Faculté des Sciences
Aix*Marseille Université

CTES
Centre de Télé-Enseignement Sciences

LICENCE 3 MATHEMATIQUES – INFORMATIQUE. MATHEMATIQUES GENERALES. L3MiMG.

Expédition dans la semaine n°	Etape	Code UE	N° d'envoi de l'UE
9	**2L3MAT**	**SMI5U4T**	**5**

Nom de l'UE : Analyse numérique et optimisation

Le cours contient 3 chapitres (systèmes linéaires, systèmes non linéaires, optimisation). Pour chaque semaine, il est proposé d'étudier une partie du cours, de faire des exercices (corrigés) et, éventuellement, de réaliser un TP en python. Les TP sont fortement conseillés mais non obligatoires. Deux devoirs sont à rendre afin de bénéficier d'une note de contrôle continu.
note finale=max(note-examen, 1/3(2 note-examen + note-contrôle-continu)).

- Contenu de l'envoi : Polycopié, chapitre 3 (optimisation)

- Guide du travail à effectuer

Semaine 1 :
Etudier les paragraphes 3.1 et 3.2 (optimisation sans contrainte) et 3.4 (optimisation avec contrainte)
Ces paragraphes font aussi partie du cours de calcul différentiel et optimisation
Exercices proposés (avec corrigés) :
110 (exemples), 112 (fonctions quadratiques) et 115 (complément de Schur)

Semaine 2 :
Etudier les paragraphes 3.3.1 (méthodes de descente) et 3.3.2 (algorithme du gradient conjugué, GC)
Exercices proposés (avec corrigés) :
117 (exemple), 118 (algorithme du gradient à pas optimal) et 119 (Jacobi et optimisation)

Semaine 3 :
Etudier le paragraphe 3.3.3 (Newton)
Exercice proposé (avec corrigé) : 127 (Polak-Ribière)

Semaine 4 :
Etudier les paragraphe 3.4 et 3.5 (optimisation avec contrainte)
Exercice proposé (avec corrigé) : 139 (Uzawa)

Le corrigé du deuxième devoir sera, à la fin du mois de mars, sur le site du télé-enseignement et sur site web indiqué ci-dessous

-Coordonnées de l'enseignant responsable de l'envoi

T. Gallouet, CMI, 39 rue Joliot Curie, 13453 marseille cedex 13
email : thierry.gallouet@univ-amu.fr
Vous pouvez aussi consulter la page web:http://www.i2m.univ-amu.fr/~gallouet/tele.d/anum.d
et me poser des questions par email

Chapitre 3

Optimisation

3.1 Définitions et rappels

3.1.1 Extrema, points critiques et points selle.

L'objectif de ce chapitre est de rechercher des extrema, c'est-à-dire des minima ou des maxima d'une fonction $f \in C(\mathbb{R}^n, \mathbb{R})$ avec ou sans contrainte. Notons que la recherche d'un minimum ou d'un maximum implique que l'on ait une relation d'ordre, pour pouvoir comparer les valeurs prises par f. On insiste donc bien sur le fait que la fonction f est à valeurs dans $\mathbb{R}$ (et non pas $\mathbb{R}^n$, comme dans le chapitre précédent). Rappelons tout d'abord quelques définitions du cours de calcul différentiel.

Définition 3.1 (Extremum d'une fonction)**.** *Soit E un espace vectoriel normé et $f : E \to \mathbb{R}$.*
On dit que $\bar{x}$ est un minimum local de f s'il existe un voisinage V de $\bar{x}$ tel que

$$f(\bar{x}) \leq f(x), \forall x \in V.$$

De même, on dit que $\bar{x}$ est un maximum local de f s'il existe un voisinage V de $\bar{x}$ tel que

$$f(\bar{x}) \geq f(x), \forall x \in V.$$

On dit que $\bar{x}$ est un extremum local de f si c'est un minimum local ou un maximum local.
On dit que $\bar{x}$ est un minimum global de f si

$$f(\bar{x}) \leq f(x), \forall x \in E.$$

De même, on dit que $\bar{x}$ est un maximum global de f si

$$f(\bar{x}) \geq f(x), \forall x \in E.$$

On dit que $\bar{x}$ est un extremum global de f si c'est un minimum global ou un maximum global.

Le problème d'optimisation sans contrainte s'écrit :

$$\begin{cases} \text{Trouver } \bar{x} \in \mathbb{R}^n \text{ tel que :} \\ f(\bar{x}) \leq f(y), \quad \forall y \in \mathbb{R}^n. \end{cases} \tag{3.1}$$

Le problème d'optimisation avec contrainte s'écrit :

$$\begin{cases} \text{Trouver } \bar{x} \in K \text{tel que :} \\ f(\bar{x}) \leq f(y), \quad \forall y \in K. \end{cases} \tag{3.2}$$

où $K \subset \mathbb{R}^n$ et $K \neq \mathbb{R}^n$ L'ensemble K où l'on recherche la solution est donc l'ensemble qui représente les contraintes. Par exemple, si l'on cherche un miminum d'une fonction f de $\mathbb{R}$ dans $\mathbb{R}$ et que l'on demande que les points qui réalisent ce minimum soient positifs, on aura $K = \mathbb{R}_+$.
Si $\bar{x}$ est solution du problème (3.1), on dit que $\bar{x} \in \arg\min_{\mathbb{R}^n} f$, et si $\bar{x}$ est solution du problème (3.2), on dit que $\bar{x} \in \arg\min_K f$.
Vous savez déjà que si un point $\bar{x}$ réalise le minimum d'une fonction f dérivable de $\mathbb{R}$ dans $\mathbb{R}$, alors $f'(\overline{x}) = 0$. On dit que c'est un point critique (voir définition 3.2). La réciproque est évidemment fausse : la fonction $x \mapsto x^3$ est dérivable sur $\mathbb{R}$, et sa dérivée s'annule en 0 qui est donc un point critique, mais 0 n'est pas un extremum (c'est un point d'inflexion). Nous verrons plus loin que de manière générale, lorsque la fonctionnelle f est différentiable, les extrema sont des points critiques de f, au sens où ils annulent le gradient.

Définition 3.2 (Point critique)**.** *Soit E un espace vectoriel normé et $f : E \to \mathbb{R}$ différentiable. On dit que $x \in E$ est un point critique de f si $Df(x) = 0$.*

Pour illustrer un cas de point critique qui n'est pas un maximum ni un minimum, prenons un exemple en dimension 2, avec

$$f(x_1, x_2) = x_1^2 - x_2^2.$$

On a alors

$$Df(x_1, x_2)(h_1, h_2) = 2(x_1 h_1 - x_2 h_2) \text{ et } Df(0,0) = 0.$$

Le point $(0,0)$ est donc un point critique de f. Si on trace la surface $x \mapsto x_1^2 - x_2^2$, on se rend compte que le point $(0,0)$ est minimal dans une direction et maximal dans une direction indépendante de la première. C'est ce qu'on appelle un point selle

Définition 3.3 (Point selle)**.** *Soit E un espace vectoriel normé et $f : E \to \mathbb{R}$. On dit que $\bar{x}$ est un point selle de f s'il existe F et G des sous espaces vectoriels de E tels que $E = F \oplus G$ et un voisinage V de $\bar{x}$ tel que*

$$f(\bar{x} + z) \leq f(\bar{x}), \forall z \in F \,; \bar{x} + z \in V,$$
$$f(\bar{x} + z) \geq f(\bar{x}), \forall z \in G \,; \bar{x} + z \in V.$$

3.1.2 Convexité

Définition 3.4 (Convexité)**.** *Soit E un espace vectoriel (sur $\mathbb{R}$) et $f : E \to \mathbb{R}$. On dit que f est convexe si*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \text{ pour tout } (x,y) \in E^2 \text{ et } t \in [0,1].$$

On dit que f est strictement convexe si

$$f(tx + (1-t)y) < tf(x) + (1-t)f(y) \text{ pour tout } (x,y) \in E^2 \text{ t.q. } x \neq y \text{ et } t \in]0,1[.$$

Proposition 3.5 (Première caractérisation de la convexité)**.** *Soit E un espace vectoriel normé (sur $\mathbb{R}$) et $f \in C^1(E, \mathbb{R})$ alors :*

1. *la fonction f est convexe si et seulement si $f(y) \geq f(x) + Df(x)(y-x)$, pour tout couple $(x,y) \in E^2$,*
2. *la fonction f est strictement convexe si et seulement si $f(y) > f(x) + Df(x)(y-x)$ pour tout couple $(x,y) \in E^2$ tel que $x \neq y$.*

DÉMONSTRATION – *Démonstration de 1.*
($\Rightarrow$) Supposons que f est convexe : soit $(x,y) \in E^2$; on veut montrer que $f(y) \geq f(x) + Df(x)(y-x)$. Soit $t \in [0,1]$, alors $f(ty + (1-t)x) \leq tf(y) + (1-t)f(x)$ grâce au fait que f est convexe. On a donc :

$$f(x + t(y-x)) - f(x) \leq t(f(y) - f(x)). \tag{3.3}$$

Comme f est différentiable, $f(x + t(y-x)) = f(x) + Df(x)(t(y-x)) + t\varepsilon(t)$ où $\varepsilon(t)$ tend vers 0 lorsque t tend vers 0. Donc en reportant dans (3.3),

$$\varepsilon(t) + Df(x)(y-x) \leq f(y) - f(x), \quad \forall t \in]0,1[.$$

En faisant tendre t vers 0, on obtient alors :

$$f(y) \geq Df(x)(y-x) + f(x).$$

($\Leftarrow$) Montrons maintenant la réciproque : Soit $(x,y) \in E^2$, et $t \in]0,1[$ (pour $t = 0$ ou $= 1$ on n'a rien à démontrer). On veut montrer que $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$. On pose $z = tx + (1-t)y$. On a alors par hypothèse :

$$\begin{array}{ll} f(y) & \geq f(z) + Df(z)(y-z), \\ \text{et } f(x) & \geq f(z) + Df(z)(x-z). \end{array}$$

En multipliant la première inégalité par $1-t$, la deuxième par t et en les additionnant, on obtient :

$$\begin{array}{ll} (1-t)f(y) + tf(x) & \geq f(z) + (1-t)Df(z)(y-z) + tDf(z)(x-z) \\ (1-t)f(y) + tf(x) & \geq f(z) + Df(z)((1-t)(y-z) + t(x-z)). \end{array}$$

Et comme $(1-t)(y-z) + t(x-z) = 0$, on a donc $(1-t)f(y) + tf(x) \geq f(z) = f(tx + (1-t)y)$.

Démonstration de 2
($\Rightarrow$) On suppose que f est strictement convexe, on veut montrer que $f(y) > f(x) + Df(x)(y-x)$ si $y \neq x$. Soit donc $(x,y) \in E^2$, $x \neq y$. On pose $z = \frac{1}{2}(y-x)$, et comme f est convexe, on peut appliquer la partie 1. du théorème et écrire que $f(x+z) \geq f(x) + Df(x)(z)$. On a donc $f(x) + Df(x)(\frac{y-x}{2}) \leq f(\frac{x+y}{2})$. Comme f est strictement convexe, ceci entraîne que $f(x) + Df(x)(\frac{y-x}{2}) < \frac{1}{2}(f(x) + f(y))$, d'où le résultat.
($\Leftarrow$) La méthode de démonstration est la même que pour le 1. ∎

Proposition 3.6 (Seconde caractérisation de la convexité)**.** *Soit $E = \mathbb{R}^n$ et $f \in C^2(E, \mathbb{R})$. Soit $H_f(x)$ la hessienne de f au point x, i.e. $(H_f(x))_{i,j} = \partial^2_{i,j} f(x)$. Alors*

1. *f est convexe si et seulement si $H_f(x)$ est symétrique et positive pour tout $x \in E$ (c.à.d. $H_f(x)^t = H_f(x)$ et $H_f(x)y \cdot y \geq 0$ pour tout $y \in \mathbb{R}^n$)*
2. *f est strictement convexe si $H_f(x)$ est symétrique définie positive pour tout $x \in E$. (Attention la réciproque est fausse.)*

DÉMONSTRATION – *Démonstration de 1.*
($\Rightarrow$) Soit f convexe, on veut montrer que $H_f(x)$ est symétrique positive. Il est clair que $H_f(x)$ est symétrique car $\partial^2_{i,j} f = \partial^2_{j,i} f$ car f est C^2. Par définition, $H_f(x) = D(\nabla f(x))$ et $\nabla f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$. Soit $(x,y) \in E^2$, comme f est convexe et de classe C^1, on a, grâce à la proposition 3.5 :

$$f(y) \geq f(x) + \nabla f(x) \cdot (y-x). \tag{3.4}$$

Soit $\varphi \in C^2(\mathbb{R}, \mathbb{R})$ définie par $\varphi(t) = f(x + t(y-x))$. Alors :

$$f(y) - f(x) = \varphi(1) - \varphi(0) = \int_0^1 \varphi'(t)dt = [\varphi'(t)(t-1)]_0^1 - \int_0^1 \varphi''(t)(t-1)dt,$$

c'est–à dire : $f(y) - f(x) = \varphi'(0) + \int_0^1 \varphi''(t)(1-t)dt$. Or $\varphi'(t) = \nabla f(x + t(y-x)) \cdot (y-x)$, et

$$\varphi''(t) = D(\nabla f(x + t(y-x))(y-x) \cdot (y-x) = H_f(x + t(y-x))(y-x) \cdot (y-x).$$

On a donc :

$$f(y) - f(x) = \nabla f(x)(y-x) + \int_0^1 H_f(x + t(y-x))(y-x) \cdot (y-x)(1-t)dt. \tag{3.5}$$

Les inégalités (3.4) et (3.5) entraînent : $\int_0^1 H_f(x+t(y-x))(y-x)\cdot(y-x)(1-t)dt \geq 0\ \forall x,y\in E$. On a donc :

$$\int_0^1 H_f(x+tz)z\cdot z(1-t)dt \geq 0 \quad \forall x,\forall z\in E. \tag{3.6}$$

En fixant $x\in E$, on écrit (3.6) avec $z=\varepsilon y$, $\varepsilon>0$, $y\in\mathbb{R}^n$. On obtient :

$$\varepsilon^2\int_0^1 H_f(x+t\varepsilon y)y\cdot y(1-t)dt\geq 0 \quad \forall x,y\in E,\ \forall\varepsilon>0, \text{ et donc :}$$

$$\int_0^1 Hf(x+t\varepsilon y)y\cdot y(1-t)dt\geq 0 \quad \forall\varepsilon>0.$$

Pour $(x,y)\in E^2$ fixé, $H_f(x+t\varepsilon y)$ tend vers $H_f(x)$ uniformément lorsque $\varepsilon\to 0$, pour $t\in[0,1]$. On a donc :

$$\int_0^1 H_f(x)y\cdot y(1-t)dt\geq 0, \text{ c.à.d. } \frac{1}{2}H_f(x)y\cdot y\geq 0.$$

Donc pour tout $(x,y)\in(\mathbb{R}^n)^2$, $H_f(x)y\cdot y\geq 0$ donc $H_f(x)$ est positive.

($\Leftarrow$) Montrons maintenant la réciproque : On suppose que $H_f(x)$ est positive pour tout $x\in E$. On veut démontrer que f est convexe ; on va pour cela utiliser la proposition 3.5 et montrer que : $f(y)\geq f(x)+\nabla f(x)\cdot(y-x)$ pour tout $(x,y)\in E^2$. Grâce à (3.5), on a :

$$f(y)-f(x)=\nabla f(x)\cdot(y-x)+\int_0^1 H_f(x+t(y-x))(y-x)\cdot(y-x)(1-t)dt.$$

Or $H_f(x+t(y-x))(y-x)\cdot(y-x)\geq 0$ pour tout couple $(x,y)\in E^2$, et $1-t\geq 0$ sur $[0,1]$. On a donc $f(y)\geq f(x)+\nabla f(x)\cdot(y-x)$ pour tout couple $(x,y)\in E^2$. La fonction f est donc bien convexe.

Démonstration de 2.

($\Leftarrow$) On suppose que $H_f(x)$ est strictement positive pour tout $x\in E$, et on veut montrer que f est strictement convexe. On va encore utiliser la caractérisation de la proposition 3.5. Soit donc $(x,y)\in E^2$ tel que $y\neq x$. Alors :

$$f(y)=f(x)+\nabla f(x)\cdot(y-x)+\int_0^1 \underbrace{H_f(x+t(y-x))(y-x)\cdot(y-x)}_{>0 \text{ si } x\neq y}\ \underbrace{(1-t)}_{\neq 0 \text{ si } t\in]0,1[}\ dt.$$

Donc $f(y)>f(x)+\nabla f(x)(y-x)$ si $x\neq y$, ce qui prouve que f est strictement convexe. ∎

Contre-exemple Pour montrer que la réciproque de 2. est fausse, on propose le contre-exemple suivant : Soit $n=1$ et $f\in C^2(\mathbb{R},\mathbb{R})$, on a alors $H_f(x)=f''(x)$. Si f est la fonction définie par $f(x)=x^4$, alors f est strictement convexe mais $f''(0)=0$.

3.1.3 Exercices (extrema, convexité)

Exercice 110 (Vrai / faux)**.** *corrigé en page 212*

1. L'application $x\mapsto\|x\|_\infty$ est convexe sur $\mathbb{R}^2$.
2. L'application $x\mapsto\|x\|_\infty$ est strictement convexe sur $\mathbb{R}^2$.
3. L'application de $\mathbb{R}^2$ dans $\mathbb{R}$ définie par $F(x,y)=x^2-2xy+3y^2+y$ admet un unique minimum.
4. Soit $A\in\mathcal{M}_{n,m}(\mathbb{R})$, $b\in\mathbb{R}^n$, l'application $x\mapsto\|Ax-b\|_2$ admet un unique minimum.

Exercice 111 (Minimisation dans $\mathbb{R}$)**.** *Corrigé en page 212*

On considère les fonctions définies de $\mathbb{R}$ dans $\mathbb{R}$ par $f_0(x)=x^2$, $f_1(x)=x^2(x-1)^2$, $f_2(x)=|x|$, $f_3(x)=\cos x$, $f_4(x)=|\cos x|$, $f_5(x)=e^x$. On pose $K=[-1,1]$. Pour chacune de ces fonctions, répondre aux questions suivantes :

1. Etudier la différentiabilité et la (stricte) convexité éventuelles de la fonction, ; donner l'allure de son graphe.
2. La fonction admet elle un minimum global sur $\mathbb{R}$; ce minimum est-il unique ? Le cas échéant, calculer ce minimum.

3. La fonction admet elle un minimum sur K ; ce minimum est-il unique ? Le cas échéant, calculer ce minimum.

Exercice 112 (Fonctions quadratiques)**.**

1. Montrer que la fonction f de $\mathbb{R}^2$ dans $\mathbb{R}$ définie par $f(x,y) = x^2 + 4xy + 3y^2$ n'admet pas de minimum en $(0,0)$.
2. Trouver la matrice symétrique S telle que $f(x) = x^t Sx$, pour $f_1(x) = 2(x_1^2 + x_2^2 + x_3^2 - x_1x_2 - x_2x_3)$, puis pour $f_2(x) = 2(x_1^2 + x_2^2 + x_3^2 - x_1x_2 - x_1x_3 - x_2x_3)$ Etudier la convexité des fonctions f_1 et f_2.
3. Calculer les matrices hessiennes de g_1 et g_2 définies par : $g_1(x,y) = \frac{1}{4}x^4 + x^2y + y^2$ et $g_2(x,y) = x^3 + xy - x$ et étudier la convexité de ces deux fonctions.

Exercice 113 (Convexité et continuité)**.** *Suggestions en page 211.*

1. Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction convexe.
   (a) Montrer que f est continue.
   (b) Montrer que f est localement lipschitzienne.

2. Soit $n \geq 1$ et $f : \mathbb{R}^n \to \mathbb{R}$. On suppose que f est convexe.
   (a) Montrer f est bornée supérieurement sur les bornés (c'est-à-dire : pour tout $R > 0$, il existe m_R t.q. $f(x) \leq m_R$ si la norme de x est inférieure ou égale à R).
   (b) Montrer que f est continue.
   (c) Montrer que f est localement lipschitzienne.
   (d) On remplace maintenant $\mathbb{R}^n$ par E, e.v.n. de dimension finie. Montrer que f est continue et que f est localement lipschitzienne.

3. Soient E un e.v.n. de dimension infinie et $f : E \to \mathbb{R}$. On suppose que f est convexe.
   (a) On suppose, dans cette question, que f est bornée supérieurement sur les bornés. Montrer que f est continue.
   (b) Donner un exemple d'e.v.n. (noté E) et de fonction convexe $f : E \to \mathbb{R}$ t.q. f soit non continue.

Suggestions pour les exercices

Exercice 113 page 211 (Convexité et continuité)

1.(a) Pour montrer la continuité en 0, soit $x \neq 0$, $|x| < 1$. On pose $a = \text{sgn}(x)$ $(= \frac{x}{|x|})$. Ecrire x comme une combinaison convexe de 0 et a et écrire 0 comme une combinaison convexe de x et $-a$. En déduire une majoration de $|f(x) - f(0)|$.
   (b) Utiliser la continuité de f et la majoration précédente.

2.(a) Faire une récurrence sur n et pour $x = (x_1, y)^t$ avec $-R < x_1 < R$ et $y \in \mathbb{R}^{n-1}$ $(n > 1)$, majorer $f(x)$ en utilisant $f(+R, y)$ et $f(-R, y)$.
   (b) Reprendre le raisonnement fait pour $n = 1$.
   (c) Se ramener à $E = \mathbb{R}^n$.

3.(a) reprendre le raisonnement fait pour $E = \mathbb{R}$.
   (b) On pourra, par exemple choisir $E = C([0,1], \mathbb{R})\ldots$

Corrigés des exercices

Exercice 110 page 210 (Minimisation dans R)

1. Vrai.
2. Faux. L'application est convexe mais pas strictement convexe. Si on fixe $v_1 = (1,0)$ et $v_2 = (1,1)$, alors pour tout $t \in [0,1]$,
$$\|tv_1 + (1-t)v_2\|_\infty = \|(1, 1-t)\|_\infty = 1 = t\|v_1\|_\infty + (1-t)\|v_2\|_\infty.$$
3. Vrai. Posons $X = (x,y)^t$, on reconnait la fonctionnelle quadratique $F(x,y) = \frac{1}{2}(AX,X) - (b,X)$ avec $A = \begin{bmatrix} 1 & -1 \\ -1 & 3 \end{bmatrix}$ et $b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. La matrice A une matrice symétrique définie positive. Le cours nous dit alors que F admet un unique minimum.
4. Contre-exemple. Soit $A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ et $b = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Alors $\|Ax - b\|_2 = x_1^2$ et toute la droite $x_1 = 0$ réalise le minimum de f.

Exercice 111 page 210 (Minimisation dans $\mathbb{R}$)

1. La fonction f_0 est différentiable sur $\mathbb{R}$, et strictement convexe. Elle admet un minimum unique sur $\mathbb{R}$ et sur K et son minimum est réalisé en $\bar{x} = 0$, et on a $f_0(\bar{x}) = 0$.
2. La fonction f_1 est différentiable sur $\mathbb{R}$, et non convexe. La fonction f_1 admet un maximum local en $\bar{x} = \frac{1}{2}$, et on a $f(\bar{x}) = \frac{1}{16}$. Elle admet un minimum global non unique, réalisé en 0 et 1, et dont la valeur est 0.
3. La fonction f_2 est différentiable sur $\mathbb{R} \setminus \{0\}$, et convexe, mais pas strictement convexe. La fonction f_2 admet un minimum unique sur $\mathbb{R}$ et sur K et son minimum est réalisé en $\bar{x} = 0$, et on a $f_2(\bar{x}) = 0$, mais la fonction f_2 n'est pas différentiable en 0.
4. La fonction f_3 est différentiable sur $\mathbb{R}$, et non convexe. La fonction f_3 admet un minimum, qui est -1, et qui n'est pas unique car il est réalisé pour les points $(2k+1)\pi$, $k \in \mathbb{Z}$.
5. La fonction f_4 est différentiable sur $\mathbb{R}$, et non convexe. La fonction f_4 admet un minimum, qui est 0, et qui n'est pas unique car il est réalisé pour les points $(2k+1)\frac{\pi}{2}$, $k \in \mathbb{Z}$. La fonction f_4 n'est pas différentiable en ces points.
6. La fonction f_5 est différentiable et strictement convexe. Elle n'admet pas de minimum. On a $f_5(x) \to 0$ lorsque $x \to -\infty$ mais $f(x) > 0$ pour tout $x \in \mathbb{R}$.

3.2 Optimisation sans contrainte

3.2.1 Définition et condition d'optimalité

Soit $f \in C(E, \mathbb{R})$ et E un espace vectoriel normé. On cherche $\bar{x}$ minimum global de f, c.à.d. :

$$\bar{x} \in E \text{ tel que } f(\bar{x}) \leq f(y) \ \forall y \in E, \tag{3.7}$$

ou un minimum local, c.à.d. :

$$\bar{x} \text{ tel que } \exists \alpha > 0 \ \ f(\bar{x}) \leq f(y) \ \forall y \in B(\bar{x}, \alpha). \tag{3.8}$$

Proposition 3.7 (Condition nécessaire d'optimalité)**.**
Soit E un espace vectoriel normé, et soient $f \in C(E, \mathbb{R})$, et $\bar{x} \in E$ tel que f est différentiable en $\bar{x}$. Si $\bar{x}$ est solution de (3.8) alors $Df(\bar{x}) = 0$.

DÉMONSTRATION – Supposons qu'il existe $\alpha > 0$ tel que $f(\bar{x}) \leq f(y)$ pour tout $y \in B(\bar{x}, \alpha)$. Soit $z \in E \setminus \{0\}$, alors si $|t| < \frac{\alpha}{\|z\|}$, on a $\bar{x} + tz \in B(\bar{x}, \alpha)$ (où $B(\bar{x}, \alpha)$ désigne la boule ouverte de centre $\bar{x}$ et de rayon α) et on a donc $f(\bar{x}) \leq f(\bar{x} + tz)$. Comme f est différentiable en $\bar{x}$, on a :

$$f(\bar{x} + tz) = f(\bar{x}) + Df(\bar{x})(tz) + |t|\varepsilon_z(t),$$

où $\varepsilon_z(t) \to 0$ lorsque $t \to 0$. On a donc $f(\bar{x}) + tDf(\bar{x})(z) + |t|\varepsilon_z(t) \geq f(\bar{x})$. Et pour $\frac{\alpha}{\|z\|} > t > 0$, on a $Df(\bar{x})(z) + \varepsilon_z(t) \geq 0$. En faisant tendre t vers 0, on obtient que

$$Df(\bar{x})(z) \geq 0, \quad \forall z \in E.$$

On a aussi $Df(\bar{x})(-z) \geq 0 \quad \forall z \in E$, et donc : $-Df(\bar{x})(z) \geq 0 \quad \forall z \in E$.
On en conclut que

$$Df(\bar{x}) = 0.$$

■

Remarque 3.8. Attention, la proposition précédente donne une condition nécessaire mais non suffisante. En effet, $Df(\bar{x}) = 0$ n'entraîne pas que f atteigne un minimum (ou un maximum) même local, en $\bar{x}$. Prendre par exemple $E = \mathbb{R}$, $\overline{x} = 0$ et la fonction f définie par : $f(x) = x^3$ pour s'en convaincre.

3.2.2 Résultats d'existence et d'unicité

Théorème 3.9 (Existence). *Soit $E = \mathbb{R}^n$ et $f : E \to \mathbb{R}$ une application telle que*

(i) *f est continue,*

(ii) *$f(x) \to +\infty$ quand $\|x\| \to +\infty$.*

Alors il existe $\bar{x} \in \mathbb{R}^n$ tel que $f(\bar{x}) \leq f(y)$ pour tout $y \in \mathbb{R}^n$.

DÉMONSTRATION – La condition (ii) peut encore s'écrire

$$\forall A \in \mathbb{R}, \quad \exists R \in \mathbb{R}; \|x\| \geq R \Rightarrow f(x) \geq A. \tag{3.9}$$

On écrit (3.9) avec $A = f(0)$. On obtient alors :

$$\exists R \in \mathbb{R} \text{ tel que } \|x\| \geq R \Rightarrow f(x) \geq f(0).$$

On en déduit que $\inf_{\mathbb{R}^n} f = \inf_{B_R} f$, où $B_R = \{x \in \mathbb{R}^n; |x| \leq R\}$. Or, B_R est un compact de $\mathbb{R}^n$ et f est continue donc il existe $\bar{x} \in B_R$ tel que $f(\bar{x}) = \inf_{B_R} f$ et donc $f(\bar{x}) = \inf_{\mathbb{R}^n} f$. ■

Remarque 3.10.

1. Le théorème est faux si E est un espace de Banach (c'est-à-dire un espace vectoriel normé complet) de dimension infinie car, dans ce cas, la boule fermée B_R n'est pas compacte.
2. L'hypothèse (ii) du théorème peut être remplacée par

$$(ii)' \qquad \exists b \in \mathbb{R}^n, \exists R > 0 \text{ tel que } \|x\| \geq R \Rightarrow f(x) \geq f(b).$$

3. Sous les hypothèses du théorème il n'y a pas toujours unicité de $\bar{x}$ même dans le cas $n = 1$, prendre pour s'en convaincre la fonction f définie de $\mathbb{R}$ dans $\mathbb{R}$ par $f(x) = x^2(x-1)(x+1)$.

Théorème 3.11 (Condition suffisante d'unicité). *Soit E un espace vectoriel normé et $f : E \to \mathbb{R}$ strictement convexe alors il existe au plus un $\bar{x} \in E$ tel que $f(\bar{x}) \leq f(y), \forall y \in E$.*

DÉMONSTRATION – Soit f strictement convexe, supposons qu'il existe $\bar{x}$ et $\bar{\bar{x}} \in E$ tels que $f(\bar{x}) = f(\bar{\bar{x}}) = \inf_{\mathbb{R}^n} f$. Comme f est strictement convexe, si $\bar{x} \neq \bar{\bar{x}}$ alors

$$f(\frac{1}{2}\bar{x} + \frac{1}{2}\bar{\bar{x}}) < \frac{1}{2}f(\bar{x}) + \frac{1}{2}f(\bar{\bar{x}}) = \inf_{\mathbb{R}^n} f,$$

ce qui est impossible ; donc $\bar{x} = \bar{\bar{x}}$. ∎

Ce théorème ne donne pas l'existence. Par exemple dans le cas $n = 1$ la fonction f définie par $f(x) = e^x$ n'atteint pas son minimumn ; en effet, $\inf_{\mathbb{R}^n} f = 0$ et $f(x) \neq 0$ pour tout $x \in \mathbb{R}$, et pourtant f est strictement convexe. Par contre, si on réunit les hypothèses des théorèmes 3.9 et 3.11, on obtient le résultat d'existence et unicité suivant :

Théorème 3.12 (Existence et unicité)**.** *Soit $E = \mathbb{R}^n$, et soit $f : E \to \mathbb{R}$. On suppose que :*

(i) f continue,

(ii) $f(x) \to +\infty$ quand $\|x\| \to +\infty$,

(iii) f est strictement convexe ;

alors il existe un unique $\bar{x} \in \mathbb{R}^n$ tel que $f(\bar{x}) = \inf_{\mathbb{R}^n} f$.

L'hypothèse (i) du théorème 3.12 est en fait inutile car une fonction convexe de $\mathbb{R}^n$ dans $\mathbb{R}$ est nécessairement continue.
Nous donnons maintenant des conditions suffisantes d'existence et d'unicité du minimum pour une fonction de classe C^1.

Proposition 3.13 (Condition suffisante d'existence et unicité)**.** *Soit $f \in C^1(\mathbb{R}^n, \mathbb{R})$. On suppose que :*

$$\exists \alpha > 0; (\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \alpha |x - y|^2, \ \forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n, \tag{3.10}$$

Alors :

1. *f est strictement convexe,*
2. *$f(x) \to +\infty$ quand $|x| \to +\infty$,*

et en conséquence, il existe un unique $\bar{x} \in \mathbb{R}^n$ tel que $f(\bar{x}) = \inf_{\mathbb{R}^n} f$.

DÉMONSTRATION –

1. Soit φ la fonction définie de $\mathbb{R}$ dans $\mathbb{R}^n$ par : $\varphi(t) = f(x + t(y - x))$. Alors

$$f(y) - f(x) = \varphi(1) - \varphi(0) = \int_0^1 \nabla f(x + t(y - x)) \cdot (y - x) dt,$$

On en déduit que

$$f(y) - f(x) - \nabla f(x) \cdot (y - x) = \int_0^1 (\nabla f(x + t(y - x)) \cdot (y - x) - \nabla f(x) \cdot (y - x)) dt,$$

c'est–à–dire :

$$f(y) - f(x) - \nabla f(x) \cdot (y - x) = \int_0^1 \underbrace{(\nabla f(x + t(y - x)) - \nabla f(x)) \cdot (y - x)}_{\geq \alpha t |y - x|^2} dt.$$

Grâce à l'hypothèse (3.10) sur f, ceci entraîne :

$$f(y) - f(x) - \nabla f(x) \cdot (y - x) \geq \alpha \int_0^1 t|y - x|^2 dt = \frac{\alpha}{2}|y - x|^2 > 0 \text{ si } y \neq x. \tag{3.11}$$

On a donc, pour tout $(x,y) \in E^2$, $f(y) > f(x) + \nabla f(x)\cdot(y-x)$; d'après la première caractérisation de la convexité, voir proposition 3.5, on en déduit que f est strictement convexe.

2. Montrons maintenant que $f(y) \to +\infty$ quand $|y| \to +\infty$. On écrit (3.11) pour $x = 0$: $f(y) \geq f(0) + \nabla f(0)\cdot y + \frac{\alpha}{2}|y|^2$. Comme $\nabla f(0)\cdot y \geq -|\nabla f(0)|(y)$, on a donc

$$f(y) \geq f(0) + |y|\left(\frac{\alpha}{2}|y| - |\nabla f(0)|\right) \to +\infty \text{ quand } |y| \to +\infty.$$

La fonction f vérifie donc bien les hypothèses du théorème 3.30, et on en déduit qu'il existe un unique $\bar{x}$ qui minimise f. ■

Remarque 3.14 (Généralisation à un espace de Hilbert)**.** Le théorème 3.12 reste vrai si E est un espace de Hilbert ; on a besoin dans ce cas pour la partie existence des hypothèses (i), (ii) et de la convexité de f.

Proposition 3.15 (Caractérisation des points tels que $f(\bar{x}) = \inf_E f$)**.** *Soit E espace vectoriel normé et f une fonction de E dans $\mathbb{R}$. On suppose que $f \in C^1(E,\mathbb{R})$ et que f est convexe. Soit $\bar{x} \in E$. Alors :*

$$f(\bar{x}) = \inf_E f \Leftrightarrow Df(\bar{x}) = 0.$$

En particulier si $E = \mathbb{R}^n$ alors $f(\bar{x}) = \inf_{x\in\mathbb{R}^n} f(x) \Leftrightarrow \nabla f(\bar{x}) = 0$.

Démonstration
($\Rightarrow$) Supposons que $f(\bar{x}) = \inf_E f$ alors on sait (voir Proposition 3.7) que $Df(\bar{x}) = 0$ (la convexité est inutile).
($\Leftarrow$) Si f est convexe et différentiable, d'après la proposition 3.5, on a : $f(y) \geq f(\bar{x}) + Df(\bar{x})(y-x)$ pour tout $y \in E$ et comme par hypothèse $Df(\bar{x}) = 0$, on en déduit que $f(y) \geq f(\bar{x})$ pour tout $y \in E$. Donc $f(\bar{x}) = \inf_E f$.

Cas d'une fonction quadratique On appelle fonction quadratique une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$ définie par

$$\boldsymbol{x} \mapsto f(\boldsymbol{x}) = \frac{1}{2}A\boldsymbol{x}\cdot\boldsymbol{x} - \boldsymbol{b}\cdot\boldsymbol{x} + c, \tag{3.12}$$

où $A \in \mathcal{M}_n(\mathbb{R})$, $\boldsymbol{b} \in \mathbb{R}^n$ et $c \in \mathbb{R}$. On peut vérifier facilement que $f \in C^\infty(\mathbb{R}^n,\mathbb{R})$. Calculons le gradient de f et sa hessienne : on a

$$\begin{aligned} f(\boldsymbol{x}+\boldsymbol{h}) &= \frac{1}{2}A(\boldsymbol{x}+\boldsymbol{h})\cdot(\boldsymbol{x}+\boldsymbol{h}) - \boldsymbol{b}\cdot(\boldsymbol{x}+\boldsymbol{h}) + c \\ &= \frac{1}{2}A\boldsymbol{x}\cdot\boldsymbol{x} + \frac{1}{2}A\boldsymbol{x}\cdot\boldsymbol{h} + \frac{1}{2}A\boldsymbol{h}\cdot\boldsymbol{x} + \frac{1}{2}A\boldsymbol{h}\cdot\boldsymbol{h} - \boldsymbol{b}\cdot\boldsymbol{x} - \boldsymbol{b}\cdot\boldsymbol{h} + c \\ &= f(\boldsymbol{x}) + \frac{1}{2}(A\boldsymbol{x}\cdot\boldsymbol{h} + A\boldsymbol{h}\cdot\boldsymbol{x}) - \boldsymbol{b}\cdot\boldsymbol{h} + \frac{1}{2}A\boldsymbol{h}\cdot\boldsymbol{h} \\ &= f(\boldsymbol{x}) + \frac{1}{2}(A\boldsymbol{x} + A^t\boldsymbol{x})\cdot\boldsymbol{h} - \boldsymbol{b}\cdot\boldsymbol{h} + \frac{1}{2}A\boldsymbol{h}\cdot\boldsymbol{h}. \end{aligned}$$

Et comme $|A\boldsymbol{h}\cdot\boldsymbol{h}| \leq \|A\|_2\,|\boldsymbol{h}|^2$, on en déduit que :

$$\nabla f(\boldsymbol{x}) = \frac{1}{2}(A\boldsymbol{x} + A^t\boldsymbol{x}) - \boldsymbol{b}. \tag{3.13}$$

Si A est symétrique, on a donc $\nabla f(\boldsymbol{x}) = A\boldsymbol{x} - \boldsymbol{b}$. Calculons maintenant la hessienne de f. D'après (3.13), on a :

$$\nabla f(\boldsymbol{x}+\boldsymbol{h}) = \frac{1}{2}(A(\boldsymbol{x}+\boldsymbol{h}) + A^t(\boldsymbol{x}+\boldsymbol{h})) - \boldsymbol{b} = \nabla f(\boldsymbol{x}) + \frac{1}{2}(A\boldsymbol{h} + A^t\boldsymbol{h})$$

et donc $H_f(\boldsymbol{x}) = D(\nabla f(\boldsymbol{x})) = \frac{1}{2}(A + A^t)$. On en déduit que si A est symétrique, $H_f(\boldsymbol{x}) = A$. Dans le cas où A est symétrique définie positive, f est donc strictement convexe.

De plus on a $f(\boldsymbol{x}) \to +\infty$ quand $|\boldsymbol{x}| \to +\infty$. (On note comme d'habitude $|\cdot|$ la norme euclidienne de $\boldsymbol{x}$.) En effet,

$$A\boldsymbol{x}\cdot\boldsymbol{x} \geq \alpha|\boldsymbol{x}|^2 \text{ où } \alpha \text{ est la plus petite valeur propre de } A, \text{ et } \alpha > 0.$$

Donc

$$f(\boldsymbol{x}) \geq \frac{\alpha}{2}|\boldsymbol{x}|^2 - |\boldsymbol{b}\cdot\boldsymbol{x}| - |c|;$$

Mais comme $|\boldsymbol{b}\cdot\boldsymbol{x}| \leq |\boldsymbol{b}||\boldsymbol{x}|$, on a

$$f(\boldsymbol{x}) \geq |\boldsymbol{x}|\left(\frac{\alpha|\boldsymbol{x}|}{2} - |\boldsymbol{b}|\right) - |c| \longrightarrow +\infty \text{ quand } |\boldsymbol{x}| \to +\infty.$$

On en déduit l'existence et l'unicité de $\bar{\boldsymbol{x}}$ qui minimise f. On a aussi :

$$\nabla f(\bar{\boldsymbol{x}}) = 0 \Leftrightarrow f(\bar{\boldsymbol{x}}) = \inf_{\mathbb{R}^n} f$$

et donc $\bar{\boldsymbol{x}}$ est l'unique solution du système $A\boldsymbol{x} = \boldsymbol{b}$.
On en déduit le théorème suivant, très important, puisqu'il va nous permettre en particulier le lien entre certains algorithmes d'optimisation et les méthodes de résolution de systèmes linéaires vues au chapitre 1.

Théorème 3.16 (Minimisation d'une fonction quadratique)**.** *Soit f une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$ définie par* (3.12) *où $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice symétrique définie positive et $\boldsymbol{b} \in \mathbb{R}^n$. Alors il existe un unique $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ qui minimise f, et $\bar{\boldsymbol{x}}$ est l'unique solution du système linéaire $A\boldsymbol{x} = \boldsymbol{b}$.*

3.2.3 Exercices (optimisation sans contrainte)

Exercice 114 (Maximisation)**.** *Suggestions en page 218*
Soit E un espace vectoriel normé et $f : E \to \mathbb{R}$. En utilisant les résultats de la section 3.2.2, répondre aux questions suivantes :

1. Donner une condition suffisante d'existence de $\bar{x} \in E$ tel que $f(\bar{x}) = \sup_{x\in E} f(x)$.
2. Donner une condition suffisante d'unicité de $\bar{x} \in E$ tel que $f(\bar{x}) = \sup_{x\in E} f(x)$.
3. Donner une condition suffisante d'existence et unicité de $\bar{x} \in E$ tel que $f(\bar{x}) = \sup_{x\in E} f(x)$.

Exercice 115 (Complément de Schur)**.** *Corrigé en page 218*

Soient n et p deux entiers naturels non nuls. Dans toute la suite, si u et v sont deux vecteurs de $\mathbb{R}^k$, $k \geq 1$, le produit scalaire de u et v est noté $u\cdot v$. Soient A une matrice carrée d'ordre n, inversible, soit B une matrice $n \times p$, C une matrice carrée d'ordre p, et soient $f \in \mathbb{R}^n$ et $g \in \mathbb{R}^p$. On considère le système linéaire suivant :

$$M\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \text{ avec } M = \begin{bmatrix} A & B \\ B^t & C \end{bmatrix}. \tag{3.14}$$

1. On suppose dans cette question seulement que $n = p = 1$, et $A = \begin{bmatrix} a \end{bmatrix}$, $B = \begin{bmatrix} b \end{bmatrix}$, $C = \begin{bmatrix} c \end{bmatrix}$
   (a) Donner une condition nécessaire et suffisante sur a, b, et c pour que M soit inversible.
   (b) Donner une condition nécessaire et suffisante sur a, b, et c pour que M soit symétrique définie positive.

On définit la matrice $S = C - B^tA^{-1}B$, qu'on appelle "complément de Schur".
2. Calculer S dans le cas $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, $B = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, $C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.
3. Montrer qu'il existe une unique solution au problème (3.14) si et seulement si la matrice S est inversible. Est-ce le cas dans la question 2 ?

On suppose maintenant que A est symétrique définie positive.
4. On suppose dans cette question que C est symétrique.

(a) Vérifier que M est symétrique.

(b) Soient $x \in \mathbb{R}^n$, $y \in \mathbb{R}^p$ et $z = (x, y) \in \mathbb{R}^{n+p}$. Calculer $Mz \cdot z$ en fonction de A, B, C, x et y.

(c) On fixe maintenant $y \in \mathbb{R}^p$, et on définit la fonction F de $\mathbb{R}^n$ dans $\mathbb{R}$ par : $x \mapsto Ax \cdot x + 2By \cdot x + Cy \cdot y$. Calculer $\nabla F(x)$, et calculer $x_0 \in \mathbb{R}^n$ tel que $\nabla F(x_0) = 0$

(d) Montrer que la fonction F définie en 3(c) admet un unique minimum, et calculer la valeur de ce mimimum.

(e) En déduire que M est définie positive si et seulement si S est définie positive.

5. On suppose dans cette question que C est la matrice (carrée d'ordre p) nulle.

(a) Montrer que la matrice $\tilde{S} = -S$ est symétrique définie positive si et seulement si $p \leq n$ et rang(B)=p. On supposera que ces deux conditions sont vérifiées dans toute la suite de la question.

(b) En déduire que la matrice $P = \begin{bmatrix} A & 0 \\ 0 & \tilde{S} \end{bmatrix}$ est symétrique définie positive.

(c) Calculer les valeurs propres de la matrice $T = P^{-1}M$ (il peut être utile de distinguer les cas Ker$B^t = \{0\}$ et Ker$B^t \neq \{0\}$).

Exercice 116 (Approximation au sens des moindres carrés)**.**

1. **Un premier exemple.** Dans le plan (s, t), on cherche la droite d'équation $t = \alpha + \beta s$ qui passe par les points $(0, 1)$, $(1, 9)$, $(3, 9)$, $(4, 21)$.

   (a) Montrer que si cette droite existait, le vecteur $x = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ serait solution d'un système linéaire $Ax = b$; on donnera explicitement la matrice A et le vecteur b.

   (b) Montrer qu'une telle droite n'existe pas. Dans la suite du problème on va trouver la droite qui passe le "plus près" possible de ces quatre points, au sens de la norme euclidienne.

2. **Un second exemple.** On cherche maintenant à déterminer les coefficients α, β et γ d'une fonction linéaire T de $\mathbb{R}^3$ dans $\mathbb{R}$, dont on ne connaît la valeur qu'en deux points : $T(1, 1, 1) = 3$ et $T(0, 1, 1) = 2$.

   (a) Montrer que les coefficients α, β et γ s'ils existent, satisfont un système linéaire $Ax = b$; on donnera explicitement la matrice A et le vecteur b.

   (b) Montrer qu'il existe une infinité de solutions au système $Ax = b$. Dans la suite du problème on va trouver les coefficients α, β et γ qui donnent un vecteur x de norme euclidienne minimale.

On considère maintenant une matrice A d'ordre $n \times m$ et $b \in \mathbb{R}^n$, et on veut résoudre dans un sens aussi "satisfaisant" que possible le système linéaire

$$Ax = b,\ x \in \mathbb{R}^m, \tag{3.15}$$

lorsque $m \neq n$ ou lorsque $m = n$ mais que A n'est pas inversible. On note $\|y\| = \left(\sum_{i=1}^p y_i^2\right)^{\frac{1}{2}}$ la norme euclidienne sur $\mathbb{R}^p$, $p = n$ ou m suivant les cas et $(\cdot \mid \cdot)$ le produit scalaire associé. Soit f la fonction définie de $\mathbb{R}^m$ dans $\mathbb{R}$ par $f(x) = \|Ax - b\|^2$. On cherche à minimiser f, c.à.d. à trouver $\bar{x} \in \mathbb{R}^n$ tel que

$$f(\bar{x}) = \min\{f(x), x \in \mathbb{R}^n\}. \tag{3.16}$$

3. Soit E un sous espace vectoriel de $\mathbb{R}^m$ tel que $\mathbb{R}^m = E \bigoplus \mathrm{Ker}A$.

   (a) Montrer que $f(z) \to +\infty$ lorsque $\|z\| \to +\infty$ avec $z \in E$.

   (b) Montrer que f est strictement convexe de E dans $\mathbb{R}$.

   (c) En déduire qu'il existe un unique $\bar{z} \in E$ tel que

$$f(\bar{z}) \leq f(z), \forall z \in E.$$

4. Soit $X_b = \{\bar{z} + y, y \in \mathrm{Ker}A\}$, où $\bar{z}$ est défini à la question précédente. Montrer que X_b est égal à l'ensemble des solutions du problème de minimisation (3.16).

5. Montrer que $x \in X_b \iff A^tAx = A^tb$, où A^t désigne la matrice transposée de A. On appelle système d'équations normales le système $A^tAx = A^tb$.
6. Ecrire les équations normales dans le cas de l'exemple de la question 1, et en déduire l'équation de la droite obtenue par moindres carrés, *i.e.* par résolution de (3.16). Tracer les quatre points donnés à la question 1 et la droite obtenue sur un graphique.
7. Ecrire les équations normales dans le cas de l'exemple de la question 2, et vérifier que le système obtenu n'est pas inversible.
8. Pour $y \in \mathrm{Ker}A$, on pose $g(y) = \|y + \bar{z}\|^2$, où $\bar{z}$ est définie à la question 3. Montrer qu'il existe un unique $\bar{y} \in \mathrm{Ker}A$ tel que $g(\bar{y}) \leq g(y)$ pour tout $y \in \mathrm{Ker}A$. En déduire qu'il existe un unique $\bar{x} \in X_b$ tel que $\|\bar{x}\|^2 \leq \|x\|^2$ pour tout $x \in X_b$. On appelle $\bar{x}$ pseudo-solution de (3.16).
9. Calculer $\bar{x}$ dans le cas des exemples des questions 1 et 2.

Dans la suite du problème, on considère, pour $\varepsilon > 0$ fixé, une version pénalisée du problème (3.16). On introduit la fonction f_ε de $\mathbb{R}^m$ dans $\mathbb{R}$, définie par $f_\varepsilon(x) = \|x\|^2 + \frac{1}{\varepsilon}\|A^tAx - A^tb\|^2$, et on cherche à trouver x_ε solution du problème de minimisation suivant :

$$f_\varepsilon(x_\varepsilon) \leq f_\varepsilon(x), \forall x \in \mathbb{R}^m. \tag{3.17}$$

10. Montrer que le problème (3.17) possède une unique solution x_ε.
11. Calculer $\nabla f_\varepsilon(x)$ et en déduire l'équation satisfaite par x_ε.
12. Montrer que x_ε converge vers $\bar{x}$ lorsque $\varepsilon \to 0$.

Suggestions pour les exercices

Exercice 114 page 216 (Maximisation) Appliquer les théorèmes du cours à $-f$.

Corrigés des exercices

Exercice 115 page 216 (Complément de Schur)
1.

(a) La matrice M est inversible si et seulement si son déterminant est non nul, c.à.d. ssi $ac - b^2 \neq 0$.

(b) La matrice M est symétrique par construction. Elle est définie positive si et seulement si ses valeurs propres sont strictement positives, c.à.d. si et seulement si $ac - b^2 > 0$ et $a > 0$.

2. $S = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix}$

3. Montrons que $\mathrm{Ker}(M) = \{0\}$ si et seulement si $\mathrm{Ker}(S) \neq \{0\}$. Comme M et S sont des matrices carrées, ceci revient à dire que le système (3.14) a une unique solution si et seulement si la matrice S est inversible. Soit $(x, y) \in \mathrm{Ker}M$. Comme A est inversible, ceci est équivalent à dire que

$$x = -A^{-1}By, \tag{3.18}$$
$$(C - B^tA^{-1}B)y = 0. \tag{3.19}$$

Ceci est équivalent à $x = 0, y = 0$ si et seulement si $\mathrm{Ker}(C - B^tA^{-1}B) = \{0\}$, c.à.d ssi S est inversible. Ce n'est pas le cas de la matrice S de la question (a).

4.

(a) Si $i, j \leq n$, $m_{i,j} = a_{i,j} = a_{j,i} = mj, i$; si $i, j \geq n$, $m_{i,j} = c_{i,j} = c_{j,i} = mj, i$; et enfin $i \leq n$ et $j \geq n$ $m_{i,j} = b_{i,j} = (B^t)_{j,i} = m_{j,i}$. Donc M est bien symétrique.

(b) $Mz \cdot z = Ax \cdot x + 2By \cdot x + Cy \cdot y$.

(c) $\nabla F(x) = 2Ax + 2By$, et $x_0 = -A^{-1}By$.

(d) Si A est définie positive, alors la fonction F définie en 3.(b) est quadratique, donc, d'après le cours, admet un unique minimum en x_0. La valeur de ce minimum est donc $F(x_0) = -AA^{-1}By \cdot (-A^{-1}By) + 2By \cdot (-A^{-1}By) + Cy \cdot y = Sy \cdot y$.

(e) Supposons A et S définies positives. Soit $z = (x, y) \in \mathbb{R}^{n+p}$. On a $Mz \cdot z = F(x) \geq Sy \cdot y \forall x \in \mathbb{R}^n$ si A est définie positive, d'après la question précédente. Donc $Mz \cdot z \geq 0$ dès que S est semi-définie positive. Supposons $Mz \cdot z = 0$, alors $F(x) = 0$ mais comme S est définie positive, $F(x) \geq F(x_0) = Sy \cdot y > 0$ sauf si $y = 0$ et $x = x_0 = -A^{-1}By = 0$, ce qui prouve que M est définie positive.
Réciproquement, si M est définie positive, alors en prenant successivement $z = (x, 0)$ puis $z = (0, y)$, on obtient facilement que A et C sont définies positives ; la matrice S est aussi définie positive, car $Sy \cdot y = Fs(x_0) = Mz \cdot z > 0$ avec $z = (x_0, y)$, et donc $Sy \cdot y > 0$ si $y \neq 0$.

5.

(a) Comme A est symétrique définie positive, A^{-1} l'est également, et $\tilde{S}$ est évidemment symétrique. On a $\tilde{S}y \cdot y = -Sy \cdot y = B^tA^{-1}By \cdot y = A^{-1}By \cdot By \geq 0$ pour tout $y \in \mathbb{R}^p$. Soit $z = By$. On a donc : $\tilde{S}y \cdot y = A^{-1}z \cdot z$. Supposons $\tilde{S}y \cdot y = 0$. On a donc $A^{-1}z \cdot z = 0$, et donc $z = 0$. Si $p \leq n$ et si le rang de B est p ceci entraîne que $y = 0$.

Réciproquement, si $p \leq n$ et si le rang de B n'est strictement inférieur à p, alors il existe $y_0 \neq 0$ élément de Ker$B$ et donc $\tilde{S}y_0 \cdot y_0 = 0$ alors que $y_0 \neq 0$.

D'autre part, si $p > n$, alors la matrice $\tilde{S}$ est une matrice de rang au plus n et de taille $p > n$; par le théorème du rang, $\dim \text{Ker}(\tilde{S}) = p - n > 0$ et la matrice $\tilde{S}$ n'est donc pas inversible.

On a donc bien montré l'équivalence souhaitée. .

(b) Soit $z = (x, y) \in \mathbb{R}^{n+p}$; on a $Pz \cdot z = Ax \cdot x - Sy \cdot y = Ax \cdot x + BA^{-1}B^ty \cdot y$. Supposons $Pz \cdot z = 0$. On déduit de la question précédente que $x = 0$ et $y = 0$, ce qui montre que P est symétrique définie positive.

(c) Soit λ une valeur propre de T et $z = \begin{bmatrix} x \\ y \end{bmatrix}$ un vecteur propre associé.

On a donc $Mz = \lambda Pz$, c.à.d. :

$$\begin{bmatrix} A & B \\ B^t & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} A & 0 \\ 0 & \tilde{S} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

c.à.d.

$$\begin{aligned} Ax + By &= \lambda Ax \\ B^tx &= \lambda \tilde{S}y \end{aligned}$$

Considérons tout d'abord le cas $\lambda = 1$. On a alors $By = 0$, et donc $y = 0$ car le rang de B est p par hypothèse. On en déduit que $B^tx = 0$, et donc il n'existe un vecteur propre associé à λ que si ker$B^t \neq \{0\}$. Supposons maintenant que $\lambda \neq 1$. Dans ce cas, on obtient que $x = \frac{1}{\lambda - 1}A^{-1}By$, et donc $\frac{1}{\lambda-1}\tilde{S}y = \lambda \tilde{S}y$. Comme on veut que $z \neq 0$, on a $y \neq 0$ et donc $\tilde{S}y \neq 0$. On en déduit que les valeurs propres sont les racines du polynôme $-\lambda^2 + \lambda + 1 = 0$, c.à.d.$\lambda = \frac{1}{2}(1 \pm \sqrt{5})$.

Exercice 116 page 217 (Approximation au sens des moindres carrés)

1. (a) Une condition nécessaire pour que la droite existe est que α et β vérifie le système linéaire

$$\begin{aligned} \text{point } (0,1) : &\quad \alpha = 1 \\ \text{point } (1,9) : &\quad \alpha + \beta = 9 \\ \text{point } (3,9) : &\quad \alpha + 3\beta = 9 \\ \text{point } (4,21) : &\quad \alpha + 4\beta = 21 \end{aligned}$$

Autrement dit $x = \begin{bmatrix}\alpha\\ \beta\end{bmatrix}$ est une solution de $Ax = b$, avec

$$A = \begin{bmatrix}1 & 0\\ 1 & 1\\ 1 & 3\\ 1 & 4\end{bmatrix} \text{ et } b = \begin{bmatrix}1\\ 9\\ 9\\ 21\end{bmatrix}.$$

(b) Montrer qu'une telle droite n'existe pas.
Si l'on on retranche la ligne 2 à la ligne 3 du système, on obtient $\beta = 0$ et si l'on retranche la ligne 1 à la ligne 2, on obtient $\beta = 8$. Donc le système n'admet pas de solution.

2. (a) Une condition nécessaire pour que la droite existe est que α, β et γ vérifie le système

$$\begin{aligned}\alpha + \beta + \gamma &= 3\\ \beta + \gamma &= 2\end{aligned}$$

Autrement dit $x = \begin{bmatrix}\alpha\\ \beta\end{bmatrix}$ est une solution de $Ax = b$, avec

$$A = \begin{bmatrix}1 & 1 & 1\\ 0 & 1 & 1\end{bmatrix} \text{ et } b = \begin{bmatrix}3\\ 2\end{bmatrix}.$$

(b) Une solution particulière de ce système est $x = \begin{bmatrix}1\\ 2\\ 0\end{bmatrix}$, et le noyau de A est engendré par $\begin{bmatrix}0\\ -1\\ 1\end{bmatrix}$.
L'ensemble des solutions est de la forme $\{\begin{bmatrix}1\\ 2\\ 0\end{bmatrix} + \gamma \begin{bmatrix}0\\ -1\\ 1\end{bmatrix}, \gamma \in \mathbb{R}\}$, qui est infini.

3. (a) On a

$$\begin{aligned}f(z) &= (Az - b)\cdot(Az - b)\\ &= Az\cdot Az - 2Az\cdot b + b\cdot b\\ &\geq \|Az\|^2 - 2\|b\|\|Az\| + \|b\|^2 \qquad \text{d'après l'inégalité de Cauchy-Schwarz}\\ &\longrightarrow +\infty \text{ lorsque } \|Az\| \to \infty\end{aligned}$$

Il reste maintenant à montrer que $\|Az\| \longrightarrow +\infty$ lorsque $\|z\| \to \infty$. Pour cela, on remarque que

$$\|Az\| = \left\|A\frac{z}{\|z\|}\right\| \|z\| \geq \inf_{w\in E, \|w\|=1} \|Aw\|\|z\| = \|A\bar{w}\|\|z\|,$$

car l'ensemble $K = \{w \in E, \|w\| = 1\}$ est un compact de $\mathbb{R}^n$ et comme la fonction $\varphi : K \to \mathbb{R}$ définie par $\varphi(w) = \|Aw\|$ est continue, elle atteint son minimum en $\bar{w} \in K$:

$$\inf_{w\in E, \|w\|=1} \|Aw\| = \|A\bar{w}\|$$

Or $A\bar{w} \neq 0$, et donc $\|A\bar{w}\|\|z\| \to +\infty$ lorsque $\|z\| \to +\infty$.

(b) Calculons ∇f.

$$\forall x \in E, \nabla f(x) = 2(A^tAx - A^tb)$$

Par conséquent, pour $x, y \in E$,

$$\begin{aligned}f(y) - \nabla f(x)\cdot(y - x) &= (Ay - b)\cdot(Ay - b) - 2(Ax - b)\cdot A(y - x),\ \forall (x, y) \in E^2; x \neq y.\\ &= |Ay|^2 - 2Ay\cdot b + |b|^2 + 2|Ax|^2 - 2Ax\cdot Ay + 2b\cdot Ay - 2b\cdot Ax\\ &= |Ay|^2 + |b|^2 + 2|Ax|^2 - 2Ax\cdot Ay - 2b\cdot Ax\\ &= |Ay - Ax|^2 + |Ax - b|^2\\ &> 0,\ \forall (x, y) \in E^2; x \neq y.\end{aligned}$$

On en déduit que f est strictement convexe par la proposition 3.5 (première caractérisation de la convexité).

(c) On applique le théorème 3.12 : f est une application continue de E dans $\mathbb{R}$, qui tend vers l'infini à l'infini et qui admet donc un minimum. L'unicité du minimum vient de la stricte convexité de cette application.

4. Soit $x \in \mathbb{R}^m$, x peut s'écrire $x = z + y$ avec $z \in E$ et $y \in \mathrm{Ker}A$, par suite

$$f(x) = \|A(z+y) - b\|^2 = \|Az - b\|^2 = f(z) \geq f(\bar{z}).$$

D'autre part,

$$f(\bar{z} + y) = f(\bar{z}) \forall y \in \mathrm{Ker}A.$$

Donc X_b est bien l'ensemble des solutions du problème de minimisation (3.16).

5. Condition nécessaire : On a déjà vu que f est différentiable et que $\nabla f(x) = 2(A^tAx - A^tb)$. Comme f est différentiable toute solution de (3.16) vérifie l'équation d'Euler $\nabla f(x) = 0$.
Condition suffisante : Soit x tel que $A^tAx = A^tb$, c'est à dire tel que $\nabla f(x) = 0$. Comme f est de classe C^1 et convexe sur $\mathbb{R}^m$, alors x est un minimum global de f. (Noter que la convexité de f peut se montrer comme à la question 3(b) en remplaçant E par $\mathbb{R}^m$.)

6. On a

$$A^tA = \begin{bmatrix} 4 & 8 \\ 8 & 26 \end{bmatrix} \text{ et } A^tb = \begin{bmatrix} 40 \\ 120 \end{bmatrix}$$

Les équations normales de ce problème s'écrivent donc

$$A^tA \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = A^tb.$$

La matrice A^tA est inversible, par conséquent il y a une unique solution à ces equations normales donnée par $\begin{bmatrix} 2 \\ 4 \end{bmatrix}$.

7. On a

$$A^tA = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 2 \end{bmatrix} \text{ et } A^tb = \begin{bmatrix} 3 \\ 5 \\ 5 \end{bmatrix}$$

Les deux dernières lignes de la matrice A^tA sont identiques donc la matrice n'est pas inversible. Comme les deux dernières lignes de A^tb sont elles aussi identiques, on en déduit que le système admet une infinité de solutions.
On peut échelonner le système :

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 1 & 2 & 2 & 5 \\ 1 & 2 & 2 & 5 \end{array}\right] \xrightarrow[T_{32}(-1), T_{21}(-1)]{} \left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{array}\right] \xrightarrow[T_{12}(-1)]{} \left[\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

On retrouve les solutions obtenues à la question 2-b : $X_b = \underbrace{\begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}}_{\bar{z}} + \mathbb{R} \underbrace{\begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}}_{\bar{u}} = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} + \mathrm{Ker}A.$

8. La fonction g est continue sur l'espace vectoriel $\mathrm{Ker}A$ et vérifie

$$g(y) \geq \|y\|^2 - 2\|y\|\|\bar{z}\| + \|\bar{z}\|^2 \longrightarrow +\infty \text{ lorsque } \|\bar{z}\| \longrightarrow +\infty;$$

par conséquent, g admet un minimum sur $\mathrm{Ker}A$. Ce minimum est unique car g est strictement convexe, car c'est le carré de la norme euclidienne. On peut le montrer directement, ou si l'on ne connaît pas ce résultat, on peut dire que c'est la composée d'une application convexe et d'une application strictement convexe et croissante : $g = q(N(x))$ avec $N : x \mapsto \|x\|$ convexe et $q : s \mapsto s^2$. On pourrait également remarquer que l'application

$$\begin{aligned} &\mathrm{Ker}A \to \mathbb{R} \\ &v \mapsto D^2g(y)(v)(v) \end{aligned}$$

est une forme quadratique définie positive, car

$$Dg(y)(w) = 2(y + \bar{z}) \cdot w, \text{ et } D^2g(y)(h)(v) = 2h \cdot v$$

L'application $v \mapsto D^2g(y)(v)(v) = 2v \cdot v$ est clairement définie positive, ce qui montre une fois de plus que g est strictement convexe.
On en déduit alors qu'il existe un unique $\bar{x} \in X_b$ de norme minimale.

9. Dans le premier exemple, les équations normales admettent une seule solution $\bar{x} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$, voir question 6.

   Pour le deuxième exemple, on calcule $\|x\|^2$ pour $x = \bar{z} + t\bar{u} \in X_b$:

   $$\|x\|^2 = \|\bar{z} + t\bar{u}\|^2 = 1 + (2-t)^2 + t^2 = 5 - 4t + 2t^2 = 2\,(t-1)^2 + 3$$

   On voit $\|x\|^2$ est minimale pour $t = 1$, autrement dit $\bar{x} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

10. La fonction f_ε est une fonction continue, infinie à l'infini car

    $$f_\varepsilon(x) \geq \|x\|^2 \longrightarrow +\infty \text{ lorsque } \|x\| \to \infty.$$

    On a donc existence d'un minimum pour f_ε. De plus, la fonction f_ε est de classe C^2 avec

    $$\nabla f_\varepsilon(x) = 2(x + \frac{1}{\varepsilon}A^tA(A^tAx - A^tb)),\ D^2f_\varepsilon(x) = 2(Id + \frac{1}{\varepsilon}(A^tA)^2)$$

    La matrice A^tA est positive, donc la matrice $(A^tA)^2$ est positive par suite la matrice $D^2f(x)$ est définie positive. La fonction f_ε est donc strictement convexe. Par conséquent, f_ε admet un unique minimum.

11. On sait que le minimum x_ε de f_ε est un zéro de ∇f_ε, soit

    $$x_\varepsilon + \frac{1}{\varepsilon}A^tA(A^tAx_\varepsilon - A^tb) = 0$$

    et donc $(Id + \frac{1}{\varepsilon}(A^tA)^2)x_\varepsilon = \frac{1}{\varepsilon}A^tAA^tb$, ce qui donne

    $$x_\varepsilon = (Id + \frac{1}{\varepsilon}(A^tA)^2)^{-1}\frac{1}{\varepsilon}A^tAA^tb.$$

12. On commence par remarquer que $\|x_\varepsilon\|^2 \leq f_\varepsilon(x_\varepsilon) \leq f_\varepsilon(\bar{x}) = \|\bar{x}\|^2$. Par conséquent, la famille$\{x_\varepsilon,\, \varepsilon > 0\}$ est bornée dans $\mathbb{R}^m$ qui est de dimension finie. Pour montrer que $x_\varepsilon \to \bar{x}$ quand $\varepsilon \to 0$, il suffit donc de montrer que $\bar{x}$ est la seule valeur d'adhérence de la famille $\{x_\varepsilon,\, \varepsilon > 0\}$. Soit $\bar{y}$ une valeur d'adhérence de la famille $\{x_\varepsilon,\, \varepsilon > 0\}$. Il existe une suite $\varepsilon_n \to 0$ pour laquelle x_{ε_n} converge vers $\bar{y} \in \mathbb{R}^m$. Montrons que $A^tA\bar{y} = A^tb$. On rappelle que

    $$\frac{1}{\varepsilon}\|A^tAx_\varepsilon - A^tb\|^2 \leq f_\varepsilon(x_\varepsilon) \leq \|\bar{x}\|^2.$$

    On en déduit que

    $$\|A^tAx_{\varepsilon_n} - A^tb\|^2 \leq \varepsilon_n\|\bar{x}\|^2 \longrightarrow 0 \text{ lorsque } n \longrightarrow \infty$$

    et en passant à la limite on obtient $\|A^tA\bar{y} - A^tb\|^2 = 0$. On a également par un argument analogue $\|\bar{y}\|^2 \leq \|\bar{x}\|^2$. Donc $\bar{y} \in X_b$ et comme $\bar{x}$ est l'unique vecteur de X_b de norme minimale, on en déduit que $\bar{y} = \bar{x}$.

3.3 Algorithmes d'optimisation sans contrainte

Soit $f \in C(\mathbb{R}^n, \mathbb{R})$. On suppose qu'il existe $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ tel que $f(\bar{\boldsymbol{x}}) = \inf_{\mathbb{R}^n} f$.
On cherche à calculer $\bar{\boldsymbol{x}}$ (si f est de classe C^1, on a nécessairement $\nabla f(\bar{\boldsymbol{x}}) = 0$). On va donc maintenant développer des algorithmes (ou méthodes de calcul) du point $\bar{\boldsymbol{x}}$ qui réalise le minimum de f. Il existe deux grandes classes de méthodes :

— Les méthodes dites "directes" ou bien "de descente", qui cherchent à construire une suite minimisante, c.à.d. une suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ telle que :

$$f(\boldsymbol{x}^{(k+1)}) \leq f(\boldsymbol{x}^{(k)}),$$
$$\boldsymbol{x}^{(k)} \to \bar{\boldsymbol{x}} \text{ quand } k \to +\infty.$$

— Les méthodes basées sur l'équation d'Euler, qui consistent à chercher une solution de l'équation (dite d'Euler) $\nabla f(\boldsymbol{x}) = 0$ (ces méthodes nécessitent donc que f soit dérivable).

3.3.1 Méthodes de descente

Définition 3.17. *Soit $f \in C(\mathbb{R}^n, \mathbb{R})$.*

1. *Soit $\boldsymbol{x} \in \mathbb{R}^n$, on dit que $\boldsymbol{w} \in \mathbb{R}^n \setminus \{0\}$ est une direction de descente en $\boldsymbol{x}$ s'il existe $\alpha_0 > 0$ tel que*
$$f(\boldsymbol{x} + \alpha\boldsymbol{w}) \leq f(\boldsymbol{x}), \ \ \forall \alpha \in [0, \alpha_0]$$
2. *Soit $\boldsymbol{x} \in \mathbb{R}^n$, on dit que $w \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ est une direction de descente stricte en $\boldsymbol{x}$ si s'il existe $\alpha_0 > 0$ tel que*
$$f(\boldsymbol{x} + \alpha\boldsymbol{w}) < f(\boldsymbol{x}), \ \ \forall \alpha \in]0, \alpha_0].$$
3. *Une "méthode de descente" pour la recherche de $\bar{\boldsymbol{x}}$ tel que $f(\bar{\boldsymbol{x}}) = \inf_{\mathbb{R}^n} f$ consiste à construire une suite $(x_k)_{k\in\mathbb{N}}$ de la manière suivante :*
   (a) *Initialisation : $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$;*
   (b) *Itération k : on suppose $\boldsymbol{x}^{(0)}, \ldots, \boldsymbol{x}^{(k)}$ connus ($k \geq 0$) ;*
      i. *On cherche $\boldsymbol{w}^{(k)}$ direction de descente stricte en $\boldsymbol{x}^{(k)}$*
      ii. *On prend $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}$ avec $\alpha_k > 0$ "bien choisi".*

Proposition 3.18 (Caractérisation des directions de descente)**.** *Soient $f \in C^1(\mathbb{R}^n, \mathbb{R})$, $\boldsymbol{x} \in \mathbb{R}^n$ et $\boldsymbol{w} \in \mathbb{R}^n \setminus \{0\}$; alors*

1. *si $\boldsymbol{w}$ direction de descente en $\boldsymbol{x}$ alors $\boldsymbol{w} \cdot \nabla f(\boldsymbol{x}) \leq 0$,*
2. *si $\boldsymbol{w} \cdot \nabla f(\boldsymbol{x}) < 0$ alors $\boldsymbol{w}$ direction de descente stricte en $\boldsymbol{x}$,*
3. *si $\nabla f(\boldsymbol{x}) \neq 0$ alors $w = -\nabla f(\boldsymbol{x})$ est une direction de descente stricte en $\boldsymbol{x}$.*

DÉMONSTRATION –

1. Soit $\boldsymbol{w} \in \mathbb{R}^n \setminus \{0\}$ une direction de descente en $\boldsymbol{x}$: alors par définition,
$$\exists \alpha_0 > 0 \text{ tel que } f(\boldsymbol{x} + \alpha\boldsymbol{w}) \leq f(\boldsymbol{w}), \ \ \forall \alpha \in [0, \alpha_0].$$
Soit φ la fonction de $\mathbb{R}$ dans $\mathbb{R}$ définie par : $\varphi(\alpha) = f(\boldsymbol{x}+\alpha\boldsymbol{w})$. On a $\varphi \in C^1(\mathbb{R}, \mathbb{R})$ et $\varphi'(\alpha) = \nabla f(\boldsymbol{x}+\alpha\boldsymbol{w})\cdot\boldsymbol{w}$. Comme $\varphi(\alpha) \leq \varphi(0)$ pour tout $\alpha \in [0, \alpha_0]$ on a
$$\forall \alpha \in]0, \alpha_0[, \quad \frac{\varphi(\alpha) - \varphi(0)}{\alpha} \leq 0;$$
en passant à la limite lorsque α tend vers 0, on déduit que $\varphi'(0) \leq 0$, c.à.d. $\nabla f(\boldsymbol{x}) \cdot \boldsymbol{w} \leq 0$.

2. On reprend les notations précédentes. Si $\nabla f(\boldsymbol{x}) \cdot \boldsymbol{w} < 0$, on a $\varphi'(0) < 0$. Par continuité de ∇f, il existe $\alpha_0 > 0$ tel que $\varphi'(\alpha) < 0$ si $\alpha \in [0, \alpha_0]$. En utilisant le théorème des accroissements finis on en déduit que $\varphi(\alpha) < \varphi(0)$ si $\alpha \in]0, \alpha_0[$ et donc que w est une direction de descente stricte.
3. Si $\nabla f(\boldsymbol{x}) \neq 0$, $w = -\nabla f(\boldsymbol{x})$ est une direction de descente stricte en $\boldsymbol{x}$ car $\nabla f(\boldsymbol{x}) \cdot \boldsymbol{w} < 0 = -|\nabla f(\boldsymbol{x})|^2 < 0$.

■

Algorithme du gradient à pas fixe Soient $f \in C^1(E, \mathbb{R})$ et $E = \mathbb{R}^n$. On se donne $\alpha > 0$.

$$\left\{ \begin{array}{ll} \text{Initialisation :} & \boldsymbol{x}^{(0)} \in E, \\ \text{Itération } k : & \boldsymbol{x}^{(k)} \text{ connu, } (k \geq 0) \\ & \boldsymbol{w}^{(k)} = -\nabla f(\boldsymbol{x}^{(k)}), \\ & \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha \boldsymbol{w}^{(k)}. \end{array} \right. \tag{3.20}$$

Théorème 3.19 (Convergence du gradient à pas fixe)**.** *Soient $E = \mathbb{R}^n$ et $f \in C^1(E, \mathbb{R})$ vérifiant les hypothèses*

$$\exists \omega > 0; (\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \omega |x - y|^2, \ \forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n, \tag{3.21a}$$

$$\exists M > 0; \|\nabla f(x) - \nabla f(y)\| \leq M |x - y|, \ \forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n. \tag{3.21b}$$

L'hypothese 3.21a est l'hypothese 3.10 de la proposition 3.13. La fonction f est donc strictement convexe et croissante à l'infini, et admet donc un unique minimum. De plus, si $0 < \alpha < \dfrac{2\omega}{M^2}$ alors la suite $(\boldsymbol{x}^{(k)})_{k \in \mathbb{N}}$ construite par (3.20) *converge vers $\bar{\boldsymbol{x}}$ lorsque $k \to +\infty$.*

DÉMONSTRATION –
Montrons la convergence de la suite construite par l'algorithme de gradient à pas fixe en nous ramenant à un algorithme de point fixe. On pose $h(\boldsymbol{x}) = \boldsymbol{x} - \alpha \nabla f(\boldsymbol{x})$. L'algorithme du gradient à pas fixe est alors un algorithme de point fixe pour h.

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha \nabla f(\boldsymbol{x}^{(k)}) = h(\boldsymbol{x}^{(k)}).$$

Grâce au théorème 2.8 page 155, on sait que h est strictement contractante si

$$0 < \alpha < \frac{2\omega}{M^2}.$$

Donc la suite $(x^{(k)})_{k \in \mathbb{N}}$ converge vers l'unique point fixe $\bar{\boldsymbol{x}}$ de h, caractérisé par

$$\bar{\boldsymbol{x}} = h(\bar{\boldsymbol{x}}) = \bar{\boldsymbol{x}} - \alpha \nabla f(\bar{\boldsymbol{x}})$$

On a donc $\nabla f(\bar{\boldsymbol{x}}) = 0$, et, comme f est strictement convexe, $f(\bar{\boldsymbol{x}}) = \inf_E f$. ■

Algorithme du gradient à pas optimal L'idée de l'algorithme du gradient à pas optimal est d'essayer de calculer à chaque itération le paramètre qui minimise la fonction dans la direction de descente donnée par le gradient. Soient $f \in C^1(E, \mathbb{R})$ et $E = \mathbb{R}^n$, cet algorithme s'écrit :

$$\left\{ \begin{array}{ll} \text{Initialisation :} & \boldsymbol{x}^{(0)} \in \mathbb{R}^n. \\ \text{Itération } n : & \boldsymbol{x}^{(k)} \text{ connu.} \\ & \text{On calcule } \boldsymbol{w}^{(k)} = -\nabla f(\boldsymbol{x}^{(k)}). \\ & \text{On choisit } \alpha_k \geq 0 \text{ tel que} \\ & \qquad f(\boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}) \leq f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{w}^{(k)}) \ \forall \alpha \geq 0. \\ & \text{On pose } \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}. \end{array} \right. \tag{3.22}$$

Les questions auxquelles on doit répondre pour s'assurer du bien fondé de ce nouvel algorithme sont les suivantes :

1. Existe–t–il α_k tel que $f(\boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}) \leq f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{w}^{(k)}), \forall \alpha \geq 0$?

2. Comment calcule–t'on α_k ?
3. La suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ construite par l'algorithme converge–t–elle ?

La réponse aux questions 1. et 3. est apportée par le théorème suivant :

Théorème 3.20 (Convergence du gradient à pas optimal)**.**
Soit $f \in C^1(\mathbb{R}^n, \mathbb{R})$ telle que $f(\boldsymbol{x}) \to +\infty$ quand $|\boldsymbol{x}| \to +\infty$. Alors :

1. *La suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ est bien définie par (3.22). On choisit $\alpha_k > 0$ tel que $f(\boldsymbol{x}^{(k)}+\alpha_k\boldsymbol{w}^{(k)}) \leq f(x_k+\alpha\boldsymbol{w}^{(k)})$ $\forall\alpha \geq 0$ (α_k existe mais n'est pas nécessairement unique).*
2. *La suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ est bornée et si $(\boldsymbol{x}^{(k_\ell)})_{\ell\in\mathbb{N}}$ est une sous suite convergente, i.e. $x^{(k_\ell)} \to x$ lorsque $\ell \to +\infty$, on a nécessairement $\nabla f(\boldsymbol{x}) = 0$. De plus si f est convexe on a $f(x) = \inf_{\mathbb{R}^n} f$*
3. *Si f est strictement convexe on a alors $\boldsymbol{x}^{(k)} \to \bar{\boldsymbol{x}}$ quand $k \to +\infty$, avec $f(\bar{\boldsymbol{x}}) = \inf_{\mathbb{R}^n} f$*

La démonstration de ce théorème fait l'objet de l'exercice 118. On en donne ici les idées principales.

1. On utilise l'hypothèse $f(\boldsymbol{x}) \to +\infty$ quand $|\boldsymbol{x}| \to +\infty$ pour montrer que la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ construite par (3.22) existe : en effet, à $\boldsymbol{x}^{(k)}$ connu,

   1er cas : si $\nabla f(\boldsymbol{x}^{(k)}) = 0$, alors $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)}$ et donc $\boldsymbol{x}^{(p)} = \boldsymbol{x}^{(k)}$ $\forall p \geq k$,

   2ème cas : si $\nabla f(\boldsymbol{x}^{(k)}) \neq 0$, alors $\boldsymbol{w}^{(k)} = \nabla f(\boldsymbol{x}^{(k)})$ est une direction de descente stricte.

   Dans ce deuxième cas, il existe donc α_0 tel que

   $$f(\boldsymbol{x}^{(k)} + \alpha\boldsymbol{w}^{(k)}) < f(\boldsymbol{x}^{(k)}),\ \forall\alpha \in]0, \alpha_0]. \tag{3.23}$$

   De plus, comme $\boldsymbol{w}^{(k)} \neq 0$, $|\boldsymbol{x}^{(k)} + \alpha\boldsymbol{w}^{(k)}| \to +\infty$ quand $\alpha \to +\infty$ et donc $f(\boldsymbol{x}^{(k)} + \alpha\boldsymbol{w}^{(k)}) \longrightarrow +\infty$ quand $\alpha \to +\infty$. Il existe donc $M > 0$ tel que si $\alpha > M$ alors $f(x_k + \alpha\boldsymbol{w}^{(k)}) \geq f(\boldsymbol{x}^{(k)})$. On a donc :

   $$\inf_{\alpha\in\mathbb{R}_+^*} f(\boldsymbol{x}^{(k)} + \alpha\boldsymbol{w}^{(k)}) = \inf_{\alpha\in[0,M]} f(\boldsymbol{x}^{(k)} + \alpha\boldsymbol{w}^{(k)}).$$

   Comme $[0, M]$ est compact, il existe $\alpha_k \in [0, M]$ tel que $f(x_k + \alpha_k\boldsymbol{w}^{(k)}) = \inf_{\alpha\in[0,M]} f(x_k + \alpha\boldsymbol{w}^{(k)})$. De plus on a grâce à (3.23) que $\alpha_k > 0$.
2. Le point 2. découle du fait que la suite $(f(\boldsymbol{x}^{(k)}))_{k\in\mathbb{N}}$ est décroissante, donc la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ est bornée (car $f(\boldsymbol{x}) \to +\infty$ quand $|\boldsymbol{x}| \to +\infty$). On montre ensuite que si $\boldsymbol{x}^{(k_\ell)} \to \boldsymbol{x}$ lorsque $\ell \to +\infty$ alors $\nabla f(\boldsymbol{x}) = 0$ (ceci est plus difficile, les étapes sont détaillées dans l'exercice 118).

Reste la question du calcul de α_k, qui est le paramètre optimal dans la direction de descente $\boldsymbol{w}^{(k)}$, c.à.d. le nombre réel qui réalise le minimum de la fonction φ de $\mathbb{R}_+$ dans $\mathbb{R}$ définie par : $\varphi(\alpha) = f(\boldsymbol{x}^{(k)} + \alpha\boldsymbol{w}^{(k)})$. Comme $\alpha_k > 0$ et $\varphi(\alpha_k) \leq \varphi(\alpha)$ pour tout $\alpha \in \mathbb{R}_+$, on a nécessairement

$$\varphi'(\alpha_k) = \nabla f(\boldsymbol{x}^{(k)} + \alpha_k\boldsymbol{w}^{(k)}) \cdot \boldsymbol{w}^{(k)} = 0.$$

Cette équation donne en général le moyen de calculer α_k.
Considérons par exemple le cas (important) d'une fonctionnelle quadratique, *i.e.* $f(\boldsymbol{x}) = \frac{1}{2}A\boldsymbol{x}\cdot\boldsymbol{x} - \boldsymbol{b}\cdot\boldsymbol{x}$, A étant une matrice symétrique définie positive. Alors $\nabla f(\boldsymbol{x}^{(k)}) = A\boldsymbol{x}^{(k)} - \boldsymbol{b}$, et donc

$$\nabla f(\boldsymbol{x}^{(k)} + \alpha_k\, \boldsymbol{w}^{(k)}) \cdot \boldsymbol{w}^{(k)} = (A\boldsymbol{x}^{(k)} + \alpha_k A\boldsymbol{w}^{(k)} - \boldsymbol{b}) \cdot \boldsymbol{w}^{(k)} = 0.$$

On a ainsi dans ce cas une expression explicite de α_k, avec $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}$,

$$\alpha_k = \frac{(\boldsymbol{b} - A\boldsymbol{x}^{(k)}) \cdot \boldsymbol{w}^{(k)}}{A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)}} = \frac{\boldsymbol{r}^{(k)} \cdot \boldsymbol{w}^{(k)}}{A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)}} \tag{3.24}$$

Remarquons que $A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)} \neq 0$ (car A est symétrique définie positive).
Dans le cas d'une fonction f générale, on n'a pas en général de formule explicite pour α_k. On peut par exemple le calculer en cherchant le zéro de f' par la méthode de la sécante ou la méthode de Newton...

L'algorithme du gradient à pas optimal est donc une méthode de minimisation dont on a prouvé la convergence. Cependant, cette convergence est lente (en général linéaire), et de plus, l'algorithme nécessite le calcul du paramètre α_k optimal.

Algorithme du gradient à pas variable Dans ce nouvel algorithme, on ne prend pas forcément le paramètre optimal pour α, mais on lui permet d'être variable d'une itération à l'autre. L'algorithme s'écrit :

$$\left\{\begin{array}{ll} \text{Initialisation :} & x^{(0)} \in \mathbb{R}^n. \\ \text{Itération :} & \text{On suppose } x^{(k)} \text{ connu ; soit } \boldsymbol{w}^{(k)} = -\nabla f(\boldsymbol{x}^{(k)}) \text{ où : } \boldsymbol{w}^{(k)} \neq 0 \\ & (\text{si } \boldsymbol{w}^{(k)} = 0 \text{ l'algorithme s'arrête}). \\ & \text{On prend } \alpha_k > 0 \text{ tel que } f(\boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}) < f(x_k). \\ & \text{On pose } \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}. \end{array}\right. \tag{3.25}$$

Théorème 3.21 (Convergence du gradient à pas variable)**.**
Soit $f \in C^1(\mathbb{R}^n, \mathbb{R})$ une fonction telle que $f(x) \to +\infty$ quand $|x| \to +\infty$, alors :

1. *On peut définir une suite $(x^{(k)})_{k\in\mathbb{N}}$ par (3.25).*
2. *La suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ est bornée. Si $\boldsymbol{x}^{(k_\ell)} \to \boldsymbol{x}$ quand $\ell \to +\infty$ et si $\nabla f(\boldsymbol{x}^{(k_\ell)}) \to 0$ quand $\ell \to +\infty$ alors $\nabla f(\boldsymbol{x}) = 0$. Si de plus f est convexe on a $f(\boldsymbol{x}) = \inf_{\mathbb{R}^n} f$*
3. *Si $\nabla f(\boldsymbol{x}^{(k)}) \to 0$ quand $k \to +\infty$ et si f est strictement convexe alors $\boldsymbol{x}^{(k)} \to \bar{\boldsymbol{x}}$ et $f(\bar{\boldsymbol{x}}) = \inf_{\mathbb{R}^n} f$.*

La démonstration s'effectue facilement à partir de la démonstration du théorème précédent : reprendre en l'adaptant l'exercice 118.

3.3.2 Algorithme du gradient conjugué

La méthode du gradient conjugué a été découverte en 1952 par Hestenes et Steifel pour la minimisation de fonctions quadratiques, c'est-à-dire de fonctions de la forme

$$f(\boldsymbol{x}) = \frac{1}{2} A\boldsymbol{x} \cdot \boldsymbol{x} - b \cdot \boldsymbol{x},$$

où $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice symétrique définie positive et $b \in \mathbb{R}^n$. On rappelle (voir le paragraphe 3.2.2) que $f(\bar{\boldsymbol{x}}) = \inf_{\mathbb{R}^n} f \Leftrightarrow A\bar{\boldsymbol{x}} = b$.
L'idée de la méthode du gradient conjugué est basée sur la remarque suivante : supposons qu'on sache construire n vecteurs (les directions de descente) $\boldsymbol{w}^{(0)}, \boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(n-1)}$ libres et tels que $\boldsymbol{r}^{(n)} \cdot \boldsymbol{w}^{(p)} = 0$ pour tout $p < n$. On a alors $\boldsymbol{r}^{(n)} = \boldsymbol{0}$: en effet la famille $(\boldsymbol{w}^{(0)}, \boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(n-1)})$ engendre $\mathbb{R}^n$; le vecteur $\boldsymbol{r}^{(n)}$ est alors orthogonal à tous les vecteurs d'une $\mathbb{R}^n$, et il est donc nul.
Pour obtenir une famille libre de directions de descente stricte, on va construire les vecteurs $\boldsymbol{w}^{(0)}$, $\boldsymbol{w}^{(1)}$, ..., $\boldsymbol{w}^{(n-1)}$ de manière à ce qu'ils soient orthogonaux pour le produit scalaire induit par A. Nous allons voir que ce choix marche (presque) magnifiquement bien. Mais avant d'expliquer pourquoi, écrivons une méthode de descente à pas optimal pour la minimisation de f, en supposant les directions de descente $\boldsymbol{w}^{(0)}$ connues.
On part de $\boldsymbol{x}^{(0)}$ dans $\mathbb{R}^n$ donné ; à l'itération k, on suppose que $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)} \neq \boldsymbol{0}$ (sinon on a $\boldsymbol{x}^{(k)} = \bar{\boldsymbol{x}}$ et on a fini). On calcule le paramètre α_k optimal dans la direction $\boldsymbol{w}^{(k)}$ par la formule (3.24). Et on calcule ensuite le nouvel itéré :

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}.$$

Notons que $\boldsymbol{r}^{(k+1)} = \boldsymbol{b} - A\boldsymbol{x}^{(k+1)}$ et donc

$$\boldsymbol{r}^{(k+1)} = \boldsymbol{r}^{(k)} - \alpha_k A\boldsymbol{w}^{(k)}. \tag{3.26}$$

De plus, par définition du paramètre optimal α_k, on a $\nabla f(\boldsymbol{x}^{(k+1)}) \cdot \boldsymbol{w}^{(k)} = 0$ et donc

$$\boldsymbol{r}^{(k+1)} \cdot \boldsymbol{w}^{(k)} = 0 \tag{3.27}$$

Ces deux dernières propriétés sont importantes pour montrer la convergence de la méthode. Mais il nous faut maintenant choisir les vecteurs $\boldsymbol{w}^{(k)}$ qui soient des directions de descente strictes et qui forment une famille libre. A l'étape 0, il est naturel de choisir la direction opposée du gradient :

$$\boldsymbol{w}^{(0)} = -\nabla f(\boldsymbol{x}^{(0)}) = \boldsymbol{r}^{(0)}.$$

A l'étape $k \geq 1$, on choisit la direction de descente $\boldsymbol{w}^{(k)}$ comme combinaison linéaire de $\boldsymbol{r}^{(k)}$ et de $\boldsymbol{w}^{(k-1)}$, de manière à ce que $\boldsymbol{w}^{(k)}$ soit orthogonal à $\boldsymbol{w}^{(k-1)}$ pour le produit scalaire associé à la matrice A.

$$\boldsymbol{w}^{(0)} = \boldsymbol{r}^{(0)}, \tag{3.28a}$$

$$\boldsymbol{w}^{(k)} = \boldsymbol{r}^{(k)} + \lambda_k \boldsymbol{w}^{(k-1)}, \text{ avec } \boldsymbol{w}^{(k)} \cdot A\boldsymbol{w}^{(k-1)} = 0, \text{ pour } k \geq 1. \tag{3.28b}$$

La contrainte d'orthogonalité $A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k-1)} = 0$ impose le choix du paramètre λ_k suivant :

$$\lambda_k = -\frac{\boldsymbol{r}^{(k)} \cdot A\boldsymbol{w}^{(k-1)}}{\boldsymbol{w}^{(k-1)} \cdot A\boldsymbol{w}^{(k-1)}}.$$

Remarquons que si $\boldsymbol{r}^{(k)} \neq \boldsymbol{0}$ alors $\boldsymbol{w}^{(k)} \cdot \boldsymbol{r}^{(k)} > 0$ car $\boldsymbol{w}^{(k)} \cdot \boldsymbol{r}^{(k)} = \boldsymbol{r}^{(k)} \cdot \boldsymbol{r}^{(k)}$ en raison de la propriété (3.27). On a donc $\boldsymbol{w}^{(k)} \cdot \nabla f(\boldsymbol{x}^{(k)}) < 0$, ce qui montre que $\boldsymbol{w}^{(k)}$ est bien une direction de descente stricte.
On a donc (on a déjà fait ce calcul pour obtenir la formule (3.24) du paramètre optimal)

$$\alpha_k = \frac{\boldsymbol{r}^{(k)} \cdot \boldsymbol{w}^{(k)}}{A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)}} = \frac{\boldsymbol{r}^{(k)} \cdot \boldsymbol{r}^{(k)}}{A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)}}. \tag{3.29}$$

On suppose que $\boldsymbol{r}^{(k)} \neq \boldsymbol{0}$ pour tout $k \in \{0, \ldots, n-1\}$. Montrons alors par récurrence que pour $k = 1, \ldots, n-1$, on a :

$(i)_k$ $\quad \boldsymbol{r}^{(k)} \cdot \boldsymbol{w}^{(p)} = 0$ si $p < k$,

$(ii)_k$ $\quad \boldsymbol{r}^{(k)} \cdot \boldsymbol{r}^{(p)} = 0$ si $p < k$,

$(iii)_k$ $\quad A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(p)} = 0$ si $p < k$,

Ces relations sont vérifiées pour $k = 1$. Supposons qu'elles le sont jusqu'au rang k, et montrons qu'elles le sont au rang $k+1$.

$(i)_{k+1}$: Pour $p = k$, la relation $(i)_{k+1}$ est verifiée au rang $k+1$ grâce à (3.27) ; pour $p < k$, on a

$$\boldsymbol{r}^{(k+1)} \cdot \boldsymbol{w}^{(p)} = \boldsymbol{r}^{(k)} \cdot \boldsymbol{w}^{(p)} - \alpha_k A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(p)} = 0$$

par (3.26) et hypothèse de récurrence.

$(ii)_{k+1}$: Par les relations (3.28b) et $(i)_{k+1}$, on a, pour $p \leq k$,

$$\boldsymbol{r}^{(k+1)} \cdot \boldsymbol{r}^{(p)} = \boldsymbol{r}^{(k+1)} \cdot (\boldsymbol{w}^{(p)} - \lambda_p \boldsymbol{w}^{(p-1)}) = 0.$$

$(iii)_{k+1}$: Pour $p = k$ la relation $(iii)_{k+1}$ est vérifiée grâce au choix de λ_{k+1}.
Pour $p < k$, on remarque que, avec (3.28b) et $(iii)_k$

$$\boldsymbol{w}^{(k+1)} \cdot A\boldsymbol{w}^{(p)} = (\boldsymbol{r}^{(k+1)} + \lambda_{k+1}\boldsymbol{w}^{(k)}) \cdot A\boldsymbol{w}^{(p)} = \boldsymbol{r}^{(k+1)} \cdot A\boldsymbol{w}^{(p)}.$$

On utilise maintenant (3.26) et $(i)_{k+1}$ pour obtenir

$$\boldsymbol{w}^{(k+1)} \cdot A\boldsymbol{w}^{(p)} = \frac{1}{\alpha_p}\boldsymbol{r}^{(k+1)} \cdot (\boldsymbol{r}^{(p)} - \boldsymbol{r}^{(p+1)}) = 0.$$

On a ainsi démontré la convergence de la méthode du gradient conjugué.

Mettons sous forme algorithmique les opérations que nous avons exposées, pour obtenir l'algorithme du gradient conjugué.

Algorithme 3.22 (Méthode du gradient conjugué)**.**

1. **Initialisation**
   Soit $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$, *et soit* $\boldsymbol{r}^{(0)} = \boldsymbol{b} - A\boldsymbol{x}^{(0)} = -\nabla f(\boldsymbol{x}^{(0)})$.

   Si $\boldsymbol{r}^{(0)} = \boldsymbol{0}$, *alors* $A\boldsymbol{x}^{(0)} = \boldsymbol{b}$ *et donc* $\boldsymbol{x}^{(0)} = \bar{\boldsymbol{x}}$, *auquel cas l'algorithme s'arrête.*

   Sinon, on pose
   $$\boldsymbol{w}^{(0)} = \boldsymbol{r}^{(0)},$$
   et on choisit α_0 *optimal dans la direction* $\boldsymbol{w}^{(0)}$. *On pose alors*
   $$\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} + \alpha_0\boldsymbol{w}^{(0)}.$$

2. **Itération** k, $1 \leq k \leq n-1$ *; on suppose* $\boldsymbol{x}^{(0)}, \ldots, \boldsymbol{x}^{(k)}$ *et* $\boldsymbol{w}^{(0)}, \ldots, \boldsymbol{w}^{(k-1)}$ *connus et on pose*
   $$\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}.$$

   Si $\boldsymbol{r}^{(k)} = \boldsymbol{0}$, *alors* $A\boldsymbol{x}^{(k)} = \boldsymbol{b}$ *et donc* $\boldsymbol{x}^{(k)} = \bar{\boldsymbol{x}}$, *auquel cas l'algorithme s'arrête.*

   Sinon on pose
   $$\boldsymbol{w}^{(k)} = \boldsymbol{r}^{(k)} + \lambda_{k-1}\boldsymbol{w}^{(k-1)},$$
   avec λ_{k-1} *tel que*
   $$\boldsymbol{w}^{(k)} \cdot A\boldsymbol{w}^{(k-1)} = 0,$$
   et on choisit α_k *optimal dans la direction* $\boldsymbol{w}^{(k)}$, *donné par* (3.24). *On pose alors*
   $$\boldsymbol{x}^{(k+1)} = x^{(k)} + \alpha_k\boldsymbol{w}^{(k)}.$$

Nous avons démontré plus haut la convergence de l'algorithme, résultat que nous énonçons dans le théorème suivant.

Théorème 3.23 (Convergence de l'algorithme du gradient conjugué)**.** *Soit* A *une symétrique définie positive,* $A \in \mathcal{M}_n(\mathbb{R})$, $\boldsymbol{b} \in \mathbb{R}^n$ *et* $f(\boldsymbol{x}) = \frac{1}{2}A\boldsymbol{x} \cdot \boldsymbol{x} - \boldsymbol{b} \cdot \boldsymbol{x}$. *L'algorithme* (3.22) *définit une suite* $(\boldsymbol{x}^{(k)})_{k=0,\ldots,p}$ *avec* $p \leq n$ *telle que* $\boldsymbol{x}^{(p)} = \bar{\boldsymbol{x}}$ *avec* $A\bar{\boldsymbol{x}} = \boldsymbol{b}$. *On obtient donc la solution exacte de la solution du système linéaire* $A\boldsymbol{x} = \boldsymbol{b}$ *en moins de* n *itérations.*

Efficacité de la méthode du gradient conjugué On peut calculer le nombre d'opérations nécessaires pour calculer $\bar{\boldsymbol{x}}$ (c.à.d. pour calculer $\boldsymbol{x}^{(n)}$, sauf dans le cas miraculeux où $\boldsymbol{x}^{(k)} = \bar{\boldsymbol{x}}$ pour $k < n$) et montrer (exercice) que :

$$N_{gc} = 2n^3 + \mathcal{O}(n^2).$$

On rappelle que le nombre d'opérations pour Choleski est $\dfrac{n^3}{6}$ donc la méthode du gradient conjugué n'est pas intéressante comme méthode directe car elle demande 12 fois plus d'opérations que Choleski.

On peut alors se demander si la méthode est intéressante comme méthode itérative, c.à.d. si on peut espérer que $\boldsymbol{x}^{(k)}$ soit "proche de $\bar{\boldsymbol{x}}$" pour "$k \ll n$". Malheureusement, si la dimension n du système est grande, ceci n'est pas le cas en raison de l'accumulation des erreurs d'arrondi. Il est même possible de devoir effectuer plus de n itérations pour se rapprocher de $\bar{\boldsymbol{x}}$. Cependant, dans les années 80, des chercheurs se sont rendus compte que ce défaut pouvait être corrigé à condition d'utiliser un "préconditionnement". Donnons par exemple le principe du préconditionnement dit de "Choleski incomplet".

Méthode du gradient conjugué préconditionné par Choleski incomplet On commence par calculer une "approximation" de la matrice de Choleski de A c.à.d. qu'on cherche L triangulaire inférieure inversible telle que A soit "proche" de LL^t, en un sens à définir. Si on pose $y = L^t\boldsymbol{x}$, alors le système $A\boldsymbol{x} = b$ peut aussi s'écrire $L^{-1}A(L^t)^{-1}y = L^{-1}b$, et le système $(L^t)^{-1}y = \boldsymbol{x}$ est facile à résoudre car L^t est triangulaire supérieure. Soit $B \in \mathcal{M}_n(\mathbb{R})$ définie par $B = L^{-1}A(L^t)^{-1}$, alors

$$B^t = ((L^t)^{-1})^t A^t (L^{-1})^t = L^{-1}A(L^t)^{-1} = B$$

et donc B est symétrique. De plus,

$$B\boldsymbol{x}\cdot\boldsymbol{x} = L^{-1}A(L^t)^{-1}\boldsymbol{x}\cdot\boldsymbol{x} = A(L^t)^{-1}\boldsymbol{x}\cdot(L^t)^{-1}\boldsymbol{x},$$

et donc $B\boldsymbol{x}\cdot\boldsymbol{x} > 0$ si $\boldsymbol{x} \neq 0$. La matrice B est donc symétrique définie positive. On peut donc appliquer l'algorithme du gradient conjugué à la recherche du minimum de la fonction f définie par

$$f(\boldsymbol{y}) = \frac{1}{2}B\boldsymbol{y}\cdot\boldsymbol{y} - L^{-1}b\cdot\boldsymbol{y}.$$

On en déduit l'expression de la suite $(\boldsymbol{y}^{(k)})_{k\in\mathbb{N}}$ et donc $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$.
On peut alors montrer (voir exercice 125) que l'algorithme du gradient conjugué préconditionné ainsi obtenu peut s'écrire directement pour la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$, de la manière suivante :

Itération k On pose $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}$,
on calcule $\boldsymbol{s}^{(k)}$ solution de $LL^t\boldsymbol{s}^{(k)} = \boldsymbol{r}^{(k)}$.
On pose alors $\lambda_{k-1} = \dfrac{\boldsymbol{s}^{(k)}\cdot\boldsymbol{r}^{(k)}}{\boldsymbol{s}^{(k-1)}\cdot\boldsymbol{r}^{(k-1)}}$ et $\boldsymbol{w}^{(k)} = \boldsymbol{s}^{(k)} + \lambda_{k-1}\boldsymbol{w}^{(k-1)}$.
Le paramètre optimal α_k a pour expression :

$$\alpha_k = \frac{\boldsymbol{s}^{(k)}\cdot\boldsymbol{r}^{(k)}}{A\boldsymbol{w}^{(k)}\cdot\boldsymbol{w}^{(k)}},$$

et on pose alors $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k\boldsymbol{w}^{(k)}$.
Le choix de la matrice L peut se faire par exemple dans le cas d'une matrice creuse, en effectuant une factorisation "LL^t" incomplète, qui consiste à ne remplir que certaines diagonales de la matrice L pendant la factorisation, et laisser les autres à 0.

Méthode du gradient conjugué pour une fonction non quadratique. On peut généraliser le principe de l'algorithme du gradient conjugué à une fonction f non quadratique. Pour cela, on reprend le même algorithme que (3.22), mais on adapte le calcul de λ_{k-1} et α_k.

Itération n :
A $\boldsymbol{x}^{(0)},\ldots,\boldsymbol{x}^{(k)}$ et $w^{(0)},\ldots,w^{(k-1)}$ connus, on calcule $\boldsymbol{r}^{(k)} = -\nabla f(\boldsymbol{x}^{(k)})$.
Si $\boldsymbol{r}^{(k)} = 0$ alors $\nabla f(\boldsymbol{x}^{(k)}) = 0$ auquel cas l'algorithme s'arrête (le point $\boldsymbol{x}^{(k)}$ est un point critique de f et il minimise f si f est convexe).
Si $\boldsymbol{r}^{(k)} \neq 0$, on pose $\boldsymbol{w}^{(k)} = \boldsymbol{r}^{(k)} + \lambda_{k-1}w^{(k-1)}$ où λ_{k-1} peut être choisi de différentes manières :

1ère méthode (Fletcher–Reeves)

$$\lambda_{k-1} = \frac{\boldsymbol{r}^{(k)}\cdot\boldsymbol{r}^{(k)}}{\boldsymbol{r}^{(k-1)}\cdot\boldsymbol{r}^{(k-1)}},$$

2ème méthode (Polak–Ribière)

$$\lambda_{k-1} = \frac{(\boldsymbol{r}^{(k)} - \boldsymbol{r}^{(k-1)}) \cdot \boldsymbol{r}^{(k)}}{\boldsymbol{r}^{(k-1)} \cdot \boldsymbol{r}^{(k-1)}}.$$

On pose alors $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}$, où α_k est choisi, si possible, optimal dans la direction $\boldsymbol{w}^{(k)}$.
La démonstration de la convergence de l'algorithme de Polak–Ribière fait l'objet de l'exercice 127 page 239.

En résumé, la méthode du gradient conjugué est très efficace dans le cas d'une fonction quadratique à condition de l'utiliser avec préconditionnement. Dans le cas d'une fonction non quadratique, le préconditionnement ne se trouve pas de manière naturelle et il vaut donc mieux réserver cette méthode dans le cas "n petit".

3.3.3 Méthodes de Newton et Quasi–Newton

Soit $f \in C^2(\mathbb{R}^n, \mathbb{R})$ et $g = \nabla f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$. On a dans ce cas :

$$f(\boldsymbol{x}) = \inf_{\mathbb{R}^n} f \Rightarrow g(\boldsymbol{x}) = 0.$$

Si de plus f est convexe alors on a $g(\boldsymbol{x}) = 0 \Rightarrow f(\boldsymbol{x}) = \inf_{\mathbb{R}^n} f$. Dans ce cas d'équivalence, on peut employer la méthode de Newton pour minimiser f en appliquant l'algorithme de Newton pour chercher un zéro de $g = \nabla f$. On a $D(\nabla f) = H_f$ où $H_f(\boldsymbol{x})$ est la matrice hessienne de f en $\boldsymbol{x}$. La méthode de Newton s'écrit dans ce cas :

$$\left\{ \begin{array}{ll} \text{Initialisation} & \boldsymbol{x}^{(0)} \in \mathbb{R}^n, \\ \text{Itération } k & H_f(\boldsymbol{x}^{(k)})(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) = -\nabla f(\boldsymbol{x}^{(k)}). \end{array} \right. \tag{3.31}$$

Remarque 3.24. La méthode de Newton pour minimiser une fonction f convexe est une méthode de descente. En effet, si $H_f(\boldsymbol{x}^{(k)})$ est inversible, on a $\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)} = [H_f(\boldsymbol{x}^{(k)})]^{-1}(-\nabla f(\boldsymbol{x}^{(k)}))$ soit encore $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}$ où $\alpha_k = 1$ et $w^{(k)} = [H_f(\boldsymbol{x}^{(k)})]^{-1}(-\nabla f(\boldsymbol{x}^{(k)}))$. Si f est convexe, H_f est une matrice symétrique positive (déjà vu). Comme on suppose $H_f(\boldsymbol{x}^{(k)})$ inversible par hypothèse, la matrice $H_f(\boldsymbol{x}^{(k)})$ est donc symétrique définie positive.
On en déduit que $w^{(k)} = 0$ si $\nabla f(\boldsymbol{x}^{(k)}) = 0$ et, si $\nabla f(\boldsymbol{x}^{(k)}) \neq 0$,

$$-\boldsymbol{w}^{(k)} \cdot \nabla f(\boldsymbol{x}^{(k)}) = [H_f(\boldsymbol{x}^{(k)})]^{-1} \nabla f(\boldsymbol{x}^{(k)}) \cdot \nabla f(\boldsymbol{x}^{(k)}) > 0,$$

ce qui est une condition suffisante pour que $\boldsymbol{w}^{(k)}$ soit une direction de descente stricte.
La méthode de Newton est donc une méthode de descente avec $\boldsymbol{w}^{(k)} = -H_f(\boldsymbol{x}^{(k)})^{-1}(\nabla f(\boldsymbol{x}^{(k)}))$ et $\alpha_k = 1$.

On peut aussi remarquer, en vertu du théorème 2.19 page 169, que si $f \in C^3(\mathbb{R}^n, \mathbb{R})$, si $\bar{\boldsymbol{x}}$ est tel que $\nabla f(\bar{\boldsymbol{x}}) = 0$ et si $H_f(\bar{\boldsymbol{x}}) = D(\nabla f)(\bar{\boldsymbol{x}})$ est inversible alors il existe $\varepsilon > 0$ tel que si $\boldsymbol{x}_0 \in B(\bar{\boldsymbol{x}}, \varepsilon)$, alors la suite $(\boldsymbol{x}^{(k)})_k$ est bien définie par (3.31) et $\boldsymbol{x}^{(k)} \to \bar{\boldsymbol{x}}$ lorsque $k \to +\infty$. De plus, d'après la proposition 2.16, il existe $\beta > 0$ tel que $|\boldsymbol{x}^{(k+1)} - \bar{\boldsymbol{x}}| \leq \beta |\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}|^2$ pour tout $k \in \mathbb{N}$.

Remarque 3.25 (Sur l'implantation numérique)**.** La convergence de la méthode de Newton est très rapide, mais nécessite en revanche le calcul de $H_f(\boldsymbol{x})$, qui peut s'avérer impossible ou trop coûteux.

On va maintenant donner des variantes de la méthode de Newton qui évitent le calcul de la matrice hessienne.

Proposition 3.26. *Soient $f \in C^1(\mathbb{R}^n, \mathbb{R})$, $\boldsymbol{x} \in \mathbb{R}^n$ tel que $\nabla f(\boldsymbol{x}) \neq 0$, et soit $B \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique définie positive ; alors $w = -B\nabla f(\boldsymbol{x})$ est une direction de descente stricte en $\boldsymbol{x}$.*

DÉMONSTRATION – On a : $w \cdot \nabla f(\boldsymbol{x}) = -B\nabla f(\boldsymbol{x}) \cdot \nabla f(\boldsymbol{x}) < 0$ car B est symétrique définie positive et $\nabla f(\boldsymbol{x}) \neq 0$ donc $\boldsymbol{w}$ est une direction de descente stricte en $\boldsymbol{x}$. En effet, soit φ la fonction de $\mathbb{R}$ dans $\mathbb{R}$ définie par $\varphi(\alpha) = f(\boldsymbol{x} + \alpha \boldsymbol{w})$. Il est clair que $\varphi \in C^1(\mathbb{R}, \mathbb{R})$, $\varphi'(\alpha) = \nabla f(\boldsymbol{x} + \alpha \boldsymbol{w}) \cdot w$ et $\varphi'(0) = \nabla f(\boldsymbol{x}) \cdot w < 0$. Donc $\exists \alpha_0 > 0$ tel que $\varphi'(\alpha) < 0$ si $\alpha \in]0, \alpha_0[$. Par le théorème des accroissements finis, $\varphi(\alpha) < \varphi(0)$ $\forall \alpha \in]0, \alpha_0[$ donc w est une direction de descente stricte. ∎

Méthode de Broyden La première idée pour construire une méthode de type quasi Newton est de prendre comme direction de descente en $\boldsymbol{x}^{(k)}$ le vecteur $\boldsymbol{w}^{(k)} = -(B^{(k)})^{-1}(\nabla f(\boldsymbol{x}^{(k)}))$ où la matrice $B^{(k)}$ est censée approcher $H_f(\boldsymbol{x}^{(k)})$ (sans calculer la dérivée seconde de f). On suppose $\boldsymbol{x}^{(k)}, \boldsymbol{x}^{(k-1)}$ et $B^{(k-1)}$ connus. Voyons comment on peut déterminer $B^{(k)}$. On peut demander par exemple que la condition suivante soit satisfaite :

$$\nabla f(\boldsymbol{x}^{(k)}) - \nabla f(\boldsymbol{x}^{(k-1)}) = B^{(k)}(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}). \tag{3.32}$$

Ceci est un système à n équations et $n \times n$ inconnues, et ne permet donc pas de déterminer entièrement la matrice $B^{(k)}$ si $n > 1$. Voici un moyen possible pour déterminer entièrement $B^{(k)}$, dû à Broyden. On pose $\boldsymbol{s}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}$, on suppose que $\boldsymbol{s}^{(k)} \neq 0$, et on pose $\boldsymbol{y}^{(k)} = \nabla f(\boldsymbol{x}^{(k)}) - \nabla f(\boldsymbol{x}^{(k-1)})$. On choisit alors $B^{(k)}$ telle que :

$$\begin{cases} B^{(k)}\boldsymbol{s}^{(k)} = \boldsymbol{y}^{(k)} \\ B^{(k)}\boldsymbol{s} = B^{(k-1)}\boldsymbol{s},\ \forall s \perp s^{(k)} \end{cases} \tag{3.33}$$

On a exactement le nombre de conditions qu'il faut avec (3.33) pour déterminer entièrement $B^{(k)}$. Ceci suggère la méthode suivante :

Initialisation Soient $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ et $B^{(0)}$ une matrice symétrique définie positive. On pose

$$\boldsymbol{w}^{(0)} = (B^{(0)})^{-1}(-\nabla f(\boldsymbol{x}^{(0)}));$$

alors $\boldsymbol{w}^{(0)}$ est une direction de descente stricte sauf si $\nabla f(\boldsymbol{x}^{(0)}) = 0$.
On pose alors

$$\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} + \alpha^{(0)}\boldsymbol{w}^{(0)},$$

où $\alpha^{(0)}$ est optimal dans la direction $\boldsymbol{w}^{(0)}$.

Itération k On suppose $\boldsymbol{x}^{(k)}$, $\boldsymbol{x}^{(k-1)}$ et $B^{(k-1)}$ connus, $(k \geq 1)$, et on calcule $B^{(k)}$ par (3.33). On pose

$$\boldsymbol{w}^{(k)} = -(B^{(k)})^{-1}(\nabla f(\boldsymbol{x}^{(k)})).$$

On choisit $\alpha^{(k)}$ optimal en $\boldsymbol{x}^{(k)}$ dans la direction $\boldsymbol{w}^{(k)}$, et on pose $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha^{(k)}\boldsymbol{w}^{(k)}$.

Le problème avec cet algorithme est que si la matrice est $B^{(k-1)}$ symétrique définie positive, la matrice $B^{(k)}$ ne l'est pas forcément, et donc $\boldsymbol{w}^{(k)}$ n'est pas forcément une direction de descente stricte. On va donc modifier cet algorithme dans ce qui suit.

Méthode de BFGS La méthode BFGS (de Broyden[1], Fletcher[2], Goldfarb[3] et Shanno[4]) cherche à construire $B^{(k)}$ proche de $B^{(k-1)}$, telle que $B^{(k)}$ vérifie (3.32) et telle que si $B^{(k-1)}$ est symétrique définie positive alors $B^{(k)}$ est symétrique définie positive. On munit $\mathcal{M}_n(\mathbb{R})$ d'une norme induite par un produit scalaire, par exemple si $A \in \mathcal{M}_n(\mathbb{R})$ et $A = (a_{i,j})_{i,j=1,\ldots,n}$ on prend $\|A\| = \left(\sum_{i,j=1}^n a_{i,j}^2\right)^{1/2}$. $\mathcal{M}_n(\mathbb{R})$ est alors un espace de Hilbert.

On suppose $\boldsymbol{x}^{(k)}$, $\boldsymbol{x}^{(k-1)}$, $B^{(k-1)}$ connus, et on définit

$$\mathcal{C}_k = \{B \in \mathcal{M}_n(\mathbb{R}) | B \text{ symétrique, vérifiant (3.32)}\},$$

qui est une partie de $\mathcal{M}_n(\mathbb{R})$ convexe fermée non vide. On choisit alors $B^{(k)} = P_{\mathcal{C}_k}B^{(k-1)}$ où $P_{\mathcal{C}_k}$ désigne la projection orthogonale sur $\mathcal{C}_k$. La matrice $B^{(k)}$ ainsi définie existe et est unique ; elle est symétrique d'après le choix de $\mathcal{C}_k$. On peut aussi montrer que si $B^{(k-1)}$ symétrique définie positive alors $B^{(k)}$ est aussi symétrique définie positive.

1. Broyden, C. G., The Convergence of a Class of Double-rank Minimization Algorithms, *Journal of the Institute of Mathematics and Its Applications* 1970, 6, 76-90
2. Fletcher, R., A New Approach to Variable Metric Algorithms, Computer Journal 1970, 13, 317-322
3. Goldfarb, D., A Family of Variable Metric Updates Derived by Variational Means, Mathematics of Computation 1970, 24, 23-26
4. Shanno, D. F.,Conditioning of Quasi-Newton Methods for Function Minimization , Mathematics of Computation 1970, 24, 647-656

Avec un choix convenable de la norme sur $\mathcal{M}_n(\mathbb{R})$, on obtient le choix suivant de $B^{(k)}$ si $s^{(k)} \neq 0$ et $\nabla f(\boldsymbol{x}^{(k)}) \neq 0$ (sinon l'algorithme s'arrête) :

$$B^{(k)} = B^{(k-1)} + \frac{\boldsymbol{y}^{(k)}(\boldsymbol{y}^{(k)})^t}{(\boldsymbol{s}^{(k)})^t \cdot y^{(k)}} - \frac{B^{(k-1)}\boldsymbol{s}^{(k)}(\boldsymbol{s}^{(k)})^t B^{(k-1)}}{(\boldsymbol{s}^{(k)})^t B^{(k-1)}\boldsymbol{s}^{(k)}}. \tag{3.34}$$

L'algorithme obtenu est l'algorithme de BFGS.

Algorithme de BFGS

$$\left\{\begin{array}{ll} \underline{\text{Initialisation}} & \text{On choisit } \boldsymbol{x}^{(0)} \in \mathbb{R}^n \text{ et} \\ & B^{(0)} \text{ symétrique définie positive} \\ & (\text{ par exemple } B^{(0)} = Id) \text{ et on pose} \\ & \boldsymbol{w}^{(0)} = -B^{(0)}\nabla f(\boldsymbol{x}^{(0)}) \\ & \text{si } \nabla f(\boldsymbol{x}^{(0)}) \neq 0, \text{ on choisit } \alpha^{(0)} \text{ optimal} \\ & \text{dans la direction } \boldsymbol{w}^{(0)}, \text{ et donc} \\ & \boldsymbol{w}^{(0)} \text{ est une direction de descente stricte.} \\ & \text{On pose } \boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} + \alpha^{(0)}\boldsymbol{w}^{(0)}. \\ \underline{\text{Itération } k} & \text{A } \boldsymbol{x}^{(k)}, \boldsymbol{x}^{(k-1)} \text{ et } B_{k-1} \text{ connus } \ (k \geq 1) \\ & \text{On pose} \\ & \boldsymbol{s}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)} \ \ \boldsymbol{y}^{(k)} = \nabla f(\boldsymbol{x}^{(k)}) - \nabla f(\boldsymbol{x}^{(k-1)}) \\ & \text{si } \boldsymbol{s}^{(k)} \neq 0 \text{ et } \nabla f(\boldsymbol{x}^{(k)}) \neq 0, \\ & \text{on choisit } B^{(k)} \text{ vérifiant (3.34)} \\ & \text{On calcule } \boldsymbol{w}^{(k)} = -(B^{(k)})^{-1}(\nabla f(\boldsymbol{x}^{(k)})) \\ & \text{(direction de descente stricte en } \boldsymbol{x}^{(k)}). \\ & \text{On calcule } \alpha^{(k)} \text{ optimal dans la direction } \boldsymbol{w}^{(k)} \\ & \text{et on pose } \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha^{(k)}\boldsymbol{w}^{(k)}. \end{array}\right. \tag{3.35}$$

On donne ici sans démonstration le théorème de convergence suivant :

Théorème 3.27 (Fletcher, 1976)**.** *Soit $f \in C^2(\mathbb{R}^n, \mathbb{R})$ telle que $f(\boldsymbol{x}) \to +\infty$ quand $|\boldsymbol{x}| \to +\infty$. On suppose de plus que f est strictement convexe (donc il existe un unique $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ tel que $f(\bar{\boldsymbol{x}}) = \inf_{\mathbb{R}^n} f$) et on suppose que la matrice hessienne $H_f(\bar{\boldsymbol{x}})$ est symétrique définie positive.*
Alors si $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ et si $B^{(0)}$ est symétrique définie positive, l'algorithme BFGS définit bien une suite $\boldsymbol{x}^{(k)}$ et on a $\boldsymbol{x}^{(k)} \to \bar{\boldsymbol{x}}$ quand $k \to +\infty$
De plus, si $\boldsymbol{x}^{(k)} \neq \bar{\boldsymbol{x}}$ pour tout k, la convergence est super linéaire i.e.

$$\left|\frac{\boldsymbol{x}^{(k+1)} - \bar{\boldsymbol{x}}}{\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}}\right| \to 0 \text{ quand } k \to +\infty.$$

Pour éviter la résolution d'un système linéaire dans BFGS, on peut choisir de travailler sur $(B^{(k)})^{-1}$ au lieu de $B^{(k)}$.

$$\left\{\begin{array}{l} \textbf{Initialisation Soit } \boldsymbol{x}^{(0)} \in \mathbb{R}^n \text{ et } K^{(0)} \text{ symétrique définie positive} \\ \text{telle que } \alpha_0 \text{ soit optimal dans la direction } -K^{(0)}\nabla f(\boldsymbol{x}^{(0)}) = \boldsymbol{w}^{(0)} \\ \boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} + \alpha_0 w^{(0)} \\ \textbf{Itération } k : \text{A } \boldsymbol{x}^{(k)}, \boldsymbol{x}^{(k-1)}, K^{(k-1)} \text{ connus, } k \geq 1, \\ \text{on pose } \boldsymbol{s}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}, \boldsymbol{y}^{(k)} = \nabla f(\boldsymbol{x}^{(k)}) - \nabla f(\boldsymbol{x}^{(k-1)}) \\ \text{et } K^{(k)} = P_{\mathcal{C}_k} K^{(k-1)}. \\ \text{On calcule } \boldsymbol{w}^{(k)} = -K^{(k)}\nabla f(\boldsymbol{x}^{(k)}) \text{ et on choisit } \alpha_k \\ \text{optimal dans la direction } \boldsymbol{w}^{(k)}. \\ \text{On pose alors } \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}. \end{array}\right. \tag{3.36}$$

Remarquons que le calcul de la projection de $P_{e_k} K^{(k-1)}$ peut s'effectuer avec la formule (3.34) où on a remplacé $B^{(k-1)}$ par $K^{(k-1)}$. Malheureusement, on obtient expérimentalement une convergence nettement moins bonne pour l'algorithme de quasi-Newton modifié (3.36) que pour l'algorithme de BFGS (3.34).

3.3.4 Résumé sur les méthodes d'optimisation

Faisons le point sur les avantages et inconvénients des méthodes qu'on a vues sur l'optimisation sans contrainte.

Méthodes de gradient : Ces méthodes nécessitent le calcul de $\nabla f(\boldsymbol{x}^{(k)})$. Leur convergence est linéaire (donc lente).

Méthode de gradient conjugué : Si f est quadratique (c.à.d. $f(\boldsymbol{x}) = \frac{1}{2} A\boldsymbol{x} \cdot \boldsymbol{x} - b \cdot \boldsymbol{x}$ avec A symétrique définie positive), la méthode est excellente si elle est utilisée avec un préconditionnement (pour n grand). Dans le cas général, elle n'est efficace que si n n'est pas trop grand.

Méthode de Newton : La convergence de la méthode de Newton est excellente (convergence localement quadratique) mais nécessite le calcul de $H_f(\boldsymbol{x}^{(k)})$ (et de $\nabla f(\boldsymbol{x}^{(k)})$). Si on peut calculer $H_f(\boldsymbol{x}^{(k)})$, cette méthode est parfaite.

Méthode de quasi Newton : L'avantage de la méthode de quasi Newton est qu'on ne calcule que $\nabla f(x^{(k)})$ et pas $H_f(x^{(k)})$). La convergence est super linéaire. Par rapport à une méthode de gradient où on calcule $\boldsymbol{w}^{(k)} = -\nabla f(\boldsymbol{x}^{(k)})$, la méthode BFGS nécessite une résolution de système linéaire :

$$B^{(k)} \boldsymbol{w}^{(k)} = -\nabla f(\boldsymbol{x}^{(k)}).$$

Quasi–Newton modifié :

Pour éviter la résolution de système linéaire dans BFGS, on peut choisir de travailler sur $(B^{(k)})^{-1}$ au lieu de $B^{(k)}$, pour obtenir l'algorithme de quasi Newton (3.36). Cependant, on perd alors en vitesse de convergence.

Comment faire si on ne veut (ou peut) pas calculer $\nabla f(\boldsymbol{x}^{(k)})$? On peut utiliser des "méthodes sans gradient", c.à.d. qu'on choisit *a priori* les directions $\boldsymbol{w}^{(k)}$. Ceci peut se faire soit par un choix déterministe, soit par un choix stochastique.

Un choix déterministe possible est de calculer $\boldsymbol{x}^{(k)}$ en résolvant n problèmes de minimisation en une dimension d'espace. Pour chaque direction $i = 1, \ldots, n$, on prend $w^{(n,i)} = \boldsymbol{e}_i$, où $\boldsymbol{e}_i$ est le i-ème vecteur de la base canonique, et pour $i = 1, \ldots, n$, on cherche $\theta \in \mathbb{R}$ tel que :

$$f(x_1^{(k)}, x_2^{(k)}, \ldots, \theta, \ldots, x_n^{(k)}) \leq f(x_1^{(k)}, \boldsymbol{x}_2^{(k)}, \ldots, t, \ldots, \boldsymbol{x}_n^{(k)}), \forall t \in \mathbb{R}.$$

Remarquons que si f est quadratique, on retrouve la méthode de Gauss Seidel.

3.3.5 Exercices (algorithmes pour l'optimisation sans contraintes)

Exercice 117 (Mise en oeuvre de GPF, GPO)**.** *Corrigé en page 242.*
On considère la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par $f(x_1, x_2) = 2x_1^2 + x_2^2 - x_1 x_2 - 3x_1 - x_2 + 4$.

1. Montrer qu'il existe un unique $\bar{x} \in \mathbb{R}^2$ tel que $\bar{x} = \min_{x \in \mathbb{R}^2} f(x)$ et le calculer.
2. Calculer le premier itéré donné par l'algorithme du gradient à pas fixe (GPF) et du gradient à pas optimal (GPO), en partant de $(x_1^{(0)}, x_2^{(0)}) = (0, 0)$, pour un pas de $\alpha = .5$ dans le cas de GPF.

Exercice 118 (Convergence de l'algorithme du gradient à pas optimal)**.** *Suggestions en page 241. Corrigé détaillé en page 243*
Soit $f \in C^2(\mathbb{R}^n, \mathbb{R})$ t.q. $f(x) \to \infty$ quand $|x| \to \infty$. Soit $x_0 \in \mathbb{R}^n$. On va démontrer dans cet exercice la convergence de l'algorithme du gradient à pas optimal.

1. Montrer qu'il existe $R > 0$ t.q. $f(x) > f(x_0)$ pour tout $x \notin B_R$, avec $B_R = \{x \in \mathbb{R}^n, |x| \leq R\}$.
2. Montrer qu'il existe $M > 0$ t.q. $|H(x)y \cdot y| \leq M|y|^2$ pour tout $y \in \mathbb{R}^n$ et tout $x \in B_{R+1}$ ($H(x)$ est la matrice hessienne de f au point x, R est donné à la question 1).

3. (Construction de "la" suite $(x_k)_{k\in\mathbb{N}}$ de l'algorithme du gradient à pas optimal.) On suppose x_k connu ($k \in \mathbb{N}$). On pose $w_k = -\nabla f(x_k)$. Si $w_k = 0$, on pose $x_{k+1} = x_k$. Si $w_k \neq 0$, montrer qu'il existe $\overline{\rho} > 0$ t.q. $f(x_k + \rho w_k) \leq f(x_k + \rho w_k)$ pour tout $\rho \geq 0$. On choisit alors un $\rho_k > 0$ t.q. $f(x_k + \rho_k w_k) \leq f(x_k + \rho w_k)$ pour tout $\rho \geq 0$ et on pose $x_{k+1} = x_k + \rho_k w_k$.
   On considère, dans les questions suivantes, la suite $(x_k)_{k\in\mathbb{N}}$ ainsi construite.
4. Montrer que (avec R et M donnés aux questions précédentes)
   (a) la suite $(f(x_k))_{k\in\mathbb{N}}$ est une suite convergente,
   (b) $x_k \in B_R$ pour tout $k \in \mathbb{N}$,
   (c) $f(x_k + \rho w_k) \leq f(x_k) - \rho|w_k|^2 + (\rho^2/2)M|w_k|^2$ pour tout $\rho \in [0, 1/|w_k|]$.
   (d) $f(x_{k+1}) \leq f(x_k) - |w_k|^2/(2M)$, si $|w_k| \leq M$.
   (e) $-f(x_{k+1}) + f(x_k) \geq |w_k|^2/(2\overline{M})$, avec $\overline{M} = \sup(M, \tilde{M})$, $\tilde{M} = \sup\{|\nabla f(x)|, x \in B_R\}$.
5. Montrer que $\nabla f(x_k) \to 0$ (quand $k \to \infty$) et qu'il existe une sous suite $(n_k)_{k\in\mathbb{N}}$ t.q. $x_{n_k} \to x$ quand $k \to \infty$ et $\nabla f(x) = 0$.
6. On suppose qu'il existe un unique $\overline{x} \in \mathbb{R}^n$ t.q. $\nabla f(\overline{x}) = 0$. Montrer que $f(\overline{x}) \leq f(x)$ pour tout $x \in \mathbb{R}^n$ et que $x_k \to \overline{x}$ quand $k \to \infty$.

Exercice 119 (Jacobi et optimisation)**.** *Corrigé détaillé en page 245*
Rappel Soit $f \in C^1(\mathbb{R}^n, \mathbb{R})$; on appelle **méthode de descente à pas fixe** $\alpha \in \mathbb{R}_+^*$ pour la minimisation de f, une suite définie par

$$\begin{aligned} &\boldsymbol{x}^{(0)} \in \mathbb{R}^n \text{ donné,} \\ &\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha \boldsymbol{w}^{(k)}, \text{ pour } k \geq 0, \end{aligned}$$

où $\boldsymbol{w}^{(k)}$ est une **direction de descente** en $\boldsymbol{x}^{(k)}$.
Dans toute la suite, on considère la fonction f de $\mathbb{R}^n$ dans $\mathbb{R}$ définie par

$$f(\boldsymbol{x}) = \frac{1}{2} A\boldsymbol{x} \cdot \boldsymbol{x} - \boldsymbol{b} \cdot \boldsymbol{x}, \tag{3.37}$$

où A une matrice carrée d'ordre n, symétrique définie positive, et $\boldsymbol{b} \in \mathbb{R}^n$. On pose $\bar{x} = A^{-1}b$.

1. Montrer que la méthode de Jacobi pour la résolution du système $A\boldsymbol{x} = \boldsymbol{b}$ peut s'écrire comme une méthode de descente à pas fixe pour la minimisation de la fonction f définie par (3.37). Donner l'expression du pas α et de la direction de descente $\boldsymbol{w}^{(k)}$ à chaque itération k et vérifier que c'est bien une direction de descente stricte si $\boldsymbol{x}^{(k)} \neq A^{-1}\boldsymbol{b}$.
2. On cherche maintenant à améliorer la méthode de Jacobi en prenant non plus un pas fixe dans l'algorithme de descente ci-dessus, mais un pas optimal qui est défini à l'itération k par

$$f(\boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}) = \min_{\alpha>0} f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{w}^{(k)}), \tag{3.38}$$

   où $\boldsymbol{w}^{(k)}$ est défini à la question précédente. On définit alors une méthode de descente à pas optimal par :

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}.$$

   On appelle cette nouvelle méthode "méthode de Jacobi à pas optimal".
   (a) Justifier l'existence et l'unicité du pas optimal défini par (3.38), et donner son expression à chaque itération.
   (b) Montrer que $|f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^{(k+1)})| = \dfrac{|\boldsymbol{r}^{(k)} \cdot \boldsymbol{w}^{(k)}|^2}{2A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)}}$ si $w^{(k)} \neq 0$.

(c) Montrer que $\boldsymbol{r}^{(k)} \to \mathbf{0}$ lorsque $k \to +\infty$, et en déduire que la suite donnée par la méthode de Jacobi à pas optimal converge vers la solution $\bar{\boldsymbol{x}}$ du système linéaire $A\boldsymbol{x} = \boldsymbol{b}$.

(d) On suppose que la diagonale extraite D de la matrice A (qui est symétrique définie positive) est de la forme $D = \alpha \mathrm{Id}$ avec $\alpha \in \mathbb{R}$.

i. Ecrire l'algorithme de descente à pas optimal dans ce cas.

ii. Comparer les algorithmes de descente obtenus par Jacobi et Jacobi à pas optimal avec les algorithmes de gradient que vous connaissez.

Exercice 120 (Fonction non croissante à l'infini)**.** *Suggestions en page 242.*

Soient $n \geq 1$, $f \in C^2(\mathbb{R}^n, \mathbb{R})$ et $a \in \mathbb{R}$. On suppose que $A = \{x \in \mathbb{R}^n;\ f(x) \leq f(a)\}$ est un ensemble borné de $\mathbb{R}^n$ et qu'il existe $M \in \mathbb{R}$ t.q. $|H(x)y \cdot y| \leq M|y|^2$ pour tout $x, y \in \mathbb{R}^n$ (où $H(x)$ désigne la matrice hessienne de f au point x).

1. Montrer qu'il existe $\overline{x} \in A$ t.q. $f(\overline{x}) = \min\{f(x),\ x \in \mathbb{R}^n\}$ (noter qu'il n'y a pas nécessairement unicité de $\overline{x}$).
2. Soit $x \in A$ t.q. $\nabla f(x) \neq 0$. On pose $T(x) = \sup\{\alpha \geq 0;\ [x, x - \alpha\nabla f(x)] \subset A\}$. Montrer que $0 < T(x) < +\infty$ et que $[x, x - T(x)\nabla f(x)] \subset A\}$ (où $[x, x - T(x)\nabla f(x)]$ désigne l'ensemble $\{tx + (1-t)(x - T(x)\nabla f(x)), t \in [0,1]\}$.
3. Pour calculer une valeur appochée de $\overline{x}$ (t.q. $f(\overline{x}) = \min\{f(x), x \in \mathbb{R}^n\}$), on propose l'algorithme suivant :
   Initialisation : $x_0 \in A$,

   Itérations : Soit $k \geq 0$.
   Si $\nabla f(x_k) = 0$, on pose $x_{k+1} = x_k$. Si $\nabla f(x_k) \neq 0$, on choisit $\alpha_k \in [0, T(x_k)]$ t.q. $f(x_k - \alpha_k \nabla f(x_k)) = \min\{f(x_k - \alpha\nabla f(x_k)),\ 0 \leq \alpha \leq T(x_k)\}$ (La fonction T est définie à la question 2) et on pose $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$.

   (a) Montrer que, pour tout $x_0 \in A$, l'algorithme précédent définit une suite $(x_k)_{k\in\mathbb{N}} \subset A$ (c'est–à–dire que, pour $x_k \in A$, il existe bien au moins un élément de $[0, T(x_k)]$, noté α_k, t.q. $f(x_k - \alpha_k \nabla f(x_k) = \min\{f(x_k - \alpha\nabla f(x_k)),\ 0 \leq \alpha \leq T(x_k)\}$).

   (b) Montrer que cet algorithme n'est pas nécessairement l'algorithme du gradient à pas optimal. [on pourra chercher un exemple avec $n = 1$.]

   (c) Montrer que $f(x_k) - f(x_{k+1}) \geq \frac{|\nabla f(x_k)|^2}{2M}$, pour tout $k \in \mathbb{N}$.
4. On montre maintenant la convergence de la suite $(x_k)_{k\in\mathbb{N}}$ construite à la question précédente.
   (a) Montrer qu'il existe une sous suite $(x_{k_\ell})_{\ell\in\mathbb{N}}$ et $x \in A$ t.q. $x_{k_\ell} \to x$, quand $\ell \to \infty$, et $\nabla f(x) = 0$.
   (b) On suppose, dans cette question, qu'il existe un et un seul élément $z \in A$ t.q. $\nabla f(z) = 0$. Montrer que $x_k \to z$, quand $k \to \infty$, et que $f(z) = \min\{f(x), x \in A\}$.

Exercice 121 (Application du GPO)**.** *Corrigé détaillé en page 245*

Soit $A \in \mathcal{M}_n(\mathbb{R})$ et J la fonction définie de $\mathbb{R}^n$ dans $\mathbb{R}$ par $J(x) = e^{\|Ax\|^2}$, où $\|\cdot\|$ désigne la norme euclidienne sur $\mathbb{R}^n$.

1. Montrer que J admet un minimum (on pourra le calculer...).
2. On suppose que la matrice A est inversible, montrer que ce minimum est unique.
3. Ecrire l'algorithme du gradient à pas optimal pour la recherche de ce minimum. [On demande de calculer le paramètre optimal α_k en fonction de A et de x_k.] A quelle condition suffisante cet algorithme converge-t-il ?

Exercice 122 (Méthode de relaxation)**.** *Corrigé détaillé en page 246*

Soit f une fonction continûment différentiable de $E = \mathbb{R}^n$ dans $\mathbb{R}$ vérifiant l'hypothèse (3.10) :
1. Justifier l'existence et l'unicité de $\overline{x} \in \mathbb{R}^n$ tel que $f(\overline{x}) = \inf_{x\in\mathbb{R}^n} f(x)$.

On propose l'algorithme de recherche de minimum de f suivant :

$$
\left\{
\begin{array}{ll}
\text{Initialisation :} & x^{(0)} \in E, \\
\text{Itération } n : & x^{(k)} \text{ connu, } (n \geq 0) \\
& \text{Calculer } x_1^{(k+1)} \text{ tel que, pour tout } \xi \in \mathbb{R}, \\
& f(x_1^{(k+1)}, x_2^{(k)}, x_3^{(k)}, \ldots, x_n^{(k)}) \leq f(\xi, x_2^{(k)}, x_3^{(k)}, \ldots, x_n^{(k)}), \\
& \text{Calculer } x_2^{(k+1)} \text{ tel que, pour tout } \xi \in \mathbb{R}, \\
& f(x_1^{(k+1)}, x_2^{(k+1)}, x_3^{(k)}, \ldots, x_n^{(k)}) \leq f(x_1^{(k+1)}, \xi, x_3^{(k)}, \ldots, x_n^{(k)}), \\
& \qquad \ldots \\
& \text{Calculer } x_k^{(k+1)} \text{ tel que, pour tout } \xi \in \mathbb{R}, \\
& f(x_1^{(k+1)}, \ldots, x_{k-1}^{(k+1)}, x_k^{(k+1)}, x_{(k+1)}^{(k)}, \ldots, x_n^{(k)}) \\
& \qquad \leq f(x_1^{(k+1)}, \ldots, x_{k-1}^{(k+1)}, \xi, x_{(k+1)}^{(k)}, \ldots, x_n^{(k)}), \\
& \qquad \ldots \\
& \text{Calculer } x_n^{(k+1)} \text{ tel que, pour tout } \xi \in \mathbb{R}, \\
& f(x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_{n-1}^{(k+1)}, x_n^{(k+1)}) \leq f(x_1^{(k+1)}, \ldots, x_{n-1}^{(k+1)}, \xi).
\end{array}
\right. \tag{3.39}
$$

2. Pour $n \in \mathbb{N}$ et $1 \leq k \leq N$, soit $\varphi_k^{(k+1)}$ la fonction de $\mathbb{R}$ dans $\mathbb{R}$ définie par :

$$\varphi_k^{(k+1)}(s) = f(x_1^{(k+1)}, \ldots, x_{k-1}^{(k+1)}, s, x_{(k+1)}^{(k)}, \ldots, x_n^{(k)}).$$

Montrer qu'il existe un unique élément $\overline{s} \in \mathbb{R}$ tel que

$$\varphi_k^{(k+1)}(\overline{s}) = \inf_{s\in\mathbb{R}} \varphi_k^{(k+1)}(s).$$

En déduire que la suite $(x^{(k)})_{n\in\mathbb{N}}$ construite par l'algorithme (3.39) est bien définie.

Dans toute la suite, on note $\|\cdot\|$ la norme euclidienne sur $\mathbb{R}^n$ et $(\cdot|\cdot)$ le produit scalaire associé. Pour $i = 1, \ldots, n$, on désigne par $\partial_i f$ la dérivée partielle de f par rapport à la i-ème variable.

3. Soit $(x^{(k)})_{n\in\mathbb{N}}$ la suite définie par l'algorithme (3.39).
Pour $n \geq 0$, on définit $x^{(n+1,0)} = x^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})^t$, et pour $1 \leq k \leq n$,

$$x^{(n+1,k)} = (x_1^{(k+1)}, \ldots, x_k^{(k+1)}, x_{k+1}^{(k)}, \ldots, x_n^{(k)})^t$$

(de sorte que $x^{(n+1,n)} = x^{(k+1)}$).

(a) Soit $n \in \mathbb{N}$. Pour $1 \leq k \leq n$, montrer que $\partial_k f(x^{(n+1,k)}) = 0$, pour $k = 1, \ldots, n$. En déduire que

$$f(x^{(n+1,k-1)}) - f(x^{(n+1,k)}) \geq \frac{\alpha}{2}\|x^{(n+1,k-1)} - x^{(n+1,k)}\|^2.$$

(b) Montrer que la suite $(x^{(k)})_{n\in\mathbb{N}}$ vérifie

$$f(x^{(k)}) - f(x^{(k+1)}) \geq \frac{\alpha}{2}\|x^{(k)} - x^{(k+1)}\|^2.$$

En déduire que $\lim_{n\to+\infty} \|x^{(k)} - x^{(k+1)}\| = 0$ et que, pour $1 \leq k \leq n$, $\lim_{n\to+\infty} \|x^{(n+1,k)} - x^{(k+1)}\| = 0$.

4. Montrer que

$$\|x^{(k+1)} - \overline{x}\| \leq \frac{1}{\alpha} \left(\sum_{k=1}^{n} |\partial_k f(x^{(k+1)})|^2 \right)^{\frac{1}{2}}.$$

5. Montrer que les suites $(x^{(k)})_{n\in\mathbb{N}}$, et $(x^{(n+1,k)})_{n\in\mathbb{N}}$, pour $k = 1, \ldots, n$, sont bornées.
Montrer que

$$|\partial_k f(x^{(k+1)})| \to 0 \text{ lorsque } n \to +\infty.$$

(On rappelle que $\partial_k f(x^{(n+1,k)}) = 0$.)
Conclure quant à la convergence de la suite$(x^{(k)})_{n\in\mathbb{N}}$ lorsque $n \to +\infty$.

6. On suppose dans cette question que $f(x) = \frac{1}{2}(Ax|x) - (b|x)$. Montrer que dans ce cas, l'algorithme (3.39) est équivalent à une méthode itérative de résolution de systèmes linéaires qu'on identifiera.

7. On suppose dans cette question que $n = 2$. Soit g la fonction définie de $\mathbb{R}^2$ dans $\mathbb{R}$ par : $g(x) = x_1^2 + x_2^2 - 2(x_1 + x_2) + 2|x_1 - x_2|$, avec $x = (x_1, x_2)^t$.

(a) Montrer qu'il existe un unique élément $\overline{x} = (\overline{x}_1, \overline{x}_2)^t$ de $\mathbb{R}^2$ tel que $g(\overline{x}) = \inf_{x\in\mathbb{R}^2} g(x)$.

(b) Montrer que $\overline{x} = (1, 1)^t$.

(c) Montrer que si $x^{(0)} = (0, 0)^t$, l'algorithme (3.39) appliqué à g ne converge pas vers $\overline{x}$. Quelle est l'hypothèse mise en défaut ici ?

Exercice 123 (Mise en oeuvre de GC)**.**

On considère la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par $f(x_1, x_2) = 2x_1^2 + x_2^2 - x_1x_2 - 3x_1 - x_2 + 4$.

1. Montrer qu'il existe un unique $\bar{x} \in \mathbb{R}^2$ tel que $\bar{x} = \min_{x\in\mathbb{R}^2} f(x)$ admet un unique minimum, et le calculer.
2. Calculer le premier itéré donné par l'algorithme du gradient conjugué, en partant de $(x_1^{(0)}, x_2^{(0)}) = (0, 0)$, pour un pas de $\alpha = .5$ dans le cas de GPF.

Exercice 124 (Gradient conjugué pour une matrice non symétrique)**.** *Corrigé détaillé en page 248*

Soit $n \in \mathbb{N}$, $n \geq 1$. On désigne par $\|\cdot\|$ la norme euclidienne sur $\mathbb{R}^n$, et on munit l'ensemble $\mathcal{M}_n(\mathbb{R})$ de la norme induite par la norme $\|\cdot\|$, $\|\cdot\|$. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. On définit $M \in \mathcal{M}_n(\mathbb{R})$ par $M = A^tA$. On se donne un vecteur $b \in \mathbb{R}^n$, et on s'intéresse à la résolution du système linéaire

$$Ax = b;. \tag{3.40}$$

1. Montrer que $x \in \mathbb{R}^n$ est solution de (1.126) si et seulement si x est solution de

$$Mx = A^tb;. \tag{3.41}$$

2. On rappelle que le conditionnement d'une matrice $C \in \mathcal{M}_n(\mathbb{R})$ inversible est défini par $\text{cond}(C) = \|C\|\|C^{-1}\|$ (et dépend donc de la norme considérée ; on rappelle qu'on a choisi ici la norme induite par la norme euclidienne).
   (a) Montrer que les valeurs propres de la matrice M sont toutes strictement positives.
   (b) Montrer que $\text{cond}(A) = \sqrt{\frac{\lambda_n}{\lambda_1}}$, où λ_n (resp. λ_1) est la plus grande (resp. plus petite) valeur propre de M.
3. Ecrire l'algorithme du gradient conjugué pour la résolution du système (3.41), en ne faisant intervenir que les matrices A et A^t (et pas la matrice M) et en essayant de minimiser le nombre de calculs. Donner une estimation du nombre d'opérations nécessaires et comparer par rapport à l'algorithme du gradient conjugué écrit dans le cas d'une matrice carré d'ordre n symétrique définie positive.

Exercice 125 (Gradient conjugué préconditionné par LL^t)**.** *Corrigé en page 249*

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique définie positive, et $b \in \mathbb{R}^n$. Soit L une matrice triangulaire inférieure inversible, soit $B = L^{-1}A(L^t)^{-1}$ et $\tilde{b} = L^{-1}b$.

1. Montrer que B est symétrique définie positive.
2. Justifier l'existence et l'unicité de $x \in \mathbb{R}^n$ tel que $Ax = b$, et de $y \in \mathbb{R}^n$ tel que $By = \tilde{b}$. Ecrire x en fonction de y.

Soit $y^{(0)} \in \mathbb{R}^n$ fixé. On pose $\tilde{r}^{(0)} = \tilde{w}^{(0)} = \tilde{b} - By^{(0)}$. Si $\tilde{r}^{(0)} \neq 0$, on pose alors $y^{(1)} = y^{(0)} + \rho_0 \tilde{w}^{(0)}$, avec $\rho_0 = \frac{\tilde{r}^{(0)} \cdot \tilde{r}^{(0)}}{\tilde{w}^{(0)} \cdot A\tilde{w}^{(0)}}$.
Pour $n > 1$, on suppose $y^{(0)}, \ldots, y^{(k)}$ et $\tilde{w}^{(0)}, \ldots, \tilde{w}^{(k-1)}$ connus, et on pose : $\tilde{\boldsymbol{r}}^{(k)} = \tilde{b} - By^{(k)}$. Si $\tilde{\boldsymbol{r}}^{(k)} \neq 0$, on calcule : $\tilde{w}^{(k)} = \tilde{\boldsymbol{r}}^{(k)} + \lambda_{k-1}\tilde{w}^{(k-1)}$ avec $\lambda_{k-1} = \frac{\tilde{\boldsymbol{r}}^{(k)} \cdot \tilde{\boldsymbol{r}}^{(k)}}{\tilde{r}^{(k-1)} \cdot \tilde{r}^{(k-1)}}$ et on pose alors : $y^{(k+1)} = y^{(k)} + \alpha_k \tilde{w}^{(k)}$ avec $\alpha_k = \frac{\tilde{\boldsymbol{r}}^{(k)} \cdot \tilde{\boldsymbol{r}}^{(k)}}{\tilde{w}^{(k)} \cdot B\tilde{w}^{(k)}}$,

3. En utilisant le cours, justifier que la famille $y^{(k)}$ ainsi construite est finie. A quoi est égale sa dernière valeur ?

Pour $n \in \mathbb{N}$, on pose : $x^{(k)} = L^{-t}y^{(k)}$ (avec $L^{-t} = (L^{-1})^t = (L^t)^{-1}$), $\boldsymbol{r}^{(k)} = b - Ax^{(k)}$, $w^{(k)} = L^{-t}\tilde{w}^{(k)}$ et $s^{(k)} = (LL^t)^{-1}\boldsymbol{r}^{(k)}$.

4. Soit $n > 0$ fixé. Montrer que :

$$(a) \quad \lambda_{k-1} = \frac{s^{(k)} \cdot \boldsymbol{r}^{(k)}}{s^{(k-1)} \cdot r^{(k-1)}}, \qquad (b) \quad \rho_n = \frac{s^{(k)} \cdot \boldsymbol{r}^{(k)}}{w^{(k)} \cdot Aw^{(k)}},$$

$$(c) \quad w^{(k)} = s^{(k)} + \lambda_n w^{(k-1)}, \qquad (d) \quad x^{(k+1)} = x^{(k)} + \alpha_k w^{(k)}.$$

5. On suppose que la matrice LL^t est une factorisation de Choleski incomplète de la matrice A. Ecrire l'algorithme du gradient conjugué préconditionné par cette factorisation, pour la résolution du système $Ax = b$.

Exercice 126 (Méthode de quasi-linéarisation)**.** *Corrigé détaillé en page 251*
Soit $f \in C^3(\mathbb{R}, \mathbb{R})$ une fonction croissante à l'infini, c. à.d. telle que $f(x) \to +\infty$ lorsque $|x| \to +\infty$; soit $d \in \mathbb{R}$ et soit J la fonction de $\mathbb{R}$ dans $\mathbb{R}$ par

$$J(x) = (f(x) - d)^2.$$

1. Montrer qu'il existe $\bar{x} \in \mathbb{R}$ tel que $J(\bar{x}) \leq J(x), \forall x \in \mathbb{R}$.
2. (Newton) On cherche ici à déterminer un minimum de J en appliquant la méthode de Newton pour trouver une solution de l'équation $J'(x) = 0$. Ecrire l'algorithme de Newton qui donne x_{k+1} en fonction de x_k et des données d, f, f' et f''.
3. L'algorithme dit de "quasi–linéarisation" consiste à remplacer, à chaque itération $k \in \mathbb{N}$, la minimisation de la fonctionnelle J par celle de la fonctionnelle J_k, définie de $\mathbb{R}$ dans $\mathbb{R}$ obtenue à partir de J en effectuant un développement limité au premier ordre de $f(x)$ en $x^{(k)}$, c.à.d.

$$J_k(x) = \left(f(x_k) + f'(x_k)(x - x_k) - d\right)^2$$

   Montrer que à k fixé, il existe un unique $\bar{x} \in \mathbb{R}$ qui minimise J_k et le calculer (on supposera que $f'(x_k) \neq 0$). On pose donc $x_{k+1} = \bar{x}$. Que vous rappelle l'expression de x_{k+1} ?
4. Ecrire l'algorithme du gradient à pas fixe pour la minimisation de J.
5. Dans cette question, on prend $f(x) = x^2$.
   (a) Donner l'ensemble des valeurs $\bar{x} \in \mathbb{R}$ qui minimisent J, selon la valeur de d. Y a t-il unicité de $\bar{x}$?
   (b) Montrer que quelque soit la valeur de d, l'algorithme de Newton converge si le choix initial x_0 est suffisamment proche de $\bar{x}$.
   (c) On suppose que $d > 0$; montrer que l'algorithme de quasi-linéarisation converge pour un choix initial x_0 dans un voisinage de 1.

(d) On suppose maintenant que $d = -1$. Montrer que l'algorithme de quasi-linéarisation ne converge que pour un ensemble dénombrable de choix initiaux x_0.

Exercice 127 (Méthode de Polak-Ribière)**.** *Suggestions en page 242, corrigé en page 252*

Dans cet exercice, on démontre la convergence de la méthode de Polak-Ribière (méthode de gradient conjugué pour une fonctionnelle non quadratique) sous des hypothèses "simples" sur f.

Soit $f \in C^2(\mathbb{R}^n, \mathbb{R})$. On suppose qu'il existe $\alpha > 0$, $\beta \geq \alpha$ tel que $\alpha|y|^2 \leq H(x)y \cdot y \leq \beta|y|^2$ pour tout x, $y \in \mathbb{R}^n$. ($H(x)$ est la matrice hessienne de f au point x.)

1. Montrer que f est strictement convexe, que $f(x) \to \infty$ quand $|x| \to \infty$ et que le spectre $\mathcal{VP}(H(x))$ de $H(x)$ est inclus dans $[\alpha, \beta]$ pour tout $x \in \mathbb{R}^n$.

   On note $\overline{x}$ l'unique point de $\mathbb{R}^n$ t.q. $f(\overline{x}) \leq f(x)$ pour tout $x \in \mathbb{R}^n$ (l'existence et l'unicité de $\overline{x}$ est donné par la question précédente). On cherche une approximation de $\overline{x}$ en utilisant l'algorithme de Polak-Ribière :

   initialisation. $x^{(0)} \in \mathbb{R}^n$. On pose $g^{(0)} = -\nabla f(x^{(0)})$. Si $g^{(0)} = 0$, l'algorithme s'arrête (on a $x^{(0)} = \overline{x}$). Si $g^{(0)} \neq 0$, on pose $w^{(0)} = g^{(0)}$ et $x^{(1)} = x^{(0)} + \rho_0 w^{(0)}$ avec ρ_0 "optimal" dans la direction $w^{(0)}$.

   itération. $x^{(k)}$, $w^{(k-1)}$ connus ($k \geq 1$). On pose $g^{(k)} = -\nabla f(x^{(k)})$. Si $g^{(k)} = 0$, l'algorithme s'arrête (on a $x^{(k)} = \overline{x}$). Si $g^{(k)} \neq 0$, on pose $\lambda_{k-1} = [g^{(k)} \cdot (g^{(k)} - g^{(k-1)})]/[g^{(k-1)} \cdot g^{(k-1)}]$, $w^{(k)} = g^{(k)} + \lambda_{k-1} w^{(k-1)}$ et $x^{(k+1)} = x^{(k)} + \alpha_k w^{(k)}$ avec α_k "optimal" dans la direction w_k. (Noter que α_k existe bien.)

   On suppose dans la suite que $g^{(k)} \neq 0$ pour tout $k \in \mathbb{N}$.

2. Montrer (par récurrence sur k) que $g^{(k+1)} \cdot w^{(k)} = 0$ et $g^{(k)} \cdot g^{(k)} = g^{(k)} \cdot w^{(k)}$, pour tout $k \in \mathbb{N}$.

3. On pose
$$J^{(k)} = \int_0^1 H(x^{(k)} + \theta \alpha_k w^{(k)}) d\theta.$$
   Montrer que $g^{(k+1)} = g^{(k)} + \alpha_k J^{(k)} w^{(k)}$ et que $\alpha_k = (-g^{(k)} \cdot w^{(k)})/(J^{(k)} w^{(k)} \cdot w^{(k)})$ (pour tout $k \in \mathbb{N}$).

4. Montrer que $|w^{(k)}| \leq (1 + \beta/\alpha)|g^{(k)}|$ pour tout $k \in \mathbb{N}$. [Utiliser, pour $k \geq 1$, la question précédente et la formule donnant λ_{k-1}.]

5. Montrer que $x^{(k)} \to \overline{x}$ quand $k \to \infty$.

Exercice 128 (Algorithme de quasi Newton)**.**
Corrigé détaillé en page 254

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique définie positive et $b \in \mathbb{R}^n$. On pose $f(x) = (1/2)Ax \cdot x - b \cdot x$ pour $x \in \mathbb{R}^n$. On rappelle que $\nabla f(x) = Ax - b$. Pour calculer $\overline{x} \in \mathbb{R}^n$ t.q. $f(\overline{x}) \leq f(x)$ pour tout $x \in \mathbb{R}^n$, on va utiliser un algorithme de quasi Newton, c'est-à-dire :

initialisation. $x^{(0)} \in \mathbb{R}^n$.

itération. $x^{(k)}$ connu ($n \geq 0$. On pose $x^{(k+1)} = x^{(k)} - \alpha_k K^{(k)} g^{(k)}$ avec $g^{(k)} = \nabla f(x^{(k)})$, $K^{(k)}$ une matrice symétrique définie positive à déterminer et α_k "optimal" dans la direction $w^{(k)} = -K^{(k)} g^{(k)}$. (Noter que α_k existe bien.)

Partie 1. Calcul de α_k. On suppose que $g^{(k)} \neq 0$.

1. Montrer que $w^{(k)}$ est une direction de descente stricte en $x^{(k)}$ et calculer la valeur de α_k (en fonction de $K^{(k)}$ et $g^{(k)}$).

2. On suppose que, pour un certain $n \in \mathbb{N}$, on a $K^{(k)} = (H(x^{(k)}))^{-1}$ (où $H(x)$ est la matrice hessienne de f en x, on a donc ici $H(x) = A$ pour tout $x \in \mathbb{R}^n$). Montrer que $\alpha_k = 1$.

3. Montrer que la méthode de Newton pour calculer $\overline{x}$ converge en une itération (mais nécessite la résolution du système linéaire $A(x^{(1)} - x^{(0)}) = b - Ax^{(0)} \ldots$)

Partie 2. Méthode de Fletcher-Powell. On prend maintenant $K^{(0)} = Id$ et

$$K^{(k+1)} = K^{(k)} + \frac{s^{(k)}(s^{(k)})^t}{s^{(k)} \cdot y^{(k)}} - \frac{(K^{(k)}y^{(k)})(K^{(k)}(y^{(k)})^t}{K^{(k)}y^{(k)} \cdot y^{(k)}},\ n \geq 0, \tag{3.42}$$

avec $s^{(k)} = x^{(k+1)} - x^{(k)}$ et $y^{(k)} = g^{(k+1)} - g^{(k)} = As^{(k)}$.

On va montrer que cet algorithme converge en au plus n itérations (c'est-à-dire qu'il existe $n \leq n+1$ t.q. $x_{N+1} = \overline{x}$.)

1. Soit $n \in \mathbb{N}$. On suppose, dans cette question, que $s^{(0)}, \ldots, s^{(k-1)}$ sont des vecteurs A-conjugués et non-nuls et que $K^{(0)}, \ldots, K^{(k)}$ sont des matrices symétriques définies positives t.q. $K^{(j)}As^{(i)} = s^{(i)}$ si $0 \leq i < j \leq n$ (pour $n = 0$ on demande seulement $K^{(0)}$ symétrique définie positive).

   (a) On suppose que $g^{(k)} \neq 0$. Montrer que $s^{(k)} \neq 0$ (cf. Partie I) et que, pour $i < n$,

   $$s^{(k)} \cdot As^{(i)} = 0 \Leftrightarrow g^{(k)} \cdot s^{(i)} = 0.$$

   Montrer que $g^{(k)} \cdot s^{(i)} = 0$ pour $i < n$. [On pourra remarquer que $g^{(i+1)} \cdot s^{(i)} = g^{(i+1)} \cdot w^{(i)} = 0$ et $(g^{(k)} - g^{(i+1)}) \cdot s^{(i)} = 0$ par l'hypothèse de conjugaison de $s^{(0)}, \ldots, s^{(k-1)}$.] En déduire que $s^{(0)}, \ldots, s^{(k)}$ sont des vecteurs A-conjugués et non-nuls.

   (b) Montrer que $K^{(k+1)}$ est symétrique.

   (c) Montrer que $K^{(k+1)}As^{(i)} = s^{(i)}$ si $0 \leq i \leq n$.

   (d) Montrer que, pour tout $x \in \mathbb{R}^n$, on a

   $$K^{(k+1)}x \cdot x = \frac{(K^{(k)}x \cdot x)(K^{(k)}y^{(k)} \cdot y^{(k)}) - (K^{(k)}y^{(k)} \cdot x)^2}{K^{(k)}y^{(k)} \cdot y^{(k)}} + \frac{(s^{(k)} \cdot x)^2}{As^{(k)} \cdot s^{(k)}}.$$

   En déduire que $K^{(k+1)}$ est symétrique définie positive. [On rappelle (inégalité de Cauchy-Schwarz) que, si K est symétrique définie positive, on a $(Kx \cdot y)^2 \leq (Kx \cdot x)(Ky \cdot y)$ et l'égalité a lieu si et seulement si x et y sont colinéaires.]

2. On suppose que $g^{(k)} \neq 0$ si $0 \leq n \leq n-1$. Montrer (par récurrence sur n, avec la question précédente) que $s^{(0)}, \ldots, s^{(n-1)}$ sont des vecteurs A-conjugués et non-nuls et que $K^{(n)}As^{(i)} = s^{(i)}$ si $i < n$. En déduire que $K^{(n)} = A^{-1}$, $\alpha_n = 1$ et $x^{(n+1)} = A^{-1}b = \overline{x}$.

Exercice 129 (Méthodes de Gauss–Newton et de quasi–linéarisation)**.** *Corrigé en page 257*

Soit $f \in C^2(\mathbb{R}^n, \mathbb{R}^p)$, avec $n, p \in \mathbb{N}^*$. Soit $C \in \mathcal{M}_p(\mathbb{R})$ une matrice réelle carrée d'ordre p, symétrique définie positive, et $\boldsymbol{d} \in \mathbb{R}^p$. Pour $\boldsymbol{x} \in \mathbb{R}^n$, on pose

$$J(\boldsymbol{x}) = (f(\boldsymbol{x}) - \boldsymbol{d}) \cdot C(f(\boldsymbol{x}) - \boldsymbol{d}).$$

On cherche à minimiser J.

I *Propriétés d'existence et d'unicité*

(a) Montrer que J est bornée inférieurement.

(b) Donner trois exemples de fonctions f pour lesquels les fonctionnelles J associées sont telles que l'on ait :

i. existence et unicité de $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ qui réalise le minimum de J, pour le premier exemple.

ii. existence et non unicité de $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ qui réalise le minimum de J, pour le second exemple.

iii. non existence de $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ qui réalise le minimum de J, pour le troisième exemple.

(On pourra prendre $n = p = 1$.)

II *Un peu de calcul différentiel*

(a) On note Df et D_2f les différentielles d'ordre 1 et 2 de f. A quels espaces appartiennent $Df(\boldsymbol{x})$, $D_2f(\boldsymbol{x})$ (pour $\boldsymbol{x} \in \mathbb{R}^n$), ainsi que Df et D_2f ? Montrer que pour tout $\boldsymbol{x} \in \mathbb{R}^n$, il existe $M(\boldsymbol{x}) \in \mathcal{M}_{p,n}(\mathbb{R})$, où $\mathcal{M}_{p,n}(\mathbb{R})$ désigne l'ensemble des matrices réelles à p lignes et n colonnes, telle que $Df(\boldsymbol{x})(\boldsymbol{y}) = M(\boldsymbol{x})\boldsymbol{y}$ pour tout $\boldsymbol{y} \in \mathbb{R}^n$.

(b) Pour $\boldsymbol{x} \in \mathbb{R}^n$, calculer $\nabla J(\boldsymbol{x})$.

(c) Pour $\boldsymbol{x} \in \mathbb{R}^n$, calculer la matrice hessienne de J en $\boldsymbol{x}$ (qu'on notera $H(\boldsymbol{x})$). On suppose maintenant que M ne dépend pas de $\boldsymbol{x}$; montrer que dans ce cas $H(\boldsymbol{x}) = 2M(\boldsymbol{x})^t CM(\boldsymbol{x})$.

III *Algorithmes d'optimisation* Dans toute cette question, on suppose qu'il existe un unique $\bar{b}fx \in \mathbb{R}^n$ qui réalise le minimum de J, qu'on cherche à calculer de manière itérative. On se donne pour cela $x_0 \in \mathbb{R}^n$, et on cherche à construire une suite $(x_k)_{k\in\mathbb{N}}$ qui converge vers $\bar{x}$.

(a) On cherche à calculer $\bar{x}$ en utilisant la méthode de Newton pour annuler ∇J. Justifier brièvement cette procédure et écrire l'algorithme obtenu.

(b) L'algorithme dit de "Gauss-Newton" est une modification de la méthode précédente, qui consiste à approcher, à chaque itération n, la matrice jacobienne de ∇J en x_k par la matrice obtenue en négligeant les dérivées secondes de f. Ecrire l'algorithme ainsi obtenu.

(c) L'algorithme dit de "quasi–linéarisation" consiste à remplacer, à chaque itération $k \in \mathbb{N}$, la minimisation de la fonctionnelle J par celle de la fonctionnelle J_k, définie de $\mathbb{R}^n$ dans $\mathbb{R}$, et obtenue à partir de J en effectuant un développement limité au premier ordre de $f(\boldsymbol{x})$ en $\boldsymbol{x}^{(k)}$, c.à.d.

$$J_k(\boldsymbol{x}) = (f(\boldsymbol{x}^{(k)}) + Df(\boldsymbol{x}^{(k)})(\boldsymbol{x} - \boldsymbol{x}^{(k)}) - d) \cdot C(f(\boldsymbol{x}^{(k)}) + Df(\boldsymbol{x}^{(k)})(\boldsymbol{x} - \boldsymbol{x}^{(k)}) - \boldsymbol{d}).$$

i. Soit $k \geq 0$, $\boldsymbol{x}^{(k)} \in \mathbb{R}^n$ connu, $M_k = M(\boldsymbol{x}^{(k)}) \in \mathcal{M}_{p,n}(\mathbb{R})$, et $\boldsymbol{x} \in \mathbb{R}^n$. On pose $\boldsymbol{h} = \boldsymbol{x} - \boldsymbol{x}^{(k)}$. Montrer que

$$J_k(\boldsymbol{x}) = J(\boldsymbol{x}^{(k)}) + M_k^t CM_k\boldsymbol{h} \cdot \boldsymbol{h} + 2M_k^t C(f(\boldsymbol{x}^{(k)}) - \boldsymbol{d}) \cdot \boldsymbol{h}.$$

ii. Montrer que la recherche du minimum de J_k est équivalente à la résolution d'un système linéaire dont on donnera l'expression.

iii. Ecrire l'algorithme de quasi–linéarisation, et le comparer avec l'algorithme de Gauss-Newton.

Suggestions pour les exercices

Exercice 118 page 233 (Algorithme du gradient à pas optimal)

2. Utiliser le fait que H est continue.

3. Etudier la fonction $\varphi : \mathbb{R}_+$ dans $\mathbb{R}$ définie par $\varphi(\rho) = f(x_k + \rho\boldsymbol{w}^{(k)})$.

4. a. Montrer que f est minorée et remarquer que la suite $(f(x_k))_{k\in\mathbb{N}}$ est décroissante.
4.b se déduit du 4.a
4.c. Utiliser la fonction φ définie plus haut, la question 4.b. et la question 2.
4.d. Utiliser le fait que le choix de α_k est optimal et le résultat de 4.c.
4.e. Etudier le polynôme du 2nd degré en ρ défini par : $P_k(\rho) = f(x_k) - \rho|\boldsymbol{w}^{(k)}|^2 + \frac{1}{2}M|\boldsymbol{w}^{(k)}|^2\rho^2$ dans les cas où $|\boldsymbol{w}^{(k)}| \leq M$ (fait la quesiton 4.c) puis dans le cas $|\boldsymbol{w}^{(k)}| \geq M$.

5. utiliser l'inégalité prouvée en 4.e. pour montrer que $|\boldsymbol{w}^{(k)}| \to 0$ lorsque $n \to +\infty$.

6. Pour montrer que toute la suite converge, utiliser l'argument d'unicité de la limite, en raisonnant par l'absurde (supposer que la suite ne converge pas et aboutir à une contradiction).

Exercice 120 page 235 (Cas où f n'est pas croissante à l'infini)

S'inspirer des techniques utilisées lors des démonstrations de la proposition 3.13 et du théorème 3.19 (il faut impérativement les avoir fait avant...).

Exercice 127 page 239 (Méthode de Polak-Ribière)

1. Utiliser la deuxième caractérisation de la convexité. Pour montrer le comportement à l'infini, introduire la fonction φ habituelle... ($\varphi(t) = f(x + ty)$).
2. Pour montrer la convergence, utiliser le fait que si $w_k \cdot \nabla f(x_k) < 0$ alors w_k est une direction de descente stricte de f en x_k, et que si α_k est optimal alors $\nabla f(x_k + \alpha_k \boldsymbol{w}^{(k)}) = 0$.
3. Utiliser la fonction φ définie par $\varphi(\theta) = \nabla f(x_k + \theta\alpha_k \boldsymbol{w}^{(k)})$.
4. C'est du calcul...
5. Montrer d'abord que $-g_k \boldsymbol{w}^{(k)} \leq -\gamma |\boldsymbol{w}^{(k)}||g_k|$. Montrer ensuite (en utilisant la bonne vieille fonction φ définie par $\varphi(t) = f(x_k + t\alpha_k)$, que $g_k \to 0$ lorsque $n \to +\infty$.

Exercice 132 page 263 (Fonctionnelle quadratique)

1. Pour montrer que K est non vide, remarquer que comme $d \neq 0$, il existe $\tilde{x} \in \mathbb{R}^n$ tel que $d \cdot \tilde{x} = \alpha \neq 0$. En déduire l'existence de $x \in \mathbb{R}^n$ tel que $d \cdot x = c$.
2. Montrer par le théorème de Lagrange que si $\bar{\boldsymbol{x}}$ est solution de (3.49), alors $y = (\bar{\boldsymbol{x}}, \lambda)^t$ est solution du système (3.58), et montrer ensuite que le système (3.58) admet une unique solution.

Corrigés des exercices

Corrigé de l'exercice 117 page 233 (Mise en oeuvre de GPF et GPO)

1. On a

$$\nabla f(x) = \begin{bmatrix} 4x_1 - x_2 - 3 \\ 2x_2 - x_1 - 1 \end{bmatrix} \text{ et } H_f = \begin{bmatrix} 4 & -1 \\ -1 & 2 \end{bmatrix}$$

La fonction f vérifie les hypothèses du théorème 3.30 d'existence et d'unicité du minimum. En particulier la hessienne $H_f = \begin{bmatrix} 4 & -1 \\ -1 & 2 \end{bmatrix}$ est s.d.p.. Le minimum est obtenu pour

$$\partial_1 f(x_1, x_2) = 4x_1 - x_2 - 3 = 0$$
$$\partial_2 f(x_1, x_2) = 2x_2 - x_1 - 1 = 0$$

c'est-à-dire $\bar{x}_1 = 1$ et $\bar{x}_2 = 1$. Ce minimum est $f(\bar{x}_1, \bar{x}_2) = 2$.

2. L'algorithme du gradient à pas fixe s'écrit :

$$\begin{cases} \text{Initialisation :} & x^{(0)} \in \mathbb{R}^2,\ \rho > 0 \\ \text{Itération } k : & x^{(k)} \text{ connu, } (k \geq 0) \\ & w^{(k)} = -\nabla f(x^{(k)}), \\ & x^{(k+1)} = x^{(k)} + \rho w^{(k)}. \end{cases}$$

A la première itération, on a $\nabla f(0,0) = (-3,-1)$ et donc $w^{(0)} = (3,1)$. On en déduit, pour $\rho = 0.5$, $x^{(1)} = (3\rho, \rho) = (3/2, 1/2)$ et $f(x^{(1)}) = 3$.

L'algorithme du gradient à pas optimal s'écrit :

$$\begin{cases} \text{Initialisation :} & x^{(0)} \in \mathbb{R}^n. \\ \text{Itération } k : & x^{(k)} \text{ connu.} \\ & \text{On calcule } w^{(k)} = -\nabla f(x^{(k)}). \\ & \text{On choisit } \rho_k \geq 0 \text{ tel que} \\ & \qquad f(x^{(k)} + \rho_k w^{(k)}) \leq f(x^{(k)} + \rho w^{(k)})\ \ \forall \rho \geq 0. \\ & \text{On pose } x^{(k+1)} = x^{(k)} + \rho_k w^{(k)}. \end{cases}$$

Calculons le ρ_0 optimal à l'itération 0. On a vu précédemment que $w^{(0)} = (3,1)$. Le ρ_0 optimal minimise la fonction $\rho \mapsto \varphi(\rho) = f(x^{(0)} + \rho w^{(0)}) = f(3\rho, \rho)$. On doit donc avoir $\varphi'(\rho_0) = 0$. Calculons $\varphi'(\rho)$. Par le théorème de dérivation des fonctions composées, on a :

$$\varphi'(\rho) = \nabla f(x^{(0)} + \rho w^{(0)}) \cdot w(0) = \begin{bmatrix} 11\rho - 3 \\ -\rho - 1 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ 1 \end{bmatrix} = 3(11\rho - 3) + (-\rho - 1) = 32\rho - 10.$$

On en déduit que $\rho_0 = \frac{5}{16}$. On obtient alors $x^{(1)} = x^{(0)} + \rho_0 w^{(0)} = (\frac{15}{16}, \frac{5}{16})$, et $f(x^{(1)}) = 2.4375$, ce qui est, comme attendu, mieux qu'avec GPF.

Corrigé de l'exercice 118 page 233 (Convergence de l'algorithme du gradient à pas optimal)

1. On sait que $f(x) \to +\infty$ lorsque $|x| \to +\infty$. Donc $\forall A > 0$, $\exists R \in \mathbb{R}_+ ; |x| > R \Rightarrow f(x) > A$. En particulier pour $A = f(x_0)$ ceci entraîne :

$$\exists R \in \mathbb{R}_+; \;\; x \in B_R \Rightarrow f(x) > f(x_0).$$

2. Comme $f \in C^2(\mathbb{R}^n, \mathbb{R})$, sa hessienne H est continue, donc $\|H\|_2$ atteint son $\max$ sur B_{R+1} qui est un fermé borné de $\mathbb{R}^n$. Soit $M = \max_{x \in B_{R+1}} \|H(x)\|_2$, on a $|H(x)y \cdot y| \leq My \cdot y \leq M|y|^2$.
3. Soit $w_k = -\nabla f(x_k)$.
   Si $w_k = 0$, on pose $x_{k+1} = x_k$.
   Si $w_k \neq 0$, montrons qu'il existe $\bar{\rho} > 0$ tel que

$$f(x_k + \bar{\rho} w_k) \leq f(x_k + \rho w_k) \;\; \forall \rho > 0.$$

   On sait que $f(x) \to +\infty$ lorsque $|x| \to +\infty$.
   Soit $\varphi : \mathbb{R}_+ \to \mathbb{R}$ définie par $\varphi(\rho) = f(x_k + \rho w_k)$. On a $\varphi(0) = f(x_k)$ et $\varphi(\rho) = f(x_k + \rho w_k) \to +\infty$ lorsque $\rho \to +\infty$.
   En effet si $\rho \to +\infty$, on a $|x_k + \rho w_k| \to +\infty$. Donc φ étant continue, φ admet un minimum, atteint en $\bar{\rho}$, et donc $\exists \bar{\rho} \in \mathbb{R}_+ ; f(x_k + \bar{\rho} w) \leq f(x_k + \rho w_k) \;\; \forall \rho > 0$.
4. a) Montrons que la suite $(f(x_k))_{n \in \mathbb{N}}$ est convergente. La suite $(f(x_k))_{n \in \mathbb{N}}$ vérifie

$$f(x_{k+1}) \leq f(x_k).$$

   De plus $f(x) \to +\infty$ lorsque $|x| \to +\infty$ donc f est bornée inférieurement. On en conclut que la suite $(f(x_k))_{n \in \mathbb{N}}$ est convergente.

   b) Montrons que $x_k \in B_R \;\; \forall k \in \mathbb{N}$. On sait que si $x \notin B_R$ alors $f(x) > f(x_0)$. Or la suite $(f(x_k))_{n \in \mathbb{R}}$ est décroissante donc $f(x_k) \leq f(x_0) \;\; \forall k$, donc $x_k \in B_R, \;\; \forall k \in \mathbb{N}$.

   c) Montrons que $f(x_k + \rho w_k) \leq f(x_k) - \rho |w_k|^2 + \frac{\rho^2}{2} M |w_k|^2, \;\; \forall \rho \in [0, \frac{1}{|w_k|}]$. Soit φ définie de $\mathbb{R}_+$ dans $\mathbb{R}$ par $\varphi(\rho) = f(x_k + \rho w_k)$. On a

$$\varphi(\rho) = \varphi(0) + \rho \varphi'(0) + \frac{\rho^2}{2} \varphi''(\tilde{\rho}), \text{ où } \;\; \tilde{\rho} \in]0, \rho[.$$

   Or $\varphi'(\rho) = \nabla f(x_k + \rho w_k) \cdot w_k$ et $\varphi''(\rho) = H(x_k + \rho w_k) w_k \cdot w_k$. Donc

$$\varphi(\rho) = \underbrace{\varphi(0)}_{f(x_k)} + \rho \underbrace{\nabla f(x_k)}_{-w_k} \cdot w_k + \frac{\rho^2}{2} H(x_k + \tilde{\rho} w_k) w_k \cdot w_k.$$

Si $\rho \in [0, \frac{1}{|w_k|}]$ on a

$$\begin{aligned} |x_k + \tilde{\rho} w_k| &\leq |x_k| + \frac{1}{|w_k|}|w_k| \\ &\leq R+1, \end{aligned}$$

donc $x_k + \tilde{\rho} w_k \in B_{R+1}$ et par la question 2,

$$H(x_k + \tilde{\rho} w_k) w_k \cdot w_k \leq M |w_k|^2.$$

On a donc bien

$$\varphi(\rho) = f(x_k + \rho w_k) \leq f(x_k) - \rho |w_k|^2 + \frac{\rho^2}{2} M |w_k|^2.$$

d) Montrons que $f(x_{k+1}) \leq f(x_k) - \dfrac{|w_k|^2}{2M}$ si $|w_k| \leq M$.

Comme le choix de α_k est optimal, on a

$$f(x_{k+1}) = f(x_k + \alpha_k w_k) \leq f(x_k + \rho w_k), \quad \forall \rho \in \mathbb{R}_+.$$

donc en particulier

$$f(x_{k+1}) \leq f(x_k + \rho w_k), \quad \forall \rho \in [0, \frac{1}{|w_k|}].$$

En utilisant la question précédente, on obtient

$$f(x_{k+1}) \leq f(x_k) - \rho |w_k|^2 + \frac{\rho^2}{2} M |w_k|^2 = \varphi(\rho), \quad \forall \rho \in [0, \frac{1}{|w_k|}]. \tag{3.43}$$

Or la fonction φ atteint son minimum pour

$$-|w_k|^2 + \rho M |w_k|^2 = 0$$

c'est–à–dire $\rho M = 1$ ou encore $\rho = \frac{1}{M}$ ce qui est possible si $\dfrac{1}{|w_k|} \geq \dfrac{1}{M}$ (puisque 3.43 est vraie si $\rho \leq \dfrac{1}{|w_k|}$).

Comme on a supposé $|w_k| \leq M$, on a donc

$$f(x_{k+1}) \leq f(x_k) - \frac{|w_k|^2}{M} + \frac{|w_k|^2}{2M} = f(x_k) - \frac{|w_k|^2}{2M}.$$

e) Montrons que $-f(x_{k+1}) + f(x_k) \geq \dfrac{|w_k|^2}{2\bar{M}}$ où $\bar{M} = \sup(M, \tilde{M})$ avec $\tilde{M} = \sup\{|\nabla f(x)|,\ x \in B_R\}$.

On sait par la question précédente que si

$$|w_k| \leq M, \text{ on a } -f(x_{k+1}) - f(x_k) \geq \frac{|w_k|^2}{2M}.$$

Montrons que si $|w_k| \geq M$, alors $-f(x_{k+1}) + f(x_k) \geq \dfrac{|w_k|^2}{2\tilde{M}}$. On aura alors le résultat souhaité.

On a

$$f(x_{k+1}) \leq f(x_k) - \rho |w_k|^2 + \frac{\rho^2}{2} M |w_k|^2, \ \forall \rho \in [0, \frac{1}{|w_k|}].$$

Donc

$$f(x_{k+1}) \leq \min_{[0, \frac{1}{|w_k|}]} \underbrace{[f(x_k) - \rho |w_k|^2 + \frac{\rho^2}{2} M |w_k|^2]}_{P_k(\rho)}$$

— 1er cas si $|w_k| \leq M$, on a calculé ce $\min$ à la question c).

— si $|w_k| \geq M$, la fonction $P_k(\rho)$ est décroissante sur $[0, \frac{1}{|w_k|}]$ et le minimum est donc atteint pour $\rho = \frac{1}{|w_k|}$.

$$\begin{aligned} \text{Or } P_k\left(\frac{1}{|w_k|}\right) &= f(x_k) - |w_k| + \frac{M}{2} \leq f(x_k) - \frac{|w_k|}{2} \\ &\leq f(x_k) - \frac{|w_k|^2}{2\tilde{M}}, \end{aligned}$$

en remarquant que $|w_k| \leq \tilde{M}$.

5. Montrons que $\nabla f(x_k) \to 0$ lorsque $k \to +\infty$. On a montré que $\forall k$, $|w_k|^2 \leq 2\bar{M}(f(x_k) - f(x_{k+1}))$. Or la suite $(f(x_k))_{n\in\mathbb{N}}$ est convergente. Donc $|w_k| \to 0$ lorsque $n \to +\infty$ et $w_k = \nabla f(x_k)$ ce qui prouve le résultat.
   La suite $(x_k)_{n\in\mathbb{N}}$ est bornée donc $\exists (n_k)_{k\in\mathbb{N}}$ et $\tilde{x} \in \mathbb{R}^n$; $x_{n_k} \to \tilde{x}$ lorsque $k \to +\infty$ et comme $\nabla f(x_{n_k}) \to 0$, on a, par continuité, $\nabla f(\tilde{x}) = 0$.
6. On suppose qu'il existe un unique $\bar{x} \in \mathbb{R}^n$ tel que $\nabla f(\bar{x}) = 0$. Comme f est croissante à l'infini, il existe un point qui réalise un minimum de f, et on sait qu'en ce point le gradient s'annule ; en utilisant l'hypothèse d'unicité, on en déduit que ce point est forcément $\bar{x}$. On remarque aussi que $\bar{x}$ est la seule valeur d'adhérence de la suite (bornée) $(x_k)_{k\in\mathbb{N}}$, et donc que $x_k \to \bar{x}$ quand $k \to +\infty$.

Corrigé de l'exercice 121 page 235 (Algorithme du gradient à pas optimal)

Corrigé en cours de rédaction...

Corrigé de l'exercice 119 page 234 (Jacobi et optimisation)

1. La méthode de Jacobi peut s'écrire

$$\begin{aligned} \boldsymbol{x}^{(k+1)} &= (\mathrm{Id} - D^{-1}A)\boldsymbol{x}^{(k)} + D^{-1}\boldsymbol{b} \\ &= \boldsymbol{x}^{(k)} + D^{-1}(\boldsymbol{b} - A\boldsymbol{x}^{(k)}) \\ &= \boldsymbol{x}^{(k)} + \boldsymbol{w}^{(k)} \end{aligned}$$

   avec $\boldsymbol{w}^{(k)} = D^{-1}(\boldsymbol{b} - A\boldsymbol{x}^{(k)}) = D^{-1}\boldsymbol{r}^{(k)}$. On a $\boldsymbol{w}^{(k)} \cdot \nabla f(\boldsymbol{x}^{(k)}) = -D^{-1}\boldsymbol{r}^{(k)} \cdot \boldsymbol{r}^{(k)}$, et comme A est s.d.p., D^{-1} l'est également, et donc $\boldsymbol{w}^{(k)} \cdot \nabla f(\boldsymbol{x}^{(k)}) < 0$ si $\boldsymbol{x}^{(k)} \neq A^{-1}\boldsymbol{b}$. Ceci montre que $\boldsymbol{w}^{(k)}$ est une direction de descente stricte en $x^{(k)}$.

2. (a) Le pas optimal α_k est celui qui minimise la fonction φ définie de $\mathbb{R}$ dans $\mathbb{R}$ par $f(\boldsymbol{x}^{(k)} + \alpha\boldsymbol{w}^{(k)})$, qui est de classe C^1, strictement convexe et croissante à l'infini, ce qui donne l'existence et l'unicité ; de plus α_k vérifie :

$$\nabla f(\boldsymbol{x}^{(k)} + \alpha_k\boldsymbol{w}^{(k)}) \cdot \boldsymbol{w}^{(k)} = 0, \text{ c.à.d. } (A\boldsymbol{x}^{(k)} + \alpha_k A\boldsymbol{w}^{(k)} + \boldsymbol{b}) \cdot \boldsymbol{w}^{(k)} = 0.$$

   On en déduit que (si $w^{(k)} \neq 0$)

$$\alpha_k = \frac{(\boldsymbol{b} - A\boldsymbol{x}^{(k)}) \cdot \boldsymbol{w}^{(k)}}{A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)}} = \frac{\boldsymbol{r}^{(k)} \cdot \boldsymbol{w}^{(k)}}{A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)}} = \frac{\boldsymbol{r}^{(k)} \cdot D^{-1}\boldsymbol{r}^{(k)}}{AD^{-1}\boldsymbol{r}^{(k)} \cdot D^{-1}\boldsymbol{r}^{(k)}}.$$

   (Si $w^{(k)} = 0$, on a alors $r^{(k)} = 0$ et $x^{(k)} = \bar{x}$, l"algorithme s'arrête.)

(b) On a :

$$f(\boldsymbol{x}^{(k+1)}) = f(\boldsymbol{x}^{(k)}) - \gamma\alpha_k + \delta\alpha_k^2,$$

avec $\gamma = \boldsymbol{r}^{(k)} \cdot \boldsymbol{w}^{(k)}$ et $\delta = \frac{1}{2}A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)}$. Comme α_k minimise ce polynôme de degré 2 en α, on a

$$\begin{aligned} f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)}) &= -\frac{\gamma^2}{4\delta} \\ &= -\frac{|\boldsymbol{r}^{(k)} \cdot \boldsymbol{w}^{(k)}|^2}{2A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)}}, \end{aligned}$$

d'où le résultat.

(c) On suppose que $w^{(k)} \neq 0$ pour tout k. La suite $(f(x^{(k)}))_{k\in\mathbb{N}}$ est décroissante et bornée inférieurement (car la fonction f est bornée inférieurement). Elle est donc convergente. Ce qui prouve que $\lim_{k\to+\infty} f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)}) = 0$.
On sait que $\boldsymbol{w}^{(k)} = D^{-1}\boldsymbol{r}^{(k)}$. On a donc, par la question précédente,

$$\frac{|\boldsymbol{r}^{(k)} \cdot \boldsymbol{w}^{(k)}|^2}{A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)}} = \frac{|\boldsymbol{r}^{(k)} \cdot D^{-1}\boldsymbol{r}^{(k)}|^2}{AD^{-1}\boldsymbol{r}^{(k)} \cdot D^{-1}\boldsymbol{r}^{(k)}} = 2|f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^{(k+1)})|.$$

Or

$$0 < AD^{-1}\boldsymbol{r}^{(k)} \cdot D^{-1}\boldsymbol{r}^{(k)} \leq \zeta|\boldsymbol{r}^{(k)}|^2$$

avec $\zeta = \|A\|_2\|D^{-1}\|_2^2$ et

$$\boldsymbol{r}^{(k)} \cdot D^{-1}\boldsymbol{r}^{(k)} \geq \theta|\boldsymbol{r}^{(k)}|^2,$$

où $\theta = \min_{i\in\{1,\ldots,n\}} 1/a_{i,i}$. (Les $a_{i,i}$ étant les termes diagonaux de A.) On en déduit que

$$\frac{\theta^2}{\zeta}|\boldsymbol{r}^{(k)}|^2 \leq |f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^{(k+1)})| \to 0 \text{ lorsque } k \to +\infty,$$

et donc $\boldsymbol{r}^{(k)} \to \mathbf{0}$ lorsque $k \to +\infty$.
Comme $\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}} = -A^{-1}\boldsymbol{r}^{(k)}$, on en déduit la convergence de la suite $\boldsymbol{x}^{(k)}$ vers la solution du système.

(d) i. Si $D = \alpha\mathrm{Id}$, on a

$$\alpha_k = \frac{(\boldsymbol{b} - A\boldsymbol{x}^{(k)}) \cdot \boldsymbol{w}^{(k)}}{A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)}} = \frac{\boldsymbol{r}^{(k)} \cdot \boldsymbol{w}^{(k)}}{A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)}} = \frac{1}{\alpha}\frac{\boldsymbol{r}^{(k)} \cdot \boldsymbol{r}^{(k)}}{A\boldsymbol{r}^{(k)} \cdot \boldsymbol{r}^{(k)}}.$$

ii. Jacobi simple= algorithme de gradient avec $\rho = \frac{1}{\alpha}$
Jacobi à pas optimal= algorithme de gradient à pas optimal.

Corrigé de l'exercice 122 page 235 (Méthode de relaxation)

1. On sait par la propostion 3.13 que si f vérifie l'hypothèse (3.10) alors f est strictement convexe et tend vers l'infini en l'infini, et donc il existe un unique $\bar{x} \in \mathbb{R}^n$ réalisant son minimum.

2. Ecrivons l'hypothèse (3.10) avec $x = se_k$ et $y = te_k$ où $(s,t) \in \mathbb{R}^2$ et e_k est le k-ième vecteur de la base canonique de $\mathbb{R}^n$; en notant $\partial_k f$ la dérivée partielle de f par rapport à la k-ième variable, il vient :

$$(\partial_k f(s) - \partial_k f(t))(s-t) \geq \alpha|s-t|^2.$$

En appliquant à nouveau la proposition 3.13 au cas $n = 1$, on en déduit l'existence et unicité de $\bar{s}$ tel que

$$\varphi_k^{(k+1)}(\overline{s}) = \inf_{s\in\mathbb{R}} \varphi_k^{(k+1)}(s).$$

Comme l'algorithme (3.39) procède à n minimisations de ce type à chaque itération, on en déduit que la suite $(x^{(k)})_{n\in\mathbb{N}}$ construite par cet algorithme est bien définie.

3.(a) Par définition, $x_k^{(k+1)}$ réalise le minimum de la fonction $\varphi_k^{(k+1)}$ sur $\mathbb{R}$. Comme de plus, $\varphi_k^{(k+1)} \in C^1(\mathbb{R},\mathbb{R})$, on a donc $(\varphi_k^{(k+1)})'(x_k^{(k+1)}) = 0$. Or $(\varphi_k^{(k+1)})'(x_k^{(k+1)}) = \partial_k f(x^{(n+1,k)})$, et donc $\partial_k f(x^{(n+1,k)}) = 0$.
D'après la démonstration de la proposition 3.13 (voir l'inégalité (3.11)), on a

$$\begin{aligned} f(x^{(n+1,k-1)}) - f(x^{(n+1,k)}) &\geq \nabla f(x^{(n+1,k)})\cdot(x^{(n+1,k-1)} - x^{(n+1,k)}) \\ &+ \frac{\alpha}{2}|x^{(n+1,k-1)} - x^{(n+1,k)}|^2. \end{aligned}$$

Or $x^{(n+1,k-1)} - x^{(n+1,k)} = -x_k^{(k+1)} e_k$ et $\nabla f(x^{(n+1,k)})\cdot e_k = \partial_k f(x^{(n+1,k)}) = 0$. On en déduit que :

$$f(x^{(n+1,k-1)}) - f(x^{(n+1,k)}) \geq \frac{\alpha}{2}|x^{(n+1,k-1)} - x^{(n+1,k)}|^2.$$

3.(b) Par définition de la suite $(x^{(k)})_{n\in\mathbb{N}}$, on a :

$$f(x^{(k)}) - f(x^{(k+1)}) = \sum_{k=1}^{n} f(x^{(n+1,k-1)}) - f(x^{(n+1,k)}).$$

Par la question précédente, on a donc :

$$f(x^{(k)}) - f(x^{(k+1)}) \geq \frac{\alpha}{2}\sum_{k=1}^{n} |x^{(n+1,k-1))} - x^{(n+1,k)}|^2.$$

Or $x^{(n+1,k-1))} - x^{(n+1,k)} = -x_k^{(k+1)} e_k$, et $(e_k)_{k\in n}$ est une base orthonormée. On peut donc écrire que

$$\begin{aligned} \sum_{k=1}^{n} |x^{(n+1,k-1))} - x^{(n+1,k)}|^2 &= \sum_{k=1}^{n} |(x_k^{(k)} - x_k^{(k+1)})e_k|^2 \\ &= |\sum_{k=1}^{n} (x_k^{(k)} - x_k^{(k+1)})e_k|^2 \\ &= |\sum_{k=1}^{n} (x^{(n+1,k-1))} - x^{(n+1,k)})|^2 \\ &= |x^{(k)} - x^{(k+1)}|^2. \end{aligned}$$

On en déduit que

$$f(x^{(k)}) - f(x^{(k+1)}) \geq \frac{\alpha}{2}|x^{(k)} - x^{(k+1)}|^2.$$

La suite $(f(x^{(k)}))_{k\in\mathbb{N}}$ est bornée inférieurement par $f(\bar{x})$; l'inégalité précédente montre qu'elle est décroissante, donc elle converge. On a donc $f(x^{(k)}) - f(x^{(k+1)} \to 0$ lorsque $n \to +\infty$, et donc par l'inégalité précédente,

$$\lim_{n\to+\infty} |x^{(k)} - x^{(k+1)}| = 0.$$

De plus, pour $1 \leq k \leq n$,

$$\begin{aligned} |x^{(n+1,k)} - x^{(k+1)}|^2 &= \sum_{\ell=k}^{n} |(x_\ell^{(k)} - x_\ell^{(k+1)})e_\ell|^2 \\ &= |\sum_{\ell=k}^{n} (x_\ell^{(k)} - x_\ell^{(k+1)})e_\ell|^2 \\ &= |\sum_{\ell=k}^{n} (x^{(n+1,\ell-1))} - x^{(n+1,\ell)})|^2 \\ &\leq |x^{(k)} - x^{(k+1)}|^2. \end{aligned}$$

d'où l'on déduit que $\lim_{n\to+\infty} |x^{(n+1,k)} - x^{(k+1)}| = 0$.

4. En prenant $x = \bar{x}$ et $y = x^{(k+1)}$ dans l'hypothèse (3.10) et en remarquant que, puisque $\bar{x}$ réalise le minimum de f, on a $\nabla f(\bar{x}) = 0$, on obtient :

$$(-\nabla f(x^{(k+1)}) \cdot (\bar{x} - x^{(k+1)}) \geq \alpha |\bar{x} - x^{(k+1)}|^2,$$

et donc, par l'inégalité de Cauchy Schwarz :

$$|x^{(k+1)} - \overline{x}| \leq \frac{1}{\alpha} \left(\sum_{k=1}^{n} |\partial_k f(x^{(k+1)})|^2 \right)^{\frac{1}{2}}.$$

5. En vertu de la proposition 3.13, on sait que la fonction f est croissante à l'infini. Donc il existe $R > 0$ tel que si $|x| > R$ alors $f(x) > f(x_0)$. Or, la suite $(f(x_k))_{k\in\mathbb{N}}$ étant décroissante, on a $f(x_k)) \leq f(x_0)$ pour tout n, et donc $|x_k| \leq R$ pour tout n. Par la question 3(b), on sait que pour tout $k \geq 1$, $\lim_{n\to+\infty} |x^{(n+1,k)} - x^{(k+1)}| = 0$, ce qui prouve que les suites $(x^{(n+1,k)})_{n\in\mathbb{N}}$, pour $k = 1, \ldots, n$, sont également bornées.

Comme $\lim_{n\to+\infty} |x^{(n+1,k)} - x^{(k+1)}| = 0$, on a pour tout $\eta > 0$, l'existence de $N_\eta \in \mathbb{N}$ tel que $|x^{(n+1,k)} - x^{(k+1)}| < \eta$ si $n \geq N_\eta$. Comme $f \in C^1(\mathbb{R}, \mathbb{R})$, la fonction $\partial_k f$ est uniformément continue sur les bornés (théorème de Heine), et donc pour tout $\varepsilon > 0$, il existe $\eta > 0$ tel que si $|x - y| < \eta$ alors $|\partial_k f(x) - \partial_k f(y)| \leq \epsilon$. On a donc, pour $n \geq N_\eta$: $|\partial_k f(x^{(n+1,k)}) - \partial_k f(x^{(k+1)})| \leq \epsilon$, ce qui démontre que :

$$|\partial_k f(x^{(k+1)})| \to 0 \text{ lorsque } n \to +\infty.$$

On en conclut par le résultat de la question 4 que $x^{(k)} \to \bar{x}$ lorsque $n \to +\infty$.

6. On a vu au paragraphe 3.2.2 que dans ce cas, $\nabla f(x) = \frac{1}{2}(A + A^t)x - b$. L'algorithme 3.39 est donc la méthode de Gauss Seidel pour la résolution du système linéaire $\frac{1}{2}(A + A^t)x = b$.

7 (a) La fonction g est strictement convexe (car somme d'une fonction strictement convexe : $(x_1, x_2) \to x_1^2 + x_2^2$, d'une fonction linéaire par morceaux : $(x_1, x_2) \mapsto -2(x_1 + x_2) + 2|x_1 - x_2|$. et croissante à l'infini grâce aux termes en puissance 2. Il existe donc un unique élément $\overline{x} = (\overline{x}_1, \overline{x}_2)^t$ de $\mathbb{R}^2$ tel que $g(\overline{x}) = \inf_{x\in\mathbb{R}^2} g(x)$.
7 (b) Soit $\epsilon > 0$. On a, pour tout $x \in \mathbb{R}$, $\phi_x(\epsilon) = g(x, x + \epsilon) = x^2 + (x + \epsilon)^2 - 4x$, qui atteint (pour tout x) son minimum pour $\epsilon = 0$. Le minimum de g se situe donc sur l'axe $x = y$. Or $\psi(x) = g(x, x) = 2x^2 - 4x$ atteint son minimum en $x = 1$.
7 (c) Si $x^{(0)} = (0, 0)^t$, on vérifie facilement que l'algorithme (3.39) appliqué à g est stationnaire. La suite ne converge donc pas vers $\overline{x}$. La fonction g n'est pas différentiable sur la droite $x_1 = x_2$.

Corrigé de l'exercice 124 page 237 (Gradient conjugué pour une matrice non symétrique)

1. Comme A est inversible, A^t l'est aussi et donc les systèmes (3.40) et (3.41) sont équivalents.

2 (a) La matrice M est symétrique définie positive, car A est inversible et $M = AA^t$ est symétrique. Donc ses valeurs propres sont strictement positives.

2 (b) On a $\text{cond}(A) = \|A\|\|A^{-1}\|$. Comme la norme est ici la norme euclidienne, on a : $\|A\| = (\rho(A^tA))^{\frac{1}{2}}$ et $\|A^{-1}\| = (\rho((A^{-1})^tA^{-1}))^{\frac{1}{2}} = (\rho(AA^t)^{-1}))^{\frac{1}{2}}$. On vérifie facilement que $M = A^tA$ et A^tA ont mêmes valeurs propres et on en déduit le résultat.

3. Ecrivons l'algorithme du gradient conjugué pour la résolution du système (3.41)

$$\left\{\begin{array}{l}
\textbf{Initialisation}\\
\text{Soit } x^{(0)} \in \mathbb{R}^n, \text{ et soit } r^{(0)} = A^tb - A^tAx^{(0)} =\\
\quad 1) \text{ Si } r^{(0)} = 0, \text{ alors } Ax^{(0)} = b \text{ et donc } x^{(0)} = \bar{\boldsymbol{x}},\\
\quad \text{auquel cas l'algorithme s'arrête.}\\
\quad 2) \text{ Si } r^{(0)} \neq 0, \text{ alors on pose } w^{(0)} = r^{(0)}, \text{ et on choisit } \rho_0 = \dfrac{r^{(0)}\cdot r^{(0)}}{A^tAw^{(0)}\cdot w^{(0)}}.\\
\quad \text{On pose alors } x^{(1)} = x^{(0)} + \rho_0 w^{(0)}.\\
\\
\textbf{Itération } 1 \leq n \leq n-1 :\\
\text{On suppose } x^{(0)}, \ldots, x^{(k)} \text{ et } w^{(0)}, \ldots, w^{(k-1)} \text{ connus et on pose}\\
\quad \boldsymbol{r}^{(k)} = A^tb - A^tAx^{(k)}.\\
\quad 1) \text{ Si } \boldsymbol{r}^{(k)} = 0 \text{ on a } Ax^{(k)} = b \text{ donc } x^{(k)} = \bar{\boldsymbol{x}}\\
\quad \text{auquel cas l'algorithme s'arrête.}\\
\quad 2) \text{ Si } \boldsymbol{r}^{(k)} \neq 0, \text{ alors on pose } w^{(k)} = \boldsymbol{r}^{(k)} + \lambda_{k-1}w^{(k-1)}\\
\quad \text{avec } \lambda_{k-1} = \frac{\boldsymbol{r}^{(k)}\cdot\boldsymbol{r}^{(k)}}{r^{(k-1)}\cdot r^{(k-1)}}. \qquad \text{et on pose } \alpha_k = \dfrac{\boldsymbol{r}^{(k)}\cdot\boldsymbol{r}^{(k)}}{A^tAw^{(k)}\cdot w^{(k)}}.\\
\quad \text{On pose alors } x^{(k+1)} = x^{(k)} + \alpha_k w^{(k)}.
\end{array}\right.$$

Si on implémente l'algorithme sous cette forme, on a intérêt à calculer d'abord $\tilde{b} = A^tb$ et $M = A^tA$ pour minimiser le nombre de mutliplications matrice matrice et matrice vecteur. Au lieu du coût de l'algorithme initial, qui est en $2n^3 + O(n^2)$, on a donc un coût en $3n^3 + O(n^2)$.
Maintenant si on est optimiste, on peut espérer converger en moins de n itérations (en fait, c'est malheureusement rarement le cas), et dans ce cas il est plus économique d'écrire l'algorithme précédent sous la forme suivante.

$$\left\{\begin{array}{l}
\textbf{Initialisation}\\
\text{Soit } x^{(0)} \in \mathbb{R}^n, \text{ et soit } s^{(0)} = b - Ax^{(0)} \text{ et soit } r^{(0)} = A^ts^{(0)}\\
\quad 1) \text{ Si } r^{(0)} = 0, \text{ alors } Ax^{(0)} = b \text{ et donc } x^{(0)} = \bar{\boldsymbol{x}},\\
\quad \text{auquel cas l'algorithme s'arrête.}\\
\quad 2) \text{ Si } r^{(0)} \neq 0, \text{ alors on pose } w^{(0)} = r^{(0)}, y^{(0)} = Aw^{(0)} \text{ et on choisit } \rho_0 = \dfrac{r^{(0)}\cdot r^{(0)}}{y^{(0)}\cdot y^{(0)}}.\\
\quad \text{On pose alors } x^{(1)} = x^{(0)} + \rho_0 w^{(0)}.\\
\\
\textbf{Itération } 1 \leq n \leq n-1 :\\
\text{On suppose } x^{(0)}, \ldots, x^{(k)} \text{ et } w^{(0)}, \ldots, w^{(k-1)} \text{ connus et on pose}\\
\quad s^{(k)} = b - Ax^{(k)} \text{ et } \boldsymbol{r}^{(k)} = A^ts^{(k)}.\\
\quad 1) \text{ Si } \boldsymbol{r}^{(k)} = 0 \text{ on a } Ax^{(k)} = b \text{ donc } x^{(k)} = \bar{\boldsymbol{x}}\\
\quad \text{auquel cas l'algorithme s'arrête.}\\
\quad 2) \text{ Si } \boldsymbol{r}^{(k)} \neq 0, \text{ alors on pose } w^{(k)} = \boldsymbol{r}^{(k)} + \lambda_{k-1}w^{(k-1)}\\
\quad \text{avec } \lambda_{k-1} = \frac{\boldsymbol{r}^{(k)}\cdot\boldsymbol{r}^{(k)}}{r^{(k-1)}\cdot r^{(k-1)}}. \qquad \text{et on pose } \alpha_k = \dfrac{\boldsymbol{r}^{(k)}\cdot\boldsymbol{r}^{(k)}}{y^{(k)}\cdot y^{(k)}} \text{ avec } y^{(k)} = Aw^{(k)}.\\
\quad \text{On pose alors } x^{(k+1)} = x^{(k)} + \alpha_k w^{(k)}.
\end{array}\right.$$

On peut facilement vérifier que dans cette version, on a un produit matrice vecteur en plus à chaque itération, donc le coût est le même pour n itérations, mais il est inférieur si on a moins de n itérations.
Remarque : Cette méthode s'appelle méthode du gradient conjugué appliquée aux équations normales. Elle est facile à comprendre et à programmer. Malheureusement, elle ne marche pas très bien dans la pratique, et on lui préfère des méthodes plus sophistiquées telles sue la méthode "BICGSTAB" ou "GMRES".

Corrigé de l'exercice 125 page 237 (Gradient conjugué préconditionné par LL^t)

1. *Soit $x \in \mathbb{R}^n$. On a $Bx\cdot x = L^{-1}A(L^t)^{-1}x\cdot x = A(L^t)^{-1}x\cdot(L^{-1})^tx = A(L^t)^{-1}x\cdot(L^t)^{-1}x = Ay\cdot y$ avec $y = (L^t)^{-1}x$. Donc $Bx\cdot x \geq 0$, et $Bx\cdot x = 0$ ssi $(L^t)^{-1}x = 0$, c.à.d., puisque L est inversible,*

si $x = 0$. De plus $B^t = (L^{-1}A(L^t)^{-1})^t = ((L^t)^{-1})^t A^t (L^{-1})^t = L^{-1}A(L^t)^{-1}$ car A est symtrique. La matrice B est donc symétrique dfinie positive.

2. *Par dfinition, A et B son s.d.p. donc inversibles, et $y = B^{-1}\tilde{b} = B^{-1}L^{-1}b = (L^{-1}A(L^t)^{-1})^{-1}L^{-1}b = L^tA^{-1}LL^{-1}b = L^tx$.*

3. *On a montré en cours que l'algorithme du gradient conjugué pour la résolution d'un système linéaire avec une matrice symétrique définie positive converge en au plus N itérations. Or la famille (y_k) est construite par cet algorithme pour la résolution du système linéaire $By = \tilde{b}$, ou pour la minimisation de la fonctionnelle J définie par $J(y) = \frac{1}{2}By \cdot y - \tilde{b} \cdot y$, car B est symétrique définie positive. On en déduit que la famille $(y^{(k)})$ est finie et que $y^{(N)} = y$.*

4. (a) $\lambda_{k-1} = \dfrac{s^{(k)} \cdot \boldsymbol{r}^{(k)}}{s^{(k-1)} \cdot r^{(k-1)}}$, *Par définition,* $\lambda_{k-1} = \frac{\tilde{\boldsymbol{r}}^{(k)} \cdot \tilde{\boldsymbol{r}}^{(k)}}{\tilde{r}^{(k-1)} \cdot \tilde{r}^{(k-1)}}$. *Or*

$$\begin{aligned} \tilde{\boldsymbol{r}}^{(k)} &= \tilde{b} - By^{(k)} \\ &= L^{-1}b - L^{-1}A(L^{-1})^t y^{(k)} \\ &= L^{-1}(b - Ax^{(k)} \\ &= L^{-1}\boldsymbol{r}^{(k)}. \end{aligned}$$

et donc $\tilde{\boldsymbol{r}}^{(k)} \cdot \tilde{\boldsymbol{r}}^{(k)} = L^{-1}\boldsymbol{r}^{(k)} \cdot L^{-1}\boldsymbol{r}^{(k)} = (LL^t)^{-1}\boldsymbol{r}^{(k)} \cdot \boldsymbol{r}^{(k)} = s^{(k)} \cdot \boldsymbol{r}^{(k)}$. On en déduit que $\lambda_{k-1} = \frac{s^{(k)} \cdot \boldsymbol{r}^{(k)}}{s^{(k-1)} \cdot r^{(k-1)}}$.

(b) $\rho_n = \dfrac{s^{(k)} \cdot \boldsymbol{r}^{(k)}}{w^{(k)} \cdot Aw^{(k)}}$,
Par définition et par ce qui précède,

$$\alpha_k = \frac{\tilde{\boldsymbol{r}}^{(k)} \cdot \tilde{\boldsymbol{r}}^{(k)}}{\tilde{w}^{(k)} \cdot B\tilde{w}^{(k)}} = \frac{s^{(k)} \cdot \boldsymbol{r}^{(k)}}{\tilde{w}^{(k)} \cdot B\tilde{w}^{(k)}}.$$

Or

$$\begin{aligned} \tilde{w}^{(k)} \cdot B\tilde{w}^{(k)} &= \tilde{w}^{(k)} \cdot L^{-1}A(L^{-1})^t\tilde{w}^{(k)} \\ &= (L^{-1})^t\tilde{w}^{(k)} \cdot A(L^{-1})^t\tilde{w}^{(k)} \\ &= w^{(k)} \cdot Aw^{(k)}. \end{aligned}$$

On en déduit l'expression de α_k.

(c) $w^{(k)} = s^{(k)} + \lambda_n w^{(k-1)}$,
Par dfinition,

$$\begin{aligned} w^{(k)} &= (L^{-1})^t\tilde{w}^{(k)} \\ &= (L^{-1})^t(\tilde{\boldsymbol{r}}^{(k)} + \lambda_{k-1}\tilde{w}^{(k-1)}) \\ &= (L^{-1})^tL^{-1}\boldsymbol{r}^{(k)} + \lambda_{k-1}w^{(k-1)} \\ &= s^{(k)} + \lambda_n w^{(k-1)}. \end{aligned}$$

(d) $x^{(k+1)} = x^{(k)} + \alpha_k w^{(k)}$. *Par dfinition,*

$$\begin{aligned} x^{(k+1)} &= (L^t)^{-1}y^{(k+1)} \\ &= (L^t)^{-1}(y^{(k)} + \rho_n\tilde{w}^{(k)}) \\ &= x^{(k)} + \rho_n w^{(k)}. \end{aligned}$$

5. *D'après les questions précédentes, l'algorithme du gradient conjugué préconditionné par Choleski incomplète s'écrit donc :*

Itération n *On pose* $\boldsymbol{r}^{(k)} = b - Ax^{(k)}$,
on calcule $s^{(k)}$ *solution de* $LL^t s^{(k)} = \boldsymbol{r}^{(k)}$.
On pose alors $\lambda_{k-1} = \dfrac{s^{(k)} \cdot \boldsymbol{r}^{(k)}}{s^{(k-1)} \cdot r^{(k-1)}}$ *et* $w^{(k)} = s^{(k)} + \lambda_{k-1} w^{(k-1)}$.
Le paramètre optimal α_k *a pour expression :* $\alpha_k = \dfrac{s^{(k)} \cdot \boldsymbol{r}^{(k)}}{Aw^{(k)} \cdot w^{(k)}}$, *et on pose alors* $x^{(k+1)} = x^{(k)} + \alpha_k w^{(k)}$.

Corrigé de l'exercice 126 page 238 (Méthode de quasi-linéarisation)

Soit $f \in C^3(\mathbb{R}, \mathbb{R})$ une fonction croissante à l'infini, c. à.d. telle que $f(x) \to +\infty$ lorsque $|x| \to +\infty$; soit $d \in \mathbb{R}$ et soit J la fonction de $\mathbb{R}$ dans $\mathbb{R}$ par

$$J(x) = (f(x) - d)^2.$$

1. La fonction J est continue et croissante à l'infini grâce aux hypothèses sur$f$, et il existe donc bien $\bar{x} \in \mathbb{R}$ tel que $J(\bar{x}) \leq J(x), \forall x \in \mathbb{R}$.
2. L'algorithme de Newton qui donne x_{k+1} en fonction de x_k s'écrit
$$\left[2(f'(x_k))^2 - 2(f(x_k) - d)f''(x_k))\right](x_{k+1} - x_k) = -2(f(x_k) - d)f'(x_k).$$
3. La fonction J_k est quadratique, et on a vu en cours qu'elle admet donc un unique minimum $x_{k+1} \in \mathbb{R}$ qui est obtenu en annulant la dérivée. Or $J_k'(x) = 2f'(x_k)\,(f(x_k) + f'(x_k)(x - x_k) - d)$, et donc
$$x_{k+1} = x_k - \frac{f(x_k) - d}{f'(x_k)}.$$
C'est l'algorithme de Newton pour la résolution de $f(x) = d$.
4. L'algorithme du gradient s'écrit :
$$\begin{aligned} x_{k+1} &= x_k - \rho J'(x_k) \\ &= x_k - 2\rho(f(x_k) - d)f'(x_k) \end{aligned}$$
5.
   (a) Dans le cas où $f(x) = x^2$, on a $J(x) = (x^2 - d)^2$. La fonction J est positive ou nulle.
   — Dans le cas où $d > 0$, le minimum est donc atteint pour $x^2 - d = 0$ c.à.d. $x = \pm\sqrt{d}$. Il n'y a pas unicité car la fonction J n'est pas strictement convexe.
   — Dans le cas où $d \leq 0$, le minimum est atteint pour $\bar{x} = 0$, et dans ce cas J est strictement convexe et la solution est unique.
   (b) La fonction J' appartient à $C^2[\mathbb{R}, \mathbb{R})$ par hypothèse sur f, et donc d'après le cours on a convergence locale quadratique.
   (c) Soit x_0 le choix initial. La suite $(x_n)_{n\in\mathbb{N}}$ construite par l'algorithme de quasi-linéarisation s'écrit :
   $$2x_k(x_{k+1} - x_k) = -(x_k^2 - d)$$
   C'est la méthode de Newton pour la recherche d'un zéro de la fonction $\psi(x) = x^2 - d$, dont on sait par le théorème du cours qu'elle converge localement (et quadratiquement).
   (d) Soit x_0 le choix initial. Soit $(x_k)_{k\in\mathbb{N}}$ la suite construite par l'algorithme de quasi-linéarisation ; elle s'écrit :
   $$2x_k(x_{k+1} - x_k) = -(x_k^2 + 1)$$
   C'est la méthode de Newton pour la recherche d'un zéro de la fonction $\psi(x) = x^2 + 1$, qui n'a pas de solution dans $\mathbb{R}$. Supposons que $x_k \neq 0$ pour tout k et que la suite converge à l'infini vers une limite ℓ. On a alors $0 = -(\ell^2 + 1) < 0$, ce qui est impossible. Le seul cas de convergence est obtenu s'il existe n tel que $x_n = 0 (= \bar{x})$. C'est le cas si $x_0 = 0$, ou bien si $x_0 = 1$ (auquel cas $x_1 = 0$ et l'algorithme s'arrête), ou bien si $x_0 = \frac{1}{2}(1 \pm \sqrt{5})$ auquel cas $x_1 = 1$ et $x_2 = 0$ etc...)

Corrigé de l'exercice 127 page 239 (Méthode de Polak-Ribière)

1. Montrons que f est strictement convexe et croissante à l'infini. Soit φ la fonction de $\mathbb{R}$ dans $\mathbb{R}$ définie par

$$\varphi(t) = f(x + t(y - x)).$$

On a $\varphi \in C^2(\mathbb{R}, \mathbb{R})$, $\varphi(0) = f(x)$ et $\varphi(1) = f(y)$, et donc :

$$f(y) - f(x) = \varphi(1) - \varphi(0) = \int_0^1 \varphi'(t)dt.$$

En intégrant par parties, ceci entraîne :

$$f(y) - f(x) = \varphi'(0) + \int_0^1 (1 - t)\varphi''(t)dt. \tag{3.44}$$

Or $\varphi'(t) = \nabla(x + t(y - x)) \cdot (y - x)$ et donc $\varphi''(t) = H(x + t(y - x))(y - x) \cdot (y - x)$. On a donc par hypothèse $\varphi''(t) \geq \alpha|y - x|^2$. On déduit alors de 3.44 que

$$f(y) \geq f(x) + \nabla f(x) \cdot (y - x) + \frac{\alpha}{2}|y - x|^2. \tag{3.45}$$

L'inégalité 3.45 entraîne la stricte convexité de f et sa croissance à l'infini (voir la démonstration de la proposition 3.13).

Il reste à montrer que l'ensemble $\mathcal{VP}(H(x))$ des valeurs propres de $H(x)$ est inclus dans $[\alpha, \beta]$. Comme $f \in C^2(\mathbb{R}, \mathbb{R})$, $H(x)$ est symétrique pour tout $x \in \mathbb{R}$, et donc diagonalisable dans $\mathbb{R}$. Soit $\lambda \in \mathcal{VP}(H(x))$; il existe donc $y \in \mathbb{R}^n$, $y \neq 0$ tel que $H(x)y = \lambda y$, et donc $\alpha y \cdot y \leq \lambda y \cdot y \leq \beta y \cdot y, \forall \lambda \in \mathcal{VP}(H)(x))$. On en déduit que $\mathcal{VP}(H(x)) \subset [\alpha, \beta]$.

2. Montrons par récurrence sur n que $g^{(k+1)} \cdot w^{(k)} = 0$ et $g^{(k)} \cdot g^{(k)} = g^{(k)} \cdot w^{(k)}$ pour tout $k \in \mathbb{N}$.

Pour $k = 0$, on a $w^{(0)} = g^{(0)} = -\nabla f(x^{(0)})$.

Si $\nabla f(x^{(0)}) = 0$ l'algorithme s'arrête. Supposons donc que $\nabla f(x^{(0)}) \neq 0$. Alors $w^{(0)} = -\nabla f(x^{(0)})$ est une direction de descente stricte. Comme $x^{(1)} = x^{(0)} + \rho_0 w^{(0)}$ où ρ_0 est optimal dans la direction $w^{(0)}$, on a $g^{(1)} \cdot w^{(0)} = -\nabla f(x^{(1)}) \cdot w^{(0)} = 0$. De plus, on a évidemment $g^{(0)} \cdot w^{(0)} = g^{(0)} \cdot g^{(0)}$.

Supposons maintenant que $g^{(k)} \cdot w^{(k-1)} = 0$ et $g^{(k-1)} \cdot g^{(k-1)} = g^{(k-1)} \cdot w^{(k-1)}$, et montrons que $g^{(k+1)} \cdot w^{(k)} = 0$ et $g^{(k)} \cdot g^{(k)} = g^{(k)} \cdot w^{(k)}$.

Par définition, on a :

$$w^{(k)} = g^{(k)} + \lambda_{k-1} w^{(k-1)}, \text{ donc}$$
$$w^{(k)} \cdot g^{(k)} = g^{(k)} \cdot g^{(k)} + \lambda_{k-1} w^{(k-1)} \cdot g^{(k)} = g^{(k)} \cdot g^{(k)}$$

par hypothèse de récurrence. On déduit de cette égalité que $w^{(k)} \cdot g^{(k)} > 0$ (car $g^{(k)} \neq 0$) et donc $w^{(k)}$ est une direction de descente stricte en $x^{(k)}$. On a donc $\nabla f(x^{(k+1)}) \cdot w^{(k)} = 0$, et finalement $g^{(k+1)} \cdot w^{(k)} = 0$.

3. Par définition, $g^{(k)} = -\nabla f(x^{(k)})$; or on veut calculer $g^{(k+1)} - g^{(k)} = -\nabla f(x^{(k+1)}) + \nabla f(x^{(k)})$. Soit φ la fonction de $\mathbb{R}$ dans $\mathbb{R}$ définie par :

$$\varphi(t) = -\nabla f(x^{(k)} + t(x^{(k+1)} - x^{(k)})).$$

On a donc :

$$\begin{aligned} \varphi(1) - \varphi(0) &= g^{(k+1)} - g^{(k)} \\ &= \int_0^1 \varphi'(t)dt. \end{aligned}$$

Calculons $\varphi' : \varphi'(t) = H(x^{(k)} + t(x^{(k+1)} - x^{(k)}))(x^{(k+1)} - x^{(k)})$. Et comme $x^{(k+1)} = x^{(k)} + \alpha_k w^{(k)}$, on a donc :

$$g^{(k+1)} - g^{(k)} = \alpha_k J^{(k)} w^{(k)}. \tag{3.46}$$

De plus, comme $g^{(k+1)} \cdot w^{(k)} = 0$ (question 1), on obtient par (3.46) que

$$\alpha_k = \frac{g^{(k)} \cdot w^{(k)}}{J^{(k)} w^{(k)} \cdot w^{(k)}}$$

(car $J^{(k)} w^{(k)} \cdot w^{(k)} \neq 0$, puisque $J^{(k)}$ est symétrique définie positive).

4. Par définition, on a $w^{(k)} = g^{(k)} + \lambda_{k-1} w^{(k-1)}$, et donc

$$|w^{(k)}| \leq |g^{(k)}| + |\lambda_{k-1}||w^{(k-1)}|. \tag{3.47}$$

Toujours par définition, on a :

$$\lambda_{k-1} = \frac{g^{(k)} \cdot (g^{(k)} - g^{(k-1)})}{g^{(k-1)} \cdot g^{(k-1)}}.$$

Donc, par la question 3, on a :

$$\lambda_{k-1} = \frac{\alpha_{k-1} g^{(k)} \cdot J^{(k-1)} w^{(k-1)}}{g^{(k-1)} \cdot g^{(k-1)}}.$$

En utilisant la question 2 et à nouveau la question 3, on a donc :

$$\lambda_{k-1} = -\frac{J^{(k-1)} w^{(k-1)} \cdot g^{(k)}}{J^{(k-1)} w^{(k-1)} \cdot w^{(k-1)}},$$

et donc

$$|\lambda_{k-1}| = \frac{|J^{(k-1)} w^{(k-1)} \cdot g^{(k)}|}{J^{(k-1)} w^{(k-1)} \cdot w^{(k-1)}},$$

car $J^{(k-1)}$ est symétrique définie positive.
De plus, en utilisant les hypothèses sur H, on vérifie facilement que

$$\alpha |x|^2 \leq J^{(k)} x \cdot x \leq \beta |x|^2 \ \ \forall x \in \mathbb{R}^n.$$

On en déduit que

$$|\lambda_{k-1}| \leq \frac{|J^{(k-1)} w^{(k-1)} \cdot g^{(k)}|}{\alpha |w^{(k-1)}|^2}.$$

On utilise alors l'inégalité de Cauchy–Schwarz :

$$\begin{aligned} |J^{(k-1)} w^{(k-1)} \cdot g^{(k)}| &\leq \|J^{(k-1)}\|_2 \ |w^{(k-1)}| \ |g^{(k)}| \\ &\leq \beta |w^{(k-1)}| \ |g^{(k)}|. \end{aligned}$$

On obtient donc que

$$|\lambda_{k-1}| \leq \frac{\beta}{\alpha} \frac{|g^{(k)}|}{|w^{(k-1)|}},$$

ce qui donne bien grâce à (3.47) :

$$|w^{(k)}| \leq |g^{(k)}|(1 + \frac{\beta}{\alpha}).$$

5. • Montrons d'abord que la suite $(f(x^{(k)}))_{n \in \mathbb{N}}$ converge. Comme $f(x^{(k+1)}) = f(x^{(k)} + \alpha_k w^{(k)}) \leq f(x^{(k)} + \rho w^{(k)}) \ \ \forall \rho \geq 0$, on a donc en particulier $f(x^{(k+1)}) \leq f(x^{(k)})$. La suite $(f(x^{(k)}))_{n \in \mathbb{N}}$ est donc décroissante. De plus, elle est minorée par $f(\bar{\boldsymbol{x}})$. Donc elle converge, vers une certaine limite $\ell \in \mathbb{R}$, lorsque k tend vers $+\infty$.

- La suite $(x^{(k)})_{k\in\mathbb{N}}$ est bornée : en effet, comme f est croissante à l'infini, il existe $R > 0$ tel que si $|x| > R$ alors $f(x) > f(x^{(0)})$. Or $f(x^{(k)}) \leq f(x^{(0)})$ pout tout $k \in \mathbb{N}$, et donc la suite $(x^{(k)})_{n\in\mathbb{N}}$ est incluse dans la boule de rayon R.
- Montrons que $\nabla f(x^{(k)}) \to 0$ lorsque $n \to +\infty$.
  On a, par définition de $x^{(k+1)}$,

$$f(x^{(k+1)}) \leq f(x^{(k)} + \rho w^{(k)}), \quad \forall \rho \geq 0.$$

  En introduisant la fonction φ définie de $\mathbb{R}$ dans $\mathbb{R}$ par $\varphi(t) = f(x^{(k)} + t\rho w^{(k)})$, on montre facilement (les calculs sont les mêmes que ceux de la question 1) que

$$f(x^{(k)} + \rho w^{(k)}) = f(x^{(k)}) + \rho \nabla f(x^{(k)}) \cdot w^{(k)} + \rho^2 \int_0^1 H(x^{(k)} + t\rho w^{(k)}) w^{(k)} \cdot w^{(k)} (1-t) dt,$$

  pour tout $\rho \geq 0$. Grâce à l'hypothèse sur H, on en déduit que

$$f(x^{(k+1)}) \leq f(x^{(k)}) + \rho \nabla f(x^{(k)}) \cdot w^{(k)} + \frac{\beta}{2}\rho^2 |w^{(k)}|^2, \ \forall \rho \geq 0.$$

  Comme $\nabla f(x^{(k)}) \cdot w^{(k)} = -g^{(k)} \cdot w^{(k)} = -|g^{(k)}|^2$ (question 2) et comme $|w^{(k)}| \leq |g^{(k)}|(1 + \frac{\beta}{\alpha})$ (question 4), on en déduit que :

$$f(x^{(k+1)}) \leq f(x^{(k)}) - \rho |g^{(k)}|^2 + \rho^2 \gamma |g^{(k)}|^2 = \psi_k(\rho), \ \forall \rho \geq 0,$$

  où $\gamma = \frac{\beta^2}{2} + (1 + \frac{\beta}{\alpha})^2$. La fonction ψ_k est un polynôme de degré 2 en ρ, qui atteint son minimum lorsque $\psi_k'(\rho) = 0$, *i.e.* pour $\rho = \dfrac{1}{2\gamma}$. On a donc, pour $\rho = \dfrac{1}{2\gamma}$,

$$f(x^{(k+1)}) \leq f(x^{(k)}) - \frac{1}{4\gamma} |g^{(k)}|^2,$$

  d'où on déduit que

$$|g^{(k)}|^2 \leq 4\gamma (f(x^{(k)}) - f(x^{(k+1)}) \underset{k\to+\infty}{\to} 0$$

  On a donc $\nabla f(x^{(k)}) \to 0$ lorsque $k \to +\infty$.
- La suite $(x^{(k)})_{n\in\mathbb{N}}$ étant bornée, il existe une sous–suite qui converge vers $x \in \mathbb{R}^n$, comme $\nabla f(x^{(k)}) \to 0$ et comme ∇f est continue, on a $\nabla f(x) = 0$. Par unicité du minimum (f est croissante à l'infini et strictement convexe) on a donc $x = \bar{\boldsymbol{x}}$.
  Enfin on conclut à la convergence de toute la suite par un argument classique (voir question 6 de l'exercice 118 page 233).

Corrigé de l'exercice 128 page 239 (Algorithme de quasi Newton)

Partie 1

1. Par définition de $w^{(k)}$, on a :

$$w^{(k)} \cdot \nabla f(x^{(k)}) = -K^{(k)} \nabla f(x^{(k)}) \cdot \nabla f(x^{(k)}) < 0$$

   car K est symétrique définie positive.
   Comme α_k est le paramètre optimal dans la direction $w^{(k)}$, on a $\nabla f(x^{(k)} + \alpha_k w^{(k)}) \cdot w^{(k)} = 0$, et donc $Ax^{(k)} \cdot w^{(k)} + \alpha_k A w^{(k)} \cdot w^{(k)} = b \cdot w^{(k)}$; on en déduit que

$$\alpha_k = -\frac{g^{(k)} \cdot w^{(k)}}{Aw^{(k)} \cdot w^{(k)}}.$$

Comme $w^{(k)} = -K^{(k)}g^{(k)}$, ceci s'écrit encore :

$$\alpha_k = \frac{g^{(k)} \cdot K^{(k)}g^{(k)}}{AK^{(k)}g^{(k)} \cdot K^{(k)}g^{(k)}}.$$

2. Si $K^{(k)} = A^{-1}$, la formule précédente donne immédiatement $\alpha_k = 1$.
3. La méthode de Newton consiste à chercher le zéro de ∇f par l'algorithme suivant (à l'itération 1) :

$$H_f(x^{(0)})(x^{(1)} - x^{(0)}) = -\nabla f(x^{(0)}),$$

(où $H_f(x)$ désigne la hessienne de f au point x) c'est–à–dire

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

On a donc $Ax^{(k)} = b$, et comme la fonction f admet un unique minimum qui vérifie $Ax = b$, on a donc $x^{(1)} = x$, et la méthode converge en une itération.

Partie 2 Méthode de Fletcher–Powell.

1. Soit $n \in \mathbb{N}$, on suppose que $g^{(k)} \neq 0$. Par définition, on a $s^{(k)} = x^{(k+1)} - x^{(k)} = -\alpha_k K^{(k)}g^{(k)}$, avec $\alpha_k > 0$. Comme $K^{(k)}$ est symétrique définie positive elle est donc inversible ; donc comme $g^{(k)} \neq 0$, on a $K^{(k)}g^{(k)} \neq 0$ et donc $s^{(k)} \neq 0$.
   Soit $i < n$, par définition de $s^{(k)}$, on a :

$$s^{(k)} \cdot As^{(i)} = -\alpha_k K^{(k)}g^{(k)} \cdot As^{(i)}.$$

Comme $K^{(k)}$ est symétrique,

$$s^{(k)} \cdot As^{(i)} = -\alpha_k g^{(k)} \cdot K^{(k)}As^{(i)}.$$

Par hypothèse, on a $K^{(k)}As^{(i)} = s^{(i)}$ pour $i < n$, donc on a bien que si $i < n$

$$s^{(k)} \cdot As^{(i)} = 0 \Leftrightarrow g^{(k)} \cdot s^{(i)} = 0.$$

Montrons maintenant que $g^{(k)} \cdot s^{(i)} = 0$ pour $i < n$.

• On a

$$\begin{aligned} g^{(i+1)} \cdot s^{(i)} &= -\rho_i g^{(i+1)} \cdot K^{(i)}g^{(i)} \\ &= -\rho_i g^{(i+1)} \cdot w^{(i)}. \end{aligned}$$

Or $g^{(i+1)} = \nabla f(x^{(i+1)})$ et ρ_i est optimal dans la direction $w^{(i)}$. Donc

$$g^{(i+1)} \cdot s^{(i)} = 0.$$

• On a

$$\begin{aligned} (g^{(k)} - g^{(i+1)}) \cdot s^{(i)} &= (Ax^{(k)} - Ax^{(i+1)}) \cdot s^{(i)} \\ &= \sum_{k=i+1}^{n-1} (Ax^{(k+1)} - Ax^{(k)}) \cdot s^{(i)} \\ &= \sum_{k=i+1}^{n-1} As^{(k)} \cdot s^{(i)}, \\ &= 0 \end{aligned}$$

Par hypothèse de A–conjugaison de la famille $(s^{(i)})_{i=1,k-1}$ on déduit alors facilement des deux égalités précédentes que $g^{(k)} \cdot s^{(i)} = 0$. Comme on a montré que $g^{(k)} \cdot s^{(i)} = 0$ si et seulement si $s^{(k)} \cdot As^{(i)} = 0$, on en conclut que la famille $(s^{(i)})_{i=1,\dots,n}$ est $A-$conjuguée, et que les vecteurs $s^{(i)}$ sont non nuls.

2. Montrons que $K^{(k+1)}$ est symétrique. On a :

$$(K^{(k+1)})^t = (K^{(k)})^t + \frac{(s^{(k)}(s^{(k)})^t)^t}{s^{(k)} \cdot y^{(k)}} - \frac{[(K^{(k)}y^{(k)})(K^{(k)}y^{(k)})^t]^t}{K^{(k)}y^{(k)} \cdot y^{(k)}} = K^{(k+1)},$$

car $K^{(k)}$ est symétrique.

3. Montrons que $K^{(k+1)}As^{(i)} = s^{(i)}$ si $0 \leq i \leq n$. On a :

$$K^{(k+1)}As^{(i)} = K^{(k)}As^{(i)} + \frac{s^{(k)}(s^{(k)})^t}{s^{(k)} \cdot y^{(k)}}As^{(i)} - \frac{(K^{(k)}y^{(k)})(K^{(k)}(y^{(k)})^t}{K^{(k)}y^{(k)} \cdot y^{(k)}}As^{(i)}. \tag{3.48}$$

— Considérons d'abord le cas $i < n$. On a

$$s^{(k)}(s^{(k)})^tAs^{(i)} = s^{(k)}[(s^{(k)})^tAs^{(i)}] = s^{(k)}[s^{(k)} \cdot As^{(i)}] = 0$$

car $s^{(k)} \cdot As^{(i)} = 0$ si $i < n$. De plus, comme $K^{(k)}$ est symétrique, on a :

$$(K^{(k)}y^{(k)})(K^{(k)}y^{(k)})^tAs^{(i)} = K^{(k)}y^{(k)}(y^{(k)})^tK^{(k)}As^{(i)}.$$

Or par la question (c), on a $K^{(k)}As^{(i)} = s^{(i)}$ si $0 \leq i \leq n$. De plus, par définition, $y^{(k)} = As^{(k)}$. On en déduit que

$$(K^{(k)}y^{(k)})(K^{(k)}y^{(k)})^tAs^{(i)} = K^{(k)}y^{(k)}(As^{(k)})^ts^{(i)} = K^{(k)}y^{(k)}(s^{(k)})^tAs^{(i)} = 0$$

puisque on a montré en (a) que les vecteurs $s^{(0)}, \ldots, s^{(k)}$ sont A-conjugués. On déduit alors de (3.48) que

$$K^{(k+1)}As^{(i)} = K^{(k)}As^{(i)} = s^{(i)}.$$

— Considérons maintenant le cas $i = n$. On a

$$K^{(k+1)}As^{(k)} = K^{(k)}As^{(k)} + \frac{s^{(k)}(s^{(k)})^t}{s^{(k)} \cdot y^{(k)}}As^{(k)} - \frac{(K^{(k)}y^{(k)})(K^{(k)}(y^{(k)})^t}{K^{(k)}y^{(k)} \cdot y^{(k)}}As^{(k)},$$

et comme $y^{(k)} = As^{(k)}$,, ceci entraîne que

$$K^{(k+1)}As^{(k)} = K^{(k)}As^{(k)} + s^{(k)} - K^{(k)}y^{(k)} = s^{(k)}.$$

4. Pour $x \in \mathbb{R}^n$, calculons $K^{(k+1)}x \cdot x$:

$$K^{(k+1)}x \cdot x = K^{(k)}x \cdot x + \frac{s^{(k)}(s^{(k)})^t}{s^{(k)} \cdot y^{(k)}}x \cdot x - \frac{(K^{(k)}y^{(k)})(K^{(k)}y^{(k)})^t}{K^{(k)}y^{(k)} \cdot y^{(k)}}x \cdot x.$$

Or $s^{(k)}(s^{(k)})^tx \cdot x = s^{(k)}(s^{(k)} \cdot x) \cdot x = (s^{(k)} \cdot x)^2$, et de même, $(K^{(k)}y^{(k)})(K^{(k)}y^{(k)})^tx \cdot x = (K^{(k)}y^{(k)} \cdot x)^2$. On en déduit que

$$K^{(k+1)}x \cdot x = K^{(k)}x \cdot x + \frac{(s^{(k)} \cdot x)^2}{s^{(k)} \cdot y^{(k)}} - \frac{(K^{(k)}y^{(k)} \cdot x)^2}{K^{(k)}y^{(k)} \cdot y^{(k)}}.$$

En remarquant que $y^{(k)} = As^{(k)}$, et en réduisant au même dénominateur, on obtient alors que

$$K^{(k+1)}x \cdot x = \frac{(K^{(k)}x \cdot x)(K^{(k)}y^{(k)} \cdot y^{(k)}) - (K^{(k)}y^{(k)} \cdot x)^2}{(K^{(k)}y^{(k)} \cdot y^{(k)})} + \frac{(s^{(k)} \cdot x)^2}{As^{(k)} \cdot s^{(k)}}.$$

Montrons maintenant que $K^{(k+1)}$ est symétrique définie positive. Comme $K^{(k)}$ est symétrique définie positive, on a grâce à l'inégalité de Cauchy-Schwarz que $(K^{(k)}y^{(k)} \cdot x)^2 \leq (K^{(k)}x \cdot x)(K^{(k)}y^{(k)})$

avec égalité si et seulement si x et $y^{(k)}$ sont colinéaires. Si x n'est pas colinéaire à $y^{(k)}$, on a donc donc clairement

$$K^{(k+1)}x \cdot x > 0.$$

Si maintenant x est colinéaire à $y^{(k)}$, *i.e.* $x = \alpha y^{(k)}$ avec $\alpha \in \mathbb{R}_+^*$, on a, grâce au fait que $y^{(k)} = As^{(k)}$,

$$\frac{(s^{(k)} \cdot x)^2}{As^{(k)} \cdot s^{(k)}} = \alpha^2 \frac{(s^{(k)} \cdot As^{(k)})^2}{As^{(k)} \cdot s^{(k)}} > 0, \text{ et donc } K^{(k+1)}x \cdot x > 0.$$

On en déduit que $K^{(k+1)}$ est symétrique définie positive.

5. On suppose que $g^{(k)} \neq 0$ si $0 \leq n \leq n-1$. On prend comme hypothèse de récurrence que les vecteurs $s^{(0)}, \ldots, s^{(k-1)}$ sont A-conjugués et non-nuls, que $K^{(j)}As^{(i)} = s^{(i)}$ si $0 \leq i < j \leq n$ et que les matrices $K^{(j)}$ sont symétriques définies positives pour $j = 0, \ldots, n$.
   Cette hypothèse est vérifiée au rang $n = 1$ grâce à la question 1 en prenant $n = 0$ et $K^{(0)}$symétrique définie positive.
   On suppose qu'elle est vraie au rang n. La question 1 prouve qu'elle est vraie au rang $n + 1$.
   Il reste maintenant à montrer que $x^{(n+1)} = A^{-1}b = \overline{x}$. On a en effet $K^{(n)}As^{(i)} = s^{(i)}$ pour $i = 0$ à $n - 1$. Or les vecteurs $s^{(0)}, \ldots, s^{(k-1)}$ sont A-conjugués et non-nuls : ils forment donc une base. On en déduit que $K^{(n)}A = Id$, ce qui prouve que $K^{(n)} = A^{-1}$, et donc, par définition de $x^{(n+1)}$, que $x^{(n+1)} = A^{-1}b = \overline{x}$.

Corrigé de l'exercice 129 page 240 (Méthodes de Gauss–Newton et de quasi–linéarisation)

I *Propriétés d'existence et d'unicité*

(a) Comme C est symétrique éfinie positive, on a $\boldsymbol{y} \cdot C\boldsymbol{y} \geq 0$ pour tout $\boldsymbol{y} \in \mathbb{R}^n$, ce qui prouve que $J(\boldsymbol{x}) \geq 0$ pour tout $\boldsymbol{x} \in \mathbb{R}^n$. Donc J est bornée inférieurement.

(b) Trois exemples

i. Si $n = p$ et $f(\boldsymbol{x}) = \boldsymbol{x}$, $J(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{d}) \cdot C(\boldsymbol{x} - \boldsymbol{d})$ qui est une fonction quadratique pour laquelle on a existence et unicité de $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ qui réalise le minimum de J.

ii. Si $f(\boldsymbol{x}) = \boldsymbol{0}$, $J(\boldsymbol{x}) = \boldsymbol{d} \cdot C\boldsymbol{d}$ et J est donc constante. Il y a donc existence et non unicité de $\bar{x} \in \mathbb{R}^n$ qui réalise le minimum de J.

iii. Pour $n = p = 1$, si $f(x) = e^x$, avec $c = 1$ et $d = 0$, $J(x) = (e^x)^2$ tend vers 0 en l'infini mais 0 n'est jamais atteint. Il ya donc non existence de $\bar{x} \in \mathbb{R}^n$ qui réalise le minimum de J.

II *Un peu de calcul différentiel*

(a) La fonction $Df(x)$ est la différentielle de f en $\boldsymbol{x}$ et c'est donc une application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}^p$. Donc il existe $M(\boldsymbol{x}) \in \mathcal{M}_{p,n}(\mathbb{R})$, où $\mathcal{M}_{p,n}(\mathbb{R})$ désigne l'ensemble des matrices réelles à p lignes et n colonnes, telle que $Df(\boldsymbol{x})(\boldsymbol{y}) = M(\boldsymbol{x})\boldsymbol{y}$ pour tout $\boldsymbol{y} \in \mathbb{R}^n$. On a ensuite $D_2f(\boldsymbol{x}) \in \mathcal{L}(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathbb{R}^p))$. Enfin, on a $Df \in C^1(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathbb{R}^p))$ et $D_2f \in \mathcal{L}(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathbb{R}^p))$.

(b) Comme C ne dépend pas de $\boldsymbol{x}$, on a $\nabla J(\boldsymbol{x}) = M(\boldsymbol{x})C(f(\boldsymbol{x}) - d) + (f(\boldsymbol{x}) - d)CM(\boldsymbol{x})$.

(c)

III *Algorithmes d'optimisation*

3.4 Optimisation sous contraintes

3.4.1 Définitions

Soit $E = \mathbb{R}^n$, soit $f \in C(E, \mathbb{R})$, et soit K un sous ensemble de E. On s'intéresse à la recherche de $\bar{u} \in K$ tel que :

$$\begin{cases} \bar{u} \in K \\ f(\bar{u}) = \inf\limits_K f \end{cases} \tag{3.49}$$

Ce problème est un problème de minimisation avec contrainte (ou “sous contrainte") au sens où l'on cherche u qui minimise f en restreignant l'étude de f aux éléments de K. Voyons quelques exemples de ces contraintes (définies par l'ensemble K), qu'on va expliciter à l'aide des p fonctions continues, $g_i \in C(E, \mathbb{R})$ $i = 1 \dots p$.

1. **Contraintes égalités.** On pose $K = \{x \in E,\, g_i(x) = 0\ i = 1 \dots p\}$. On verra plus loin que le problème de minimisation de f peut alors être résolu grâce au théorème des multiplicateurs de Lagrange (voir théorème 3.34).
2. **Contraintes inégalités.** On pose $K = \{x \in E,\, g_i(x) \leq 0\ i = 1 \dots, p\}$. On verra plus loin que le problème de minimisation de f peut alors être résolu grâce au théorème de Kuhn–Tucker (voir théorème 3.38).
   - — *Programmation linéaire.* Avec un tel ensemble de contraintes K, si de plus f est linéaire, c'est-à-dire qu'il existe $b \in \mathbb{R}^n$ tel que $f(x) = b{\cdot}x$, et les fonctions g_i sont affines, c'est-à-dire qu'il existe $b_i \in \mathbb{R}^n$ et $c_i \in \mathbb{R}$ tels que $g_i(x) = b_i \cdot x + c_i$, alors on dit qu'on a affaire à un problème de “programmation linéaire”. Ces problèmes sont souvent résolus numériquement à l'aide de l'algorithme de Dantzig, inventé vers 1950.
   - — *Programmation quadratique.* Avec le même ensemble de contraintes K, si de plus f est quadratique, c'est-à-dire si f est de la forme $f(x) = \frac{1}{2}Ax \cdot x - b \cdot x$, et les fonctions g_i sont affines, alors on dit qu'on a affaire à un problème de “programmation quadratique”.
3. **Programmation convexe.** Dans le cas où f est convexe et K est convexe, on dit qu'on a affaire à un problème de “programmation convexe”.

3.4.2 Existence – Unicité – Conditions d'optimalité simple

Théorème 3.28 (Existence)**.** *Soit $E = \mathbb{R}^n$ et $f \in C(E, \mathbb{R})$.*

1. *Si K est un sous-ensemble fermé borné non vide de E, alors il existe $\bar{\boldsymbol{x}} \in K$ tel que $f(\bar{\boldsymbol{x}}) = \inf_K f$.*
2. *Si K est un sous-ensemble fermé non vide de E, et si f est croissante à l'infini, c'est–à–dire que $f(x) \to +\infty$ quand $|x| \to +\infty$, alors $\exists \bar{\boldsymbol{x}} \in K$ tel que $f(\bar{\boldsymbol{x}}) = \inf_K f$*

DÉMONSTRATION –

1. Si K est un sous-ensemble fermé borné non vide de E, comme f est continue, elle atteint ses bornes sur K, d'où l'existence de $\bar{\boldsymbol{x}}$.
2. Soit $x_0 \in K$. Si f est croissante à l'infini, alors il existe $R > 0$ tel que si $\|x - x_0\| > R$ alors $f(x) > f(x_0)$; donc $\inf_K f = \inf_{K \cap B(x_0, R)} f$, où $B(x_0, R)$ désigne la boule (fermé) de centre x_0 et de rayon R. L'ensemble $K \cap B(x_0, R)$ est compact, car intersection d'un fermé et d'un compact. Donc, par ce qui précède, il existe $\bar{\boldsymbol{x}} \in K$ tel que $f(\bar{\boldsymbol{x}}) = \inf_{K \cap B(x_0, R)} f = \inf_K f$.

■

Théorème 3.29 (Unicité)**.** *Soit $E = \mathbb{R}^n$ et $f \in C(E, \mathbb{R})$. On suppose que f est strictement convexe et que K est convexe. Alors il existe au plus un élément $\bar{\boldsymbol{x}}$ de K tel que $f(\bar{\boldsymbol{x}}) = \inf_K f$.*

DÉMONSTRATION – Supposons que $\bar{\boldsymbol{x}}$ et $\bar{\bar{\boldsymbol{x}}}$ soient deux solutions du problème (3.49), avec $\bar{\boldsymbol{x}} \neq \bar{\bar{\boldsymbol{x}}}$
Alors $f(\frac{1}{2}\bar{\boldsymbol{x}} + \frac{1}{2}\bar{\bar{\boldsymbol{x}}}) < \dfrac{1}{2}f(\bar{\boldsymbol{x}}) + \dfrac{1}{2}f(\bar{\bar{\boldsymbol{x}}}) = \inf\limits_K f$. On aboutit donc à une contradiction. ∎

Des théorèmes d'existence 3.28 et d'unicité 3.29 on déduit immédiatement le théorème d'existence et d'unicité suivant :

Théorème 3.30 (Existence et unicité)**.** *Soient $E = \mathbb{R}^n$, $f \in C(E, \mathbb{R}^n)$ une fonction strictement convexe et K un sous ensemble convexe fermé de E. Si K est borné ou si f est croissante à l'infini, c'est–à–dire si $f(x) \to +\infty$ quand $\|x\| \to +\infty$, alors il existe un unique élément $\bar{\boldsymbol{x}}$ de K solution du problème de minimisation* (3.49)*, i.e. tel que $f(\bar{\boldsymbol{x}}) = \inf\limits_K f$*

Remarque 3.31. *On peut remplacer $E = \mathbb{R}^n$ par E espace de Hilbert de dimension infinie dans le dernier théorème, mais on a besoin dans ce cas de l'hypothèse de convexité de f pour assurer l'existence de la solution (voir cours de maîtrise).*

Proposition 3.32 (Condition simple d'optimalité)**.** *Soient $E = \mathbb{R}^n$, $f \in C(E, \mathbb{R})$ et $\bar{\boldsymbol{x}} \in K$ tel que $f(\bar{\boldsymbol{x}}) = \inf\limits_K f$. On suppose que f est différentiable en $\bar{\boldsymbol{x}}$*

1. *Si $\bar{\boldsymbol{x}} \in \overset{\circ}{K}$ alors $\nabla f(\bar{\boldsymbol{x}}) = 0$.*
2. *Si K est convexe, alors $\nabla f(\bar{\boldsymbol{x}}) \cdot (x - \bar{\boldsymbol{x}}) \geq 0$ pour tout $x \in K$.*

DÉMONSTRATION –
1. Si $\bar{\boldsymbol{x}} \in \overset{\circ}{K}$, alors il existe $\varepsilon > 0$ tel que $B(\bar{\boldsymbol{x}}, \varepsilon) \subset K$ et $f(\bar{\boldsymbol{x}}) \leq f(x)\ \forall x \in B(\bar{\boldsymbol{x}}, \varepsilon)$. Alors on a déjà vu (voir preuve de la Proposition 3.7 page 212) que ceci implique $\nabla f(\bar{\boldsymbol{x}}) = 0$.
2. Soit $x \in K$. Comme $\bar{\boldsymbol{x}}$ réalise le minimum de f sur K, on a : $f(\bar{\boldsymbol{x}} + t(x - \bar{\boldsymbol{x}})) = f(tx + (1-t)\bar{\boldsymbol{x}}) \geq f(\bar{\boldsymbol{x}})$ pour tout $t \in]0, 1]$, par convexité de K. On en déduit que
$$\frac{f(\bar{\boldsymbol{x}} + t(x - \bar{\boldsymbol{x}})) - f(\bar{\boldsymbol{x}})}{t} \geq 0 \text{ pour tout } t \in]0, 1].$$
En passant à la limite lorsque t tend vers 0 dans cette dernière inégalité, on obtient : $\nabla f(\bar{\boldsymbol{x}}) \cdot (x - \bar{\boldsymbol{x}}) \geq 0$.

∎

3.4.3 Conditions d'optimalité dans le cas de contraintes égalité

Dans tout ce paragraphe, on considèrera les hypothèses et notations suivantes :

$$\begin{array}{l} f \in C(\mathbb{R}^n, \mathbb{R}),\quad g_i \in C^1(\mathbb{R}^n, \mathbb{R}),\ \ i = 1 \dots p\,; \\ K = \{u \in \mathbb{R}^n,\ \ g_i(u) = 0\ \ \forall i = 1 \dots p\}\,; \\ g = (g_1, \dots, g_p)^t \in C^1(\mathbb{R}^n, \mathbb{R}^p) \end{array} \tag{3.50}$$

Remarque 3.33 (Quelques rappels de calcul différentiel)**.**
Comme $g \in C^1(\mathbb{R}^n, \mathbb{R}^p)$, si $u \in \mathbb{R}^n$, alors $Dg(u) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^p)$, ce qui revient à dire, en confondant l'application linéaire $Dg(u)$ avec sa matrice, que $Dg(u) \in \mathcal{M}_{p,n}(\mathbb{R})$. Par définition, $Im(Dg(u)) = \{Dg(u)z,\ \ z \in \mathbb{R}^n\} \subset \mathbb{R}^p$, et $\text{rang}(Dg(u)) = \dim(Im(Dg(u))) \leq p$. On rappelle de plus que

$$Dg(u) = \begin{pmatrix} \dfrac{\partial g_1}{\partial x_1}, & \dots, & \dfrac{\partial g_1}{\partial x_n} \\ \dots, & \ddots, & \dots \\ \dfrac{\partial g_p}{\partial x_1}, & \dots, & \dfrac{\partial g_p}{\partial x_n} \end{pmatrix},$$

et que rang $(Dg(u)) \leq \min(n, p)$. De plus, si rang $(Dg(u)) = p$, alors les vecteurs $(Dg_i(u))_{i=1\dots p}$ sont linéairement indépendants dans $\mathbb{R}^n$.

Théorème 3.34 (Multiplicateurs de Lagrange)**.** *Soit $\bar{u} \in K$ tel que $f(\bar{u}) = \inf_K f$. On suppose que f est différentiable en $\bar{u}$ et $\dim(Im(Dg(\bar{u})) = p$ (ou rang $(Dg(\bar{u})) = p$), alors :*

$$\text{il existe } (\lambda_1, \ldots, \lambda_p)^t \in \mathbb{R}^p \text{ tels que } \nabla f(\bar{u}) + \sum_{i=1}^{p} \lambda_i \nabla g_i(\bar{u}) = 0.$$

(Cette dernière égalité a lieu dans $\mathbb{R}^n$)

DÉMONSTRATION – Pour plus de clarté, donnons d'abord une idée "géométrique" de la démonstration dans le cas $n = 2$ et $p = 1$. On a dans ce cas $f \in C^1(\mathbb{R}^2, \mathbb{R})$ et $K = \{(x, y) \in \mathbb{R}^2 g(x, y) = 0\}$, et on cherche $u \in K$ tel que $f(u) = \inf_K f$. Traçons dans le repère (x, y) la courbe $g(x, y) = 0$, ainsi que les courbes de niveau de f. Si on se "promène" sur la courbe $g(x, y) = 0$, en partant du point P_0 vers la droite (voir figure 3.1), on rencontre les courbes de niveau successives de f et on se rend compte sur le dessin que la valeur minimale que prend f sur la courbe $g(x, y) = 0$ est atteinte lorsque cette courbe est tangente à la courbe de niveau de f : sur le dessin, ceci correspond au point P_1 où la courbe $g(x, y) = 0$ est tangente à la courbe $f(x, y) = 3$. Une fois qu'on a passé ce point de tangence, on peut remarquer que f augmente.

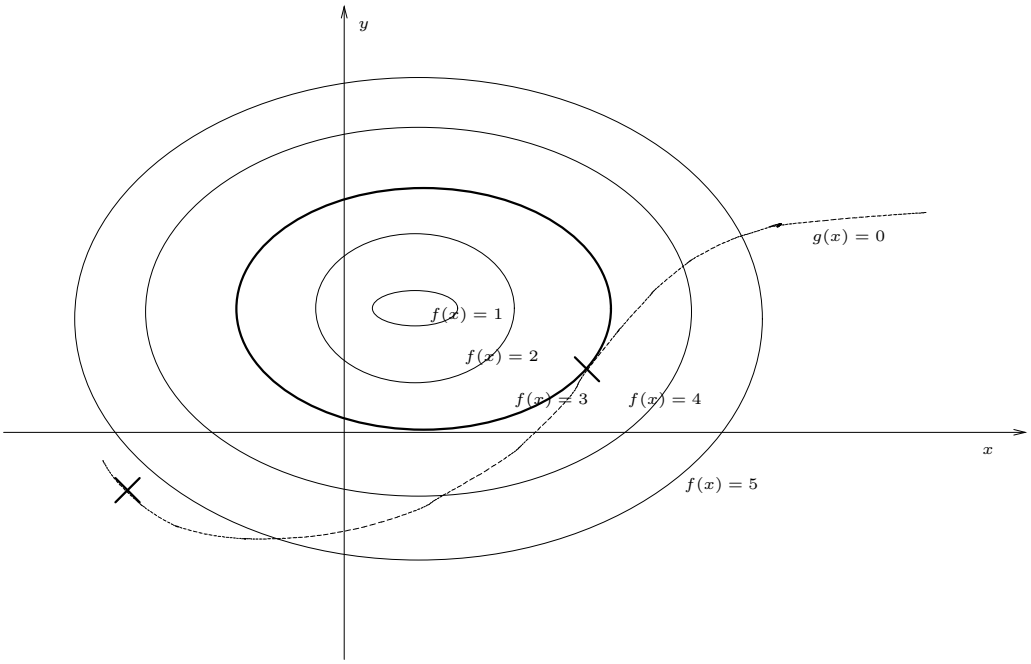

FIGURE 3.1: Interprétation géométrique des multiplicateurs de Lagrange

On utilise alors le fait que si φ est une fonction continûment différentiable de $\mathbb{R}^2$ dans $\mathbb{R}$, le gradient de φ est orthogonal à toute courbe de niveau de φ, c'est-à-dire toute courbe de la forme $\varphi(x, y) = c$, où $c \in \mathbb{R}$. (En effet, soit $(x(t), y(t))$, $t \in \mathbb{R}$ un paramétrage de la courbe $g(x, y) = c$, en dérivant par rapport à t, on obtient : $\nabla g(x(t), y(t)) \cdot (x'(t), y'(t))^t = 0$). En appliquant ceci à f et g, on en déduit qu'au point de tangence entre une courbe de niveau de f et la courbe $g(x, y) = 0$, les gradients de f et g sont colinéaires. Et donc si $\nabla g(u) \neq 0$, il existe $\lambda \neq 0$ tel que $\nabla f(u) = \lambda \nabla g(u)$.

Passons maintenant à la démonstration rigoureuse du théorème dans laquelle on utilise le théorème des fonctions implicites [5].

5. **Théorème des fonctions implicites** Soient p et q des entiers naturels, soit $h \in C^1(\mathbb{R}^q \times \mathbb{R}^p, \mathbb{R}^p)$, et soient $(\bar{x}, \bar{y}) \in \mathbb{R}^q \times \mathbb{R}^p$ et $c \in \mathbb{R}^p$ tels que $h(\bar{x}, \bar{y}) = c$. On suppose que la matrice de la différentielle $D_2 h(\bar{x}, \bar{y}) (\in \mathcal{M}_p(\mathbb{R}))$ est inversible. Alors il existe $\varepsilon > 0$ et

Par hypothèse, $Dg(\bar{u}) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^p)$ et $Im(Dg(\bar{u})) = \mathbb{R}^p$. Donc il existe un sous espace vectoriel F de $\mathbb{R}^n$ de dimension p, tel que $Dg(\bar{u})$ soit bijective de F dans $\mathbb{R}^p$. En effet, soit $(e_1 \dots e_p)$ la base canonique de $\mathbb{R}^p$, alors pour tout $i \in \{1, \dots p\}$, il existe $y_i \in \mathbb{R}^n$ tel que $Dg(\bar{x})y_i = e_i$. Soit F le sous espace engendré par la famille $\{y_1 \dots y_p\}$; on remarque que cette famille est libre, car si $\sum_{i=1}^p \lambda_i y_i = 0$, alors $\sum_{i=1}^p \lambda_i e_i = 0$, et donc $\lambda_i = 0$ pour tout $i = 1, \dots p$. On a ainsi montré l'existence d'un sous espace F de dimension p telle que $Dg(\bar{x})$ soit bijective (car surjective) de F dans $\mathbb{R}^p$.

Il existe un sous espace vectoriel G de $\mathbb{R}^n$, tel que $\mathbb{R}^n = F \bigoplus G$. Pour $v \in F$ et $w \in G$; on pose $\bar{g}(w, v) = g(v + w)$ et $\bar{f}(w, v) = f(v + w)$. On a donc $\bar{f} \in C(G \times F,\ \mathbb{R})$ et $\bar{g} \in C^1(G \times F,\ \mathbb{R})$. De plus, $D_2\bar{g}(w, v) \in \mathcal{L}(F,\ \mathbb{R}^p)$, et pour tout $z \in F$, on a $D_2\bar{g}(w, v)z = Dg(v + w)z$.

Soit $(\bar{v}, \bar{w}) \in F \times G$ tel que $\bar{u} = \bar{v} + \bar{w}$. Alors $D_2\bar{g}(\bar{w}, \bar{v})z = Dg(\bar{u})z$ pour tout $z \in F$. L'application $D_2\bar{g}(\bar{w}, \bar{v})$ est une bijection de F sur $\mathbb{R}^p$, car, par définition de F, $Dg(\bar{u})$ est bijective de F sur $\mathbb{R}^p$.

On rappelle que $K = \{u \in \mathbb{R}^n : g(u) = 0\}$ et on définit $\overline{K} = \{(w, v) \in G \times F,\ \bar{g}(w, v) = 0\}$. Par définition de $\bar{f}$ et de $\bar{g}$, on a

$$\begin{cases} (\bar{w}, \bar{v}) \in \overline{K} \\ \bar{f}(\bar{w}, \bar{v}) \leq f(w, v)\ \ \forall (w, v) \in \overline{K} \end{cases} \tag{3.51}$$

D'autre part, le théorème des fonctions implicites (voir note de bas de page 260) entraîne l'existence de $\varepsilon > 0$ et $\nu > 0$ tels que pour tout $w \in B_G(\bar{w}, \varepsilon)$ il existe un unique $v \in B_F(\bar{v}, \nu)$ tel que $\bar{g}(w, v) = 0$. On note $v = \phi(w)$ et on définit ainsi une application $\phi \in C^1(B_G(\bar{w}, \varepsilon), B_F(\bar{v}, \nu))$.

On déduit alors de (3.51) que :

$$\bar{f}(\bar{w}, \phi(\bar{w})) \leq \bar{f}(w, \phi(w))),\ \forall w \in B_G(\bar{w}, \varepsilon),$$

et donc

$$f(\bar{u}) = f(\bar{w} + \phi(\bar{w})) \leq f(w + \phi(w)),\ \forall w \in B_G(\bar{w}, \varepsilon).$$

En posant $\psi(w) = \bar{f}(w, \phi(w))$, on peut donc écrire

$$\psi(\bar{w}) = \bar{f}(\bar{w}, \phi(\bar{w})) \leq \psi(w),\ \forall w \in B_G(\bar{w}, \varepsilon).$$

On a donc, grâce à la proposition 3.32,

$$D\psi(\bar{w}) = 0. \tag{3.52}$$

Par définition de ψ, de $\bar{f}$ et de $\bar{g}$, on a :

$$D\psi(\bar{w}) = D_1\bar{f}(\bar{w}, \phi((\bar{w})) + D_2\bar{f}(\bar{w}, \phi(\bar{w}))D\phi(\bar{w}).$$

D'après le théorème des fonctions implicites,

$$D\phi(\bar{w}) = -[D_2\bar{g}(\bar{w}, \phi((\bar{w}))]^{-1} D_1\bar{g}(\bar{w}, \phi((\bar{w})).$$

On déduit donc de (3.52) que

$$D_1\bar{f}(\bar{w}, \phi((\bar{w}))w - [D_2\bar{g}(\bar{w}, \phi((\bar{w}))]^{-1} D_1\bar{g}(\bar{w}, \phi((\bar{w}))w = 0,\ \text{pour tout } w \in G. \tag{3.53}$$

De plus, comme $D_2\bar{g}(\bar{w}, \phi((\bar{w}))]^{-1} D_2\bar{g}(\bar{w}, \phi((\bar{w})) = Id$, on a :

$$D_2\bar{f}(\bar{w}, \phi((\bar{w}))z - D_2\bar{f}(\bar{w}, \phi((\bar{w}))\,[D_2\bar{g}(\bar{w}, \phi((\bar{w}))]^{-1}\, D_2\bar{g}(\bar{w}, \phi((\bar{w}))z = 0,\ \forall z \in F. \tag{3.54}$$

Soit $x \in \mathbb{R}^n$, et $(z, w) \in F \times G$ tel que $x = z + w$. En additionant (3.53) et (3.54), et en notant

$$\Lambda = -D_2\bar{f}(\bar{w}, \phi((\bar{w}))\,[D_2\bar{g}(\bar{w}, \phi((\bar{w}))]^{-1}\,,$$

on obtient :

$$Df(\bar{u})x + \Lambda Dg(\bar{u})x = 0,$$

ce qui donne, en transposant : $\nabla f(\bar{u}) + \sum_{i=1}^p \lambda_i \nabla g_i(\bar{u}) = 0$, avec $\Lambda = (\lambda_1, \dots, \lambda_p)$.

■

Remarque 3.35 (Utilisation pratique du théorème de Lagrange)**.** Soit $f \in C^1(\mathbb{R}^n, \mathbb{R})$, $g = (g_1, \dots, g_p)^t$ avec $g_i \in C(\mathbb{R}^n, \mathbb{R})$ pour $i = 1, \dots, p.$, et soit $K = \{u \in \mathbb{R}^n,\ \ g_i(u) = 0,\ \ i = 1, \dots, p\}$.
Le problème qu'on cherche à résoudre est le problème de minimisation (3.49) qu'on rappelle ici :

$$\begin{cases} \bar{u} \in K \\ f(\bar{u}) = \inf\limits_K f \end{cases}$$

$\nu > 0$ tels que pour tout $x \in B(\bar{x}, \varepsilon)$, il existe un unique $y \in B(\bar{y}, \nu)$ tel que $h(x, y) = c$. on peut ainsi définir une application ϕ de $B(\bar{x}, \varepsilon)$ dans $B(\bar{y}, \nu)$ par $\phi(x) = y$. On a $\phi(\bar{x}) = \bar{y}$, $\phi \in C^1(\mathbb{R}^p, \mathbb{R}^p)$ et $D\phi(x) = -[D_2h(x, \phi(x))]^{-1} \cdot D_1h(x, \phi(x))$.

D'après le théorème des multiplicateurs de Lagrange, si $\bar{u}$ est solution de (3.49) et $Im(Dg(\bar{u})) = \mathbb{R}^p$, alors il existe $(\lambda_1, \dots, \lambda_p) \in \mathbb{R}^p$ tel que $\bar{u}$ est solution du problème

$$\begin{cases} \dfrac{\partial f}{\partial x_j}(\bar{u}) + \displaystyle\sum_{i=1}^{p} \lambda_i \frac{\partial g_i}{\partial x_j} = 0, j = 1, \dots n, \\ g_i(\bar{u}) = 0, \quad i = 1, \dots, p. \end{cases} \tag{3.55}$$

Le système (3.55) est un système non linéaire de de $(n+p)$ équations et à $(n+p)$ inconnues $(\bar{x}, \dots, \bar{x}_n, \lambda_i \dots \lambda_p)$. Ce système sera résolu par une méthode de résolution de système non linéaire (Newton par exemple).

Remarque 3.36. *On vient de montrer que si $\bar{x}$ solution de* (3.49) *et $Im(Dg(\bar{x})) = \mathbb{R}^p$, alors $\bar{x}$ solution de* (3.55)*. Par contre, si $\bar{x}$ est solution de* (3.55)*, ceci n'entraîne pas que $\bar{x}$ est solution de* (3.49)*.*

Des exemples d'application du théorème des multiplicateurs de Lagrange sont donnés dans les exercices 131 page 263 et 132 page 263.

3.4.4 Contraintes inégalités

Soit $f \in C(\mathbb{R}^n, \mathbb{R})$ et $g_i \in C^1(\mathbb{R}^n, \mathbb{R}) \quad i = 1, \dots, p$, on considère maintenant un ensemble K de la forme : $K = \{x \in \mathbb{R}^n, \ g_i(x) \leq 0 \ \forall i = 1 \dots p\}$, et on cherche à résoudre le problème de minimisation (3.49) qui sécrit :

$$\begin{cases} \bar{x} \in K \\ f(\bar{x}) \leq f(x), \quad \forall x \in K. \end{cases}$$

Remarque 3.37. *Soit $\bar{x}$ une solution de* (3.49) *et supposons que $g_i(\bar{x}) < 0$, pour tout $i \in \{1, \dots, p\}$. Il existe alors $\varepsilon > 0$ tel que si $x \in B(\bar{x}, \varepsilon)$ alors $g_i(x) < 0$ pour tout $i = 1, \dots, p$.*
On a donc $f(\bar{x}) \leq f(x) \quad \forall x \in B(\bar{x}, \varepsilon)$. On est alors ramené à un problème de minimisation sans contrainte, et si f est différentiable en $\bar{x}$, on a donc $\nabla f(\bar{x}) = 0$.

On donne maintenant sans démonstration le théorème de Kuhn-Tücker qui donne une caractérisation de la solution du problème (3.49).

Théorème 3.38 (Kuhn–Tucker)**.** *Soit $f \in C(\mathbb{R}^n, \mathbb{R})$, soit $g_i \in C^1(\mathbb{R}^n, \mathbb{R})$, pour $i = 1, \dots, p$, et soit $K = \{\boldsymbol{x} \in \mathbb{R}^n, \ g_i(\boldsymbol{x}) \leq 0 \ \forall i = 1 \dots p\}$. On suppose qu'il existe $\bar{x}$ solution de* (3.49)*, et on pose $I(\bar{\boldsymbol{x}}) = \{i \in \{1, \dots, p\}; |g_i(\bar{\boldsymbol{x}}) = 0\}$. On suppose que f est différentiable en $\bar{\boldsymbol{x}}$ et que la famille (de $\mathbb{R}^n$) $\{\nabla g_i(\bar{\boldsymbol{x}}), i \in I(\bar{\boldsymbol{x}})\}$ est libre. Alors il existe une famille $(\lambda_i)_{i \in I(\bar{\boldsymbol{x}})} \subset \mathbb{R}_+$ telle que*

$$\nabla f(\bar{\boldsymbol{x}}) + \sum_{i \in I(\bar{\boldsymbol{x}})} \lambda_i \nabla g_i(\bar{\boldsymbol{x}}) = 0.$$

Remarque 3.39.

1. Le théorème de Kuhn-Tucker s'applique pour des ensembles de contrainte de type inégalité. Si on a une contraite de type égalité, on peut évidemment se ramener à deux contraintes de type inégalité en remarquant que $\{h(x) = 0\} = \{h(x) \leq 0)\} \cap \{-h(x) \leq 0\}$. Cependant, si on pose $g_1 = h$ et $g_2 = -h$, on remarque que la famille $\{\nabla g_1(\bar{x}), \nabla g_2(\bar{x})\} = \{\nabla h(\bar{x}), -\nabla h(\bar{x})\}$ n'est pas libre. On ne peut donc pas appliquer le théorème de Kuhn-Tucker sous la forme donnée précédemment dans ce cas (mais on peut il existe des versions du théorème de Kuhn-Tucker permettant de traiter ce cas, voir Bonans-Saguez).
2. Dans la pratique, on a intérêt à écrire la conclusion du théorème de Kuhn-Tucker (i.e. l'existence de la famille $(\lambda_i)_{i \in I(\bar{x})}$) sous la forme du système de $n + p$ équations et $2p$ inéquations à résoudre suivant :

$$\left\{\begin{array}{l} \nabla f(\bar{x}) + \sum_{i=1}^{p} \lambda_i \nabla g_i(\bar{x}) = 0, \\ \lambda_i g_i(\bar{x}) = 0, \quad \forall i = 1, \ldots, p, \\ g_i(\bar{x}) \leq 0, \quad \forall i = 1, \ldots, p, \\ \lambda_i \geq 0, \quad \forall i = 1, \ldots, p. \end{array}\right.$$

3.4.5 Exercices (optimisation avec contraintes)

Exercice 130 (Sur l'existence et l'unicité)**.** *Corrigé en page 264*
Etudier l'existence et l'unicité des solutions du problème (3.49), avec les données suivantes : $E = \mathbb{R}$, $f : \mathbb{R} \to \mathbb{R}$ est définie par $f(x) = x^2$, et pour les quatre différents ensembles K suivants :

$$\begin{array}{llll} (i) & K = \{|x| \leq 1\}\,; & (ii) & K = \{|x| = 1\} \\ (iii) & K = \{|x| \geq 1\}\,; & (iv) & K = \{|x| > 1\}. \end{array} \tag{3.56}$$

Exercice 131 (Aire maximale d'un rectangle à périmètre donné)**.** *Corrigé en page 265*

1. On cherche à maximiser l'aire d'un rectangle de périmètre donné égal à 2. Montrer que ce problème peut se formuler comme un problème de minimisation de la forme (3.49), où K est de la forme $K = \{x \in \mathbb{R}^2; g(x) = 0\}$. On donnera f et g de manière explicite.
2. Montrer que le problème de minimisation ainsi obtenu est équivalent au problème

$$\left\{\begin{array}{l} \bar{x} = (\bar{x}_1, \bar{x}_2)^t \in \tilde{K} \\ f(\bar{x}_1, \bar{x}_2) \leq f(x_1, x_2), \quad \forall\, (x_1, x_2)^t \in \tilde{K}, \end{array}\right. \tag{3.57}$$

   où $\tilde{K} = K \cap [0,1]^2$, K et f étant obtenus à la question 1. En déduire que le problème de minimisation de l'aire admet au moins une solution.
3. Calculer $Dg(x)$ pour $x \in K$ et en déduire que si x est solution de (3.57) alors $x = (1/2, 1/2)$. En déduire que le problème (3.57) admet une unique solution donnée par $\bar{x} = (1/2, 1/2)$.

Exercice 132 (Fonctionnelle quadratique)**.** *Suggestions en page 242, corrigé en page 265*
Soit f une fonction quadratique, *i.e.* $f(x) = \frac{1}{2}Ax \cdot x - b \cdot x$, où $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice symétrique définie positive et $b \in \mathbb{R}^n$. On suppose que la contrainte g est une fonction linéaire de $\mathbb{R}^n$ dans $\mathbb{R}$, c'est-à-dire $g(x) = d \cdot x - c$ où $c \in \mathbb{R}$ et $d \in \mathbb{R}^n$, et que $d \neq 0$. On pose $K = \{x \in \mathbb{R}^n,\ \ g(x) = 0\}$ et on cherche à résoudre le problème de minimisation (3.49).

1. Montrer que l'ensemble K est non vide, fermé et convexe. En déduire que le problème (3.49) admet une unique solution.
2. Montrer que si $\bar{x}$ est solution de (3.49), alors il existe $\lambda \in \mathbb{R}$ tel que $y = (\bar{x}, \lambda)^t$ soit l'unique solution du système :

$$\left[\begin{array}{c|c} A & d \\ \hline d^t & 0 \end{array}\right] \left[\begin{array}{c} \bar{x} \\ \hline \lambda \end{array}\right] = \left[\begin{array}{c} b \\ \hline c \end{array}\right] \tag{3.58}$$

Exercice 133 (Minimisation sans dérivabilité)**.**
Soient $A \in \mathcal{M}_n(\mathbb{R})$ une matrice s.d.p., $b \in \mathbb{R}^n$, $j : \mathbb{R}^n \to \mathbb{R}$ une fonction continue et convexe, à valeurs positives ou nulles (mais non nécessairement dérivable, par exemple $j(v) = \sum_{j=1}^n \alpha_i |v_i|$, avec $\alpha_i \geq 0$ pour tout $i \in \{1, \ldots, n\}$). Soit U une partie non vide, fermée convexe de $\mathbb{R}^n$. Pour $v \in \mathbb{R}^n$, on pose $J(v) = (1/2)Av \cdot v - b \cdot v + j(v)$.

1. Montrer qu'il existe un et un seul u tel que :

$$u \in U,\ J(u) \leq J(v),\ \forall v \in U. \tag{3.59}$$

2. Soit $u \in U$, montrer que u est solution de (3.59) si et seulement si $(Au-b)\cdot(v-u)+j(v)-j(u)\geq 0$, pour tout $v \in U$.

Exercice 134 (Utilisation du théorème de Lagrange)**.**

1. Pour $(x,y) \in \mathbb{R}^2$, on pose : $f(x,y) = -y$, $g(x,y) = x^2+y^2-1$. Chercher le(s) point(s) où f atteint son maximum ou son minimum sous la contrainte $g = 0$.
2. Soit $a = (a_1,\ldots,a_n) \in \mathbb{R}^n$, $a \neq 0$. Pour $x = (x_1,\ldots,x_n) \in \mathbb{R}^n$, on pose : $f(x) = \sum_{i=1}^n |x_i - a_i|^2$, $g(x) = \sum_{i=1}^n |x_i|^2$. Chercher le(s) point(s) où f atteint son maximum ou son minimum sous la contrainte $g = 1$.
3. Soient $A \in \mathcal{M}_n(\mathbb{R})$ symétrique, $B \in \mathcal{M}_n(\mathbb{R})$ s.d.p. et $b \in \mathbb{R}^n$. Pour $v \in \mathbb{R}^n$, on pose $f(v) = (1/2)Av\cdot v - b\cdot v$ et $g(v) = Bv\cdot v$. Peut-on appliquer le théorème de Lagrange et quelle condition donne-t-il sur u si $f(u) = \min\{f(v), v \in K\}$ avec $K = \{v \in \mathbb{R}^n; g(v) = 1\}$?

Exercice 135 (Contre exemple aux multiplicateurs de Lagrange)**.**
Soient f et $g : \mathbb{R}^2 \to \mathbb{R}$, définies par : $f(x,y) = y$, et $g(x,y) = y^3 - x^2$. On pose $K = \{(x,y) \in \mathbb{R}^2; g(x,y) = 0\}$.

1. Calculer le minimum de f sur K et le point $(\overline{x},\overline{y})$ où ce minimum est atteint.
2. Existe-t-il λ tel que $Df(\overline{x},\overline{y}) = \lambda Dg(\overline{x},\overline{y})$?
3. Pourquoi ne peut-on pas appliquer le théorème des multiplicateurs de Lagrange ?
4. Que trouve-t-on lorsqu'on applique la méthode dite "de Lagrange" pour trouver $(\overline{x},\overline{y})$?

Exercice 136 (Application simple du théorème de Kuhn-Tucker)**.** *Corrigé en page 266*
Soit f la fonction définie de $E = \mathbb{R}^2$ dans $\mathbb{R}$ par $f(x) = x^2+y^2$ et $K = \{(x,y) \in \mathbb{R}^2; x+y \geq 1\}$. Justifier l'existence et l'unicité de la solution du problème (3.49) et appliquer le théorème de Kuhn-Tucker pour la détermination de cette solution.

Exercice 137 (Exemple d'opérateur de projection)**.** *Correction en page 266*

1. Soit $K = C^+ = \{x \in \mathbb{R}^n,\ x = (x_1,\ldots,x_k)^t,\ x_i \geq 0,\ \forall i = 1,\ldots,N\}$.
   (a) Montrer que K est un convexe fermé non vide.
   (b) Montrer que pour tout $y \in \mathbb{R}^n$, on a : $(p_K(y))_i = \max(y_i, 0)$.
2. Soit $(\alpha_i)_{i=1,\ldots,n} \subset \mathbb{R}^n$ et $(\beta_i)_{i=1,\ldots,n} \subset \mathbb{R}^n$ tels que $\alpha_i \leq \beta_i$ pour tout $i = 1,\ldots,n$. Soit $K = \{x = (x_1,\ldots,x_n)^t; \alpha_i \leq \beta_i, i = 1,\ldots,n\}$.
   (a) Montrer que K est un convexe fermé non vide.
   (b) Soit p_K l'opérateur de projection définie à la proposition 3.40 page 267. Montrer que pour tout $y \in \mathbb{R}^n$, on a :
   $$(p_K(y))_i = \max(\alpha_i, \min(y_i, \beta_i)), \quad \forall i = 1,\ldots,n.$$

Corrigés

Exercice 130 page 263 (Sur l'existence et l'unicité)

La fonction $f : \mathbb{R} \to \mathbb{R}$ définie par $f(x) = x^2$ est continue, strictement convexe, et croissante à l'infini. Etudions maintenant les propriétés de K dans les quatre cas proposés :

(i) L'ensemble $K = \{|x| \leq 1\}$ est fermé borné et convexe. On peut donc appliquer le théorème d'existence et d'unicité 3.30 page 259. En remarquant que $f(x) \geq 0$ pour tout $x \in \mathbb{R}$ et que $f(0) = 0$, on en déduit que l'unique solution du problème (3.49) est donc $\bar{x} = 0$.

(ii) L'ensemble $K = \{|x| = 1\}$ est fermé borné mais non convexe. Le théorème d'existence 3.28 page 258 s'applique donc, mais pas le théorème d'unicité 3.29 page 258. De fait, on peut remarquer que $K = \{-1, 1\}$, et donc $\{f(x), x \in K\} = \{1\}$. Il existe donc deux solutions du problème (3.49) : $\bar{x}_1 = 1$ et $\bar{x}_1 = -1$.

(iii) L'ensemble $K = \{|x| \geq 1\}$ est fermé, non borné et non convexe. Cependant, on peut écrire $K = K_1 \cup K_2$ où $K_1 = [1, +\infty[$ et $K_2 =]-\infty, -1]$ sont des ensembles convexes fermés. On peut donc appliquer le théorème 3.30 page 259 : il existe un unique $\bar{x}_1 \in \mathbb{R}$ et un unique $\bar{x}_2 \in \mathbb{R}$ solution de (3.49) pour $K = K_1$ et $K = K_2$ respectivement. Il suffit ensuite de comparer $\bar{x}_1$ et $\bar{x}_2$. Comme $\bar{x}_1 = -1$ et $\bar{x}_2 = 1$, on a existence mais pas unicité.

(iv) L'ensemble $K = \{|x| > 1\}$ n'est pas fermé, donc le théorème 3.28 page 258 ne s'applique pas. De fait, il n'existe pas de solution dans ce cas, car on a $lim_{x \to 1^+} f(x) = 1$, et donc $\inf_K f = 1$, mais cet infimum n'est pas atteint.

Exercice 131 page 263 (Maximisation de l'aire d'un rectangle à périmètre donné)

1. On peut se ramener sans perte de généralité au cas du rectangle $[0, x_1] \times [0, x_2]$, dont l'aire est égale à $x_1 x_2$ et de périmètre $2(x_1 + x_2)$. On veut donc maximiser $x_1 x_2$, ou encore minimiser $-x_1 x_2$. Pour$x = (x_1, x_2)^t \in \mathbb{R}^2$, posons $f(x_1, x_2) = -x_1 x_2$ et $g(x_1, x_2) = x_1 + x_2$. Définissons

$$K = \left\{ x = (x_1, x_2)^t \in (\mathbb{R}_+)^2 \text{ tel que } x_1 + x_2 = 1 \right\}.$$

Le problème de minimisation de l'aire du rectangle de périmètre donné et égal à 2 s'écrit alors :

$$\left\{ \begin{array}{l} \left[\begin{array}{c} x_1 \\ x_2 \end{array} \right] \in K \\ f(\bar{x}_1, \bar{x}_2) \leq f(x_1, x_2) \quad \forall (x_1, x_2) \in K \end{array} \right. \tag{3.60}$$

2. Comme x_1 et x_2 sont tous deux positifs, puisque leur somme doit être égale à 1, ils sont forcément tous deux inférieurs à 1. Il est donc équivalent de résoudre (3.60) ou (3.57). L'ensemble $\tilde{K}$ est un convexe ferme borné, la fonction f est continue, et donc par le théorème 3.28 page 258, il existe au moins une solution du problème (3.57) (ou (3.60)).

3. Calculons ∇g : $\nabla g(x) = (1, 1)^t$, donc rang $Dg(x, y) = 1$. Par le théorème de Lagrange, si $x = (x_1, x_2)^t$ est solution de (3.60), il existe $\lambda \in \mathbb{R}$ tel que

$$\left\{ \begin{array}{l} \nabla f(\bar{x}, \bar{y}) + \lambda \nabla g(\bar{x}, \bar{y}) = 0, \\ \bar{x} + \bar{y} = 1. \end{array} \right.$$

Or $\nabla f(\bar{x}, \bar{y}) = (-\bar{x}, -\bar{y})^t$, et $\nabla g(\bar{x}, \bar{y}) = (1, 1)^t$. Le système précédent s'écrit donc :

$$\begin{aligned} -\bar{y} + \lambda &= 0 \\ -\bar{x} + \lambda &= 0 \\ \bar{x} + \bar{y} &= 1. \end{aligned}$$

On a donc

$$\bar{x} = \bar{y} = \frac{1}{2}.$$

Exercice 132 page 263 (Fonctionnelle quadratique)

1. Comme $d \neq 0$, il existe $\tilde{x} \in \mathbb{R}^n$ tel que $d \cdot \tilde{x} = \alpha \neq 0$. Soit $x = \frac{c}{\alpha} \tilde{x}$ alors $d \cdot x = c$. Donc l'ensemble K est non vide. L'ensemble K est fermé car noyau d'une forme linéaire continue de $\mathbb{R}^n$ dans $\mathbb{R}$, et K est évidemment convexe. La fonction f est strictement convexe et $f(x) \to +\infty$ quand $|x| \to +\infty$, et donc par les théorèmes 3.28 et 3.29 il existe un unique $\bar{x}$ solution de (3.49).

2. On veut calculer $\bar{x}$. On a : $Dg(x)z = d \cdot z$, et donc $Dg(x) = d^t$. Comme $d \neq 0$ on a $rang(Dg(x)) = 1$, ou encore $Im(Dg(x)) = \mathbb{R}$ pour tout x. Donc le théorème de Lagrange s'applique. Il existe donc $\lambda \in \mathbb{R}$ tel que $\nabla f(\bar{x}) + \lambda \nabla g(\bar{x}) = 0$, c'est-à-dire $A\bar{x} - b + \lambda d = 0$. Le couple $(\bar{x}, \lambda)$ est donc solution du problème suivant' :

$$\left\{ \begin{array}{l} A\bar{x} - b + \lambda d = 0, \\ d \cdot \bar{x} = c \end{array} \right. , \tag{3.61}$$

qui s'écrit sous forme matricielle : $By = e$, avec $B = \left[\begin{array}{c|c} A & d \\ \hline d^t & 0 \end{array} \right] \in \mathcal{M}_{n+1}(\mathbb{R})$, $y = \left[\begin{array}{c} \bar{x} \\ \lambda \end{array} \right] \in \mathbb{R}^{n+1}$ et $e = \left[\begin{array}{c} b \\ \hline \\ c \end{array} \right] \in \mathbb{R}^{n+1}$. Montrons maintenant que B est inversible. En effet, soit $z \left[\begin{array}{c} x \\ \mu \end{array} \right] \in \mathbb{R}^{n+1}$, avec $x \in \mathbb{R}^n$ et $\mu \in \mathbb{R}$ tel que $Bz = 0$. Alors

$$\left[\begin{array}{c|c} A & d \\ \hline d^t & 0 \end{array} \right] \left[\begin{array}{c} x \\ \mu \end{array} \right] = 0.$$

Ceci entraîne $Ax - d\mu = 0$ et $d^t x = d \cdot x = 0$. On a donc $Ax \cdot x - (d \cdot x)\mu = 0$. On en déduit que $x = 0$, et comme $d \neq 0$, que $\mu = 0$. On a donc finalement $z = 0$.
On en conclut que B est inversible, et qu'il existe un unique $(x, \lambda)^t \in \mathbb{R}^{n+1}$ solution de (3.61) et et $\bar{x}$ est solution de (3.49).

Exercice 136 page 264 (Application simple du théorème de Kuhn-Tucker

La fonction f définie de $E = \mathbb{R}^2$ dans $\mathbb{R}$ par $f(x) = x^2 + y^2$ est continue, strictement convexe et croissante à l'infini. L'ensemble K qui peut aussi être défini par : $K = \{(x, y) \in \mathbb{R}^2; g(x, y) \leq 0\}$, avec $g(x, y) = 1 - x - y$ est convexe et fermé. Par le théorème 3.30 page 259, il y a donc existence et unicité de la solution du problème (3.49). Appliquons le théorème de Kuhn-Tucker pour la détermination de cette solution. On a :

$$\nabla g(x, y) = \left(\begin{array}{c} -1 \\ -1 \end{array} \right) \text{ et } \nabla f(x, y) = \left(\begin{array}{c} 2x \\ 2y \end{array} \right).$$

Il existe donc $\lambda \in \mathbb{R}_+$ tel que :

$$\left\{ \begin{array}{r} 2x - \lambda = 0, \\ 2y - \lambda = 0, \\ \lambda(1 - x - y) = 0, \\ 1 - x - y \leq 0, \\ \lambda \geq 0. \end{array} \right.$$

Par la troisième équation de ce système, on déduit que $\lambda = 0$ ou $1 - x - y = 0$. Or si $\lambda = 0$, on a $x = y = 0$ par les première et deuxième équations, ce qui est impossible en raison de la quatrième. On en déduit que $1 - x - y = 0$, et donc, par les première et deuxième équations, $x = y = \frac{1}{2}$.

Exercice 137 page 264 (Exemple d'opérateur de projection)

2. Soit p_K l'opérateur de projection définie à la proposition 3.40 page 267, il est facile de montrer que, pour tout $i = 1, \ldots, n$, :

$(p_K(y))_i = y_i$ si $y_i \in [\alpha_i, \beta_i]$,
$(p_K(y))_i = \alpha_i$ si $y_i < \alpha_i$, ce qui entraîne
$(p_K(y))_i = \beta_i$ si $y_i > \beta_i$,

$$(p_K(y))_i = \max(\alpha_i, \min(y_i, \beta_i)) \text{ pour tout } i = 1, \ldots, n.$$

3.5 Algorithmes d'optimisation sous contraintes

3.5.1 Méthodes de gradient avec projection

On rappelle le résultat suivant de projection sur un convexe fermé :

Proposition 3.40 (Projection sur un convexe fermé)**.** *Soit E un espace de Hilbert, muni d'une norme $\|.\|$ induite par un produit scalaire $(.,.)$, et soit K un convexe fermé non vide de E. Alors, tout $x \in E$, il existe un unique $x_0 \in K$ tel que $\|x - x_0\| \leq \|x - y\|$ pour tout $y \in K$. On note $x_0 = p_K(x)$ la projection orthogonale de x sur K. Soient $x \in E$ et $x_0 \in K$. On a également :*

$$x_0 = p_K(x) \text{ si et seulement si } (x - x_0, x_0 - y) \geq 0, \quad \forall y \in K.$$

Dans le cadre des algorithmes de minimisation avec contraintes que nous allons développer maintenant, nous considèrerons $E = \mathbb{R}^n$, $f \in C^1(\mathbb{R}^n, \mathbb{R})$ une fonction convexe, et K fermé convexe non vide. On cherche à calculer une solution approchée de $\bar{x}$, solution du problème (3.49).

Algorithme du gradient à pas fixe avec projection sur K (GPFK) Soit $\rho > 0$ donné, on considère l'algorithme suivant :

Algorithme (GPFK)

Initialisation : $x_0 \in K$

Itération :

x_k connu $\qquad\qquad x_{k+1} = p_K(x_k - \rho \nabla f(x_k))$

où p_K est la projection sur K définie par la proposition 3.40.

Lemme 3.41. *Soit $(x_k)_k$ construite par l'algorithme (GPFK). On suppose que $x_k \to x$ quand $n + \infty$. Alors x est solution de* (3.49)*.*

DÉMONSTRATION – Soit $p_K : \mathbb{R}^n \to K \subset \mathbb{R}^n$ la projection sur K définie par la proposition 3.40. Alors p_K est continue. Donc si

$x_k \to x$ quand $n \to +\infty$ alors $x = p_K(x - \rho \nabla f(x))$ et $x \in K$ (car $x_k \in K$ et K est fermé).

La caractérisation de $p_K(x - \rho \nabla f(x))$ donnée dans la proposition 3.40 donne alors :

$(x - \rho \nabla f(x) - x / x - y) \geq 0$ pour tout $y \in K$, et comme $\rho > 0$, ceci entraîne $(\nabla f(x) / x - y) \leq 0$ pour tout $y \in K$.

Or f est convexe donc $f(y) \geq f(x) + \nabla f(x)(y - x)$ pour tout $y \in K$, et donc $f(y) \geq f(x)$ pour tout $y \in K$, ce qui termine la démonstration.

■

Théorème 3.42 (Convergence de l'algorithme GPFK)**.**
Soit $f \in C^1(\mathbb{R}^n, \mathbb{R})$, et K convexe fermé non vide. On suppose que :

1. *il existe $\alpha > 0$ tel que $(\nabla f(x) - \nabla f(y) | x - y) \geq \alpha |x - y|^2$, pour tout $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$,*
2. *il existe $M > 0$ tel que $|\nabla f(x) - \nabla f(y)| \leq M |x - y|$ pour tout $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$,*

alors :

1. *il existe un unique élément $\bar{x} \in K$ solution de* (3.49)*,*
2. *si $0 < \rho < \dfrac{2\alpha}{M^2}$, la suite (x_k) définie par l'algorithme $(GPFK)$ converge vers $\bar{x}$ lorsque $n \to +\infty$.*

DÉMONSTRATION –

1. La condition 1. donne que f est strictement convexe et que $f(x) \to +\infty$ quand $|x| \to +\infty$. Comme K est convexe fermé non vide, il existe donc un unique $\bar{x}$ solution de (3.49).
2. On pose, pour $x \in \mathbb{R}^n$, $h(x) = p_K(x - \rho\nabla f(x))$. On a donc $x_{k+1} = h(x_k)$. Soit $0 < \rho < \frac{2\alpha}{M^2}$. Pour montrer que la suite $(x_k)_{k\in\mathbb{N}}$ converge, il suffit donc de montrer que h est strictement contractante. Grâce au lemme 3.43, on sait que p_K est contractante. Or h est définie par :
$$h(x) = p_K(\bar{h}(x)) \quad \text{où } \bar{h}(x) = x - \rho\nabla f(x).$$
On a déjà vu que $\bar{h}$ est strictement contractante (car $0 < \rho < \frac{2\alpha}{M^2}$, voir le théorème 3.19 page 224). Plus précisément, on a :
$$|\bar{h}(x) - \bar{h}(y)| \leq (1 - 2\alpha\rho + M^2\rho^2)|x - y|^2.$$
On en déduit que :
$$|h(x) - h(y)|^2 \leq |p_K(\bar{h}(x)) - p_K(\bar{h}(y))|^2 \leq |\bar{h}(x) - \bar{h}(y))|^2 \leq (1 - 2\alpha\rho + \rho^2M^2)|x - y|^2.$$
L'application h est donc strictement contractante. La suite $(x_k)_{k\in\mathbb{N}}$ est donc convergente. on note $\tilde{x}$ sa limite. il reste à montrer que $\tilde{x} = \bar{x}$. On remarque tout d'abord que $\tilde{x} \in K$ (car K est fermé). Puis, comme $\tilde{x}$ est un point fixe de h, on a $\tilde{x} = p_K(\tilde{x} - \rho\nabla f(\tilde{x}))$. La caractérisation de p_K donnée dans la proposition 3.40 donne alors
$$(\tilde{x} - \rho\nabla f(\tilde{x}) - \tilde{x}) \cdot (\tilde{x} - y) \geq 0 \text{ pour tout } y \in K,$$
ce qui donne $\nabla f(\tilde{x}) \cdot (y - \tilde{x}) \geq 0$ pour tout $y \in K$ et donc, comme f est convexe, $f(y) \geq f(\tilde{x}) + \nabla f(\tilde{x}) \cdot (y - \tilde{x}) \geq f(\tilde{x})$ pour tout $y \in K$. Ceci montre bien que $\tilde{x} = \bar{x}$.

■

Lemme 3.43 (Propriété de contraction de la projection orthogonale)**.** *Soit E un espace de Hilbert, $\|\cdot\|$ la norme et $(\cdot,\cdot)$ le produit scalaire, K un convexe fermé non vide de E et p_K la projection orthogonale sur K définie par la proposition 3.40, alors $\|p_K(x) - p_K(y)\| \leq \|x - y\|$ pour tout $(x, y) \in E^2$.*

DÉMONSTRATION – Comme E est un espace de Hilbert,
$$\|p_K(x) - p_K(y)\|^2 = (p_K(x) - p_K(y)|p_K(x) - p_K(y)).$$
On a donc
$$\begin{aligned}\|p_K(x) - p_K(y)\|^2 = &\quad (p_K(x) - x + x - y + y - p_K(y)|p_K(x) - p_K(y)) \\ = &\quad (p_K(x) - x|p_K(x) - p_K(y))_E + (x - y|p_K(x) - p_K(y)) + \\ &\quad (y - p_K(y)|p_K(x) - p_K(y)).\end{aligned}$$
Or $(p_K(x) - x|p_K(x) - p_K(y)) \leq 0$ et $(y - p_K(y)|p_K(x) - p_K(y)) \leq 0$, d'où :
$$\|p_K(x) - p_K(y)\|^2 \leq (x - y|p_K(x) - p_K(y)),$$
et donc, grâce à l'inégalité de Cauchy-Schwarz,
$$\|p_K(x) - p_K(y)\|^2 \leq \|x - y\| \, \|p_K(x) - p_K(y)\|,$$
ce qui permet de conclure. ■

Algorithme du gradient à pas optimal avec projection sur K (GPOK)
L'algorithme du gradient à pas optimal avec projection sur K s'écrit :

Initialisation	$x_0 \in K$
Itération	x_k connu
	$w_k = -\nabla f(x_k)$; calculer α_k optimal dans la direction w_k
	$x_{k+1} = p_K(x_k + \alpha_k \boldsymbol{w}^{(k)})$

La démonstration de convergence de cet algorithme se déduit de celle de l'algorithme à pas fixe.

Remarque 3.44. *On pourrait aussi utiliser un algorithme de type Quasi–Newton avec projection sur K.*

Les algorithmes de projection sont simples à décrire, mais ils soulèvent deux questions :

1. Comment calcule-t-on p_K ?
2. Que faire si K n'est pas convexe ?

On peut donner une réponse à la première question dans les cas simples :

Cas 1. On suppose ici que $K = C^+ = \{x \in \mathbb{R}^n,\ x = (x_1, \ldots, x_k)^t\ x_i \geq 0\ \forall i\}$.

Si $y \in \mathbb{R}^n\ y = (y_1 \ldots y_n)^t$, on peut montrer (exercice 137 page 264) que

$$(p_K(y))_i = y_i^+ = \max(y_i, 0), \quad \forall i \in \{1, \ldots, n\}$$

Cas 2. Soit $(\alpha_i)_{i=1,\ldots,n} \subset \mathbb{R}^n$ et $(\beta_i)_{i=1,\ldots,n} \subset \mathbb{R}^n$ tels que $\alpha_i \leq \beta_i$ pour tout $i = 1, \ldots, n$. Si

$$K = \prod_{i=1,n} [\alpha_i, \beta_i],$$

alors

$$(p_K(y))_i = \max(\alpha_i, \min(y_i, \beta_i)), \quad \forall i = 1, \ldots, n$$

Dans le cas d'un convexe K plus "compliqué", ou dans le cas où K n'est pas convexe, on peut utiliser des méthodes de dualité introduites dans le paragraphe suivant.

3.5.2 Méthodes de dualité

Supposons que les hypothèses suivantes sont vérifiées :

$$\left\{ \begin{array}{l} f \in C^1(\mathbb{R}^n, \mathbb{R}), \\ g_i \in C^1(\mathbb{R}^n, \mathbb{R}), \\ K = \{x \in \mathbb{R}^n,\ g_i(x) \leq 0\ i = 1, \ldots, p\},\ \text{et } K \text{ est non vide.} \end{array} \right. \tag{3.62}$$

On définit un problème "primal" comme étant le problème de minimisation d'origine, c'est–à–dire

$$\left\{ \begin{array}{l} \bar{x} \in K, \\ f(\bar{x}) \leq f(x),\ \text{pour tout } x \in K, \end{array} \right. \tag{3.63}$$

On définit le "lagrangien" comme étant la fonction L définie de $\mathbb{R}^n \times \mathbb{R}^p$ dans $\mathbb{R}$ par :

$$L(x, \lambda) = f(x) + \lambda \cdot g(x) = f(x) + \sum_{i=1}^{p} \lambda_i g_i(x), \tag{3.64}$$

avec $g(x) = (g_1(x), \ldots, g_p(x))^t$ et $\lambda = (\lambda_1, \ldots, \lambda_p)^t$.
On note C^+ l'ensemble défini par

$$C^+ = \{\lambda \in \mathbb{R}^p,\ \lambda = (\lambda_1, \ldots, \lambda_p)^t, \lambda_i \geq 0 \text{ pour tout } i = 1, \ldots, p\}.$$

Remarque 3.45. *Sous les hypothèses du théorème de Kuhn-Tucker, si $\bar{x}$ est solution du problème primal* (3.63) *alors il existe $\lambda \in C^+$ tel que $D_1 L(\bar{x}, \lambda) = 0$ (c'est–à–dire $Df(\bar{x}) + \lambda \cdot Dg(\bar{x}) = 0$) et $\lambda \cdot g(\bar{x}) = 0$.*

On définit alors l'application M de $\mathbb{R}^p$ dans $\mathbb{R}$ par :

$$M(\lambda) = \inf_{x \in \mathbb{R}^n} L(x, \lambda),\ \text{pour tout } \lambda \in \mathbb{R}^p. \tag{3.65}$$

On peut donc remarquer que $M(\lambda)$ réalise le minimum (en x) du problème sans contrainte, qui s'écrit, pour $\lambda \in \mathbb{R}^p$ fixé :

$$\left\{ \begin{array}{l} x \in \mathbb{R}^n \\ L(x, \lambda) \leq L(y, \lambda) \text{ pour tout } x \in \mathbb{R}^n, \end{array} \right. \tag{3.66}$$

Lemme 3.46. *L'application M de $\mathbb{R}^p$ dans $\mathbb{R}$ définie par* (3.65) *est concave (ou encore l'application -M est convexe), c'est–à–dire que pour tous $\lambda, \mu \in \mathbb{R}^p$ et pour tout $t \in]0,1[$ on a $M(t\lambda + (1-t)\mu) \geq tM(\lambda) + (1-t)\mathcal{M}(u)$*

DÉMONSTRATION – Soit $\lambda, \mu \in \mathbb{R}^p$ et $t \in]0,1[$; on veut montrer que $M(t\lambda + (1-t)\mu) \geq tM(\lambda) + (1-t)M(\mu)$.
Soit $x \in \mathbb{R}^n$, alors :

$$\begin{aligned} L(x, t\lambda + (1-t)\mu) &= f(x) + (t\lambda + (1-t)\mu)g(x) \\ &= tf(x) + (1-t)f(x) + (t\lambda + (1-t)\mu)g(x). \end{aligned}$$

On a donc $L(x, t\lambda + (1-t)\mu) = tL(x,\lambda) + (1-t)L(x,\mu)$. Par définition de M, on en déduit que pour tout $x \in \mathbb{R}^n$,

$$L(x, t\lambda + (1-t)\mu) \geq tM(\lambda) + (1-t)M(\mu)$$

Or, toujours par définition de M,

$$M(t\lambda + (1-t)\mu) = \inf_{x \in \mathbb{R}^n} L(x, t\lambda + (1-t)\mu) \geq tM(\lambda) + (1-t)M(\mu).$$

■

On considère maintenant le problème d'optimisation dit "dual" suivant :

$$\begin{cases} \mu \in C^+, \\ M(\mu) \geq M(\lambda) \quad \forall \lambda \in C^+. \end{cases} \tag{3.67}$$

Définition 3.47. *Soit $L : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ et $(x, \mu) \in \mathbb{R}^n \times C^+$. On dit que (x, μ) est un point selle de L sur $\mathbb{R}^n \times C^+$ si*

$$L(x,\lambda) \leq L(x,\mu) \leq L(y,\mu) \text{ pour tout } y \in \mathbb{R}^n \text{ et pour tout } \lambda \in C^+.$$

Proposition 3.48. *Sous les hypothèses* (3.62)*, soit L définie par $L(x,\lambda) = f(x) + \lambda g(x)$ et $(\bar{x}, \mu) \in \mathbb{R}^n \times C^+$ un point selle de L sur $\mathbb{R}^n \times C^+$.*
alors

1. *$\bar{x}$ est solution du problème* (3.63)*,*
2. *μ est solution de* (3.67)*,*
3. *$\bar{x}$ est solution du problème* (3.66) *avec $\lambda = \mu$.*

On admettra cette proposition.

Réciproquement, on peut montrer que (sous des hypothèses convenables sur f et g), si μ est solution de (3.67), et si $\bar{x}$ solution de (3.66) avec $\lambda = \mu$, alors $(\bar{x}, \mu)$ est un point selle de L, et donc $\bar{x}$ est solution de (3.63).

De ces résultats découle l'idée de base des méthodes de dualité : on cherche μ solution de (3.67). On obtient ensuite une solution $\bar{x}$ du problème (3.63), en cherchant $\bar{x}$ comme solution du problème (3.66) avec $\lambda = \mu$ (qui est un problème de minimisation sans contraintes). La recherche de la solution μ du problème dual (3.67) peut se faire par exemple par l'algorithme très classique d'Uzawa, que nous décrivons maintenant.

Algorithme d'Uzawa L'algorithme d'Uzawa consiste à utiliser l'algorithme du gradient à pas fixe avec projection (qu'on a appelé "GPFK", voir page 267) pour résoudre de manière itérative le problème dual (3.67). On cherche donc $\mu \in C^+$ tel que $M(\mu) \geq M(\lambda)$ pour tout $\lambda \in C^+$. On se donne $\rho > 0$, et on note p_{C^+} la projection sur le convexe C^+ (voir proposition 3.40 page 267). L'algorithme (GPFK) pour la recherche de μ s'écrit donc :

Initialisation : $\mu_0 \in C_+$
Itération : $\mu_{k+1} = p_{C_+}(\mu_k + \rho \nabla M(\mu_k))$

Pour définir complètement l'algorithme d'Uzawa, il reste à préciser les points suivants :

1. Calcul de $\nabla M(\mu_k)$,
2. calcul de $p_{C^+}(\lambda)$ pour λ dans $\mathbb{R}^n$.

On peut également s'intéresser aux propriétés de convergence de l'algorithme.

La réponse au point 2 est simple (voir exercice 137 page 264) : pour $\lambda \in \mathbb{R}^p$, on calcule $p_{C_+}(\lambda) = \gamma$ avec $\gamma = (\gamma_1, \ldots, \gamma_p)^t$ en posant $\gamma_i = \max(0, \lambda_i)$ pour $i = 1, \ldots, p$, où $\lambda = (\lambda_1, \ldots, \lambda_p)^t$.

La réponse au point 1. est une conséquence de la proposition suivante (qu'on admettra ici) :

Proposition 3.49. *Sous les hypothèses* (3.62)*, on suppose que pour tout $\lambda \in \mathbb{R}^n$, le problème* (3.66) *admet une solution unique, notée x_λ et on suppose que l'application définie de $\mathbb{R}^p$ dans $\mathbb{R}^n$ par $\lambda \mapsto x_\lambda$ est différentiable. Alors $M(\lambda) = L(x_\lambda, \lambda)$, M est différentiable en λ pour tout λ, et $\nabla M(\lambda) = g(x_\lambda)$.*

En conséquence, pour calculer $\nabla M(\lambda)$, on est ramené à chercher x_λ solution du problème de minimisation sans contrainte (3.66). On peut dont maintenant donner le détail de l'itération générale de l'algorithme d'Uzawa :

Itération de l'algorithme d'Uzawa. Soit $\mu_k \in C^+$ connu ;

1. On cherche $x_k \in \mathbb{R}^n$ solution de $\left\{ \begin{array}{l} x_k \in \mathbb{R}^n, \\ L(x_k, \mu_k) \leq L(x, \mu_k), \ \forall x \in \mathbb{R}^n \end{array} \right.$ (On a donc $x_k = x_{\mu_k}$)
2. On calcule $\nabla M(\mu_k) = g(x_k)$
3. $\overline{\mu}_{k+1} = \mu_k + \rho \nabla M(\mu_k) = \mu_k + \rho g(x_k) = ((\overline{\mu}_{k+1})_1, \ldots, (\overline{\mu}_{k+1})_p)^t$
4. $\mu_{k+1} = p_{C^+}(\overline{\mu}_{k+1})$, c'est–à-dire $\mu_{k+1} = ((\mu_{k+1})_1, \ldots, (\mu_{k+1})_p)^t$ avec $(\mu_{k+1})_i = \max(0, (\overline{\mu}_{k+1})_i)$ pour tout $i = 1, \ldots, p$.

L'exercice 139 donne un résulat de convergence contenant en particulier le cas très intéressant d'une fonctionnelle quadratique avec des contraintes affines "suffisamment" indépendantes (pour pouvoir appliquer le théorème de Kuhn-Tucker).

Remarque 3.50 (Sur l'algorithme d'Uzawa)**.**

1. *L'algorithme est très efficace si les contraintes sont affines : (i.e. si $g_i(x) = \alpha_i \cdot x + \beta_i$ pour tout $i = 1, \ldots, p$, avec $\alpha_i \in \mathbb{R}^n$ et $\beta_i \in \mathbb{R}$).*
2. *Pour avoir l'hypothèse 3 du théorème, il suffit que les fonctions g_i soient convexes. (On a dans ce cas existence et unicité de la solution x_λ du problème* (3.66) *et existence et unicité de la solution $\bar{x}$ du problème* (3.63)*.)*

3.5.3 Exercices (algorithmes pour l'optimisation avec contraintes)

Exercice 138 (Méthode de pénalisation)**.**
Soit f une fonction continue et strictement convexe de $\mathbb{R}^n$ dans $\mathbb{R}$, satisfaisant de plus :

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

Soit K un sous ensemble non vide, convexe (c'est–à–dire tel que $\forall (x, y) \in K^2$, $tx + (1-t)y \in K$, $\forall t \in]0, 1[$), et fermé de $\mathbb{R}^n$. Soit ψ une fonction continue de $\mathbb{R}^n$ dans $[0, +\infty[$ telle que $\psi(x) = 0$ si et seulement si $x \in K$. Pour $n \in \mathbb{N}$, on définit la fonction f_k par $f_k(x) = f(x) + n\psi(x)$.

1. Montrer qu'il existe au moins un élément $\bar{x}_k \in \mathbb{R}^n$ tel que $f_k(\bar{x}_k) = \inf_{x \in \mathbb{R}^n} f_k(x)$, et qu'il existe un unique élément $\bar{x}_K \in K$ tel que $f(\bar{x}_K) = \inf_{x \in K} f(x)$.
2. Montrer que pour tout $n \in \mathbb{N}$,
$$f(\bar{x}_n) \leq f_k(\bar{x}_n) \leq f(\bar{x}_K).$$
3. En déduire qu'il existe une sous-suite $(\bar{x}_{n_k})_{k \in \mathbb{N}}$ et $y \in K$ tels que $\bar{x}_{n_k} \to y$ lorsque $k \to +\infty$.

4. Montrer que $y = \bar{x}_K$. En déduire que toute la suite $(\bar{x}_k)_{n\in\mathbb{N}}$ converge vers $\bar{x}_K$.
5. Déduire de ces questions un algorithme (dit "de pénalisation") de résolution du problème de minimisation suivant :

$$\begin{cases} \text{Trouver } \bar{x}_K \in K; \\ f(\bar{x}_K) \leq f(x),\ \forall x \in K, \end{cases}$$

en donnant un exemple de fonction ψ.

Exercice 139 (Convergence de l'algorithme d'Uzawa)**.** *Corrigé en page 274*
Soient $n \geq 1$ $p \in \mathbb{N}^\star$. Soit $f \in C^1(\mathbb{R}^n, \mathbb{R})$ une fonction telle que

$$\exists \alpha > 0,\ (\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \alpha |x - y|^2,\ \forall x, y \in \mathbb{R}^n.$$

Soit $C \in M_{p,n}(\mathbb{R})$ (C est donc une matrice, à éléments réels, ayant p lignes et n colonnes) et $d \in \mathbb{R}^p$. On note $D = \{x \in \mathbb{R}^n, Cx \leq d\}$ et $\mathcal{C}^+ = \{u \in \mathbb{R}^p, u \geq 0\}$.
On suppose $D \neq \emptyset$ et on s'intéresse au problème suivant :

$$x \in D,\ \ f(x) \leq f(y),\ \forall y \in D. \tag{3.68}$$

1. Montrer que $f(y) \geq f(x) + \nabla f(x) \cdot (y - x) + \frac{\alpha}{2}|x - y|^2$ pour tout $x, y \in \mathbb{R}^n$.
2. Montrer que f est strictement convexe et que $f(x) \to \infty$ quand $|x| \to \infty$. En déduire qu'il existe une et une seule solution au problème (3.68).

Dans la suite, on note $\overline{x}$ cette solution.
Pour $u \in \mathbb{R}^p$ et $x \in \mathbb{R}^n$, on pose $L(x, u) = f(x) + u \cdot (Cx - d)$.

3. Soit $u \in \mathbb{R}^p$ (dans cette question, u est fixé). Montrer que l'application $x \to L(x, u)$ est strictement convexe (de $\mathbb{R}^n$ dans $\mathbb{R}$) et que $L(x, u) \to \infty$ quand $|x| \to \infty$ [Utiliser la question 1]. En déduire qu'il existe une et une seule solution au problème suivant :

$$x \in \mathbb{R}^n,\ \ L(x, u) \leq L(y, u),\ \forall y \in \mathbb{R}^n. \tag{3.69}$$

Dans la suite, on note x_u cette solution. Montrer que x_u est aussi l'unique élément de $\mathbb{R}^n$ t.q. $\nabla f(x_u) + C^t u = 0$.

4. On admet que le théorème de Kuhn-Tucker s'applique ici (cf. cours). Il existe donc $\overline{u} \in \mathcal{C}^+$ t.q. $\nabla f(\overline{x}) + C^t \overline{u} = 0$ et $\overline{u} \cdot (C\overline{x} - d) = 0$. Montrer que $(\overline{x}, \overline{u})$ est un point selle de L sur $\mathbb{R}^n \times \mathcal{C}^+$, c'est-à-dire :

$$L(\overline{x}, v) \leq L(\overline{x}, \overline{u}) \leq L(y, \overline{u}),\ \forall (y, v) \in \mathbb{R}^n \times \mathcal{C}^+. \tag{3.70}$$

Pour $u \in \mathbb{R}^p$, on pose $M(u) = L(x_u, u)$ (de sorte que $M(u) = \inf\{L(x, u),\ x \in \mathbb{R}^n\}$). On considère alors le problème suivant :

$$u \in \mathcal{C}^+,\ \ M(u) \geq M(v),\ \forall v \in \mathcal{C}^+. \tag{3.71}$$

5. Soit $(x, u) \in \mathbb{R}^n \times \mathcal{C}^+$ un point selle de L sur $\mathbb{R}^n \times \mathcal{C}^+$ (c'est-à-dire $L(x, v) \leq L(x, u) \leq L(y, u)$, pour tout $(y, v) \in \mathbb{R}^n \times \mathcal{C}^+$). Montrer que $x = \overline{x} = x_u$ (on rappelle que $\overline{x}$ est l'unique solution de (3.68) et x_u est l'unique solution de (3.69)) et que u est solution de (3.71). [On pourra commencer par montrer, en utilisant la première inégalité, que $x \in D$ et $u \cdot (Cx - d) = 0$.]
Montrer que $\nabla f(\overline{x}) + C^t u = 0$ et que $u = P_{\mathcal{C}^+}(u + \rho(C\overline{x} - d))$, pour tout $\rho > 0$, où $P_{\mathcal{C}^+}$ désigne l'opérateur de projection orthogonale sur $\mathcal{C}^+$. [on rappelle que si $v \in \mathbb{R}^p$ et $w \in \mathcal{C}^+$, on a $w = P_{\mathcal{C}^+} v \iff ((v - w) \cdot (w - z) \geq 0, \forall z \in \mathcal{C}^+)$.]

6. Déduire des questions 2, 4 et 5 que le problème (3.71) admet au moins une solution.

7. On admet que l'application $u \mapsto x_u$ est dérivable. Montrer que l'algorithme du gradient à pas fixe avec projection pour trouver la solution de (3.71) s'écrit (on désigne par $\rho > 0$ le pas de l'algorithme) :

   Initialisation. $u_0 \in \mathcal{C}^+$.

   Itérations. Pour $u_k \in \mathcal{C}^+$ connu ($k \geq 0$). On calcule $x_k \in \mathbb{R}^n$ t.q. $\nabla f(x_k) + C^t u_k = 0$ (montrer qu'un tel x_k existe et est unique) et on pose $u_{k+1} = P_{\mathcal{C}^+}(u_k + \rho(Cx_k - d))$.

Dans la suite, on s'intéresse à la convergence de la suite $(x_k, u_k)_{k\in\mathbb{N}}$ donnée par cet algorithme.

8. Soit ρ t.q. $0 < \rho < 2\alpha/\|C\|^2$ avec $\|C\| = \sup\{|Cx|,\ x \in \mathbb{R}^n \text{ t.q. } |x| = 1\}$. Soit $(\overline{x}, \overline{u}) \in \mathbb{R}^n \times \mathcal{C}^+$ un point selle de L sur $\mathbb{R}^n \times \mathcal{C}^+$ (c'est-à-dire vérifiant (3.70)) et $(x_k, u_k)_{k\in\mathbb{N}}$ la suite donnée par l'algorithme de la question précédente. Montrer que

$$|u_{k+1} - \overline{u}|^2 \leq |u_k - \overline{u}|^2 - \rho(2\alpha - \rho\|C\|^2)|x_k - \overline{x}|^2,\ \forall k \in \mathbb{R}^n.$$

   En déduire que $x_k \to \overline{x}$ quand $k \to \infty$.

   Montrer que la suite $(u_k)_{k\in\mathbb{N}}$ est bornée et que, si $\tilde{u}$ est une valeur d'adhérence de la suite $(u_k)_{k\in\mathbb{N}}$, on a $\nabla f(\overline{x}) + C^t\tilde{u} = 0$. En déduire que, si rang(C)=p, on a $u_k \to \overline{u}$ quand $k \to \infty$ et que $\overline{u}$ est *l'unique* élément de $\mathcal{C}^+$ t.q. $\nabla f(\overline{x}) + C^t\overline{u} = 0$.

Exercice 140 (Méthode de relaxation avec Newton problèmes sous contrainte)**.**
On considère le problème :

$$\begin{cases} \overline{x} \in K, \\ f(\overline{x}) \leq f(x), \forall x \in K, \end{cases} \tag{3.72}$$

où $K \subset \mathbb{R}^n$.

(a) On prend ici $K = \prod_{i=1,n}[a_i, b_i]$, où $(a_i, b_i) \in \mathbb{R}^2$ est tel que $a_i \leq b_i$. On considère l'algorithme suivant :

$$\left\{\begin{array}{ll} \text{Initialisation :} & x^{(0)} \in E, \\ \text{Itération } n : & x^{(k)} \text{ connu, } (n \geq 0) \\ & \text{Calculer } x_1^{(k+1)} \in [a_1, b_1] \text{ tel que :} \\ & f(x_1^{(k+1)}, x_2^{(k)}, x_3^{(k)}, \ldots, x_n^{(k)}) \leq f(\xi, x_2^{(k)}, x_3^{(k)}, \ldots, x_n^{(k)}), \text{ pour tout } \xi \in [a_1, b_1], \\ & \text{Calculer } x_2^{(k+1)} \in [a_2, b_2] \text{ tel que :} \\ & f(x_1^{(k+1)}, x_2^{(k+1)}, x_3^{(k)}, \ldots, x_n^{(k)}) \leq f(x_1^{(k+1)}, \xi, x_3^{(k)}, \ldots, x_n^{(k)}), \\ & \qquad \text{pour tout } \xi \in [a_2, b_2], \\ & \qquad\qquad \ldots \\ & \text{Calculer } x_k^{(k+1)} \in [a_k, b_k], \text{ tel que :} \\ & f(x_1^{(k+1)}, \ldots, x_{k-1}^{(k+1)}, x_k^{(k+1)}, x_{(k+1)}^{(k)}, \ldots, x_n^{(k)}) \\ & \qquad \leq f(x_1^{(k+1)}, \ldots, x_{k-1}^{(k+1)}, \xi, x_{(k+1)}^{(k)}, \ldots, x_n^{(k)}), \text{ pour tout } \xi \in [a_k, b_k], \\ & \qquad\qquad \ldots \\ & \text{Calculer } x_n^{(k+1)} \in [a_n, b_n] \text{ tel que :} \\ & f(x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_{n-1}^{(k+1)}, x_n^{(k+1)}) \leq f(x_1^{(k+1)}, \ldots, x_{n-1}^{(k+1)}, \xi), \\ & \qquad \text{pour tout } \xi \in [a_n, b_n]. \end{array}\right. \tag{3.73}$$

Montrer que la suite $x^{(k)}$ construite par l'algorithme (3.73) est bien définie et converge vers $\overline{x}$ lorsque n tend vers $+\infty$, où $\overline{x} \in K$ est tel que $f(\overline{x}) \leq f(x)$ pour tout $x \in K$.

(b) On prend maintenant $n = 2$, f la fonction de $\mathbb{R}^2$ dans $\mathbb{R}$ définie par $f(x) = x_1^2 + x_2^2$, et $K = \{(x_1, x_2)^t \in \mathbb{R}^2; x_1 + x_2 \geq 2\}$. Montrer qu'il existe un unique élément $\overline{x} = (\overline{x}_1, \overline{x}_2)^t$ de K tel que $f(\overline{x}) = \inf_{x \in \mathbb{R}^2} f(x)$. Déterminer $\overline{x}$.

On considère l'algorithme suivant pour la recherche de $\overline{x}$:

$$\left\{ \begin{array}{ll} \text{Initialisation :} & x^{(0)} \in E, \\ \text{Itération } n : & x^{(k)} \text{ connu, } (n \geq 0) \\ & \text{Calculer } x_1^{(k+1)} \geq 2 - x_2^{(k)} \text{ tel que :} \\ & f(x_1^{(k+1)}, x_2^{(k)}) \leq f(\xi, x_2^{(k)}), \text{ pour tout } \xi \geq 2 - x_2^{(k)}, \\ & \text{Calculer } x_2^{(k+1)} \geq 2 - x_1^{(k)} \text{ tel que :} \\ & f(x_1^{(k+1)}, x_2^{(k+1)}) \leq f(x_1^{(k+1)}, \xi), \text{ pour tout } \xi \geq 2 - x_1^{(k)}. \end{array} \right. \tag{3.74}$$

Montrer (éventuellement graphiquement) que la suite construite par l'algorithme ci-dessus ne converge vers $\overline{x}$ que si l'une des composantes de $x^{(0)}$ vaut 1.

3.5.4 Corrigés

Exercice 139 page 272 (Convergence de l'algorithme d'Uzawa)

1. Cette question a déjà été corrigée. Soit $x, y \in \mathbb{R}^n$. En posant $\varphi(t) = f(ty + (1-t)x)$, on remarque que

$$f(y) - f(x) = \varphi(1) - \varphi(0) = \int_0^1 \varphi'(t)dt = \int_0^1 \nabla f(x + t(y-x)) \cdot (y-x)dt,$$

et donc

$$f(y) - f(x) - \nabla f(x) \cdot (y-x) = \int_0^1 (\nabla f(x + t(y-x)) - \nabla f(x)) \cdot t(y-x) \frac{1}{t} dt \geq \alpha \int_0^1 t|y-x|^2 dt = \frac{\alpha}{2}|y-x|^2.$$

2. Montrer que f est strictement convexe et que $f(x) \to \infty$ quand $|x| \to \infty$. En déduire qu'il existe une et une seule solution au problème (3.68).

Cette question a aussi déjà été corrigée. La question précédente donne $f(y) \geq f(x) + \nabla f(x) \cdot (y-x)$ pour tout $x, y \in \mathbb{R}^n$. Ce qui montre que f est strictement convexe. Elle donne aussi, pour tout $x \in \mathbb{R}^n$,

$$f(x) \geq f(0) + \nabla f(0) \cdot x + \frac{\alpha}{2}|x|^2 \to +\infty \text{ quand } |x| \to +\infty.$$

De ces deux propriétés de f on déduit l'existence et l'unicité de la solution au problème (3.68).

3. L'application $x \mapsto u \cdot (Cx - d)$ est affine et donc convexe. Comme f est strictement convexe, on en déduit que $x \mapsto L(x, u)$ est aussi strictement convexe.

La question précédente donne $L(x, u) \geq f(0) + \nabla f(0) \cdot x + u \cdot (Cx - d) + \frac{\alpha}{2}|x|^2$. On en déduit que $L(x, u) \to +\infty$ quand $|x| \to +\infty$.

De ces deux propriétés de $L(\cdot, u)$ on déduit l'existence et l'unicité de la solution au problème (3.69).

Comme $L(\cdot, u)$ est strictement convexe, x_u est aussi l'unique point qui annule $\nabla L(\cdot, u)$ (c'est-à-dire le gradient de l'application $x \mapsto L(x, u)$). Ceci donne bien que x_u est l'unique point de $\mathbb{R}^n$ tel que $\nabla f(x_u) + C^t u = 0$.

4. La question précédente nous dit que $x_{\bar{u}}$ est l'unique point de $\mathbb{R}^n$ tel que $\nabla f(x_{\bar{u}}) + C^t \bar{u} = 0$. Comme $\nabla f(\overline{x}) + C^t \overline{u} = 0$, on a donc $\bar{x} = x_{\bar{u}}$ et donc

$$L(\overline{x}, \overline{u}) \leq L(y, \overline{u}) \text{ pour tout } y \in \mathbb{R}^n.$$

Soit maintenant $v \in \mathcal{C}^+$. On a, comme $C\overline{x} \leq d$ (car $\overline{x} \in D$) et $\overline{u} \cdot (C\overline{x} - d) = 0$,

$$L(\overline{x}, v) = f(\overline{x}) + v \cdot (C\overline{x} - d) \leq f(\overline{x}) = f(\overline{x}) + \overline{u} \cdot (C\overline{x} - d) = L(\overline{x}, \overline{u}).$$

5. On a $L(x,v) \leq L(x,u)$ pour tout $v \in \mathcal{C}^+$ et donc

$$(v-u)\cdot(Cx-d) \leq 0 \text{ pour tout } v \in \mathcal{C}^+.$$

On note $u = (u_1,\dots,u_p)^t$. Soit $i \in \{1,\dots,p\}$. en prenant $v = (v_1,\dots,v_p)^t$ avec $v_j = u_j$ si $j \neq i$ et $v_i = u_i + 1$ (on a bien $v \in \mathcal{C}^+$), la formule précédente nous montre que la i-ieme composante de $(Cx-d)$ est négative. On a donc $x \in D$.

Soit maintenant $i \in \{1,\dots,p\}$ tel que $u_i > 0$. En prenant $v = (v_1,\dots,v_p)^t$ avec $v_j = u_j$ si $j \neq i$ et $v_i = 0$, la formule précédente nous montre que la i-ieme composante de $(Cx-d)$ est positive. Elle donc nécessairement nulle. Ceci nous donne bien que $u\cdot(Cx-d) = 0$.

On utilise maintenant le fait que $L(x,u) \leq L(y,u)$ pour tout $y \in \mathbb{R}^n$. Ceci donne, bien sûr, que $x = x_u$. Cela donne aussi que

$$f(x) + u\cdot(Cx-d) \leq f(y) + u\cdot(Cy-d).$$

Comme on sait que $u\cdot(Cx-d) = 0$ et comme $u\cdot(Cy-d) \leq 0$ si $y \in D$, on en déduit que $f(x) \leq f(y)$ pour tout $y \in D$, et donc $x = \overline{x}$.

Enfin, $L(x,u) = L(x_u,u) = M(u)$ et $L(x,v) \geq L(x_v,v) = M(v)$. Comme $L(x,v) \leq L(x,u)$ pour tout $v \in \mathcal{C}^+$, on a donc $M(v) \leq M(u)$ pour tout $v \in \mathcal{C}^+$.

On passe maintenant à la seconde partie de cette question. On a vu à la question 3 que $\nabla f(x_u) + C^t u = 0$. Comme $\overline{x} = x_u$, on a donc $\nabla f(\overline{x}) + C^t u = 0$.

Puis pour montrer que $u = P_{\mathcal{C}^+}(u + \rho(C\overline{x} - d))$, on utilise le rappel. Pour tout $z \in \mathcal{C}^+$ on a, en utilisant $u\cdot(C\overline{x}-d) = 0$ et $\overline{x} \in D$,

$$(u+\rho(C\overline{x}-d)-u)\cdot(u-z) = \rho(C\overline{x}-d)\cdot(u-z) = -\rho(C\overline{x}-d)\cdot z \geq 0.$$

Ceci donne bien que $u = P_{\mathcal{C}^+}(u+\rho(C\overline{x}-d))$.

6. La question 2 donne l'existence de $\overline{x}$ solution de (3.68). Puis la question 4 donne l'existence de $\overline{u}$ tel que $(\overline{x},\overline{u})$ est solution de (3.70). Enfin, la question 5 donne alors que $\overline{u}$ est solution de (3.71).

7. Les itérations de l'algorithme du gradient à pas fixe avec projection pour trouver la solution de (3.71) s'écrivent

$$u_{k+1} = P_{\mathcal{C}^+}(u_k + \rho\nabla M(u_k)).$$

Comme $M(u) = L(x_u,u)$ et x_u annule le gradient de l'application $x \mapsto L(x,u)$, la dérivation de fonctions composées (que l'on peut appliquer car l'application $u \mapsto x_u$ est supposée dérivable) nous donne $\nabla M(u) = Cx_u - d$. Comme $x_{u_k} = x_k$, on en déduit que

$$u_{k+1} = P_{\mathcal{C}^+}(u_k + \rho(Cx_k - d)).$$

8. Comme $(\overline{x},\overline{u})$ est un point selle de L sur $\mathbb{R}^N \times \mathcal{C}^+$, la question 5 nous donne

$$\overline{u} = P_{\mathcal{C}^+}(\overline{u} + \rho(C\overline{x} - d)).$$

L'opérateur $P_{\mathcal{C}^+}$ étant contractant, on obtient, avec la question précédente, pour tout k,

$$\begin{aligned}|u_{k+1}-\overline{u}|^2 &\leq |u_k + \rho(Cx_k - d) - (\overline{u} + \rho(C\overline{x}-d))|^2 \\ &= |u_k - \overline{u}|^2 + 2\rho(u_k-\overline{u})\cdot C(x_k-\overline{x}) + \rho^2|C(x_k-\overline{x})|^2.\end{aligned}$$

Comme $C^t u_k = -\nabla f(x_k)$ et $C^t\overline{u} = -\nabla f(\overline{x})$, on obtient (avec l'hypothèse sur ∇f)

$$\begin{aligned}|u_{k+1}-\overline{u}|^2 &\leq |u_k-\overline{u}|^2 - 2\rho(\nabla f(x_k) - \nabla f(\overline{x}))\cdot(x_k-\overline{x}) + \rho^2|C(x_k-\overline{x})|^2 \\ &\leq |u_k-\overline{u}|^2 - 2\rho\alpha|x_k-\overline{x}|^2 + \rho^2|C(x_k-\overline{x})|^2 \leq |u_k-\overline{u}|^2 - \rho(2\alpha - \rho\|C\|^2)|x_k-\overline{x}|^2.\end{aligned}$$

Comme $2\alpha - \rho\|C\|^2 > 0$, ceci montre que la suite $(u_k - \overline{u})_{k\in\mathbb{N}}$ est décroissante (positive) et donc convergente. Il suffit alors de remarquer que

$$|x_k - \overline{x}|^2 \leq \frac{1}{\rho(2\alpha - \rho\|C\|^2)}(|u_k - \overline{u}|^2 - |u_{k+1} - \overline{u}|^2)$$

pour en déduire que $x_k \to \overline{x}$ quand $k \to +\infty$.

La suite $(u_k - \overline{u})_{k\in\mathbb{N}}$ est convergente. La suite $(u_k)_{k\in\mathbb{N}}$ est donc bornée. Si $\tilde{u}$ est une valeur d'adhérence de la suite $(u_k)_{k\in\mathbb{N}}$, en passant à la limite sur l'équation $\nabla f(x_k) + C^t u_k = 0$ on obtient $\nabla f(\overline{x}) + C^t\tilde{u} = 0$.

Si rang(C)=p, on a aussi rang(C^t)=p. L'application $u \mapsto C^t u$ est de $\mathbb{R}^p$ dans $\mathbb{R}^n$, on a donc $\dim(\mathrm{Ker}C^t) = p-\mathrm{rang}(C^t) = 0$. Ceci prouve qu'il existe un unique $\tilde{u}$ tel que $C^t\tilde{u} = -\nabla f(\overline{x})$. La suite $(u_k)_{k\in\mathbb{N}}$ n'a alors qu'une seule valeur d'adhérence et elle est donc convergente vers $\overline{u}$ et $\overline{u}$ est l'unique élément de $\mathcal{C}^+$ t.q. $\nabla f(\overline{x}) + C^t\overline{u} = 0$.