle + t^2 \|h\|^2 \varepsilon(th)$$

avec $\lim_{t \to 0} \varepsilon(th) = 0$. Donc

$$0 \leqslant f(x+th) - f(x) - t\langle \nabla f(x), h \rangle = \frac{t^2}{2} \left(\langle \nabla^2 f(x)h, h \rangle + 2\|h\|^2 \varepsilon(th) \right)$$

et on en déduit que $\langle \nabla^2 f(x)h, h \rangle \geqslant 0$ avec le raisonnement habituel.

Le cas où $U = \Omega$ est une conséquence directe. On peut aussi le démontrer rapidement en étudiant la différence entre f et ses approximations au premier ordre. En effet, pour $x \in \Omega$,

$$g(y) = f(y) - f(x) - \langle \nabla f(x), y - x \rangle$$

est une fonction convexe (en tant que somme de fonctions convexes), deux fois différentiable et telle que $\nabla^2 f(y) = \nabla^2 g(y)$. Comme $g(y) \geqslant 0$ et que $g(x) = 0$, x est un minimum global de f et donc nécessairement pour tout $h \in \mathbb{R}^n$, $\langle \nabla^2 f(x)h, h \rangle \geqslant 0$. □

2.5.3 Problèmes d'optimisation convexes

On rappelle qu'une problème d'optimisation

$$\text{Trouver } x^\star \text{ tel que } x^\star \in U \text{ et } \quad f(x^\star) = \min_{x \in U} f(x),$$

est dit *convexe* si U et f sont convexes. Vis-à-vis de l'optimisation, la convexité joue un rôle crucial puisqu'elle permet d'assurer qu'un minimum local est en fait un minimum global, comme précisé par le résultat suivant.

Théorème 2.16 (Minimum de fonctions convexes)**.** Soit $U \subset \mathbb{R}^n$ un ensemble convexe.

(a) Si une fonction convexe $f : U \to \mathbb{R}$ admet un minimum local en un point x, elle y admet en fait un minimum global sur U.

(b) Une fonction $f : U \to \mathbb{R}$ strictement convexe admet au plus un minimum local qui est en fait un minimum global strict.

(c) Soit $f : \Omega \to \mathbb{R}$ une fonction convexe définie sur un ouvert convexe $\Omega \subset \mathbb{R}^n$. Alors un point $x \in \Omega$ est un minimum global de f si et seulement si $\nabla f(x) = 0$ (ou encore $\mathrm{d}f(x) = 0$).

(d) Soit $f : \Omega \to \mathbb{R}$ une fonction définie sur un ouvert Ω contenant U et telle que f est convexe sur U. Alors $x \in U$ est un minimum de f sur U si et seulement si pour tout $y \in U$,

$$\langle \nabla f(x), y - x \rangle \geqslant 0 \quad \text{(ou encore } \mathrm{d}f(x)(y-x) \geqslant 0.$$

En particulier, si $U = x + F$ est un sous-espace affine, alors $x \in U$ est un minimum de f sur U si et seulement si pour tout $y \in U$,

$$\langle \nabla f(x), y - x \rangle = 0 \quad \text{(ou encore } \mathrm{d}f(x)(y-x) = 0.$$

Preuve. (a) Soit y un point quelconque de U. Comme précédemment, la convexité entraîne que

$$f(y) - f(x) \geqslant \frac{f(x + \theta(y-x)) - f(x)}{\theta}$$

pour tout $\theta \in]0, 1[$. Comme x est un minimum local, il existe un θ_0 assez petit tel que $f(x + \theta_0(y-x)) - f(x) \geqslant 0$. Mais alors, $f(y) - f(x) \geqslant 0$, donc x est bien un minimum global.

(b) Si f est strictement convexe et que x est un minimum local de f, alors pour $y \neq x$ le raisonnement précédent donne l'existence d'un $\theta_0 > 0$ tel que

$$f(y) - f(x) > \frac{f(x + \theta_0(y-x)) - f(x)}{\theta_0} \geqslant 0.$$

Donc $y \neq x$ implique $f(y) > f(x)$. x est donc bien un minimum strict qui est global et unique.

(c) On sait que $\nabla f(x) = 0$ est une condition nécessaire pour être un minimum global. Montrons que c'est une condition suffisante si f est convexe. D'après le Théorème 2.14, si $x \in \Omega$ est tel que $\nabla f(x) = 0$, alors pour tout $y \in \Omega$, $f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle = f(x)$, donc x est bien un minimum global.

(d) C'est le même raisonnement. La condition est nécessaire d'après le Théorème 2.9, et si elle est vérifiée, alors d'après le Théorème 2.14 $f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle \geqslant f(x)$ donc x est bien un minimum global sur U.

□

Remarque.
— Une fonction non strictement convexe peut admettre plusieurs minimums locaux. Cependant, comme on l'a vu, l'ensemble des minimums globaux forme un ensemble convexe.
— Le théorème précédent est fondamental pour la suite de ce cours. Sauf exception, en pratique on ne s'intéressera qu'à des problèmes d'optimisation convexes.

2.6 Etude des fonctionnelles quadratiques

On appelle *fonctionnelle quadratique* toute fonction $f : \mathbb{R}^n \to \mathbb{R}$ de la forme

$$f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle + c,$$

où $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice carrée symétrique, $b \in \mathbb{R}^n$ et $c \in \mathbb{R}$. La proposition suivante résume les propriétés des fonctionnelles quadratiques.

Proposition 2.17 (Propriétés des fonctionnelles quadratiques)**.** Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonctionnelle quadratique de la forme $f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle + c$. Alors,

(a) f est $\mathcal{C}^2$ sur $\mathbb{R}^n$ (et même $\mathcal{C}^\infty$).

(b) Pour tout $x \in \mathbb{R}^n$,
$$\nabla f(x) = Ax - b \quad \text{et} \quad \nabla^2 f(x) = A.$$

(c) f est convexe si et seulement si A est positive.

(d) f est strictement convexe si et seulement si A est définie positive.

(e) $\inf_{x\in\mathbb{R}^n} f(x)$ est fini si et seulement si A est positive et telle que le système linéaire $Ax = b$ admet (au moins) une solution, et alors l'ensemble de solutions de $Ax = b$ est l'ensemble des minimums globaux de f. Ainsi résoudre le problème d'optimisation associé à f revient à résoudre le système linéaire $Ax = b$.

La preuve de cette proposition est l'objet de l'exercice suivant.

Exercice 2.18. On reprend les notations de la Proposition 2.17 et on pose $S = \{x \in \mathbb{R}^n,\ Ax = b\}$.

1. Montrer que l'on peut toujours supposer que la matrice A est symétrique.
2. Montrer que f est différentiable sur $\mathbb{R}^n$ et que pour tout $x \in \mathbb{R}^n$, $\nabla f(x) = Ax - b$.
3. En déduire que $f \in \mathcal{C}^\infty(\mathbb{R}^n)$ et que pour tout $x \in \mathbb{R}^n$, $\nabla^2 f(x) = A$.
4. Justifier que f est convexe si et seulement si A est positive.
5. Montrer que f est strictement convexe si et seulement si A est définie positive.
6. Montrer qu'il existe $x \in \mathbb{R}^n$ tel que pour tout $y \in \mathbb{R}^n$, $y \neq x$, $f(x) < f(y)$ si et seulement si A est définie positive.
7. Montrer qu'il existe $x \in \mathbb{R}^n$ tel que pour tout $y \in \mathbb{R}^n$, $f(x) \leqslant f(y)$ si et seulement si A est positive et $S \neq \emptyset$.
8. Montrer que si A n'est pas positive, alors $\inf_{x\in\mathbb{R}^n} f(x) = -\infty$.
9. En considérant le projeté orthogonal de b sur le sous-espace vectoriel $\ker A$, montrer que si A est positive et que $S = \emptyset$ alors $\inf_{x\in\mathbb{R}^n} f(x) = -\infty$.

Solution de l'exercice 2.18.

1. Si A est quelconque, on remarque que remplacer A par sa partie symétrique $\frac{A+A^T}{2}$ ne change pas la valeur de la fonctionnelle. En effet
$$\langle Ax, x\rangle = \langle x, A^T x\rangle = \langle A^T x, x\rangle$$
et donc
$$\langle Ax, x\rangle = \frac{1}{2}\langle Ax, x\rangle + \frac{1}{2}\langle A^T x, x\rangle = \langle \frac{A + A^T}{2} x, x\rangle.$$

2. On a
$$f(x+h) = f(x) + \langle Ax - b, h\rangle + \frac{1}{2}\langle Ah, h\rangle.$$
Donc f est différentiable en x et $\nabla f(x) = Ax - b$.

3. En déduire que $f \in \mathcal{C}^\infty(\mathbb{R}^n)$ et que pour tout $x \in \mathbb{R}^n$, $\nabla^2 f(x) = A$. Comme $x \mapsto \nabla f(x)$ est une fonction affine c'est une application $\mathcal{C}^\infty(\mathbb{R}^n)$, et donc f est aussi $\mathcal{C}^\infty(\mathbb{R}^n)$. On a bien $\nabla^2 f(x) = A$.

4. En utilisant la caractérisation des fonctions convexes deux fois dérivables, f est convexe si et seulement si pour tout $x \in \mathbb{R}^n$, $\nabla^2 f(x) = A$ est positive, donc si et seulement si A est positive.

5. Le fait que pour tout $x \in \mathbb{R}^n$, $\nabla^2 f(x) = A$ soit définie positive implique que f est strictement convexe. En revanche la réciproque est fausse en générale, une fonction peut être strictement convexe sans avoir partout une matrice hessienne définie positive (voir par exemple $x \mapsto x^4$ sur $\mathbb{R}$). En revanche, on a une caractérisation pour les fonctions strictement convexes dérivables.
$$f \text{ est strictement convexe} \Leftrightarrow \forall x \neq y,\ f(y) > f(x) + \langle \nabla f(x), y - x\rangle$$
Or, $f(y) = (x) + \langle Ax - b, x - y\rangle + \frac{1}{2}\langle A(y-x), (y-x)\rangle$, donc $f(y) > f(x) + \langle \nabla f(x), y - x\rangle$ si et seulement si $\langle A(y-x), (y-x)\rangle > 0$. Comme cela doit être valable pour tout $y \neq x$, c'est bien équivalent à A définie positive.

6. Si A est définie positive, alors f est strictement convexe. f admet donc au plus un minimum qui est stricte. Il est caractérisé par $\nabla f(x) = 0$, soit $Ax = b$. Comme A est inversible, ce système linéaire admet une unique solution $x = A^{-1}b$. Donc f admet un unique minimum stricte en $x = A^{-1}b$. Réciproquement, si f admet un minimum strict $x \in \mathbb{R}^n$, alors $\nabla f(x) = Ax - b = 0$ et donc pour tout $h \neq 0$,
$$f(x+h) = f(x) + \frac{1}{2}\langle Ah, h\rangle > f(x).$$
Ainsi, pour tout $h \neq 0$, $\langle Ah, h\rangle > 0$, donc A est bien définie positive (en donc f est strictement convexe).

7. Si A est positive alors f est convexe et les minimums globaux de f sont caractérisés par $\nabla f(x) = 0$, soit $Ax = b$. Comme $S \neq \emptyset$, f admet bien un minimum global en tout $x \in S$. Réciproquement, si f admet un minimum global x, alors $\nabla f(x) = Ax - b = 0$, donc S est non vide. De plus, pour tout h,
$$f(x+h) = f(x) + \frac{1}{2}\langle Ah, h\rangle \geqslant f(x),$$
donc $\langle Ah, h\rangle \geqslant 0$ pour tout h, A est bien positive.

8. Si A n'est pas positive il existe h tel que $\langle Ah, h\rangle < 0$. Alors,
$$f(th) = \frac{t^2}{2}\langle Ah, h\rangle - t\langle b, h\rangle + c$$
est un trinôme du second degré à coefficient dominant strictement négatif, donc $\lim_{t\to\pm\infty} f(th) = -\infty$.

9. En considérant le projeté orthogonal de b sur le sous-espace vectoriel $\ker A$, montrer que si A est positive et que $S = \emptyset$ alors $\inf_{x\in\mathbb{R}^n} f(x) = -\infty$.

   Comme A est symétrique réelle, on a $\mathbb{R}^n = \ker(A) \oplus \mathrm{im}(A)$ où la somme est orthogonale. Soit $b = p + q$ avec $p \in \ker(A)$ et $q \in \mathrm{im}(A)$ la décomposition de b par rapport à cette somme. Comme $S = \emptyset$, $b \notin \mathrm{im}(A)$, donc $p \neq 0$. On a alors

$$f(tp) = -t\langle b, p\rangle + c = -t\|p\|^2 + c$$

   qui tend vers $-\infty$ quand t tend vers $+\infty$.

2.7 Exercices supplémentaires

Exercice 2.19. Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction. On appelle épigraphe de f le sous-ensemble de $\mathbb{R}^n \times \mathbb{R}$ défini par

$$\mathrm{epi}\, f = \{(x, \lambda) \in \mathbb{R}^n \times \mathbb{R},\ f(x) \leqslant \lambda\}.$$

1. Dans le cas où $n = 1$, donner une interprétation géométrique de l'épigraphe.
2. Montrer que f est une fonction convexe si et seulement si $\mathrm{epi}\, f$ est un ensemble convexe.

Solution de l'exercice 2.19.

1. Ensemble au-dessus du graphe de f.
2. On suppose que f est convexe. Soient (x_1, λ_1) et (x_2, λ_2) deux points de $\mathrm{epi}\, f$ et $\theta \in [0, 1]$. Montrons que $\theta(x_1, \lambda_1) + (1 - \theta)(x_2, \lambda_2) \in \mathrm{epi}\, f$. Par définition,

$$\theta(x_1, \lambda_1) + (1 - \theta)(x_2, \lambda_2) \in \mathrm{epi}\, f \Leftrightarrow f(\theta x_1 + (1 - \theta)x_2) \leqslant \theta\lambda_1 + (1 - \theta)\lambda_2$$

   Or, comme f est convexe,

$$f(\theta x_1 + (1 - \theta)x_2) \leqslant \theta f(x_1) + (1 - \theta)f(x_2) \leqslant \theta\lambda_1 + (1 - \theta)\lambda_2.$$

   Donc on a bien $\theta(x_1, \lambda_1) + (1 - \theta)(x_2, \lambda_2) \in \mathrm{epi}\, f$.

   Réciproquement, supposons que $\mathrm{epi}\, f$ soit convexe et montrons que f est convexe. Soient $x_1, x_2 \in V$ et $\theta \in [0, 1]$. $(x_1, f(x_1))$ et $(x_2, f(x_2))$ sont deux points de l'épigraphe. Comme $\mathrm{epi}\, f$ est convexe, $\theta(x_1, f(x_1)) + (1 - \theta)(x_2, f(x_2)) \in \mathrm{epi}\, f$ et donc,

$$f(\theta x_1 + (1 - \theta)x_2) \leqslant \theta f(x_1) + (1 - \theta)f(x_2).$$

   f est bien une fonction convexe.

Exercice 2.20. Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction à la fois convexe et concave, c'est-à-dire telle que

$$\forall x, y \in \mathbb{R}^n,\ \forall \theta \in [0, 1], \quad f(\theta x + (1 - \theta)y) = \theta f(x) + (1 - \theta)f(y).$$

Le but cet exercice est de montrer que f est une fonction affine : il existe $a \in \mathbb{R}^n$ et $b \in \mathbb{R}$ tels que $f(x) = \langle a, x\rangle + b$. On pose $g(x) = f(x) - f(0)$.

1. Montrer que g est impaire : $\forall x \in \mathbb{R}^n$, $g(-x) = -g(x)$.
2. Montrer que g est 1-homogène : $\forall x \in \mathbb{R}^n$, $\forall \lambda > 0$, $f(\lambda x) = \lambda f(x)$.
3. Montrer que g est linéaire et conclure.

Solution de l'exercice 2.20.

1. g est également convexe et concave. On a

$$0 = g(0) = \frac{1}{2}g(x) + \frac{1}{2}g(-x) \Rightarrow g(-x) = -g(x).$$

2. Soit $x \in \mathbb{R}^n$ et $\lambda > 0$. Si $\lambda = 1$, il n'y a rien à montrer. Si $\lambda \in]0, 1[$, alors

$$g(\lambda x) = \lambda g(x) + (1 - \lambda)g(0) = \lambda g(x).$$

   Si $\lambda > 1$, alors

$$g(x) = \frac{1}{\lambda}g(\lambda x) + \left(1 - \frac{1}{\lambda}\right) g(0) = \frac{1}{\lambda}g(\lambda x).$$

   Donc on a bien toujours $g(\lambda x) = \lambda g(x)$.

3. D'après les deux premières questions on a $g(\lambda x) = \lambda g(x)$ pour tout $x \in \mathbb{R}^n$ et $\lambda \in \mathbb{R}$. Il reste à montrer que $\forall x, y \in \mathbb{R}^n$, $g(x + y) = g(x) + g(y)$. Soient $x, y \in \mathbb{R}^n$. Alors, en utilisant l'homogénéité,

$$g(x + y) = \frac{1}{2}g(2x) + \frac{1}{2}g(2y) = g(x) + g(y).$$

   Ainsi g est une forme linéaire. On a donc $g(x) = \langle a, x\rangle$ avec $a \in \mathbb{R}^n$, et donc $f(x) = g(x) + f(0) = \langle a, x\rangle + f(0)$ est bien une fonction affine.

Chapitre 3

Algorithmes de descente pour des problèmes sans contraintes

La référence principale pour ce chapitre est le chapitre 9 de [BOYD & VANDENBERGHE].

Dans ce chapitre on s'intéresse à résoudre un problème d'optimisation convexe sans contrainte de la forme

$$\text{Trouver } x^\star \in \mathbb{R}^n \text{ tel que} \quad f(x^\star) = \min_{x\in\mathbb{R}^n} f(x),$$

où $f : \mathbb{R}^n \to \mathbb{R}$ est une fonction convexe et deux fois différentiable (sur $\mathbb{R}^n$). Si le problème admet des solutions $x^\star$, on note

$$p^\star = f(x^\star) = \min_{x\in\mathbb{R}^n} f(x).$$

Sous ces hypothèses on sait que résoudre le problème d'optimisation revient à résoudre les n équations non linéaires

$$\nabla f(x) = \begin{pmatrix} \dfrac{\partial f}{\partial x_1}(x_1,\dots,x_n) \\ \dfrac{\partial f}{\partial x_2}(x_1,\dots,x_n) \\ \vdots \\ \dfrac{\partial f}{\partial x_n}(x_1,\dots,x_n) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Afin d'assurer l'existence et l'unicité d'une telle solution, nous allons supposer que f est *fortement convexe*.

3.1 Forte convexité

Définition 3.1. Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction deux fois différentiable. On dit que f est *fortement convexe* si il existe une constante $m > 0$ telle que

$$\forall x, h \in \mathbb{R}^n, \quad \langle \nabla^2 f(x)h, h\rangle \geqslant m\|h\|^2.$$

Proposition 3.2 (Propriétés des fonctions fortement convexes)**.** Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction fortement convexe pour la constante $m > 0$. Alors f vérifie les propriétés suivantes :

(a) f est strictement convexe.

(b) Pour tous $x, y \in \mathbb{R}^n$,

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2}\|y - x\|^2.$$

(c) f est coercive.

(d) f admet un unique minimum global $x^\star$.

(e) Pour tout $x \in \mathbb{R}^n$,

$$p^\star \geqslant f(x) - \frac{1}{2m}\|\nabla f(x)\|^2 \quad \text{et} \quad \|x - x^\star\| \leqslant \frac{2}{m}\|\nabla f(x)\|.$$

Preuve. (a) Comme pour tout $x \in \mathbb{R}^n$, la matrice hessienne $\nabla^2 f(x)$ est définie positive, f est strictement convexe d'après le Théorème 2.15 sur les fonctions convexes deux fois différentiables.

(b) Soient $x, y \in \mathbb{R}^n$. D'après la formule de Taylor-Maclaurin, il existe $z \in]x, y[$ tel que

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\langle \nabla^2 f(z)(y - x), y - x \rangle.$$

En appliquant l'inégalité de forte convexité au dernier terme on obtient la minoration annoncée.

(c) En prenant $x = 0$ dans l'inégalité précédente on a

$$f(y) \geqslant f(x) + \langle \nabla f(0), y \rangle + \frac{m}{2}\|y\|^2.$$

Cette fonction minorante est coercive, donc f est elle aussi coercive.

(d) f est strictement convexe et coercive, elle admet donc un unique minimum global.

(e) Soit $x \in \mathbb{R}^n$. Alors, pour tout $y \in \mathbb{R}^n$,

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2}\|y - x\|^2.$$

Le terme de droite est une fonction quadratique de la variable y qui est minimale en $y^\star = x - \frac{1}{m}\nabla f(y)$ et cette valeur minimale vaut $f(x) - \frac{1}{2m}\|\nabla f(x)\|^2$ (à reprendre en détails dans l'exercice qui suit). Ainsi, on a pour tout $y \in \mathbb{R}^n$,

$$f(y) \geqslant f(x) - \frac{1}{2m}\|\nabla f(x)\|^2.$$

En prenant $y = x^\star$ on a

$$p^\star \geqslant f(x) - \frac{1}{2m}\|\nabla f(x)\|^2.$$

□

Exercice 3.3. Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction convexe deux fois différentiable. On suppose qu'il existe M tel que

$$\forall x, h \in \mathbb{R}^n, \quad \langle \nabla^2 f(x)h, h \rangle \leqslant M\|h\|^2,$$

autrement dit $\|\nabla^2 f(x)\|_{\mathcal{M}_n(\mathbb{R})} \leqslant M$.

1. Montrer que pour tous $x, y \in \mathbb{R}^n$,

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{M}{2}\|y - x\|^2.$$

2. Pour x fixé on pose $g(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{M}{2}\|y - x\|^2$. Montrer que g admet un unique minimum $y^\star$ sur $\mathbb{R}^n$ et calculer la valeur minimale de g.
3. On suppose que f admet un minimum global en $x^\star$ qui vaut $p^\star$. Montrer que pour tout $x \in \mathbb{R}^n$,

$$p^\star \leqslant f(x) - \frac{1}{2M}\|\nabla f(x)\|^2.$$

Solution de l'exercice 3.3.

1. Comme f est deux fois différentiable la formule de Taylor-Maclaurin assure qu'il existe $z \in]x, y[$ tel que

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\langle \nabla^2 f(z)(y - x), y - x \rangle.$$

Or par hypothèse

$$\langle \nabla^2 f(z)(y - x), y - x \rangle \geqslant M\|y - x\|^2,$$

d'où la majoration annoncée.

2. On a

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{M}{2}\|y - x\|^2.$$

Pour x fixé, on considère la fonction $g(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{M}{2}\|y - x\|^2$. On remarque que g est une fonction quadratique. Plus précisément,

$$g(y) = \frac{M}{2}\|y\|^2 - \langle Mx - \nabla f(x), y \rangle + f(x) - \langle \nabla f(x), x \rangle + \frac{M}{2}\|x\|^2.$$

La matrice A de cette application quadratique est MI_n qui est définie positive. Donc g est strictement convexe et atteint son minimum au point $y^\star$ solution de $\nabla g(y) = 0$. Or

$$\nabla g(y) = 0 \Leftrightarrow My - (Mx - \nabla f(x)) = 0 \Leftrightarrow y = x - \frac{1}{M}\nabla f(x).$$

La valeur minimale de g est donc

$$g(y^\star) = g\left(x - \frac{1}{M}\nabla f(x)\right) = f(x) - \frac{1}{M}\|\nabla f(x)\|^2 + \frac{M}{2}\|\frac{1}{M}\nabla f(x)\|^2 = f(x) - \frac{1}{2M}\|\nabla f(x)\|^2.$$

3. Pour $y = y^\star = x - \frac{1}{M}\nabla f(x)$ on a

$$f(y^\star) \leqslant f(x) - \frac{1}{2M}\|\nabla f(x)\|^2.$$

Or comme $p^\star = \min_{y \in \mathbb{R}^n} f(y)$, on en déduit que

$$p^\star \leqslant f(y^\star) \leqslant f(x) - \frac{1}{2M}\|\nabla f(x)\|^2.$$

Remarque. La formule des accroissements finis (pour les fonctions de $\mathbb{R}^n \to \mathbb{R}^n$, voir [CIARLET]) permet de montrer que si pour tous $x, h \in \mathbb{R}^n$, $\langle \nabla^2 f(x)h, h\rangle \leqslant M\|h\|^2$, alors le gradient de f est M-Lipschitz, c'est-à-dire que

$$\forall x, y \in \mathbb{R}^n, \quad \|\nabla f(y) - \nabla f(x)\| \leqslant M\|x - y\|.$$

Remarque. Dans les preuves des théorèmes de convergence on supposera qu'il existe des constantes $0 < m \leqslant M$ telles que

$$\forall x \in S,\ \forall h \in \mathbb{R}^n, \quad m\|h\|^2 \leqslant \langle \nabla^2 f(x)h, h\rangle \leqslant M\|h\|^2.$$

En général, m et M ne sont pas connues, donc les bornes de la proposition précédente ne sont pas explicites en pratique. Cependant, elles sont très importantes. En effet, les inégalités

$$f(x) - \frac{1}{2m}\|\nabla f(x)\|^2 \leqslant p^\star \leqslant f(x) - \frac{1}{2M}\|\nabla f(x)\|^2 \quad \text{et} \quad \|x - x^\star\| \leqslant \frac{2}{m}\|\nabla f(x)\|$$

montrent que si la norme du gradient $\|\nabla f(x)\|$ est faible alors x est proche de la solution $x^\star$.

3.2 Généralités sur les algorithmes de descente

3.2.1 Forme générale d'un algorithme de descente

Un algorithme de descente prend la forme générale décrite par l'Algorithme 1.

Algorithme 1 : Algorithme de descente général

Données : Un point initial $x^{(0)} \in \mathbb{R}^n$, un seuil de tolérance $\varepsilon > 0$
Résultat : Un point $x \in \mathbb{R}^n$ proche de $x^\star$
Initialiser x :
  $x \leftarrow x^{(0)}$;
  $k \leftarrow 0$;
tant que $\|\nabla f(x)\| > \varepsilon$ **faire**
  1. Déterminer une direction de descente $d^{(k)} \in \mathbb{R}^n$.
  2. Déterminer un pas de descente $t^{(k)} > 0$ tel que $f(x^{(k)} + t^{(k)}d^{(k)}) < f(x^{(k)})$.
  3. Mettre à jour x :
     $x \leftarrow x^{(k+1)} = x^{(k)} + t^{(k)}d^{(k)}$;
     $k \leftarrow k + 1$;
fin

Comme f est convexe,

$$f(x^{(k)} + t^{(k)}d^{(k)}) \geqslant f(x^{(k)}) + t^{(k)}\langle \nabla f(x^{(k)}), d^{(k)}\rangle,$$

donc, comme $t^{(k)} > 0$, pour avoir $f(x^{(k)} + t^{(k)}d^{(k)}) < f(x^{(k)})$ on doit nécessairement avoir $\langle \nabla f(x^{(k)}), d^{(k)}\rangle < 0$. On dira que $d \in \mathbb{R}^n$ est une *direction de descente* au point x si

$$\langle \nabla f(x), d\rangle < 0.$$

L'ensemble des directions de descentes au point x est ainsi un demi-espace ouvert.

Convergence : Si on fait abstraction du critère d'arrêt, un algorithme de descente produit une suite de points $(x^{(k)})_{k\in\mathbb{N}}$ définie par la relation de récurrence

$$x^{(k+1)} = x^{(k)} + t^{(k)}d^{(k)}$$

et telle que $f(x^{(k+1)}) < f(x^{(k)})$ (sauf si $x^{(k)} = x^\star$ à partir d'un certain rang). L'étude de la convergence d'un tel algorithme de descente consiste donc à savoir si la suite $(x^{(k)})_{k\in\mathbb{N}}$ converge vers $x^\star$. On rappelle que si f est fortement convexe on a $\|x - x^\star\| \leqslant \frac{2}{m}\|\nabla f(x)\|$, donc le critère d'arrêt $\|\nabla f(x)\| \leqslant \varepsilon$ implique $\|x - x^\star\| \leqslant \dfrac{2\varepsilon}{m}$.

On verra que l'on s'intéresse également à la convergence de la suite $(f(x^{(k)}) - p^\star)_{k\in\mathbb{N}}$. On parle alors de convergence pour la fonction objectif.

Exercice 3.4. Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction fortement convexe telle que

$$\forall x, h \in \mathbb{R}^n, \quad m\|h\|^2 \leqslant \langle \nabla^2 f(x)h, h\rangle \leqslant M\|h\|^2$$

avec $0 < m \leqslant M$.

1. Montrer que pour tout $x \in \mathbb{R}^n$,

$$\frac{2}{m}(f(x) - p^\star) \leqslant \|x - x^\star\|^2 \leqslant \frac{2}{M}(f(x) - p^\star).$$

2. En déduire que $(f(x^{(k)}))_{k\in\mathbb{N}}$ converge vers $p^\star$ si et seulement si $(x^{(k)})_{k\in\mathbb{N}}$ converge vers $x^\star$.

Solution de l'exercice 3.4.

1. Par forte convexité on a

$$\forall x, y \in \mathbb{R}^n, \quad f(y) \leqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{m}{2}\|y - x\|^2.$$

En prenant $y = x$ et $x = x^\star$ on en déduit que

$$f(x) \leqslant p^\star + \frac{m}{2}\|x - x^\star\|^2,$$

d'où l'inégalité

$$\frac{2}{m}(f(x) - p^\star) \leqslant \|x - x^\star\|^2.$$

De même, en utilisant la majoration

$$\forall x, y \in \mathbb{R}^n, \quad f(y) \geqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{M}{2}\|y - x\|^2.$$

on obtient que $\frac{2}{M}(f(x) - p^\star) \leqslant \|x - x^\star\|^2$. Finalement on a l'encadrement,

$$\frac{2}{m}(f(x) - p^\star) \leqslant \|x - x^\star\|^2 \leqslant \frac{2}{M}(f(x) - p^\star).$$

2. Par continuité de f si $(x^{(k)})_{k\in\mathbb{N}}$ converge vers $x^\star$ alors $(f(x^{(k)}))_{k\in\mathbb{N}}$ vers $p^\star = f(x^\star)$. L'inégalité $\|x - x^\star\|^2 \leqslant \frac{2}{M}(f(x) - p^\star)$ que la réciproque est vraie.

Vitesse de convergence : Une fois qu'un algorithme est prouvé être convergent (*i.e.* $(x^{(k)})_{k\in\mathbb{N}}$ converge vers $x^\star$), on s'intéresse à sa *vitesse de convergence*. On dit que la méthode est d'ordre $r \geqslant 1$, s'il existe une constante $C > 0$ telle que, pour k suffisamment grand

$$\frac{\|x^{(k+1)} - x^\star\|}{\|x^{(k)} - x^\star\|^r} \leqslant C.$$

— Si $r = 1$, il faut $C \in]0, 1[$ pour avoir convergence et on a alors alors *convergence linéaire*.
— Si $r = 2$, on a une *convergence quadratique*.
— Si $\lim \dfrac{\|x^{(k+1)} - x^\star\|}{\|x^{(k)} - x^\star\|^r} = 0$ alors on dit que l'on a *convergence superlinéaire* (ce qui est le cas pour toutes les méthodes d'ordre $r > 1$).

La quantité $-\log_{10} \|x^{(k)} - x^\star\|$ mesure le nombre de décimales exactes dans l'approximation de $x^\star$ par $x^{(k)}$. En cas de convergence linéaire on a

$$-\log_{10} \|x^{(k+1)} - x^\star\| \geqslant -\log_{10} \|x^{(k)} - x^\star\| - \log_{10} C$$

donc on gagne au moins $-\log_{10} C$ décimales à chaque itérations. Si on a une convergence d'ordre $r > 1$, on a

$$-\log_{10} \|x^{(k+1)} - x^\star\| \geqslant -r \log_{10} \|x^{(k)} - x^\star\| - \log_{10} C$$

donc $x^{(k+1)}$ a r fois plus de décimales exactes que $x^{(k)}$. En particulier si on a convergence quadratique, alors la précision double à chaque itération.

On parle également alors de convergence linéaire, quadratique, etc. pour la fonction objectif, c'est-à-dire la vitesse de convergence vers 0 de la suite $(f(x^{(k)}) - p^\star)_{k\in\mathbb{N}}$.

3.2.2 Algorithmes de recherche de pas de descente

Dans l'Algorithme 1, l'étape

"Déterminer un pas de descente $t^{(k)} > 0$ tel que $f(x^{(k)} + t^{(k)} d^{(k)}) < f(x^{(k)})$"

est restée volontairement floue. Il existe de nombreuses méthodes de recherche de pas avec différents critères (on parle de *line search* et des conditions de Wolfe). Comme dans le chapitre 9 de [BOYD & VANDENBERGHE], nous allons nous limiter à deux méthodes, la méthode de pas de descente optimal (dite aussi exacte) et la méthode de pas de descente par rebroussement.

Dans les deux cas, les données du problème sont une fonction f fortement convexe, un point actuel $x = x^{(k)} \neq x^\star$, une direction de descente $d = d^{(k)}$ pour le point x, et on cherche un pas de descente $t = t^{(k)} > 0$ tel que $f(x + td)$ soit "suffisamment plus petit que" $f(x)$.

Pas de descente optimal : En théorie comme en pratique, il est utile de considérer la méthode qui donne le pas de descente pour lequel $f(x + td)$ est minimal, à savoir

$$t^\star = \operatorname*{argmin}_{t>0} f(x + td).$$

Exercice 3.5.

1. Justifier que pour f une fonction fortement convexe, $x \in \mathbb{R}^n$ et $d \in \mathbb{R}^n$ une direction de descente pour f en x, le pas de descente optimal est bien défini, c'est-à-dire que $t \mapsto f(x+td)$ admet un unique minimum global $t^\star$ sur $]0, +\infty[$.
2. Montrer que $t = t^\star$ si et seulement si $\langle \nabla f(x+td), d\rangle = 0$.
3. On considère le cas particulier d'une fonctionnelle quadratique

$$f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle + c,$$

où $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice carrée symétrique définie positive, $b \in \mathbb{R}^n$ et $c \in \mathbb{R}$. En considérant un point $x \in \mathbb{R}^n$, $x \neq x^\star$, et une direction de descente d pour x, montrer que le pas de descente optimal est donné par

$$t^\star = -\frac{\langle Ax - b, d\rangle}{\langle Ad, d\rangle}.$$

Solution de l'exercice 3.5.

1. La fonction $\varphi : \mathbb{R} \to \mathbb{R}$ définie par $\varphi(t) = f(x+td)$ est strictement convexe et coercive en tant que restriction sur une droite d'une fonction f strictement convexe et coercive. Donc φ admet un unique minimum global $t^\star$ sur $\mathbb{R}$. Il reste à justifier que $t^\star$ est positif. Or, par définition d'une direction de descente $\langle \nabla f(x), d\rangle < 0$, donc

$$f(x) > f(x+t^\star d) > f(x) + t^\star \langle \nabla f(x), d\rangle \Rightarrow t^\star > 0.$$

2. Comme $\varphi : t \mapsto f(x+td)$ est strictement convexe on a

$$t = t^\star \Leftrightarrow \varphi'(t) = 0 \Leftrightarrow \langle \nabla f(x+td), d\rangle = 0.$$

3. On a montré à l'Exercice 2.18 que $\nabla f(x) = Ax - b$. Ainsi,

$$\langle \nabla f(x+td), d\rangle = 0 \Leftrightarrow t\langle Ad, d\rangle + \langle Ax - b, d\rangle = 0 \Leftrightarrow \boxed{t = -\frac{\langle Ax - b, d\rangle}{\langle Ad, d\rangle}}$$

On peut également résoudre le problème de minimisation directement : On a montré à l'Exercice 2.18 que

$$f(x+h) = f(x) + \langle Ax - b, h\rangle + \frac{1}{2}\langle Ah, h\rangle.$$

Donc, pour $x, d \in \mathbb{R}^n$ et $t \in \mathbb{R}$,

$$f(x+td) = f(x) + t\langle Ax - b, d\rangle + t^2\frac{1}{2}\langle Ad, d\rangle.$$

Ainsi $\varphi(t) = f(x+td)$ est un trinôme du second degré dont le coefficient dominant $\frac{1}{2}\langle Ad, d\rangle$ est strictement positif. φ admet donc un minimum au point $t^\star$ tel que $\varphi'(t) = 0$, soit

$$\langle Ax - b, d\rangle + t^\star\langle Ad, d\rangle = 0 \Rightarrow \boxed{t^\star = -\frac{\langle Ax - b, d\rangle}{\langle Ad, d\rangle}}.$$

Calcul du pas de descente par méthode de rebroussement : On détaille maintenant la méthode de rebroussement (*backtracking* an anglais) qui permet de calculer un pas de descente lorsque l'on ne sait pas minimiser la fonction f sur la demi-droite affine $\{x + td,\ t > 0\}$. Cette méthode est décrite par l'Algorithme 2.

Algorithme 2 : Algorithme de calcul du pas de descente par méthode de rebroussement

Données : Un point $x \in \mathbb{R}^n$, une direction de descente associée $d \in \mathbb{R}^n$, deux réels $\alpha \in]0, \frac{1}{2}[$ et $\beta \in]0, 1[$
Résultat : Un pas de descente $t > 0$
Initialiser t :
  $t \leftarrow 1$;
tant que $f(x + td) > f(x) + \alpha t \langle \nabla f(x), d \rangle$ **faire**
  Réduire t d'un facteur β :
    $t \leftarrow \beta t$;
fin

Comme f est convexe, on sait que

$$f(x + td) \geqslant f(x) + t\langle \nabla f(x), d \rangle.$$

L'Algorithme 2 cherche donc à trouver un point t pour lequel cette borne inférieure réduite par un facteur α soit une borne supérieure. En effet, l'algorithme s'arrête dès lors que

$$f(x + td) \leqslant f(x) + \alpha t \langle \nabla f(x), d \rangle.$$

Comme d'après la définition du gradient de f en x, pour t proche de 0, $f(x + td)$ est proche de $f(x) + t\langle \nabla f(x), d \rangle < f(x) + \alpha t \langle \nabla f(x), d \rangle$, on est assuré que l'algorithme converge.

En pratique on choisira $\alpha \in [0.01, 0.3]$ et $\beta \in [0.1, 0.8]$.

3.3 Algorithmes de descente de gradient

On s'intéresse maintenant à l'étude d'un algorithme de descente particulier appelé algorithme de descente de gradient. Cet algorithme utilise comme direction de descente $d = d^{(k)}$ au point $x = x^{(k)}$ le vecteur opposé du gradient, soit

$$d^{(k)} = -\nabla f(x^{(k)}).$$

Algorithme 3 : Algorithme de descente de gradient

Données : Un point initial $x^{(0)} \in \mathbb{R}^n$, un seuil de tolérance $\varepsilon > 0$
Résultat : Un point $x \in \mathbb{R}^n$ proche de $x^\star$
Initialiser x :
$x \leftarrow x^{(0)}$;
$k \leftarrow 0$;
tant que $\|\nabla f(x)\| > \varepsilon$ **faire**

1. Calculer $d^{(k)} = -\nabla f(x) \quad (d^{(k)} = -\nabla f(x^{(k)}))$.
2. Déterminer un pas de descente $t^{(k)} > 0$ par la méthode exacte (ou par la méthode de rebroussement).
3. Mettre à jour x :
   $x \leftarrow x^{(k+1)} = x^{(k)} + t^{(k)} d^{(k)}$;
   $k \leftarrow k + 1$;

fin

Attention, en pratique, on teste plutôt le critère d'arrêt après l'étape 1. afin de ne pas calculer deux fois $\nabla f(x^{(k)})$. La convergence de l'algorithme est assurée par le théorème suivant.

Théorème 3.6 (Convergence de l'algorithme de descente de gradient)**.** Soient $f : \mathbb{R}^n \to \mathbb{R}$ une fonction fortement convexe telle que

$$\forall x, h \in \mathbb{R}^n, \quad m\|h\|^2 \leqslant \langle \nabla^2 f(x)h, h\rangle \leqslant M\|h\|^2$$

avec $0 < m \leqslant M$ et $x^{(0)}$ un point quelconque de $\mathbb{R}^n$. Alors l'algorithme de descente de gradient converge et on a convergence linéaire de la suite $(f(x^{(k)}) - p^\star)_{k\in\mathbb{N}}$:

$$\forall k \in \mathbb{N}, \quad f(x^{(k)}) - p^\star \leqslant c^k (f(x^{(0)}) - p^\star),$$

où $c \in [0, 1[$ dépend de la méthode de recherche de pas de descente et est donnée par

— $c = 1 - \dfrac{m}{N}$ pour la méthode exacte/optimale,
— $c = 1 - \min\left(2m\alpha, 2\beta\alpha\dfrac{m}{M}\right)$ pour la méthode de rebroussement utilisant les constantes $\alpha \in]0, \frac{1}{2}[$ et $\beta \in]0, 1[$.

Preuve. On rappelle que sous les hypothèses du théorème on a démontré à l'Exercice 3.4 l'encadrement

$$\frac{2}{m}(f(x^{(k)}) - p^\star) \leqslant \|x^{(k)} - x^\star\|^2 \leqslant \frac{2}{M}(f(x^{(k)}) - p^\star).$$

qui montre que la convergence de $(f(x^{(k)}))_{k\in\mathbb{N}}$ vers $p^\star$ entraîne la convergence de $(x^{(k)})_{k\in\mathbb{N}}$ vers $x^\star$. On donne maintenant la preuve de la convergence linéaire de $(f(x^{(k)}) - p^\star)_{k\in\mathbb{N}}$ dans le cas de la méthode optimale pour le calcul du pas de descente. La preuve pour le cas de la méthode de rebroussement fait l'objet de l'Exercice 3.7.

On suppose que l'algorithme de gradient est à l'itération k et que $x^{(k)} \neq x^\star$ (sinon l'algorithme a convergé en un nombre fini $l \leqslant k$ d'itérations et la suite $(x^{(k)})$ est constante à $x^\star$ à partir du rang l). Le point $x^{(k+1)}$ est de la forme $x^{(k+1)} = x^{(k)} + t^{(k)}\nabla f(x^{(k)})$ avec

$$t^{(k)} = \operatorname*{argmin}_{t>0} f(x^{(k)} - t\nabla f(x^{(k)})).$$

On a pour tous $x, y \in \mathbb{R}^n$,

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{M}{2}\|y - x\|^2.$$

En prenant $x = x^{(k)}$ et $y = x^{(k)} - t\nabla f(x^{(k)})$ pour tout $t > 0$ on a

$$f(x^{(k)} - t\nabla f(x^{(k)})) \leqslant f(x^{(k)}) - t\|\nabla f(x^{(k)})\|^2 + t^2\frac{M}{2}\|\nabla f(x^{(k)})\|^2.$$

Par définition, le membre de gauche est minimal en $t = t^{(k)}$ et vaut alors $f(x^{(k)} - t^{(k)}\nabla f(x^{(k)})) = f(x^{(k+1)})$. Ainsi pour tout $t > 0$, on a

$$f(x^{(k+1)}) \leqslant f(x^{(k)}) + \left(\frac{M}{2}t^2 - t\right)\|\nabla f(x^{(k)})\|^2.$$

Le membre de droite est minimal en $t = \frac{1}{M}$, et pour cette valeur de t on obtient la majoration

$$f(x^{(k+1)}) \leqslant f(x^{(k)}) - \frac{1}{2M}\|\nabla f(x^{(k)})\|^2.$$

On soustrait ensuite $p^\star$ à cette inégalité

$$f(x^{(k+1)}) - p^\star \leqslant f(x^{(k)}) - p^\star - \frac{1}{2M}\|\nabla f(x^{(k)})\|^2.$$

Enfin, d'après la Proposition 3.2, pour tout $x \in \mathbb{R}^n$,

$$p^\star \geqslant f(x) - \frac{1}{2m}\|\nabla f(x)\|^2$$

et donc pour $x = x^{(k)}$,

$$\|\nabla f(x^{(k)})\|^2 \geqslant 2m(f(x^{(k)}) - p^\star).$$

Ainsi,

$$f(x^{(k+1)}) - p^\star \leqslant \left(1 - \frac{m}{M}\right)\left(f(x^{(k)}) - p^\star\right).$$

Par récurrence, on a donc bien $f(x^{(k)}) - p^\star \leqslant c^k(f(x^{(0)}) - p^\star)$ avec $c = 1 - \frac{m}{M} \in [0, 1[$. □

Exercice 3.7 (Convergence de l'algorithme de descente de gradient pour la méthode de rebroussement)**.** On reprend les notations de la preuve du Théorème 3.6.

1. Montrer que pour tout $t \in [0, \frac{1}{M}]$ on a $\frac{M}{2}t^2 - t \leqslant -\frac{t}{2}$ *(on pourra par exemple utiliser la convexité de* $h(t) = \frac{M}{2}t^2 - t$*).*
2. En déduire que pour tout $t \in [0, \frac{1}{M}]$,

$$f(x^{(k)} - t\nabla f(x^{(k)})) \leqslant f(x^{(k)}) - \alpha t\|\nabla f(x^{(k)})\|^2.$$

3. En déduire que la méthode de rebroussement s'arrête soit en $t^{(k)} = 1$ soit pour une valeur $t^{(k)} \geqslant \frac{\beta}{M}$.
4. En déduire que

$$f(x^{(k+1)}) \leqslant f(x^{(k)}) - \min\left(\alpha, \frac{\alpha\beta}{M}\right)\|\nabla f(x^{(k)})\|^2$$

et conclure la preuve comme pour le cas du pas optimal.

Solution de l'exercice 3.7.

1. La fonction $h(t) = \frac{M}{2}t^2 - t$ est convexe en tant que somme de fonctions convexes. Pour tout $t \in [0, \frac{1}{M}]$,

$$h(t) \leqslant (1-t)Mh(0) + tMh(\frac{1}{M}) = tM\frac{-1}{2M} = -\frac{t}{2}.$$

2. En utilisant l'inégalité précédente, on a pour tout $t \in [0, \frac{1}{M}]$,

$$f(x^{(k)} - t\nabla f(x^{(k)})) \leqslant f(x^{(k)}) + \left(\frac{M}{2}t^2 - t\right)\|\nabla f(x^{(k)})\|^2 \leqslant f(x^{(k)}) - \frac{t}{2}\|\nabla f(x^{(k)})\|^2.$$

   Comme $\alpha < \frac{1}{2}$ on a bien

$$f(x^{(k)} - t\nabla f(x^{(k)})) \leqslant f(x^{(k)}) - \alpha t\|\nabla f(x^{(k)})\|^2.$$

3. Si la méthode de rebroussement ne s'arrête pas en $t^{(k)} = 1$, alors elle s'arrête en $t^{(k)} = \beta^m$ tel que

$$f(x^{(k)} - \beta^m\nabla f(x^{(k)})) \leqslant f(x^{(k)}) - \alpha t\|\nabla f(x^{(k)})\|^2$$

   et

$$f(x^{(k)} - \beta^{m-1}\nabla f(x^{(k)})) > f(x^{(k)}) - \alpha t\|\nabla f(x^{(k)})\|^2$$

   Ainsi, vu la question précédente, on a nécessairement $\beta^{m-1} > \frac{1}{M}$, donc $t^{(k)} = \beta^m > \frac{\beta}{M}$.

4. D'après la question précédente, on a la majoration $t^{(k)} \leqslant \min\left(1, \frac{\beta}{M}\right)$. Ainsi, on a

$$\begin{aligned} f(x^{(k+1)}) &= f(x^{(k)} - t^{(k)}\nabla f(x^{(k)})) \\ &\leqslant f(x^{(k)}) - \alpha t^{(k)}\|\nabla f(x^{(k)})\|^2 \\ &\leqslant f(x^{(k)}) - \min\left(\alpha, \frac{\alpha\beta}{M}\right)\|\nabla f(x^{(k)})\|^2. \end{aligned}$$

   On soustrait ensuite par $p^\star$ et on utilise l'inégalité

$$\|\nabla f(x^{(k)})\|^2 \geqslant 2m(f(x^{(k)}) - p^\star).$$

   pour avoir

$$f(x^{(k+1)} - p^\star \leqslant \left(1 - \min\left(2m\alpha, \frac{2\alpha\beta m}{M}\right)\right)(f(x^{(k)}) - p^\star)$$

   et conclure par récurrence.

L'algorithme de descente de gradient avec la recherche de pas de descente exacte est souvent appelé *algorithme de gradient à pas optimal* [CIARLET]. Attention, c'est le pas qui est optimal, et non l'algorithme ! On étudiera des algorithmes plus "optimaux", c'est-à-dire qui convergent plus rapidement. En terme de fonction objectif, la convergence de cet algorithme est donc linéaire avec la constante $c = 1 - \frac{m}{M}$. On rappelle que m et M sont respectivement des bornes inférieures et supérieures sur les valeurs propres des matrices hessiennes $\nabla^2 f(x)$, $x \in S$. En particulier se sont des bornes sur les valeurs propres de la matrice hessienne au point optimal $x^\star$. Cela suggère que la convergence de l'algorithme de descente de gradient est d'autant plus rapide

si la matrice hessienne $\nabla^2 f(x^\star)$ est bien conditionnée (on rappelle que pour une matrice réelle symétrique A le conditionnement $\mathrm{cond}(A)$ correspond au rapport des plus grande et plus petite valeur propre, et donc $\mathrm{cond}(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \leqslant \frac{M}{m}$). On verra qu'en pratique cette observation est vérifiée. Plus rigoureusement, pour une fonctionnelle quadratique $f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle + c$ on a $\nabla^2 f(x) = A$ pour tout x, et donc on peut prendre $m = \lambda_{\min}(A)$, $M = \lambda_{\max}(A)$, et alors on a $\mathrm{cond}(A) = \frac{m}{M}$. En particulier, si $\mathrm{cond}(A) = 1$, c'est-à-dire si $A = \lambda I$ (avec $\lambda > 0$) est une matrice scalaire, alors l'algorithme du gradient à pas optimal converge en une itération !

Exercice 3.8. Vérifier que si f est une fonctionnelle quadratique avec $A = \lambda I$, $\lambda > 0$, alors l'algorithme du gradient à pas optimal converge en une itération.

Solution de l'exercice 3.8. On sait que $x^\star$ est l'unique solution du système $Ax = b$, soit ici $x^\star = \frac{1}{\lambda} b$. D'après l'Exercice 3.5, le pas de descente $t^{(0)}$ est donné par

$$t^{(0)} = -\frac{\langle Ax^{(0)} - b, d\rangle}{\langle Ad, d\rangle}.$$

Ici $d = -\nabla f(x^{(0)}) = -Ax^{(0)} + b = -\lambda x^{(0)} + b$ d'où

$$t^{(0)} = -\frac{\|\lambda x^{(0)} - b\|^2}{\lambda\| - \lambda x^{(0)} + b\|^2} = -\frac{1}{\lambda},$$

d'où

$$x^{(1)} = x^{(0)} - \frac{1}{\lambda}(\lambda x^{(0)} - b) = \frac{1}{\lambda} b = x^\star.$$

3.4 Méthode de Newton

La méthode de Newton est un algorithme de descente pour lequel le pas de descente $d^{(k)}$ au point $x^{(k)}$ est donné par

$$d^{(k)} = -\nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)}).$$

Le calcul de ce pas de descente nécessite donc la résolution d'un système linéaire de taille $n \times n$.

Remarque (Résolution de système linéaire)**.** On rappelle que l'évaluation numérique d'un vecteur x de la forme

$$x = A^{-1} b$$

ne doit en général jamais s'effectuer en calculant la matrice inverse A^{-1} puis en multipliant par le vecteur b mais en résolvant le système linéaire

$$Ax = b.$$

En Octave cela s'écrit `x = A\b` (et surtout pas `x = A^(-1)*b` !).

Avant de poursuivre l'étude de l'algorithme de Newton, justifions le choix de ce pas de descente. Pour une fonction f deux fois différentiable en x, la formule de Taylor-Young assure que

$$f(x+h) = f(x) + \langle \nabla f(x), h\rangle + \frac{1}{2}\langle \nabla^2 f(x) h, h\rangle + \|h\|^2 \varepsilon(h)$$

avec $\lim_{h\to 0} \varepsilon(h) = 0$. La fonction

$$g(x+h) = f(x) + \langle \nabla f(x), h\rangle + \frac{1}{2}\langle \nabla^2 f(x)h, h\rangle$$

est donc l'approximation d'ordre deux de f au voisinage de x. Cette fonction g est une fonctionnelle quadratique (en la variable h) avec $A = \nabla^2 f(x)$ et $b = -\nabla f(x)$. Elle est donc minimale pour le vecteur

$$h^\star = -\nabla^2 f(x)^{-1}\nabla f(x)$$

qui est le pas de Newton. Autrement dit le pas de Newton $d^{(k)} = -\nabla^2 f(x^{(k)})^{-1}\nabla f(x^{(k)})$ est choisi de sorte à ce que $x^{(k)} + d^{(k)}$ minimise l'approximation à l'ordre deux au point $x^{(k)}$ de la fonction f.

Proposition 3.9. La pas de Newton $d = -\nabla^2 f(x^{(k)})^{-1}\nabla f(x^{(k)})$ est invariant par changement de variable affine.

La preuve de cette proposition fait l'objet de l'exercice suivant.

Exercice 3.10. Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction deux fois différentiable. Soient $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée inversible et $b \in \mathbb{R}^n$. On pose $g(x) = f(Ax + b)$.

1. Exprimer le gradient et la matrice hessienne de g en fonction du gradient et de la matrice hessienne de f.
2. Soit $y \in \mathbb{R}^n$. Soit d le pas de descente de Newton de f au point $x = Ay + b$ et d' le pas de descente de Newton de g au point y. Montrer que $d = Ad'$ et $x + d = A(y + d') + b$.

Solution de l'exercice 3.10.

1. On a
$$\nabla g(x) = A^T \nabla f(Ax + b) \quad \text{et} \quad \nabla^2 g(x) = A^T \nabla^2 f(Ax + b)A$$
2. On a
$$\begin{aligned} d' &= -\nabla g^2(y)^{-1}\nabla g(y) \\ &= -\left(A^T\nabla^2 f(Ay + b)A\right)^{-1} A^T \nabla f(Ay + b) \\ &= -A^{-1}\nabla^2 f(x)^{-1}(A^T)^{-1}A^T\nabla f(x) \\ &= -A^{-1}\nabla^2 f(x)^{-1}\nabla f(x) \\ &= A^{-1}d. \end{aligned}$$
D'où
$$A(y + d') + b = Ay + b + AA^{-1}d = x + d.$$

Cette proposition est fondamentale. Alors que l'algorithme de descente de gradient est très influencé par le conditionnement de la matrice hessienne, l'algorithme de descente de Newton est invariant par changement de variable affine.

Critère d'arrêt invariant par changement de variable affine : Comme pour toute les méthodes de descente, le critère d'arrêt $\|\nabla f(x)\|^2 \leqslant \varepsilon^2$ est valide pour la méthode de Newton, mais il n'est pas invariant par changement de variable affine. Pour cela on préfère utiliser le critère $\Lambda(x) \leqslant \varepsilon^2$ où

$$\Lambda(x) = \langle \nabla^2 f(x)^{-1}\nabla f(x), \nabla f(x)\rangle = -\langle d, \nabla f(x)\rangle$$

est la norme au carré de $\nabla f(x)$ pour la norme associée à la matrice symétrique définie positive $\nabla^2 f(x)^{-1}$. On remarque que le produit scalaire $\langle d, \nabla f(x)\rangle = -\Lambda(x)$ est calculé par ailleurs dans la méthode de rebroussement pour le calcul du pas de descente, donc ce critère d'arrêt n'ajoute aucun coût de calcul.

L'algorithme de Newton est donné par l'Algorithme 4.

Algorithme 4 : Algorithme de descente de Newton

Données : Un point initial $x^{(0)} \in \mathbb{R}^n$, un seuil de tolérance $\varepsilon > 0$, des paramètres $\alpha \in]0, \frac{1}{2}[$ et $\beta \in]0, 1[$ pour la méthode de rebroussement

Résultat : Un point $x \in \mathbb{R}^n$ proche de $x^\star$

Initialiser x :

$x \leftarrow x^{(0)}$;

$k \leftarrow 0$;

Calculer la première direction de descente :

$d^{(0)} = -\nabla^2 f(x^{(0)})^{-1} \nabla f(x^{(0)})$;

$\Lambda^{(0)} = -\langle d^{(0)}, \nabla f(x^{(0)})\rangle$;

tant que $\Lambda^{(k)} > \varepsilon^2$ **faire**

1. Déterminer un pas de descente $t^{(k)} > 0$ au point $x^{(k)}$ selon la direction $d^{(k)}$ par la méthode de rebroussement avec les paramètres α et β.
2. Mettre à jour x :
   $x \leftarrow x^{(k+1)} = x^{(k)} + t^{(k)} d^{(k)}$;
   $k \leftarrow k + 1$;
3. Calculer la nouvelle direction de descente :
   $d^{(k)} = -\nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)})$;
   $\Lambda^{(k)} = -\langle d^{(k)}, \nabla f(x^{(k)})\rangle$;

fin

Théorème 3.11 (Convergence de la méthode de Newton)**.** Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction fortement convexe telle que

$$\forall x, h \in \mathbb{R}^n, \quad m\|h\|^2 \leqslant \langle \nabla^2 f(x)h, h\rangle \leqslant M\|h\|^2$$

avec $0 < m \leqslant M$ et dont la matrice hessienne est lipschitzienne pour la constante $L > 0$

$$\forall x, y \in \mathbb{R}^n, \quad \|\nabla^2 f(x) - \nabla^2 f(y)\|_{\mathcal{M}_n(\mathbb{R})} \leqslant L\|x - y\|.$$

Soit $x^{(0)}$ un point quelconque de $\mathbb{R}^n$. On pose

$$\eta = \min(1, 3(1 - 2\alpha))\frac{m^2}{L} \quad \text{et} \quad \gamma = \alpha\beta\eta^2 \frac{m}{M^2}.$$

Alors on a :

— Si $\|\nabla f(x^{(k)})\| \geqslant \eta$, alors

$$f(x^{(k+1)}) - f(x^{(k)}) \leqslant -\gamma.$$

— Si $\|\nabla f(x^{(k)})\| < \eta$, alors la méthode de rebroussement retourne le pas $t^{(k)} = 1$ et

$$\frac{L}{2m^2}\|\nabla f(x^{(k+1)})\| \leqslant \left(\frac{L}{2m^2}\|\nabla f(x^{(k)})\|\right)^2.$$

En particulier, l'algorithme de Newton converge et atteint un régime de convergence quadratique au bout d'un nombre fini d'itérations.

Le théorème est admis. On renvoie à [BOYD & VANDENBERGHE, pp. 488-491] pour une preuve détaillée.

Exercice 3.12. Vérifier que si $f : \mathbb{R}^n \to \mathbb{R}$ est une fonctionnelle quadratique, alors l'algorithme de Newton converge en une seule itération.

Solution de l'exercice 3.12. Vérifions que $t^{(0)} = 1$ est accepté (VOIR BROUILLON A RECOPIER).

Ensuite,

$$x^{(1)} = x^{(0)} - A^{-1}(Ax^{(0)} - b) = A^{-1}b.$$

On retiendra que cela n'a pas de sens d'utiliser l'algorithme de Newton pour minimiser une fonctionnelle quadratique. L'algorithme de Newton est utile pour minimiser des fonctionnelles non quadratiques, et il consiste à minimiser une fonctionnelle quadratique à chaque itération, ce qui implique la résolution d'un système linéaire de taille $n \times n$. Chaque itération a donc un coût de calcul non négligeable, mais en revanche l'algorithme converge très rapidement et nécessite un faible nombre d'itérations pour atteindre une grande précision numérique.

A l'opposé, l'objet du prochain chapitre est d'introduire un algorithme plus performant que l'algorithme de descente de gradient pour minimiser des fonctionnelles quadratiques.

Chapitre 4

Méthode du gradient conjugué

Les références principales pour ce chapitre sont la section 8.5 de [CIARLET] et le chapitre 9 de [ALLAIRE & KABER].

Le but de la méthode du gradient conjugué est de minimiser une fonctionnelle quadratique $f : \mathbb{R}^n \to \mathbb{R}$ de la forme

$$f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle + c,$$

où $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice carrée symétrique définie positive, $b \in \mathbb{R}^n$ et $c \in \mathbb{R}$. On rappelle que ce problème équivaut à résoudre le système linéaire

$$Ax = b.$$

On a étudié au chapitre précédent l'algorithme du gradient à pas optimal qui permet de minimiser ces fonctionnelles avec une vitesse de convergence linéaire. Pour cela, à chaque itération l'algorithme de descente de gradient à pas optimal minimise la fonctionnelle quadratique f sur la droite affine

$$\{x^{(k)} - t\nabla f(x^{(k)}),\ t \in \mathbb{R}\}.$$

L'idée principale de l'algorithme du gradient conjugué est d'élargir à chaque itération l'espace sur lequel on minimise la fonctionnelle quadratique : à chaque itération k on va minimiser la fonctionnelle sur un espace affine de dimension $k + 1$. Le corollaire immédiat de cette démarche est que l'algorithme converge nécessairement en au plus n itérations. La difficulté principale réside dans le fait de montrer que tous ces problèmes intermédiaires de minimisation sur des espaces de dimension croissantes sont facilement résolubles.

4.1 Description de l'algorithme et preuve de sa convergence

Définissons plus précisément l'algorithme du gradient conjugué. Comme pour toutes les méthodes de descente, on se donne un vecteur initial $x^{(0)}$ quelconque de $\mathbb{R}^n$. La première itération correspond exactement à celle du gradient à pas optimal : On se donne la direction de descente $d^{(0)} = -\nabla f(x^{(0)})$ et on minimise f sur la droite affine

$$x^{(0)} + G_0 = x^{(0)} + \mathrm{Vect}(\nabla f(x^{(0)})).$$

Pour les itérations suivantes, on suppose que l'on a déjà construit les k premiers vecteurs $x^{(1)}$, $x^{(2)}, \ldots, x^{(k)}$. On suppose de plus que pour tout $0 \leqslant i \leqslant k$,

$$\nabla f(x^{(i)}) \neq 0$$

(sinon on aurait déjà trouvé la solution $x^\star$ et l'algorithme serait terminé). Pour chaque $i \in \{0, 1, \ldots, k\}$ on définit l'espace

$$G_i = \text{Vect}\{\nabla f(x^{(j)}) \mid j \in \{0, 1, ...i\}\} = \left\{ \sum_{j=0}^{i} \eta_j \nabla f(x^{(j)}) \mid (\eta_0, \eta_1, ..., \eta_i) \in \mathbb{R}^{i+1} \right\}.$$

Les espaces G_i sont a priori de dimension inférieure à $i+1$ mais on va montrer qu'ils sont exactement de dimension $i+1$, c'est-à-dire que les vecteurs $\nabla f(x^{(j)})$ sont linéairement indépendants.

L'idée essentielle de l'algorithme du gradient conjugué est de définir $x^{(k+1)}$ comme le minimum de f sur tout l'espace affine $x^{(k)} + G_k$, c'est-à-dire

$$x^{(k+1)} \in (x^{(k)} + G_k) \quad \text{et} \quad f(x^{(k+1)}) = \min_{y \in (x^{(k)} + G_k)} f(y) \tag{4.1}$$

où $(x^{(k)} + G_k) = \{x^{(k)} + x \mid x \in G_k\}$.

Théorème 4.2 (Convergence de l'algorithme du gradient conjugué)**.** L'algorithme du gradient conjugué converge en au plus n itérations.

La preuve de ce théorème est l'objet de l'exercice suivant.

Exercice 4.3.

1. Justifier que la restriction d'une fonctionnelle quadratique à un sous-espace affine est encore une fonctionnelle quadratique. Pour cela on considérera un sous-espace affine $z + G$ de dimension l, $1 \leqslant l \leqslant n$, une base orthonormale $e_1, \ldots, e_l$ de G, et la restriction

$$\begin{array}{rcl} g : \mathbb{R}^l & \to & \mathbb{R} \\ y & \mapsto & f(z + Dy) = f\left(z + \sum_{j=1}^{l} y_j e_j\right) \end{array}$$

   où la matrice $D \in \mathcal{M}_{n,l}(\mathbb{R})$ est la matrice dont les colonnes sont les vecteurs $e_1, \ldots, e_l$.
2. En déduire que le problème (4.1) admet une unique solution $x^{(k+1)}$.
3. En utilisant le théorème 2.16 sur les minimums des fonctions convexes, montrer que pour tout $y \in G_k$,
$$\langle \nabla f(x^{(k+1)}), y \rangle = 0,$$
   i.e. $\nabla f(x^{(k+1)}) \in G_k^\perp$.
4. En déduire que soit $x^{(k+1)} = x^\star$, soit le vecteur $\nabla f(x^{(k+1)})$ est orthogonal à chacun des vecteurs $\nabla f(x^{(0)}), \nabla f(x^{(1)}), \ldots \nabla f(x^{(k)})$.
5. Montrer que l'algorithme du gradient conjugué converge en au plus n itérations.

Solution de l'exercice 4.3.

1. Il s'agit de montrer que $y \mapsto f(Dy + z)$ est quadratique en développant, avec D représentant une base de l'espace affine.
   On a
$$\min_{x \in z+G} f(x) = \min_{y \in \mathbb{R}^l} f(z + Dy) = \min_{y \in \mathbb{R}^l} g(y).$$

Or

$$\begin{aligned} g(y) &= f(z+Dy) \\ &= \frac{1}{2}\langle A(z+Dy),(z+Dy)\rangle - \langle b,(z+Dy)\rangle + c \\ &= \frac{1}{2}\langle ADy,Dy\rangle + \frac{1}{2}\langle ADy,z\rangle + \frac{1}{2}\langle Az,Dy\rangle - \langle b,Dy\rangle + f(z) \\ &= \frac{1}{2}\langle D^TADy,y\rangle - \langle D^T(b-Az),y\rangle + f(z). \end{aligned}$$

On a bien une fonctionnelle quadratique pour la variable y.

2. La fonctionnelle est quadratique. Vérifions que la matrice symétrique D^TAD est définie positive. Soit y tel que $\langle D^TADy,y\rangle = 0$. $\langle D^TADy,y\rangle = \langle ADy,Dy\rangle = 0$ et A est définie positive donc $Dy = 0$. Mais D correspond aux vecteurs de base, donc la matrice est de rang l ou encore $\ker(D) = \{0\}$ et on a bien $y = 0$. Donc la fonctionnelle est quadratique avec une matrice symétrique définie positive, le problème admet une unique solution.
3. Le théorème 2.16 assure que $x^{(k+1)}$ est solution du problème de minimisation sous contrainte (4.1) si et seulement si pour tout $y \in G_k$ $\langle \nabla f(x^{(k+1)}), y\rangle = 0$.
4. Soit $\nabla f(x^{(k+1)}) = 0$ et alors $x^{(k+1)} = x^\star$, soit $\nabla f(x^{(k+1)}) \neq 0$ et $\nabla f(x^{(k+1)})$ est orthogonal à chaque vecteur $\nabla f(x^{(0)}), \nabla f(x^{(1)}), \dots \nabla f(x^{(k)}) \in G_k$.
5. Par récurrence immédiate, les vecteurs $\nabla f(x^{(i)})$ sont deux à deux orthogonaux. Donc les vecteurs $(\nabla f(x^{(0)}), \nabla f(x^{(1)}), \dots \nabla f(x^{(k+1)}))$ forment une famille orthogonale de $\mathbb{R}^n$. Si à l'itération $k = n$ on avait encore $\nabla f(x^{(n)}) \neq 0$, alors on aurait une famille de $n+1$ vecteurs orthogonaux de $\mathbb{R}^n$.

Remarque. On peut déjà dire que cette méthode est supérieure à celle du gradient à pas optimal car la droite $\{x^{(k)} - t\nabla f(x^{(k)}),\ t \in \mathbb{R}\}$ est incluse strictement dans $(x^{(k)} + G_k)$. La convergence est acquise, mais il reste à montrer qu'il existe des relations simples pour implémenter l'algorithme du gradient conjugué, c'est-à-dire des relations qui permettent de calculer directement les vecteurs $x^{(k+1)} \in (x^{(k)} + G_k)$ à chaque étape sans avoir recours à des algorithmes complexes de minimisation.

4.2 Implémentation de l'algorithme du gradient conjugué

Le problème du calcul des itérés $x^{(k)}$ est resté en suspend dans la partie précédente. Cependant c'est un point essentiel pour l'implémentation de l'algorithme. On va montrer en utilisant le caractère quadratique de la fonctionnelle F que l'on peut trouver des expressions explicites pour ces vecteurs.

On définit les $k+1$ vecteurs des différences

$$d^{(i)} = x^{(i+1)} - x^{(i)} = \sum_{j=0}^{i} d_j^{(i)} \nabla f(x^{(j)}), \quad 0 \leqslant i \leqslant k.$$

Comme F est quadratique on a

$$\nabla f(x+y) = A(x+y) - b = \nabla f(x) + Ay, \quad \forall x, y \in \mathbb{R}^n.$$

On a donc en particulier

$$\nabla f(x^{(i+1)}) = \nabla f(x^{(i)} + d^{(i)}) = \nabla f(x^{(i)}) + Ad^{(i)}, \quad 0 \leqslant i \leqslant k.$$

De l'orthogonalité des vecteurs gradients successifs $\nabla f(x^{(i)})$, $0 \leqslant i \leqslant k$, on déduit d'une part

$$0 = \langle \nabla f(x^{(i+1)}), \nabla f(x^{(i)}) \rangle = \|\nabla f(x^{(i)})\|^2 + \langle Ad^{(i)}, \nabla f(x^{(i)}) \rangle, \quad 0 \leqslant i \leqslant k$$

et donc comme on a supposé $\nabla f(x^{(i)}) \neq 0,\ 0 \leqslant i \leqslant k$ on obtient

$$d^{(i)} \neq 0,\ 0 \leqslant i \leqslant k, \tag{4.4}$$

d'autre part pour $k \geqslant 1$ et $0 \leqslant j < i \leqslant k$:

$$0 = \langle \nabla f(x^{(i+1)}), \nabla f(x^{(j)}) \rangle = \langle f(x^{(i)}), \nabla f(x^{(j)}) \rangle + \langle Ad^{(i)}, \nabla f(x^{(j)}) \rangle = \langle Ad^{(i)}, \nabla f(x^{(j)}) \rangle. \tag{4.5}$$

De plus comme chaque vecteur $d^{(j)}$ est une combinaison linéaire des vecteurs $\nabla f(x_l),\ 1 \leqslant l \leqslant j$, (4.5) entraîne que

$$\langle Ad^{(i)}, d^{(j)} \rangle = 0,\ 0 \leqslant j < i \leqslant k. \tag{4.6}$$

On dit alors que les vecteurs $d^{(i)}$ sont *conjugués* par rapport à la matrice A, c'est à dire qu'il sont orthogonaux par rapport au produit scalaire défini par la matrice A qui est symétrique définie positive. Ce point de vue nous montre directement que des vecteurs non nuls et conjugués par rapport à A sont linéairement indépendants (car orthogonaux par rapport à un produit scalaire).

Les vecteurs $d^{(i)}$ sont non nuls d'après (4.4) et conjugués par rapport à A, donc les vecteurs $d^{(i)}$ sont linéairement indépendants.

Les vecteurs $\nabla f(x^{(i)}),\ 0 \leqslant i \leqslant k$, et les vecteurs $d^{(i)} = \sum_{j=0}^{i} d_j^{(i)} \nabla f(x^{(j)}),\ 0 \leqslant i \leqslant k$, sont linéairement indépendants, l'égalité entre les matrices de rang $(k+1)$:

$$\left(\ d^{(0)}|d^{(1)}|...|d^{(k)}\ \right) = \left(\ \nabla f(x^{(0)})|\nabla f(x^{(1)})|...|\nabla f(x^{(k)})\ \right) \begin{pmatrix} d_0^{(0)} & d_0^{(1)} & \dots & d_0^{(k)} \\ 0 & d_1^{(1)} & \dots & d_1^{(k)} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & d_k^{(k)} \end{pmatrix}$$

impose que les éléments diagonaux $d_i^{(i)}$ sont non nuls pour $0 \leqslant i \leqslant k$.

On doit maintenant calculer les composantes des ces vecteurs $d^{(i)}$. En repensant au formalisme $f(x^{(k+1)}) = f(x^{(k)} - t^{(k)} d^{(k)})$, on voit que l'on peut tout d'abord s'intéresser à la direction de $d^{(k)}$ et ensuite déterminer le $t^{(k)}$ qui minimise $f(x^{(k)} - t^{(k)} d^{(k)})$.

A partir de maintenant on renormalise chaque vecteur $d^{(i)}$ afin d'imposer que $d_i^{(i)} = 1,\ 0 \leqslant i \leqslant k$, c'est-à-dire que ces nouveaux vecteurs (que l'on appelle toujours $d^{(i)}$) sont de la forme :

$$d^{(i)} = \sum_{j=0}^{i-1} d_j^{(i)} \nabla f(x^{(j)}) + \nabla f(x^{(i)}),\ 0 \leqslant i \leqslant k. \tag{4.7}$$

On a d'après (4.6) pour $0 \leqslant j \leqslant k-1$:

$$0 = \langle d^{(k)}, Ad^{(j)} \rangle = \langle d^{(k)}, t^{(k)} Ad^{(j)} \rangle = \langle d^{(k)}, \nabla f(x^{(j+1)}) - \nabla f(x^{(j)}) \rangle$$

d'où en utilisant la relation de décomposition (4.7) et celle d'orthogonalité (4.5) on obtient

— pour $0 \leqslant j \leqslant k-2$:

$$d_{j+1}^{(k)} \|\nabla f(x^{(j+1)})\|^2 - d_j^{(k)} \|\nabla f(x^{(j)})\|^2 = 0$$

— pour $j = k-1$:

$$\|\nabla f(x^{(k)})\|^2 - d_{k-1}^{(k)} \|\nabla f(x^{(k-1)})\|^2 = 0$$

d'où par récurrence descendante,

$$\forall 0 \leqslant j \leqslant k-1,\ d_j^{(k)} = \frac{\|\nabla f(x^{(k)})\|^2}{\|\nabla f(x^{(j)})\|^2}.$$

On sait ainsi calculer tous les $d_j^{(k)}$ pour trouver (la direction de) $d^{(k)}$ à l'étape k de l'algorithme. Leur nombre étant croissant à chaque étape, le calcul de $d^{(k)}$ devrait prendre de plus en plus de temps. Cependant il existe une relation de récurrence très simple entre les $d^{(k)}$. En effet on a

$$\begin{aligned} d^{(k)} &= \nabla f(x^{(k)}) + \sum_{i=0}^{k-1} \frac{\|\nabla f(x^{(k)})\|^2}{\|\nabla f(x_i)\|^2} \nabla f(x^{(i)}) \\ &= \nabla f(x^{(k)}) + \frac{\|\nabla f(x^{(k)})\|^2}{\|\nabla f(x^{(k-1)})\|^2} \underbrace{\left(\nabla f(x^{(k-1)}) + \sum_{i=0}^{k-2} \frac{\|\nabla f(x^{(k-1)})\|^2}{\|\nabla f(x_i)\|^2} \nabla f(x^{(i)}) \right)}_{d^{(k-1)}} \end{aligned}$$

c'est-à-dire

$$d^{(k)} = \nabla f(x^{(k)}) + \frac{\|\nabla f(x^{(k)})\|^2}{\|\nabla f(x^{(k-1)})\|^2} d^{(k-1)}.$$

Cette relation est doublement heureuse car non seulement elle permet de calculer $d^{(k)}$ sans calculer toutes ses coordonnées $d_i^{(k)}$, mais en plus elle montre que pour calculer $d^{(k)}$ on n'a pas besoin de garder en mémoire toutes les valeurs $\nabla f(x^{(i)})$ mais seulement les deux dernières $\nabla f(x^{(k-1)})$ et $\nabla f(x^{(k)})$.

Pour finir d'expliciter les calculs il ne reste plus qu'à calculer $t^{(k)}$ défini par

$$f(x^{(k+1)}) = f(x^{(k)} - t^{(k)} d^{(k)}) = \inf_{t \in \mathbb{R}} f(x^{(k)} - t d^{(k)}).$$

On a déjà démontré que ceci revient à minimiser un trinôme du second degré et que la solution est donnée par

$$t^{(k)} = \frac{\langle \nabla f(x^{(k)}), d^{(k)} \rangle}{\langle A d^{(k)}, d^{(k)} \rangle}.$$

Nous avons démontré dans cette section plusieurs formules de récurrence qui nous permettent désormais d'expliciter chaque itération de l'algorithme du gradient conjugué d'un point de vue numérique.

On part d'un vecteur arbitraire $x^{(0)}$. Si $\nabla f(x^{(0)}) = 0$ alors $x^{(0)} = x^\star$ et l'algorithme s'arrête, sinon on pose

$$d^{(0)} = \nabla f(x^{(0)})$$

et

$$t^{(0)} = \frac{\langle \nabla f(x^{(0)}), d^{(0)} \rangle}{\langle A d^{(0)}, d^{(0)} \rangle}$$

puis le vecteur

$$x^{(1)} = x^{(0)} - t^{(0)} d^{(0)}.$$

Si on suppose construits de proche en proche les vecteurs $x^{(1)}, d^{(1)}, ..., x^{(k-1)}, d^{(k-1)}, x^{(k)}$, (ce qui sous-entend que les $\nabla f(x^{(i)})$ sont tous non nuls), deux cas se présentent :

— soit $\nabla f(x^{(k)}) = 0$ et alors $x^{(k)} = x^\star$ et l'algorithme est terminé,
— soit $\nabla f(x^{(k)}) \neq 0$ et alors on définit successivement

$$d^{(k)} = \nabla f(x^{(k)}) + \frac{\|\nabla f(x^{(k)})\|^2}{\|\nabla f(x^{(k-1)})\|^2} d^{(k-1)},$$

$$t^{(k)} = \frac{\langle \nabla f(x^{(k)}), d^{(k)} \rangle}{\langle A d^{(k)}, d^{(k)} \rangle}$$

et

$$x^{(k+1)} = x^{(k)} - t^{(k)} d^{(k)},$$

et on recommence cette étape avec ce vecteur $x^{(k+1)}$.

On voit finalement que cet algorithme n'effectue que des calculs élémentaires alors que l'on utilisait théoriquement des minimisations de fonctionnelles sur des espaces de dimension de plus en plus grande. Toutes ces simplifications ont été obtenues en utilisant le caractère quadratique de la fonctionnelle F. On remarque cependant que cet algorithme n'entre pas dans la définition générique des algorithmes de descente car le calcul de la direction de descente $d^{(k)}$ dépend de $x^{(k)}$ mais aussi de $x^{(k-1)}$ et $d^{(k-1)}$.

Exercice 4.8.

1. Donner un pseudo-code complet pour l'algorithme du gradient conjugué (on utilisera le critère d'arrêt usuel $\|\nabla f(x)\| \leqslant \varepsilon$ même si l'algorithme converge en n itérations).
2. Donner le code d'une fonction Octave

   ```
   function x = gradient_conjugue(A,b,x0,eps)
   ```

   qui applique la méthode du gradient conjugué pour résoudre le système $Ax = b$ en partant du point $x^{(0)}$ et avec le test d'arrêt de paramètre ε.

Solution de l'exercice 4.8.

1. On a l'algorithme suivant :

Algorithme 5 : Algorithme du gradient conjugué

Données : Un point initial $x^{(0)} \in \mathbb{R}^n$, un seuil de tolérance $\varepsilon > 0$
Résultat : Un point $x \in \mathbb{R}^n$ proche de $x^\star$
Initialiser x :
$x \leftarrow x^{(0)}$;
$k \leftarrow 0$;
Première itération :
(a) Calculer $d^{(0)} = \nabla f(x) \quad (d^{(0)} = \nabla f(x^{(k)}))$
(b) Calculer le pas de descente optimal : $t^{(0)} = \frac{\langle \nabla f(x^{(0)}), d^{(0)} \rangle}{\langle Ad^{(0)}, d^{(0)} \rangle} = \frac{\|d^{(0)}\|^2}{\langle Ad^{(0)}, d^{(0)} \rangle}$
(c) Mettre à jour x :
$x^{(1)} = x^{(0)} - t^{(0)} d^{(0)}$; $\quad k \leftarrow k + 1$;

tant que $\|\nabla f(x)\|^2 > \varepsilon^2$ **faire**

(a) Calculer $d^{(k)} = \nabla f(x^{(k)}) + \frac{\|\nabla f(x^{(k)})\|^2}{\|\nabla f(x^{(k-1)})\|^2} d^{(k-1)}$.
(b) Déterminer un pas de descente optimal $t^{(k)} > 0$ dans la direction $d^{(k)}$:
$t^{(k)} = \frac{\langle \nabla f(x^{(k)}), d^{(k)} \rangle}{\langle Ad^{(k)}, d^{(k)} \rangle}$
(c) Mettre à jour x :
$x \leftarrow x^{(k+1)} = x^{(k)} - t^{(k)} d^{(k)}$; (attention c'est bien un signe $-$)
$k \leftarrow k + 1$;

fin

2.

```
function x = gradient_conjugue(A,b,x0,eps)

x = x0;
k = 0;
gfx = A*x-b;
sqngfx = gfx'*gfx;
d = gfx;
t = sqngfx/((A*d)'*d);
x = x - t*d;
sqngfxold = sqngfx;
gfx = A*x - b;
sqngfx = gfx'*gfx;
k = k+1;

while(sqngfx>eps^2 && k < 1000)
   d = gfx + sqngfx/sqngfxold*d;
   t = gfx'*d/((A*d)'*d);
   x = x - t*d;
   sqngfxold = sqngfx;
   gfx = A*x - b;
   sqngfx = gfx'*gfx;
```

```
    k = k+1;
    disp([k,sqngfx]);
end

end

% test :
n = 1000;
A = toeplitz([2, -1, zeros(1, n-2)])*(n+1)^2;
b = rand(n,1);
xstar = A\b;
x0 = zeros(size(b));
eps = 10^(-4);
x = gradient_conjugue(A,b,x0,eps);
norm(x-xstar,'inf')
% stop toujours a n/2 en partant de b=ones(n,1), surement une symetrie
```

4.3 Algorithme du gradient conjugué comme méthode itérative

On a vu que l'algorithme du gradient conjugué convergeait en un nombre fini d'itérations inférieur ou égal à la dimension n de l'espace $\mathbb{R}^n$. On doit cependant faire deux observations.

La première est que du fait des erreurs numériques, il se peut que l'algorithme ne converge pas exactement vers la solution au bout de n itérations.

La deuxième est qu'en pratique, lorsque l'on résout de grands systèmes linéaires, la suite de vecteurs $x^{(k)}$ est très rapidement proche de $x^\star$ et on peut stopper l'algorithme bien avant que $k = n$.

Finalement, l'algorithme du gradient conjugué peut être vu comme une méthode directe de résolution de système linéaire (au même titre que l'utilisation de la factorisation LU ou la méthode de Gauss-Siedel, voir [ALLAIRE & KABER]), mais aussi comme un algorithme itératif de type descente de gradient. La proposition suivante quantifie la vitesse de convergence de l'algorithme du gradient conjugué.

Proposition 4.9 (Vitesse de convergence du gradient conjugué)**.** Soit A une matrice symétrique définie positive. Soit $x^\star$ la solution du système linéaire $Ax = b$. Soit $(x^{(k)})_k$ la suite de solutions approchées générée par l'algorithme du gradient conjugué. On a

$$\|x^{(k)} - x^\star\| \leqslant 2\sqrt{\mathrm{cond}(A)} \left(\frac{\sqrt{\mathrm{cond}(A)} - 1}{\sqrt{\mathrm{cond}(A)} + 1} \right)^k \|x^{(0)} - x^\star\|.$$

La preuve de cette proposition est admise. On renvoie à la section 9.5.2 de [ALLAIRE & KABER].

Bibliographie

[ALLAIRE & KABER] Grégoire ALLAIRE et Sidi Mahmoud KABER, *Algèbre linéaire numérique*, Ellipses, 2002

[CIARLET] Philippe G. CIARLET, *Introduction à l'analyse numérique matricielle et à l'optimisation*, cinquième édition, Dunod, 1998

[BOYD & VANDENBERGHE] Stephen BOYD and Lieven VANDENBERGHE *Convex Optimization*, Cambridge University Press, 2004