Optimisation convexe

Bernhard Beckermann
Laboratoire Paul Painlevé UMR 8524
Université de Lille
59655 Villeneuve d'Ascq CEDEX
e-mail : Bernhard.Beckermann@univ-lille.fr

Chapitre 1 avec exercices, version du 30/08/2022

Table des matières

Chapitre 1

Introduction

1.1 A propos de ce texte

Ce document contient le cours et les exercices du cours ''Optimisation convexe" dispensé en automne 2022 dans le Master 1 de Mathématiques et Applications de l'Université de Lille.

Dans le chapitre 1, nous commencerons par rappeler quelques propriétés élémentaires des ensembles convexes et des fonctions convexes, puis nous reviendrons sur le calcul différentiel des fonctions à plusieurs variables. Enfin, nous poserons le problème de minimisation sous contraintes, illustré par des exemples. Le chapitre 2 aura comme objectif de caractériser l'optimalité à l'aide d'un système d'équations non-linéaires dit de KKT. Dans le chapitre 3 nous aborderons le problème de minimisation d'une fonction convexe sur $\mathbb{R}^d$, c'est-à-dire, sans contraintes. Nous analyserons finalement la convergence de divers algorithmes de type quasi-Newton, en particulier la fameuse méthode de plus forte descente. ...

1.2 Ensembles convexes et fonctions convexes

Dans la suite on travaillera dans l'espace vectoriel $\mathbb{R}^d$.

1.2.1. Définition d'un convexe=ensemble convexe

$C \subset \mathbb{R}^d$ est dit convexe si $\forall x, y \in C \forall t \in [0,1] : tx + (1-t)y \in C$.
Exemples :

(a) *la boule fermée $\{x \in \mathbb{R}^d : \|x\| \leq r\}$ (pour toute norme), aussi la boule ouverte ;*

(b) *le segment $[x,y] = \{tx + (1-t)y : t \in [0,1]\}$, l'enveloppe convexe $conv(a_1, ..., a_m) = \{\sum_{j=1}^m t_j a_j : \sum_{j=1}^m t_j = 1, \forall j : t_j \geq 0\}$ (par exemple le triangle $conv(a_1, a_2, a_3)$) ;*

(c) *un sous-espace vectoriel ;*

(d) *un sous-espace affine $C : \forall x, y \in C \forall t \in \mathbb{R} : tx + (1-t)y \in C$ (par exemple une droite) ;*

(e) *toute intersection (finie ou infinie) de convexes ;*

(f) *si C est convexe alors aussi sa fermeture $Clos(C)$ et son intérieur $Int(C)$;*

(g) *si C_1, X_2 sont convexes et $t \in \mathbb{R}$ alors $C_1 + C_2 = \{x + y : x \in C_1, y \in C_2\}$ (somme de Minkowski), $C_1 \times C_2 = \{\begin{pmatrix} x \\ y \end{pmatrix} : x \in C_1, y \in C_2\}$ (produit cartésien) et tC_1 sont convexes ;*

(g) *si $C \in \mathbb{R}^p \times \mathbb{R}^d$ est convexe alors aussi sa projection $\{x \in \mathbb{R}^p : \begin{pmatrix} x \\ y \end{pmatrix} \in C$ pour un $y \in \mathbb{R}^d\}$ est convexe.*

Voici une classe particulière de convexes fermés qui jouent un rôle important en optimisation linéaire.

1.2.2. Définition d'un polyèdre

Soient $A \in \mathbb{R}^{1\times d}$ (un vecteur ligne) et $a \in \mathbb{R}$, alors l'hyper-plan $H(A,a) = \{x \in \mathbb{R}^d : Ax = a\}$ et le demi-espace $H^+(A,a) = \{x \in \mathbb{R}^d : Ax \geq a\}$ sont des convexes fermés. Toute intersection finie de demi-espaces est aussi un convexe fermé, dit **polyèdre**. *Nous adaptons l'écriture*

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} \leq y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \quad ssi \quad \forall j = 1,2,...,m : \quad x_j \leq y_j$$

(un ordre partiel sur $\mathbb{R}^m$), permettant d'écrire un polyèdre comme

$$\bigcap_{j=1}^{m} H^+(A_j, a_j) = \{x \in \mathbb{R}^d : Ax \geq a\}, \quad avec \quad A = \begin{bmatrix} A_1 \\ \vdots \\ A_m \end{bmatrix} \in \mathbb{R}^{m\times d}, \quad a = \begin{bmatrix} a_1 \\ \vdots \\ a_m \end{bmatrix} \in \mathbb{R}^m.$$

1.2.3. Définition d'une fonction (strictement) convexe

Soit $C \subset \mathbb{R}^d$ convexe alors $f : C \mapsto \mathbb{R}$ est dite convexe si

$$\forall x, y \in C \forall t \in]0,1[: \quad f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

(sur le segment, le graphe de la sécante en x, y de f est au-dessus du graphe de la fonction). f est dite strictement convexe si on a l'inégalité stricte

$$\forall x, y \in C, x \neq y, \forall t \in]0,1[: \quad f(tx + (1-t)y) < tf(x) + (1-t)f(y)$$

f est dite concave si $-f$ est convexe.
Exemples :

(a) *$f(x) = x^2$, $f(x) = |x|$, $f(x) = \exp(x)$, $f(x) = \exp(-x)$ sont convexes sur $\mathbb{R}$, $f(x) = 1/x$, $f(x) = \log(1/x)$ sont convexes sur $(0, +\infty)$;*

(b) *si $f_1, f_2 : C \mapsto \mathbb{R}$ sont convexes alors aussi $\max(f_1, f_2)$ et $f_1 + f_2$ (par exemple $|x| = \max(-x, x)$). Plus généralement, le supremum et la somme d'un nombre fini ou infini de fonctions convexes est convexe ;*

(c) *une fonction affine $h(x) = Ax + a$ avec $A \in \mathbb{R}^{1\times d}, a \in \mathbb{R}$ est convexe et concave sur $\mathbb{R}^d$;*

(d) *si $g : \mathbb{R}^m \mapsto \mathbb{R}$ est convexe sur $\mathbb{R}^m$ et $A \in \mathbb{R}^{m\times d}$, $a \in \mathbb{R}^m$ alors $f(x) = g(Ax + a)$ est convexe sur $\mathbb{R}^d$;*

(e) *si $h : C \mapsto C_2 \subset \mathbb{R}$ est convexe et $g : C_2 \mapsto \mathbb{R}$ est convexe et croissant sur C_2 alors la composition $f = g \circ h$ est convexe (N.B. : C_2 est forcement un intervalle (fermé ou pas)) ;*

(f) *toute norme sur $\mathbb{R}^d$ est convexe sur $\mathbb{R}^d$, mais aucune norme est strictement connexe ;*

(g) *la norme euclidienne au carré $f(x) = \|x\|^2 = x_1^2 + x_2^2 + ... + x_d^2$ est strictement convexe sur $\mathbb{R}^d$;*

(h) *si $f : \mathbb{R}^d \mapsto \mathbb{R}$ est convexe et $^y \in \mathbb{R}^d$ alors* **l'ensemble de niveau** *$C(y) = \{x \in \mathbb{R}^d : f(x) \leq f(y)\}$ est convexe.*

Parfois on définit aussi des fonctions convexes $f : C \mapsto \overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ (par exemple par prolongation d'une fonction $f : C_1 \mapsto \mathbb{R}$ convexe en posant $f(x) = +\infty$

pour tout $x \in C \setminus C_1$), ici il suffit de vérifier notre inégalité pour $x, y \in dom(f) = \{z \in C : f(z) < \infty\}$ (notons que $dom(f)$ doit être convexe).

1.2.4. Définition et Lemme

On dira que une matrice $H \in \mathbb{R}^{d\times d}$ est ssdp=symétrique semi-définie positive (ou sdp=symétrique définie positive) si $A = A^T$, et si $\forall x \in \mathbb{R}^d$ nous avons $x^T Hx \geq 0$ (et $x^T Hx \neq 0$ si $x \neq 0$, respectivement). On sait que une matrice ssdp se factorise $H = B^T B$ avec $B \in \mathbb{R}^{d\times d}$ (et B inversible si H est sdp).
Alors une forme quadratique $f(x) = x^T Hx + hx$ est convexe sur $\mathbb{R}^d$ si H est ssdp, et f est strictement convexe si H est sdp.

Démonstration. Comme $x \mapsto hx$ est une fonction affine, il suffit de considérer le cas $h = 0$. Soient $x, y \in \mathbb{R}^d$ et $t \in]0, 1[$. Avec $B \in \mathbb{R}^{d\times d}$ comme ci-dessus, posons $\widetilde{x} := Bx, \widetilde{y} = By$, alors

$$\begin{aligned} &tf(x) + (1-t)f(y) - f(tx + (1-t)y) = t\|\widetilde{x}\|^2 + (1-t)\|\widetilde{y}\|^2 - \|t\widetilde{x} + (1-t)\widetilde{y}\|^2 \\ &= t\|\widetilde{x}\|^2 + (1-t)\|\widetilde{y}\|^2 - t^2\|\widetilde{x}\|^2 - 2t(1-t)\widetilde{x}^T\widetilde{y} - (1-t)^2\|\widetilde{y}\|^2 \\ &= t(1-t)\|\widetilde{x} - \widetilde{y}\|^2 = t(1-t)\|B(x-y)\|^2 \end{aligned}$$

est ≥ 0, et > 0 si H est spd et alors B inversible, et $x \neq y$. □

1.2.5. Exercice

Démontrer toutes les propriétés énoncées dans 1.2.1 et 1.2.3.

1.2.6. Exercices :

(a) *Soit $C \subset \mathbb{R}^d$ un convexe , $x_1, x_2, \cdots, x_k \in C$, $\theta_1, \cdots, \theta_k \geq 0$, $\sum_{i=1}^k \theta_i = 1$.*
Monter que $\theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k \in C$.

(b) *Montrer que C est un convexe ssi l'intersection avec toute droite est un convexe.*

(c) *Avec A une matrice symétrique définie positive, un ellipsoïde centré en $x_0 \in \mathbb{R}^d$ s'écrit comme $\mathcal{E} = \{x \in \mathbb{R}^d : (x - x_0)^T A(x - x_0) \leq 1\}$. Vérifier que $\mathcal{E} = \{By + x_0 : \|y\| \leq 1\}$ pour une matrice B appropriée, et que $\mathcal{E}$ est convexe.*

(d) *Supposons que C vérifie la propriété de convexité du point milieu, i.e., $\forall a, b \in C \quad \frac{1}{2}(a+b) \in C$. Montrer que si C est de plus fermé alors C est convexe.*

1.2.7. Exercices :

(a) *Montrer l'équivalence*

$$\begin{aligned} & C \text{ est un sous-espace affine} \\ \iff & \exists x \in C : C - \{x\} \text{ est un sous-espace vectoriel} \\ \iff & \forall x \in C : C - \{x\} \text{ est un sous-espace vectoriel.} \end{aligned}$$

En déduire que un sous-espace affine C est un sous(espace vectoriel ssi $0 \in C$.

(b) *M.q. tout sous-espace affine (ou sous-espace vectoriel) du $\mathbb{R}^d$ peut s'écrire comme $\{By + b : y \in \mathbb{R}^d\}$ ou alors comme $\{x \in \mathbb{R}^d : Ax = a\}$, avec des matrices A, B, a, b à déterminer.*

1.2.8. Exercice

Soit $C \subset \mathbb{R}^d$ convexe, $f : C \mapsto \mathbb{R}$. L'épigraphe de f est défini par

$$epi(f) = \left\{ \begin{bmatrix} x \\ r \end{bmatrix} : x \in C, r \in \mathbb{R}, f(x) \leq r \right\}.$$

Montrer que f est convexe ssi $epi(f)$ est convexe.

1.3 Position du problème et exemples

1.3.1. Position du problème et exemples

Pour $C \subset \mathbb{R}^d$ convexe et fermé et $f : C \mapsto \mathbb{R}$ convexe, un **programme convexe** *(CP) consiste à chercher la valeur dite* **valeur optimale**

$$(CP) : \quad \inf\{f(x) : x \in C\}$$

et, si possible, trouver un $\underline{x}$ réalisant l'infimum dit **solution optimale**.[1] *Souvent,*

$$C = \{x \in \mathbb{R}^d : \underbrace{g_j(x) \leq 0}_{\text{contraintes unilaterales}} \quad \text{pour } j = 1, ..., p\} \cap \{x \in \mathbb{R}^d : \underbrace{Ax = a}_{\text{contraintes bilaterales}} \}$$

qui est un convexe fermé si les $g_j : \mathbb{R}^d \mapsto \mathbb{R}$ sont convexes et continues, et $A \in \mathbb{R}^{m \times d}, a \in \mathbb{R}^m$.
Un élément de C est dit **solution réalisable** *ou* **réalisable**.
La fonction f est dite **objectif**.
Une contrainte $g_j(x) \leq 0$ est dite active en un $\underline{x}$ réalisable si $g_j(\underline{x}) = 0$.

Regardons quelques cas particuliers, les programmes linéaires (étudiés plus en détail au module OLD eu S2), et les programmes quadratiques.

---

1. Ceci est possible par exemple d'après le théorème de Bolzano-Weierstrass si de plus f est continue et C est compact, mais pas toujours, voir par exemple $C = \mathbb{R}$ et $f(x) = \exp(x)$.

1.3.2. Exemple : programme linéaire

$$(LP): \quad \min\{fx : Ax \geq a\}, \quad f \in \mathbb{R}^{1\times d}, x \in \mathbb{R}^d, A \in \mathbb{R}^{m\times d}, a \in \mathbb{R}^m$$

est dit programme linéaire (on minimise une fonction affine sur un polyèdre). Voici trois exemples

(a)

$$\min\{3x_1+4x_2 : x_1+2x_2 = 7, x_1 \geq 0, x_2 \geq 0\} = \min\{\begin{bmatrix}3\\4\end{bmatrix}^T x : \begin{bmatrix}1 & 2\\ -1 & -2\\ 1 & 0\\ 0 & 1\end{bmatrix} x \geq \begin{bmatrix}7\\ -7\\ 0\\ 0\end{bmatrix}.$$

On cherche à trouver un plan de production $x \in \mathbb{R}^2$ minimisant le temps d'utilisation d'une machine sachant que les deux objets utilisent une autre ressource commune (7 unités disponibles).

(b) Problème de Chebyshev : étant donné $a_j \in \mathbb{R}^d$ et $b_j \in \mathbb{R}$ pour $j = 1, ..., m$, trouver

$$\min_{x\in\mathbb{R}^d} \max_{j=1,...,m} |a_j^T x - b_j|,$$

un problème de régression ou le plus grand écart est minimisé (et pas la somme des carrées des écarts). Ce problème n'est a priori pas linéaire, mais peut être reformulé comme un programme linéaire

$$\min_{x,\gamma}\{\gamma = (0, ..., 0, 1)\begin{bmatrix}x\\ \gamma\end{bmatrix} : \forall j = 1, ..., m : -\gamma \leq a_j^T x - b_j \leq \gamma\}.$$

(c) *Trouver la plus grande boule contenue dans le polyèdre* $\{x \in \mathbb{R}^d : Ax \geq b\}$. *Une boule centrée en* $z \in \mathbb{R}$ *de rayon* r *est incluse dans un demi-plan* $H^+(A_j, b_j)$ *ssi* $A_j(z - r\frac{A_j^T}{\|A_j^T\|}) = A_j z - \|A_j^T\| r \geq b_j$. *Donc on se ramène au problème*

$$\min_{r,z}\{-r = (0, ..., 0, -1)\begin{bmatrix} z \\ r \end{bmatrix} : \forall j = 1, ..., m : [A_j, -\|A_j^T\|]\begin{bmatrix} z \\ r \end{bmatrix} \geq b_j\}.$$

(d) *Un programme linéaire sous la forme* $\min\{fx : Ax = a, x \geq 0\}$ *est dit* **sous forme standard**.

1.3.3. Exemple : programme quadratique

$$(QP) : \quad \min\{x^T Hx + hx : Ax \geq a\}, \quad H \in \mathbb{R}^{d\times d} \text{ sdp}, h \in \mathbb{R}^{1\times d}, x \in \mathbb{R}^d, A \in \mathbb{R}^{m\times d}, a \in \mathbb{R}^m.$$

est dit programme quadratique (on minimise une fonction quadratique convexe sur un polyèdre).
Exemple : le problème des moindres carrés $\min_x\{\|Ax - b\| : x \in \mathbb{R}^d\}$, *parfois on ajoute la contrainte de positivité* $x \geq 0$.
Un (QPQC) est un programme quadratique ou on ajoute des contraintes quadratiques $\forall j = 1, ..., m : x^T H_j x + h_j x \leq \gamma_j$ *avec* $H_j \in \mathbb{R}^{d\times d}$ *sdp (géométriquement,* x *doit apartenir à un certain ellipsoide). Par exemple, pour résoudre un système linéaire* $Ax = b$ *avec* A *mal conditionné, on ajoute parfois une contrainte de régularisation* $\|x\| \leq \gamma$ *ou alors* $\|Px\| \leq \gamma$ *avec* P *et* $\gamma > 0$ *bien choisi (en traitement d'image, on limite la variation entre deux pixels voisins).*

1.3.4. Exemple : gestion de portefeuille

Soit $Y = (Y_1, ..., Y_d)^T$ *un vecteur de variables aléatoires, avec* Y_j *décrivant les bénéfices annuels pour l'action* j*. On suppose un modèle* $Y \sim \mathcal{N}(\mu, \Sigma)$*, avec* $\mu = \mathcal{E}(Y) \in \mathbb{R}^d$ *le vecteur des espérances, et* $\Sigma = Var(Y)$ *la matrice (ssdp) de covariance, tous les deux supposés connus (c'est pas toute la vérité, voir les modules de stat et proba, notamment pour l'estimation des valeurs numériques de* μ *et* Σ *sachant les bénéfices dans le passé).*
But : trouver une répartition de portefeuille $X = \sum_{j=1}^{n} x_j Y_j$ *avec* $x_1, ..., x_d \geq 0$ *et* $\sum_{j=1}^{d} x_j = 1$ *tout en maîtrisant les bénéfices moyennes* $\mathcal{E}(X) = \mu^T x$ *et le risque moyen* $Var(X) = x^T \Sigma x$*. Ici nos inconnues sont les* x_j*=pourcentage de l'action* j *dans notre portefeuille.*
Minimiser les risques moyens tout en limitant les bénéfices moyens à au moins γ *nous amène à résoudre le (QP)*

$$\min\{x^T \Sigma x : \mu^T x \geq \gamma, x \geq 0, (1, ..., 1)x = 1\}.$$

Maximiser les bénéfices moyens tout en limitant les risques moyens a au plus γ *nous laisse résoudre le (QPQC)*

$$\max\{\mu^T x : x^T \Sigma x \leq \gamma, x \geq 0, (1, ..., 1)x = 1\}.$$

Notons que $u(b)$*, la satisfaction en fonction des bénéfices annuelles* b*, n'est pas vraiment affine en* b*, mais cherche plutôt à s'approcher d'un seuil de richesse maximale, mathématiquement parlant* $u(b) = 1 - \exp(-kb)$ *avec* $k \geq 0$ *un paramètre scalaire connu suivant la nationalité et le sexe du client. On peut*

montrer que[2]

$$\frac{1}{k}\log\Big(1-\mathcal{E}(u(X))\Big)) = -\mathcal{E}(X)+\frac{k}{2}Var(X),$$

ainsi maximiser la satisfaction moyenne revient à résoudre le problème "compromis"

$$\max\{\mu^T x - \frac{k}{2}x^T\Sigma x : x\geq 0, (1,...,1)x = 1\}$$

(qui ressemble au Lagrangien du (QPQC) vu dans le capitre 3, avec $k/2$ la variable duale).

1.3.5. Exemple : projection sur un convexe

Si $C\subset\mathbb{R}^d$ est un convexe fermé alors le projeté $\Pi_C(x)$ d'un $x\in\mathbb{R}^d$ sur C est solution optimale du (QP) : $\inf\{\|y-x\| : y\in C\} =: dist(x,C)$.
Faites un petit dessin dans $\mathbb{R}^2$ pour voir que le projeté sur un triangle C (ou plus généralement d'un polyèdre C) est soit le sommet le plus proche soit le projeté orthogonal sur un des cotés de C.

2.

$$\begin{aligned}
1-\mathcal{E}(u(X)) &= 1-\mathcal{E}(u((x_1,...,x_d)Y)) \\
&= \int_{\mathbb{R}^d}\exp(-k(x_1,...,x_d)y)\exp(\frac{-(y-\mu)^T\Sigma^{-1}(y-\mu)}{2})\frac{dm(y)}{(2\pi)^{d/2}\sqrt{det\ \Sigma}} \\
&= \exp(-k\mu^T x)\int_{\mathbb{R}^d}\exp(-k(x_1,...,x_d)y)\exp(\frac{-y^T\Sigma^{-1}y}{2})\frac{dm(y)}{(2\pi)^{d/2}\sqrt{det\ \Sigma}} \\
&= \exp(-k\mu^T x)\exp(\frac{k^2}{2}x^T\Sigma x)
\end{aligned}$$

1.3.6. Exemple : pari sportif

Dans une course à d chevaux, l'organisateur distribue 90% des sommes pariées à ceux qui ont parié sur le cheval gagnant. On se demande comment investir γ euros, connaissant pour $j = 1, ..., d$ la probabilité p_j que le cheval j gagne, ainsi que la quantité $s_j > 0$ investie par d'autres joueurs sur le cheval j.

Posons x_j la quantité en euro pariée par notre joueur sur le cheval j. Nous obtenons alors les contraintes $x \geq 0$, et $(1, ..., 1)x = \gamma$. Si le cheval j gagne, on aura les bénéfices (à partager avec les autres joueurs)

$$0.9\,(\gamma + \sum_{k=1}^{d} s_k)\frac{x_j}{x_j + s_j}.$$

Donc pour maximiser les bénéfices moyens on doit maximiser l'espérance

$$\max\{\sum_{j=1}^{d} g_j(x_j) : x \geq 0, (1, ..., 1)x = \gamma\}, \quad g_j(x_j) = p_j\frac{x_j}{x_j + s_j}.$$

On parle d'un **objectif séparable**=*somme de fonctions d'une variable scalaire, ici des fonctions concaves et continues sur $C = \{x \in \mathbb{R}^d : x \geq 0, (1, ..., 1)x = \gamma\}$, un polyèdre compact.*

En changeant le signe dans l'objectif, on se ramène bien à un programme convexe

$$\inf\{-\sum_{j=1}^{d} g_j(x_j) : x \in C\},$$

qui admet alors une solution optimale. Le lecteur intéressé pourrait vérifier que cette solution est aussi unique (et même calculable).

1.4 Rappel sur le calcul différentiel

1.4.1. Définition : dérivée directionnelle

Soit $f : C \mapsto \mathbb{R}^m$. On dira que f admet une dérivée directionnelle en x dans la direction z si le segment $[x, x+z] \subset C$, et si la limite suivante existe

$$\lim_{t \to 0+} \frac{f(x+tz) - f(x)}{t} =: f'(x; z).$$

On dira que

$$f(h) = o(g(h)_{h \to 0} \quad \text{ssi} \quad \lim_{h \to 0} \frac{f(h)}{g(h)} = 0$$

$$\text{et} \quad f(h) = \mathcal{O}(g(h)_{h \to 0} \quad \text{ssi} \quad \limsup_{h \to 0} \frac{f(h)}{g(h)} < \infty.$$

où $g : U \mapsto (0, +\infty)$, $f : U \mapsto \mathbb{R}^m$, avec $U \subset \mathbb{R}^d$ un voisinage de l'origine.

1.4.2. Définition : fonctions continues et fonctions différentiables

Soit $f : C \mapsto \mathbb{R}^m$ avec $C \subset \mathbb{R}^d$ ouvert, et $x \in C$.
On dira que f est continue en x si

$$f(x+h) - f(x) = o(1)_{h \to 0}.$$

f est dite continue si elle est continue en tout $x \in C$.
On dira que f est différentiable en x s'il existe une matrice $\nabla f(x) \in \mathbb{R}^{m \times d}$ dite **Jacobienne** *de sorte que*

$$f(x+h) - f(x) - \nabla f(x)h = o(\|h\|)_{h \to 0}.$$

f est dite différentiable si elle est différentiable en tout $x \in C$.

(a) *Si $f = (f_1, ..., f_m)^T$ est différentiable en x alors toutes les dérivées partielles $\frac{\partial f_j}{\partial x_k}(x)$ existent, et*

$$\nabla f(x) = \left[\frac{\partial f_j}{\partial x_k}(x)\right]_{j=1,...,m}^{k=1,...,d}.$$

(j est l'indice ligne, k est l'indice colonne ! !). Réciproquement, si toutes ces dérivées partielles existent et sont continues dans un voisinage de x alors f est différentiable en x.

(b) *Dans le cas particulier $m = 1$ d'une fonctions à valeurs réelles, sa Jacobienne (aussi appelé* **gradient***) est un vecteur ligne.*

(c) *Si f est différentiable en x alors toute dérivée directionnelle existe, et vaut $f'(x; z) = \nabla f(x)z \in \mathbb{R}^m$.*

(d) *Parfois on aura aussi besoin de la continuité/différentiabilité de f en $x \notin Int(C)$ (si C n'est pas ouvert). Ici on supposera tacitement que f est continue/différentiable en x s'il existe un voisinage ouvert U de x et une fonction $\widetilde{f} : U \mapsto \mathbb{R}^m$ continue/différentiable en x qui est une extension de f : $\forall y \in U \cap C : \widetilde{f}(y) = f(y)$.*

1.4.3. Lemme : Jacobienne d'une fonction composite

Soient $C_1 \subset \mathbb{R}^d$ et $C_2 \subset \mathbb{R}^m$ des ouverts, $g : C_1 \mapsto C_2$, et $h : C_2 \mapsto \mathbb{R}^p$. Si g est différentiable en $x \in C_1$ et h est différentiable en $g(x)$ alors $f = h \circ g$ est différentiable en x, avec (attention à l'ordre des deux Jacobiennes)

$$\nabla f(x) = \nabla h(g(x))\nabla g(x).$$

Dans 1.4.2 on a défini la Jacobienne via $f(x+h) \approx f(x) + \nabla f(x)h$, le terme à droite aussi appelé le linéarisé de f en x (le plan tangent si $m = 1$) ou alors le développement de Taylor de f en x d'ordre 1. Ceci peut être généralisé pour le Hessien.

1.4.4. Dérivée seconde, le Hessien et Taylor d'ordre 2
Soit $f : \mathbb{R}^d \supset C \mapsto \mathbb{R}$. *Si la fonction* $g = (\nabla f)^T : C \mapsto \mathbb{R}^d$ *définie par* $g(x) = \nabla f(x)^T$ *est différentiable en* x *alors on note le Hessien de* f

$$\nabla^2 f(x) := \nabla g(x) = \Big[\frac{\partial^2 f}{\partial x_\ell \partial x_k}(x)\Big]_{\ell,k=1,\dots,d}.$$

Dans ce cas,

$$f(x+h) = f(x) + \nabla f(x)h + \frac{1}{2}h^T \nabla^2 f(x)h + o(\|h\|^2)_{h\to 0}.$$

1.4.5. Corollaire *Sous les hypothèses du 1.4.2, si* f *est différentiable en* $x \in C$ *alors pour tout* $y \in C$

$$f'(x; y-x) = \nabla f(x)(y-x).$$

Démonstration. Conséquence immédiate de la définition de différentiabilité. □

1.4.6. Exemple *Considérons la fonction convexe* $f : \mathbb{R} \ni x \mapsto |x|$. *Le corollaire précédant nous dit que* $f'(x;1) = f'(x) = 1$ *pour* $x > 0$ *et* $f'(x;1) = f'(x) = -1$ *pour* $x < 0$, *les dérivées à droite de* f. *Par contre, pour* $y - x = -1$ *on obtient* $f'(x;-1) = -f'(x) = -1$ *pour* $x > 0$ *et* $f'(x;-1) = -f'(x) = 1$ *pour* $x < 0$, *la derivée à gauche à un changement de signe près. Finalement,*

$f'(0;1) = 1$ *(la dérivée à droite) et* $f'(0;-1) = 1$ *(moins la dérivée à gauche). Observons aussi que* $f'(x;z) + f'(x;-z) \geq 0$*, avec égalité si* f *est différentiable en* x*.*

Le précédent exemple montre que une fonction convexe n'est pas forcément différentiable. Néanmoins, toutes les dérivées directionnelles existent.

1.4.7. Théorème sur dérivées directionnelles

Soit $C \subset \mathbb{R}^d$ *un convexe, et* $f : C \mapsto \mathbb{R}$*. Alors, pour tout* $x, y \in C$*,* $x \neq y$

$$f'(x; y-x) \in \mathbb{R} \cup \{-\infty\} \quad \textit{existe, et} \quad f(y) \geq f(x) + f'(x; y-x). \tag{1.1}$$

De plus, si $[x-z, x+z] \subset C$ *alors* $f'(x;z) + f'(x;-z) \geq 0$*.*

Démonstration. Montrons dans un premier temps que la fonction

$$\phi :]0,1] \to \mathbb{R}, \quad \phi(t) = \frac{f(x + t(y-x)) - f(x)}{t}$$

est croissante. Soient $0 < t_1 < t_2 \leq 1$, alors

$$\phi(t_2) - \phi(t_1) = \frac{1}{t_1}\Big((1 - \frac{t_1}{t_2})f(x) + \frac{t_1}{t_2}f(x + t_2(y-x)) - f(x + \frac{t_1}{t_2}t_2(y-x))\Big) \geq 0$$

par convexité de f. Donc, la limite $\phi(0+)$ existe, et vaut $-\infty$ (si ϕ n'est pas bornée inférieurement) ou appartient à $\mathbb{R}$. La pémière inégalité provient du fait que, par monotonie, $f'(x, y-x) = \phi(0+) \leq \phi(1) = f(y) - f(x)$. Finalement, pour tout $t \in]0,1]$ par convexité de f

$$\frac{1}{2}\Big(f(x+tz) + f(x-tz) - 2f(x)\Big),$$

et division par $t/2 > 0$ et passage à la limite $t \to 0$ donne $f'(x;z) + f'(x,z) \geq 0$. □

Dans le théorème 1.4.7, le cas $C = [0, +\infty)$ et $f(x) = -\sqrt{x}$ montre que $f'(0; 1) = -\infty$ n'est pas exclus. On peut montrer que la réciproque du théorème est aussi valable, si $x, y \in C, x \neq y$ nous avons (1.1) alors f est convexe. Dans le cas particulier $d = 1$ des fonctions d'une seule variable réelle, il découle du résultat précédent que une fonction convexe est continue, et même presque partout différentiable.

1.4.8. Corollaire : formule de la moyenne

Soit $C \subset \mathbb{R}^d$ un convexe, et $f : C \mapsto \mathbb{R}$ différentiable en $x \in Int(C)$. Alors pour tout $y \in C$ nous avons la **formule de la moyenne** $f(y) \geq f(x) + \nabla f(x)(y - x)$, *autrement dit, le plan tangent de f en x reste en dessous du graphe de la fonction f.*

1.4.9. Exercices : erreur dans Taylor pour les fonctions d'une variable réelle

(a) Soit $U \subset \mathbb{R}$ un convexe ouvert contenant l'origine, et soit q de classe $\mathcal{C}^1(U)$. M.q.

$$\forall t \in U \exists \eta \in [0, 1] \quad t.q. \quad q(t) = q(0) + tq'(\eta t).$$

(b) Soit q de classe $\mathcal{C}^2(U)$. En observant que

$$q(t) = q(0) + tq'(0) + \int_0^t (t - s)q''(s)\, ds$$

m.q.

$$\forall t \in U \exists \eta \in [0, 1] \quad t.q. \quad q(t) = q(0) + tq'(0) + t^2 \frac{q''(\eta t)}{2}.$$

1.4.10. Exercices

(a) Soit $f : \mathbb{R}^d \mapsto \mathbb{R}$ définie par $f(x) = Ax+b$. M.q. $\nabla f(x) = A$, $\nabla^2 f(x) = 0$.

(b) Soit $f(x) = x^T Hx+hx$. M.q. $\nabla f(x) = x^T(H+H^T)+h$, $\nabla^2 f(x) = H+H^T$.

(c) Soit $f(x) = g(Ax + b)$. . M.q. $\nabla f(x) = \nabla g(Ax + b)A$.

1.4.11. Exercices

Soit $f : C \mapsto \mathbb{R}$, avec $C \subset \mathbb{R}^d$ un convexe ouvert.

(a) Soit $f \in \mathcal{C}^2(C)$. Nous souhaitons montrer que $\nabla^2 f(x)$ est ssdp pour tout $x \in C$ ssi f est convexe.

(a1) Pour établir $\boxed{\Longrightarrow}$, soient $x, y \in C$, et $g : [0, 1]$ défini par $g(t) = f(tx + (1 - t)y)$. Notons par p la droite interpolant g aux points 0 et 1. En utilisant la formule de Cauchy, montrer que

$$\forall t \in [0, 1] \exists \eta \in [0, 1] : \quad g(t) - p(t) = \frac{g''(\eta)}{2} t(t - 1).$$

En faisant le lien entre g'' et le Hessien de f, conclure que f est convexe.

(a2) Pour établir $\boxed{\Longleftarrow}$, soit f convexe, et $z \in \mathbb{R}^d$. Montrer que $y = x+sz \in C$ pour $s > 0$ assez petit. En revenant à (a1), vérifier que $\exists \eta \in [0, 1]$ de sorte que

$$z^T \nabla^2 f(\eta x + (1 - \eta)y)z \geq 0.$$

Par un passage à la limite, conclure que $\nabla^2 f(x)$ est ssdp.

(b) Si $f \in \mathcal{C}^2(C)$ et si $\nabla^2 f(x)$ est sdp pour tout $x \in C$, m.q. f est strictement convexe. M.q. la réciproque est fausse.

(c) *Si* $f \in \mathcal{C}^1(C)$ *alors montrer que* f *est convexe ssi* $\forall x, y \in C$ *nous avons* $\nabla f(y)(x-y) \leq f(x) - f(y)$.

1.5 Caractérisation d'optimalité

1.5.1. Théorème : CNS pour optimalité
Considérons $(CP) : \inf\{f(x) : x \in C\}$ *avec* $C \subset \mathbb{R}^d$ *un convexe non vide, et* $f : C \mapsto \mathbb{R}$ *une fonction convexe.*
Alors nous avons pour un $\underline{x} \in C$ *l'équivalence*

$$\underline{x} \text{ est solution optimale de } (CP) \quad \Longleftrightarrow \quad \forall y \in C : \quad f'(\underline{x}; y - \underline{x}) \geq 0.$$

Si de plus f *est différentiable en* $\underline{x}$ *alors*

$$\underline{x} \text{ est solution optimale de } (CP) \quad \Longleftrightarrow \quad \forall y \in C : \quad \nabla f(\underline{x})(y - \underline{x}) \geq 0.$$

Démonstration. Pour montrer $\Longleftarrow$, nous appliquons le théorème 1.4.7 pour conclure que, pour tout $y \in C$, nous avons $f(y) \geq f(\underline{x}) + f'(\underline{x}; y - \underline{x}) \geq f(\underline{x})$.
Pour établir implication $\Longrightarrow$, raisonnons par absurde et supposons qu'il existe un $y \in C$ et un $\epsilon > 0$ tels que $f'(\underline{x}; y - \underline{x}) \leq -2\epsilon$. Par définition de la dérivée directionnelle, nous trouvons alors un $t > 0$ de sorte que $\phi(t) \leq -\epsilon$, et alors $f(\underline{x} + t(y - x)) < f(\underline{x})$, en contradiction avec l'hypothèse sur $\underline{x}$. □

1.5.2. Corollaire
Sous les hypothèses du théorème 1.5.1, et f *différentiable en* $\underline{x}$ *:*

(a) *Si* C *est un espace affine alors* $\underline{x}$ *est solution optimale de* (CP) *ssi* $\forall y \in C$ *:* $\nabla f(\underline{x})(y - \underline{x}) = 0$.

(b) *En particulier, si* $C = \{x \in \mathbb{R}^d : Ax = b\}$, *alors* $\underline{x}$ *est solution optimale de* (CP) *ssi* $\exists \lambda \in \mathbb{R}^{1\times m}$ *tel que* $\nabla f(\underline{x}) = \lambda A$ *(à comparer avec le théorème des extrema liés).*

(c) *Si* $\underline{x} \in Int(C)$ *alors* $\underline{x}$ *est solution optimale de* (CP) *ssi* $\nabla f(\underline{x}) = 0$ *(tout point stationnaire est un minimum global).*

Démonstration. Pour montrer (a), rappelons d'abord que C espace affine implique que $V := C - \underline{x}$ est un sous-espace vectoriel. Donc d'après 1.5.1 : $\underline{x}$ est solution optimale ssi $\forall z \in V$ nous avons $\nabla f(\underline{x})z \geq 0$ ssi $\forall z \in V$ nous avons $\nabla f(\underline{x})z = 0$ (car $z \in V$ implique que $-z \in V$).
Dans la partie (b), $C = \underline{x} + \mathrm{Ker}(A)$, et alors d'après (a) : $\underline{x}$ est solution optimale ssi $\nabla f(\underline{x})^T \in \mathrm{Ker}(A)^\perp = \mathrm{Im}(A^T)$.
Finalement, pour montrer (c), par hypothèse il existe un $\epsilon > 0$ de sorte que $y := \underline{x} - \epsilon \nabla f(\underline{x})^T \in C$, et alors

$$-\epsilon \|\nabla f(\underline{x})^T\|^2 = \nabla f(\underline{x})(y - x) \geq 0$$

par 1.5.1, ce qui implique $\nabla f(\underline{x}) = 0$. L'implication réciproque découle directement du théorème 1.5.1. □

1.5.3. Théorème : caractérisation de l'ensemble des solutions optimales

Sous les hypothèses du théorème 1.5.1, et C *fermé :*

(a) *L'ensemble* $\mathcal{S}$ *des solutions optimales de* (CP) *est convexe. Si de plus* f *est continue sur* C *alors* $\mathcal{S}$ *est fermé.*

(b) *Si* f *est strictement convexe alors il existe au plus une solution optimale.*

Démonstration. Avec M la valeur optimale de (CP), nous avons $\mathcal{S} = \{x \in C : f(x) = M\} = \{x \in C : f(x) \leq M\}$. Il n'y a rien a montrer si $\mathcal{S}$ est vide. Sinon, soient $x, y \in \mathcal{S}$, et $t \in [0, 1]$. Alors $x, y, tx + (1-t)y \in C$ par convexité de C, et $f(tx+(1-t)y) \leq tf(x)+(1-t)f(y) = M$ par convexité de f. Donc $\mathcal{S}$ est convexe. Pour montrer la fermeture, soit (x_n) une suite d'éléments de $\mathcal{S}$ qui converge vers un $x \in \mathbb{R}^d$. Alors $x \in C$ par fermeture de C, et $f(x_n) = M$ pour tout n. Donc par continuité de f, $f(x) = \lim_{n\to\infty} f(x_n) = M$, et alors $x \in \mathcal{S}$.
Pour montrer (b), soient $x, y \in \mathcal{S}$. Si $x \neq y$, alors $f(\frac{x+y}{2}) < \frac{f(x)+f(y)}{2} = M$ par convexité stricte, en contradiction avec la définition de M. Donc une solution optimale est unique. □

1.5.4. Théorème : existence des solutions optimales

Sous les hypothèses du théorème 1.5.1, et f continue sur C, il existe au moins une solution optimale de (CP) si au moins une des conditions suivantes est valable :

(a) C est compact ;

(b) $\exists x_0 \in C$ de sorte que l'ensemble niveau $C_0 = \{x \in C : f(x) \leq f(x_0)\}$ est compact.

(c) C est fermé, et $f(x) \to +\infty$ pour $\|x\| \to \infty$.

Démonstration. Comme mentionné avant, la partie (a) découle du théorème de Bolzano-Weierstrass. Comme pour $y \in C \setminus C_0$ nous avons $f(y) > f(x_0)$, alors $\inf\{f(x) : x \in C\} = \inf\{f(x) : x \in C_0\}$ et la partie (b) découle de la partie (a). Finalement, l'hypothèse de la partie (c) implique que C_0 est fermé et borné, et donc compact. □

On termine ce chapitre en revenant sur l'exemple 1.3.5 de projection sur un convexe fermé.

1.5.5. Retour sur l'exemple 1.3.5 de projection sur un convexe fermé

Etant donné $C \subset \mathbb{R}^d$ un convexe fermé, comment trouver pour un $y \in \mathbb{R}$ l'élément le plus proche $\Pi(y)$ de y dans C par rapport à la norme euclidienne ?
D'après 1.3.5, $\Pi(y)$ est solution optimale de

$$\min\{f(x) : x \in C\}, \quad f(x) = \frac{1}{2}\|x-y\|^2 = \frac{1}{2}(x-y)^T(x-y).$$

D'après 1.4.10et 1.4.11, nous avons $\nabla f(x) = (x-y)^T$, $\nabla^2 f(x) = I$ spd, et donc f est strictement convexe. Il découle de 1.5.3(b) et 1.5.4(c) que $\Pi(y)$ existe et est unique. Comme $f \geq 0$ si $y \in C$ alors $\Pi(y) = y$, mais pour $y \notin C$?
D'après 1.5.1, $\Pi(y)$ est caractérisé par

$$\forall x \in C : \quad (\Pi(y) - y)^T(x - \Pi(y)) \geq 0,$$

autrement dit, l'angle entre l'erreur de projection $y - \Pi(y)$ et tout vecteur de la forme $\Pi(y) - x$ pour $x \in C$ est de module $\leq \pi/2$. Exemples pour $y \notin C$:

- *(a) Si C est un espace affine alors $\Pi(y)$ est bien le projeté orthogonal sur C : pour tout $x \in C$, $y - \Pi(y) \perp x - y$.*
- *(b) Si C est un demi-espace alors $\Pi(y)$ est bien le projeté orthogonal sur le bord de C (un hyper-plan).*
- *(c) Si $C = \{x : \|x\| \leq 1\}$ la boule fermée d'unité, alors $\Pi(y) = y/\|y\|$.*
- *(d) Si $C = \{x \in \mathbb{R}^d : x \geq 0\}$* **l'orthant positif**, *alors la $j$ième composante de $\Pi(y)$ est égale à y_j si $y_j \geq 0$, et égale à 0 si $y_j < 0$.*

1.5.6. Exercice

On considère deux droites affines de $\mathbb{R}^n$

$$D_1 = \{a_1 + tu_1, \quad t \in \mathbb{R}\}, \quad D_2 = \{a_2 + tu_2, \quad t \in \mathbb{R}\}.$$

On veut déterminer les points $p_1 \in D_1$, $p_2 \in D_2$ *qui minimisent la distance euclidienne de* p_1 *à* p_2.

(a) *Formuler ce problème comme un problème de minimisation sans contraintes.*

(b) *Montrer qu'il s'agit d'un problème convexe.*

(c) *Ce problème a-t-il une solution optimale ? Comment la caractérise-t-on ?*

(d) *Dans quelles conditions la solution est-elle unique ? Calculer la (les) solutions.*

(e) *Montrer que dans le cas d'unicité* $p_2 - p_1$ *est orthogonal à* u_1, u_2.

1.5.7. Exercice

Soit $g : \mathbb{R}^n \longrightarrow \mathbb{R}$ *une fonction convexe et différentiable sur* $\mathbb{R}^n$ *et* $h : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$ *une fonction convexe sur* $\mathbb{R}^n$, *finie en au moins un point. On pose* $f = g + h$ *et considère le problème* $\min_{x \in \mathbb{R}^n} f(x)$.

(a) *Montrer que* $\underline{x}$ *minimise* f *sur* $\mathbb{R}^n$ *ssi* $\forall x \in \mathbb{R}^n : (\nabla g(\underline{x}, x - \underline{x}) + h(x) - h(\overline{x}) \geq 0$

(b) *Dans le cas où* C *est un convexe fermé non vide et* h *la fonction :*

$$h(x) = \begin{cases} 0 & si\ x \in C \\ +\infty & sinon \end{cases},$$

que peut-on conclure du résultat précédent ?

1.5.8. Exercice

Soit A une matrice symétrique d'ordre d, et $f : \mathbb{R}^d \setminus \{0\} \ni x \mapsto \frac{x^T Ax}{x^T x}$. Calculer $\nabla f(x)$, et le comparer à la projection orthogonale de Ax sur H, l'ensemble des vecteurs orthogonaux à x.

1.5.9. Exercice : l'entropie

Considérons la fonction $\varphi(t) :]0,1] \ni t \mapsto t\log(t)$.

(a) M.q. $\varphi \in \mathcal{C}^\infty(]0,1])$ et, en posant $\varphi(0) = 0$, nous avons $\varphi \in \mathcal{C}([0,1])$.

(b) En discutant séparément l'ensemble $]0,2[$, montrer que φ est strictement convexe sur $[0,1]$. Calculer la dérivée directionnelle $\varphi'(0;1)$.

(c) Pourquoi φ admet-elle un unique minimiseur $\underline{t}$ sur $[0,1]$? Calculer $\underline{t}$.

(d) Pour $n \geq 1$, considérons le problème

$$\inf\{f(x) : x \in C\}, \quad C = \{x \in \mathbb{R}^n : x \geq 0, \sum_{j=1}^n x_j = 1\}, \quad f(x) = \sum_{j=1}^n x_j \log(x_j)$$

(on maximise l'entropie sur C l'ensemble des vecteurs de probabilités). Pourquoi f est-elle continue et strictement convexe sur $[0,1]^n$?

(e) Par élimination $x_n = 1 - x_1 - ... - x_{n-1}$ et en résolvant un problème à $n-1$ variables, montrer que l'entropie est maximisée pour la loi uniforme.

Chapitre 2

Le Lagrangien et KKT

2.1 Énoncé du théorème KKT et exemples

Dans ce chapitre on se donne des entiers $d \geq 1$, $m, p \geq 0$, $A \in \mathbb{R}^{m\times d}$, $a \in \mathbb{R}^m$, $S = \{x \in \mathbb{R}^d : Ax = a\}$ (si $m = 0$ on posera $S = \mathbb{R}^d$), des fonctions $f, g_1, ..., g_p : S \mapsto \mathbb{R}$ convexes et continûment différentiables sur S, $g = (g_1, ..., g_p)^T : S \mapsto \mathbb{R}^p$. Posons

$$(P) : \inf\{f(x) : x \in C\}, \quad C = \{x \in S : g(x) \leq 0\}.$$

Notons que C est bien un convexe fermé. Le théorème suivant caractérise si un $\underline{x} \in \mathbb{R}^d$ est une solution optimale de (P). Sa preuve sera donnée plus tard dans 2.3.9.

2.1.1. Théorème KKT=Karush, Kuhn et Tucker

Supposons que

$$\exists \widetilde{x} \in S \quad t.q. \quad \forall j = 1, ..., p : g_j(\widetilde{x}) < 0. \tag{2.1}$$

Alors $\underline{x} \in \mathbb{R}^d$ *est solution optimale de* (P) *ssi* $\exists \lambda = (\lambda_1, ..., \lambda_p) \in \mathbb{R}^{1\times p}$ *de sorte que le couple vérifie le système*

$$(KKT) : \begin{cases} \underline{x} \in S, \quad g(\underline{x}) \leq 0 \quad (\underline{x} \text{ est réalisable pour } (P)) \\ \lambda \geq 0 \quad (\lambda \text{ est réalisable pour un problème "dual" } (D)) \\ \forall j = 1, ..., p : \quad \lambda_j g_j(\underline{x}) = 0 \quad (\text{relation dite de complémentarité}) \\ \exists \mu \in \mathbb{R}^{1\times m} : \quad \nabla f(\underline{x}) + \lambda \nabla g(\underline{x}) = \mu A. \end{cases}$$

2.1.2. Remarques sur le théorème KKT

(a) *Pour trouver une solution optimale, on doit alors résoudre un système non linéaire à* $d + m + p$ *inconnues (les composantes de* $\underline{x}, \lambda, \mu$*) et à* $m + p + d$ *équations, auquel il s'ajoute* $2p$ *inégalités.*

(b) *Le théorème KKT a aussi un sens dans le cas* $m = 0$ *(sans contraintes d'égalités) : ici la dernière condition prend*[1] *la forme* $\nabla f(\underline{x}) + \lambda \nabla g(\underline{x}) = 0$.

(c) *Le théorème KKT a aussi un sens dans le cas* $p = 0$ *(sans contraintes d'inégalités) : ici les conditions 2 et 3 sont trivialement valables, et on supprime* $\lambda g(\underline{x})$ *dans la dernière condition. Pour ce cas* $p = 0$*, notre théorème a été déjà montré, voir 1.5.2(c) pour le cas* $m = 0$ *sans contraintes, et 1.5.2(b) pour le cas* $m > 0$ *de contraintes affines d'égalité.*

(d) *La condition de complémentarité nous dit que les gradients en* $\underline{x}$ *des contraintes non actives en* $\underline{x}$ *n'apparaissent pas dans* (KKT) *(car* $g_j(\overline{x}) \neq 0$ *implique que* $\lambda_j = 0$*).*

1. En analogie avec la convention qu'une somme vide vaut 0.

Les premières deux conditions $g(\underline{x}) \leq 0$ et $\lambda \geq 0$ nous disent que l'on peut réécrire notre relation de complémentarité d'une manière équivalente comme $\lambda g(\underline{x}) = 0$.

(e) (2.1) *est dit* **condition de qualification de Slater**, *il existe dans la littérature d'autres conditions de qualification. En TD on verra que, dans certains cas, on peut supprimer*[2] *la condition* (2.1).

Avant de se lancer dans la preuve, considérons trois exemples.

2.1.3. Exemple

Considérons (P) $\inf\{\|x\|^2 : x = (x_1, x_2)^T \in \mathbb{R}^2, 5-2x_1-x_2 \leq 0\}$ *(on cherche le point le plus proche de l'origine dans un demi-plan). Ici* $f(x) = \|x\|^2$, $\nabla f(x) = (2x_1, 2x_2)$, $p = 1$, $g(x) = g_1(x) = 5 - 2x_1 - x_2$, $\nabla g(x) = (-2, -1)$, $m = 0$, *et la condition de Slater est valable pour* $\widetilde{x} = (4, 4)^T$ *(par exemple). Le point* x *est alors solution optimale de* (P) *sssi il existe* $\lambda \in \mathbb{R}$ *tel que*

$$5 - 2x_1 - x_2 \leq 0, \quad \lambda \geq 0, \quad \lambda(5 - 2x_1 - x_2) = 0, \quad (2x_1, 2x_2) + \lambda(-2, -1) = 0.$$

La 4ième relation nous donne $x = (\lambda, \lambda/2)$ *en fonction de* λ *et les premiers deux que* $\lambda \geq 2$. *La troisième relation peut être écrite comme* $\lambda = 0$ *ou* $5 - 2x_1 - x_2 = 0$, *mais on vient de voir que le premier cas est exclu. Insérant l'expression pour* x *dans la troisième relation donne alors* $5 - 2\lambda - \lambda/2 = 0$ *ou* $\lambda = 2$ *et alors* $x = (2, 1)^T$. *Réciproquement, on vérifie également que* $\lambda = 2$ *et* $\underline{x} = (2, 1)^T$ *vérifient* (KKT). *Donc par le théorème 2.1.1,* $(2, 1)^T$ *est l'unique*

2. Rappelons que dans (KKT) apparaissent seulement les gradients en $\underline{x}$ des contraintes actives en $\underline{x}$. Un domaine est dit qualifié en $\underline{x}$ ssi on obtient la même conclusion en $\underline{x}$ en remplaçant les contraintes non linéaires $g_j(x) \leq 0$ par leurs linéarisées $\widetilde{g}_j(x) := g(\underline{x}) + \nabla f(\underline{x})(x - \underline{x}) \leq 0$.

solution optimale (on savait déjà l'existence et unicité d'une solution, voir 1.5.5).

2.1.4. Exemple

Avec f strictement convexe sur $S = \mathbb{R}^d$ et $B \in \mathbb{R}^{p\times d}$: $x \in \mathbb{R}^d$ est solution optimale de $\inf\{f(x) : x \in S : b - Bx \leq 0\}$ *ssi* $\exists \lambda \in \mathbb{R}^{1\times p}$ *avec*

$$\lambda \geq 0, \quad Bx - b \geq 0, \quad \lambda(Bx - b) = 0, \quad \nabla f(x) = \lambda B.$$

Dans le cas où f est une forme quadratique comme dans le 1.2.4, la relation $\nabla f(x) = \lambda B$ permet d'exprimer x en fonction de λ, et donc d'éliminer x du système (KKT). Si $f(x) = hx$ avec $h \in \mathbb{R}^{1\times d}$ un objectif linéaire, on devrait trouver dans l'ensemble des solutions du système linéaire $\lambda B = h$ (à p inconnues et d équations) celles qui vérifient aussi les autres trois contraintes. Une relation $\lambda_j \neq 0$ donnera une équation supplémentaire $(Bx - b)_j = 0$, ce qui devrait permettre de calculer x (ou éventuellement l'ensemble des solutions optimales).

Pour l'exemple numérique 1.3.2(a), $d = 2$, $p = 4$, et

$$B = \begin{bmatrix} 1 & 2 \\ -1 & -2 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 7 \\ -7 \\ 0 \\ 0 \end{bmatrix}, \quad h = [3, 4].$$

Ici la condition de Slater (2.1) n'est pas valable.[3] *Parmi les solutions* $[\lambda_1, \lambda_2, \lambda_3, \lambda_4]B =$

3. Ceci est lié au fait que l'on a transformé une contrainte d'égalité $x_1+2x_2 = 7$ en deux contraines d'inégalités. Pour rectifier, on devrait choisir $m = 1$ et $p = 2$, avec $A = [1, 2], a = [7]$, $S = \{x \in \mathbb{R}^2 : Ax = a\}$, $g(x) = (-x_1, -x_2)^T$, vérifiant Slater pour par exemple $\widetilde{x} = [1, 3]^T$. Bien entendu, ici le système (KKT) prendra une autre forme.

h on trouve par exemple celui ou $\lambda_2 = \lambda_4 = 0$ *et alors* $\lambda_1 = 2 \neq 0$, $\lambda_3 = 1 \neq 0$, *ce qui implique que* $(Bx-b)_1 = 0 = x_1 + 2x_2 = 7$ *et* $(Bx-b)_3 = 0 = x_1$ *ou alors* $x = (0, 7/2)^T$, *ce qui vérifie les contraintes* $Bx - b \geq 0$, $\lambda \geq 0$ *et* $\lambda(Bx-b) = 0$. *Donc* $x = (0, 7/2)^T$ *est une solution optimale. En examinant les autres* 5 *choix d'annulation de deux composantes de* λ, *on voit que cette solution optimale est unique.*

2.1.5. Exemple : retour à la régularisation 1.3.3

Avec $\gamma > 0$ *et* B *une matrice inversible, considérons* $\inf\{\|Bx-b\|^2 : \|x\|^2 - \gamma \leq 0\}$, *ici* $f(x) = \|Bx-b\|^2$, $\nabla f(x) = 2(Bx-b)^T B$, $m = 0$, $g(x) = \|x\|^2 - \gamma$, $\nabla g(x) = 2x^T$, *et la condition de Slater est valable pour* $\widetilde{x} = 0$. *Donc* x *est solution optimale ssi il existe* $\lambda \in \mathbb{R}$ *de sorte que*

$$\|x\|^2 - \gamma \leq 0, \quad , \lambda \geq 0, \quad \lambda(\|x\|^2 - \gamma) = 0, \quad 2(Bx-b)^T B + 2\lambda x^T = 0.$$

La dernière relation peut être écrite comme $(B^T B + \lambda I)x = B^T b$. *Comme* $B^T B + \lambda I$ *est sdp pour* $\lambda \geq 0$, *nous pouvons alors résoudre pour* $x = x(\lambda) = (B^T B + \lambda I)^{-1} B^T b$. *Posons* $\phi(\lambda) = \|x(\lambda)\|^2$, *une fonction qui s'avère différentiable et strictement décroissante sur* $[0, \infty)$, *et qui tend vers* 0 *pour* $\lambda \to \infty$. *Pour voir ceci, on notera* (λ_j, v_j) *les éléments propres de* $B^T B$, $\lambda_j > 0$, *et on exprimera* $B^T b$ *dans la base orthonormée des vecteurs propres de* $B^T B$,

$$B^T b = \sum_{j=1}^{d} \alpha_j v_j, \quad x(\lambda) = \sum_{j=1}^{d} \frac{\alpha_j v_j}{\lambda + \lambda_j}, \quad \phi(\lambda) = \sum_{j=1}^{d} \Big(\frac{\alpha_j}{\lambda + \lambda_j} \Big)^2.$$

Nous devons distinguer 2 *cas : si* $\gamma > \phi(0)$ *alors la contrainte* $g(x) \leq 0$ *n'est pas active en* λ *pour aucun* $\lambda \geq 0$, *donc forcément* $\lambda = 0$ *et* $x(0) = B^{-1} b$

est l'unique solution optimale (et la contrainte n'est pas active en $x(0)$). Si par contre $\gamma \leq \phi(0)$ alors il existe un unique $\lambda \geq 0$ avec $\phi(\lambda) = \gamma$ et alors $g(x(\lambda)) = 0$ (la contrainte est active en $x(\lambda)$) ce qui implique que $x(\lambda)$ est une solution optimale. L'unicité provient de la convexité stricte de f.

2.1.6. Exercice :

En résolvant (KKT), résoudre le problème d'optimisation convexe suivant :

$$\min\Big\{-\sum_{i=1}^{n}\log(\alpha_i + x_i) : \quad \sum_{i=1}^{n} x_i = 1, x \geq 0\Big\} \quad \text{où } \alpha_i > 0 \text{ pour } i = 1, \ldots, n.$$

2.1.7. Exercice : Minimisation sur le simplexe unité

Soit

$$W\Lambda_n = \{x \in \mathbb{R}^n : \forall i = 1, ..., n : x_i \geq 0, \ \text{et} \ \sum_{i=1}^{n} x_i = 1\}$$

et $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ convexe différentiable. Montrer qu'une condition necessaire et suffisante pour que $\underline{x} \in \Lambda_n$ minimise f sur Λ_n est qu'il existe un $c \in \mathbb{R}$ de sorte que

$$\forall i \notin E(\underline{x}) : \quad \frac{\partial f}{\partial x_i}(\underline{x}) = c, \quad \forall i \in E(\underline{x}) : \quad \frac{\partial f}{\partial x_i}(\underline{x}) \geq c,$$

avec l'ensemble $E(\underline{x}) = \{i \in \{1, ..., n\} : \underline{x}_j = 0\}$ des indices des contraintes actives.

2.1.8. Exercice : objectif séparable

On se donne n fonctions dérivables $f_j : \mathbb{R} \to \mathbb{R}$ et on considère le problème

$$\begin{cases} \text{minimiser } \sum_{j=1}^{n} f_j(x_j) \\ \sum_{j=1}^{n} x_j = 1, \quad \forall j = 1, ..., n : x_j \geq 0 \end{cases}$$

Montrer que si $\underline{x}$ est solution de ce problème alors $\exists \mu \in \mathbb{R}$ tel que pour tout $j = 1, ..., n$:

$$\underline{x}_j > 0 \;\Rightarrow\; f_j'(\underline{x}_j) = \mu,$$
$$\underline{x}_j = 0 \;\Rightarrow\; f_j'(\underline{x}_j) \geq \mu.$$

Résoudre $\min\{x_1^3 + 3x_1 + x_2^2 : x_1 + x_2 = 1, x_1, x_2 \geq 0\}$.

2.1.9. Exercice : retour sur la maximisation de l'entropie
Résoudre le problème 1.1.13(d) en écrivant le système (KKT).

2.1.10. Exercice : *Pour $a \in \mathbb{R}^n$, on considère la fonction f_a définie sur $C = \{x \in \mathbb{R}^n : \|x\| < 1\}$ par $f_a(x) = -\log(1 - \|x\|^2) + (a, x)$.*

(a) Montrer que f_a est strictement convexe.

(b) On considère le problème de minimisation suivant

$$(P_a) \quad \min\Big\{f_a(x) : x \in C_a\Big\}, \text{ avec } C_a = \Big\{x \in \mathbb{R}^n : \|x\| \leq \frac{1}{2} \text{ et } (a, x) \leq 0\Big\}.$$

(b1) Résoudre (P_a) pour $a = 0$.

(b2) On suppose $a \neq 0$ et on désigne par $\underline{x}$ la solution optimale de (P_a). Montrer que $(a, \underline{x}) < 0$ et déterminer $\underline{x}$. Conclusion pour $a \to 0$?

2.1.11. Exercice : *Étant donné $u = (u_1, u_2, \cdots, u_n)^T \in \mathbb{R}^n$, on cherche un élément $x = (x_1, x_2, \cdots, x_n)^T \in \mathbb{R}^n$ vérifiant $x_1 \leq x_2 \leq \cdots \leq x_n$ le plus proche de u au sens de la norme euclidienne.*

(a) Formaliser cette question comme un problème de minimisation convexe.

(b) *Décrire les conditions caractérisant l'unique solution* $\underline{x} = (\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n)^T$ *du problème précédent.*

(c) *Résoudre dans le cas particulier de* $u = (2, 1, 5, 4)^T$.

2.1.12. Exercice : condition suffisante pour Slater
Pour notre problème (CP), *supposons qu'il existe un point* y *réalisable pour* (CP) *de sorte que l'ensemble des lignes de* A *et des gradients* $\nabla g_j(y)$ *des contraintes actives en* y *(veut dire, pour tout* j *t.q.* $g_j(y) = 0$*) soient libres. Déduire la condition de Slater.*
Indication : considérer $\widetilde{x} = y + td$ avec $Ad = 0$, et $\nabla g_j(y)d = -1$ pour toutes les contraintes actives.

2.2 Théorèmes de séparation

Pour démontrer notre théorème 2.1.1 nous avons besoin des deux théorèmes suivants.

2.2.1. Théorème de séparation stricte
Soit $C \subset \mathbb{R}^d$ *un convexe fermé, et* $y \in \mathbb{R}^d \setminus C$. *Alors* $\exists \phi \in \mathbb{R}^{1\times d}$ $\exists \beta \in \mathbb{R}$ *de sorte que*[4]

$$\forall x \in C : \quad \phi x \geq \beta > \phi y$$

Démonstration. Notons par $\Pi(y)$ l'élément le plus proche de y dans C, voir 1.5.5. Nous avons $y \notin C$ et alors $\Pi(y) \neq y$. Posons $\phi = (\Pi(y) - y)^T / \|\Pi(y) - y\|$, alors

4. Autrement dit, C est un sous-ensemble d'un demi-espace qui ne contient pas y. On peut alors séparer C et y strictement par un hyperplan.

$\|\phi\| = 1$ et, pour tout $x \in C$,

$$\begin{aligned}\phi x &= \phi y + \phi(\Pi(y) - y) + \phi(x - \Pi(y)) \\ &= \phi y + \|\Pi(y) - y\| + \frac{(\Pi(y) - y)^T (x - \Pi(y))}{\|\Pi(y) - y\|} \geq \phi y + \|\Pi(y) - y\| =: \beta\end{aligned}$$

et alors $\beta > \phi y$. □

2.2.2. Théorème de séparation faible

Soit $C \subset \mathbb{R}^d$ un convexe, et $y \in \mathbb{R}^d \setminus Int(C)$. Alors $\exists \phi \in \mathbb{R}^{1\times d} \setminus \{0\}$ de sorte que[5]

$$\forall x \in C : \quad \phi x \geq \phi y.$$

Démonstration. Par hypothèse sur y, il existe une suite $(y^{(n)})$ d'éléments de $\mathbb{R}^d \setminus \text{Clos}(C)$ qui converge vers y. Sachant que $\text{Clos}(C)$ est convexe, par le théorème de séparation forte nous trouvons pour tout n un $\phi^{(n)} \in \mathbb{R}^{1\times d}$ de sorte que

$$\forall x \in \text{Clos}(C) : \quad \phi^{(n)} x > \phi^{(n)} y^{(n)}.$$

Par conséquent, $\phi^{(n)} \neq 0$ et, en divisant par sa norme, nous pouvons supposer que $\|\phi^{(n)}\| = 1$. Comme la boule d'unité dans $\mathbb{R}^d$ est compacte, nous pouvons extraire une sous-suite aussi nommée $(\phi^{(n)})$ qui tend vers $\phi \in \mathbb{R}^{1\times d}$. Pour tout $x \in C$, en passant à la limite $n \to \infty$, nous obtenons bien l'inégalité désirée $\phi x \geq \phi y$. □

2.2.3. Exercice : Séparation faible de deux convexes

Soient C et D deux ensembles convexes du $\mathbb{R}^n$ d'intersection vide.

5. Si de plus y appartient au bord de C alors l'hyperplan $\{x \in \mathbb{R}^d : \phi x = \phi y\}$ est dit hyperplan d'appui de C en y.

(a) *En considérant $C - D$, montrer que C et D peuvent être faiblement séparés : $\exists \phi \in \mathbb{R}^{1\times n} \setminus \{0\}$ de sorte $\forall x \in C \forall y \in D : \phi x \geq \phi y$.*

(b) *Soit de plus D un sous-espace affine de la forme $D = \{Fu + g : u \in \mathbb{R}^m\}$, avec $F \in \mathbb{R}^{n\times m}$, $g \in \mathbb{R}^n$. Conclure que*

$$\exists a \in \mathbb{R}^n \setminus \{0\} \quad t.q. \quad : \quad F^T a = 0, \quad \forall x \in C a^T x \leq a^T g.$$

2.2.4. Exercice : Lemme de Farkas

Soient $A \in \mathbb{R}^{m\times n}$ et $b \in \mathbb{R}^m$. En se servant de l'exercice précédent, montrer qu'une condition nécessaire et suffisante pour que le système

$$Ax = b, \quad x \geq 0$$

ait une solution dans $\mathbb{R}^n$ est que

$$\forall u = (u_1, u_2, \cdots, u_m) \in \mathbb{R}^{1\times m} \text{ tel que } uA \geq 0 \text{ on ait } ub \geq 0.$$

2.2.5. Exercice : Séparation stricte de deux convexes

Soient C et D deux ensembles convexes du $\mathbb{R}^n$ d'intersection vide.

(a) *Si $0 \notin Clos(C - D)$, montrer que C et D sont strictement séparés : $\exists \phi \in \mathbb{R}^{1\times n} \exists \beta \in \mathbb{R}$ t.q. $\forall x \in C \forall y \in D : \phi x - \phi y \geq \beta > 0$.*

(b) *Si de plus C est fermé et D compact, montrer que ces deux ensembles sont strictement séparés.*

(c) *Montrer par un exemple que le résultat dans (b) peut devenir faux si D n'est pas borné.*

2.2.6. Exercice : *Montrer qu'un convexe fermé C est l'intersection de tous les demi-espaces qui le contiennent.*

2.3 Le Lagrangien

2.3.1. Définition du Lagrangien

Avec les notations comme avant le 2.1.1, $p \geq 1$, *et* $K := \{x \in \mathbb{R}^p, x \geq 0\}$, $K^* := \{\lambda \in \mathbb{R}^{1\times p}, \lambda \geq 0\}$, *on définit le Lagrangien*

$$L : S \times K^* \mapsto \mathbb{R}, \quad L(x,\lambda) = f(x) + \lambda g(x).$$

On introduit le problème dual de (P)

$$(D) \quad \sup\{w(\lambda) : \lambda \in K^*\}, \quad w(\lambda) := \inf\{L(x,\lambda) : x \in S\} \in \mathbb{R} \cup \{-\infty\}.$$

2.3.2. Rémarques sur le Lagrangien

(a) Pour tout $\lambda \in K^*$ *fixé, la fonction* $x \mapsto L(x,\lambda)$ *est une somme de fonctions convexes. Donc le programme auxiliaire* $\inf\{L(x,\lambda) : x \in S\}$ *est un programme convexe. Ce programme auxiliaire est bien plus simple à résoudre car il nous restent que des contraintes affines d'égalité. Plus précisément, d'après 1.5.2(b),(c), ce problème admet une solution optimale* $x(\lambda)$ *ssi* $\exists \mu$ *t.q.* $\nabla_x L(x(\lambda),\lambda) = \mu A$ *(ici* $m \geq 1$, *pour* $m = 0$ *nous obtenons* $\nabla_x L(x(\lambda),\lambda) = 0$*).*

(b) Pour tout $x \in S$ *fixé, la fonction* $\lambda \mapsto -L(x,\lambda)$ *est affine et alors convexe, et donc* $-w$ *est convexe par 1.2.4(b). Par conséquent,* $\inf\{-w(\lambda) : \lambda \in K^*\} = -\sup\{w(\lambda) : \lambda \in K^*\}$, *ou encore le problème* (D), *à un changement de signe près, est un programme convexe. Encore une fois on a des contraintes très simples pour* (D) *mais, généralement,* w *n'est pas*

différentiable... [6]

Le résultat suivant permet de comparer les valeurs optimales des problèmes (P) et (D).

2.3.3. Lemme de dualité

Pour tout $x \in \mathbb{R}^d$ réalisable pour (P) pour tout $\lambda \in K^*$ nous avons

$$f(x) \geq L(x, \lambda) \geq w(\lambda),$$

en particulier nous obtenons un **saut de dualité** $opt(P) - opt(D) \geq 0$.

Démonstration. Comme x est réalisable pour (P) et λ réalisable pour (D), nous avons $\lambda g(x) \leq 0$, et alors $f(x) \geq f(x) + \lambda g(x) = L(x, \lambda)$. L'inégalité $L(x, \lambda) \geq w(\lambda)$ provient de la définition de $w(\lambda)$. □

Il serait intéressant d'identifier des classes de problèmes où le saut de dualité vaut zéro, car ceci permettrait de trouver la valeur optimale (et parfois une solution optimale) de (P) en résolvant (D), ce dernier ayant des contraintes bien plus simples, voir 2.3.8(b).

2.3.4. Exemple : programme linéaire

(a) Soit $(P) : \inf\{hx : b - Bx \leq 0\}$ avec $S = \mathbb{R}^d$, alors $L(x, \lambda) = hx + \lambda(b - Bx) = (h - \lambda B)x + \lambda b$. Donc $w(\lambda) = -\infty$ si $h - \lambda B \neq 0$, et $w(\lambda) = \lambda b$ sinon. Par conséquent, nous obtenons le problème dual

$$(D) : \sup\{\lambda b : \lambda \geq 0, h - \lambda B = 0\}.$$

6. Donc notre théorème KKT ne s'applique pas pour (D). En effet, l'optimisation convexe non différentiable avec des notations comme des sous-gradients dépasse largement le cadre de ce cours.

(b) Soit $(P) : \inf\{gx : Ax = a, x \geq 0\}$, *ici* $S = \{x \in \mathbb{R}^d : Ax = a\}$, $L(x, \lambda) = fx - \lambda x$. *Si* $w(\lambda) > -\infty$ *alors* $x(\lambda)$ *est caractérisé par :* $\exists \mu : f - \lambda = \mu A$, *et donc* $w(\lambda) = (f - \lambda)x = \mu Ax = \mu a$, *donc le problème dual prend la forme* $(D) : \sup\{\mu a : f - \mu A \geq 0\}$. *En comparant avec la partie (a), à une transposée et un changement de signe près, le dual du dual est le primal.*

2.3.5. Exemple : un programme quadratique

Revenons à l'exemple 1.2.3 $(P) : \inf\{\|x\|^2 : x \in S = \mathbb{R}^2, g(x) = 5 - 2x_1 - x_2 \leq 0\}$. *Ici* $L(x, \lambda) = x_1^2 + x_2^2 + \lambda(5 - 2x_1 - x_2)$, *et alors* $\nabla_x L(x, \lambda) = 0 = (2x_1 - 2\lambda, 2x_2 - \lambda)$, *donc* $x(\lambda) = (\lambda, \lambda/2)^T$, *qui est d'ailleurs réalisable pour* (P) *seulement pour* $\lambda \geq 2$. *Nous obtenons* $w(\lambda) = L(x(\lambda), \lambda) = -\frac{5}{4}(\lambda - 2)^2 + 5$, *et donc le programme dual* $(D) = \sup\{-\frac{5}{4}(\lambda - 2)^2 + 5 : \lambda \in [0, +\infty)\}$ *avec solution optimale* $\underline{\lambda} = 2$, $\underline{x} = x(2) = (2, 1)^T$, *et un saut de dualité qui vaut zéro.*

2.3.6. Exemple qui ne vérifie pas la condition (2.1) de Slater.

Considérons $\inf\{f(x) = (x + 1)^2 : x \in S = \mathbb{R}, g(x) \leq 0\}$, *avec* $g(x) = x^2$ *si* $x < 0$ *et* $g(x) = 0$ *pour* $x \geq 0$. *Notons que* $g(x) \leq 0$ *ssi* $x \in [0, +\infty)$. *On vérifie que* g *est bien continûment différentiable et que dans le cas* $x \geq 0$ *nous avons* $\nabla_x f(x, \lambda) = 2(x + 1) = 0$ *n'admettant pas de solution. Par contre, pour* $x < 0$ *l'équation* $\nabla_x f(x, \lambda) = 2(x + 1) + 2\lambda x = 0$ *admet la solution* $x(\lambda) = -1/(1 + \lambda)$, *avec* $w(\lambda) = L(x(\lambda), \lambda) = \lambda/(1 + \lambda)$, *donnant lieu au problème dual* $(D) : \sup\{\lambda/(1 + \lambda) : \lambda \in [0, +\infty)\}$, *sans solution optimale dans* $[0, +\infty)$. *Néanmoins, les deux valeurs optimales sont bien les mêmes.*

2.3.7. Définition d'un point-col

Le couple $(\underline{x}, \underline{\lambda})$ *est dit point-col si* $(\underline{x}, \underline{\lambda}) \in S \times K^*$, *et*

$$\forall x \in S \, \forall \lambda \in K^* : \quad L(x, \underline{\lambda}) \geq L(\underline{x}, \underline{\lambda}) \geq L(\underline{x}, \lambda).$$

2.3.8. Lemme : Caractérisation d'un point-col

(a) *Le couple* $(\underline{x}, \underline{\lambda})$ *est point-col ssi il vérifie le système* (KKT) *introduit dans 2.1.1 (avec* λ *remplacé par* $\underline{\lambda}$*).*

(b) *Si* L *admet un point-col* $(\underline{x}, \underline{\lambda})$, *alors* $\underline{x}$ *est solution optimale de* (P) *et* $\underline{\lambda}$ *est solution optimale de* (D), *avec un saut de dualité qui vaut* 0.

(c) *Réciproquement, si* $\underline{x}$ *est réalisable pour* (P) *et* $\underline{\lambda}$ *est réalisable pour* (D) *avec* $f(\underline{x}) = w(\underline{x})$ *alors le couple* $(\underline{x}, \underline{\lambda})$ *est un point-col.*

Démonstration. Pour démontrer (a), soit $(\underline{x}, \underline{\lambda})$ un point-col, alors $\underline{x} \in S$, et $\underline{\lambda} \in K^*$. Par définition de K^*, $t\lambda \in K^*$ si $\lambda \in K^*$ et $t \in [0, +\infty)$. Par conséquent,

$$\forall t \in [0, +\infty) \, \forall \lambda \in K^* : \quad L(\underline{x}, t\lambda) = f(\underline{x}) + t\lambda g(\underline{x}) \leq L(\underline{x}, \underline{\lambda})$$

par définition d'un point-col. En faisant tendre $t \to \infty$, nous concluons que, pour tout $\lambda \in K^*$, nous avons $\lambda g(\underline{x}) \leq 0$, et alors $g(\underline{x}) \leq 0$, ce qui donne les deux premières relations de (KKT). En prenant $t = 0$, nous avons également $L(\underline{x}, 0) - L(\underline{x}, \underline{\lambda}) \leq 0$ et alors $\underline{\lambda} g(\underline{x}) \geq 0$, ce qui donne la troisième relation de (KKT). La quatrième relation provient du fait que $L(x, \underline{\lambda}) \geq L(\underline{x}, \underline{\lambda})$ pour tout $x \in S$, autrement dit, le programme auxiliaire définissant $w(\underline{\lambda})$ admet $\underline{x}$ comme solution optimale. Il reste à appliquer 1.5.2(c) pour le cas $m = 0$ sans contraintes, et 1.5.2(b) pour le cas $m > 0$ de contraintes affines d'égalité.

Réciproquement, supposons que $(\underline{x}, \underline{\lambda})$ vérifie (KKT) (avec λ remplacé par $\underline{\lambda}$). Il en suit que $\underline{x} \in S$ et $\underline{\lambda} \in K^*$, de plus, $g(\underline{x}) \leq 0$. Alors la troisième relation de (KKT) donne $L(\underline{x}, \underline{\lambda}) = f(\underline{x}) \geq f(\underline{x}) + \lambda g(\underline{x}) = L(\underline{x}, \lambda)$ pour tout $\lambda \in K^*$. Aussi, la quatrième relation de (KKT) nous dit que $\underline{x}$ est solution optimale du problème auxiliaire associé à $w(\underline{\lambda})$, et alors $L(x, \underline{\lambda}) \geq L(\underline{x}, \underline{\lambda})$ pour tout $x \in S$. Donc $(\underline{x}, \underline{\lambda})$ est un point-col.
Pour démontrer (b), soit $(\underline{x}, \underline{\lambda})$ un point-col, $\underline{x}$ est alors réalisable pour (P), et $\underline{\lambda}$ est réalisable pour (D) selon les premières deux relations de (KKT). Les deux autres nous disent que

$$f(\underline{x}) = f(\underline{x}) + \underline{\lambda} g(\underline{x}) = L(\underline{x}, \underline{\lambda}) = w(\underline{\lambda}).$$

Donc d'après le lemme de dualité 2.3.3, pour tout x réalisable pour (P) pour tout λ réalisable pour (D)

$$f(x) \geq L(x, \underline{\lambda}) \geq L(\underline{x}, \underline{\lambda}) = w(\underline{\lambda}) \geq w(\lambda),$$

et la partie (b) en découle.
Finalement, pour démontrer (c), rappelons que, par le lemme de dualité 2.3.3 et par l'hypothèse de la partie (c)

$$\inf_{x \in S} L(x, \underline{\lambda}) = w(\underline{\lambda}) \geq L(\underline{x}, \underline{\lambda}) \geq f(\underline{x}) = w(\underline{\lambda}).$$

Nous avons donc égalité partout, ce qui implique que $\underline{\lambda} g(\underline{x}) = 0$. Comme $g(\underline{x}) \leq 0$, on en déduit que

$$\forall \lambda \in K^* : \quad L(\underline{x}, \underline{\lambda}) - L(\underline{x}, \lambda) = (\underline{\lambda} - \lambda) g(\underline{x}) = -\lambda g(\underline{x}) \geq 0,$$

et alors $(\underline{x}, \underline{\lambda})$ est un point-col. □

2.3.9. Le théorème KKT reformulé

Sous les hypothèses du théorème 2.1.1, $\underline{x}$ est solution optimale de (P) ssi $\exists \underline{\lambda}$ de sorte que $(\underline{x}, \underline{\lambda})$ est un point-col.

Démonstration. L'implication "$\Longleftarrow$" a été démontré déjà dans le lemme 2.3.8(b). L'autre implication est naturellement plus difficile car il nous faut construire ce vecteur $\underline{\lambda}$, à l'aide de la condition (2.1) de Slater, et des théorèmes de séparation. Soit $\underline{x}$ une solution optimale de (P), et posons

$$M = \left\{ \begin{bmatrix} r \\ z \end{bmatrix} \in \mathbb{R} \times \mathbb{R}^p : \exists x \in S \text{ t.q. } f(x) \leq r, g(x) + z \leq 0 \right\},$$
$$N = \left\{ \begin{bmatrix} \underline{r} \\ \underline{z} \end{bmatrix} \in \mathbb{R} \times \mathbb{R}^p : \underline{r} \leq f(\underline{x}), \underline{z} \geq 0 \right\}.$$

Nous allons couper la preuve en 9 parties.

(a) M.q. $N - M$ est convexe : Notons que N est un polyèdre et donc convexe, voir 1.2.2. D'après 1.2.1(g) appliqué à $N - M = N + (-) M$, il suffit alors de montrer que M est convexe. Soient $\begin{bmatrix} r_1 \\ z_1 \end{bmatrix}, \begin{bmatrix} r_2 \\ z_2 \end{bmatrix} \in M$ et $t \in [0, 1]$, c'est-à-dire, il existent $x_1, x_2 \in S$ t.q. $f(x_j) \leq r_j$ et $-g(x_j) - z_j \geq 0$ pour $j = 1, 2$. Alors $x := tx_1 + (1-t)x_2 \in S$ par convexité de S, voir 1.2.1(d). Aussi,

$$f(x) \leq tf(x_1) + (1-t)x_2) \leq tr_1 + (1-t)r_2 =: r$$

par convexité de f, et pour $z := tz_1 + (1-t)z_2$

$$-g(z) - z = \Big(-g(x) + tg(x_1) + (1-t)g(x_2)\Big) + t\Big(-g(x_1) - z_1\Big) + (1-t)\Big(-g(x_2) - z_2\Big) \geq 0$$

car les trois expressions entre parenthèses sont ≥ 0 par convexité des composantes de g et par définition de M. Donc $\begin{bmatrix} r \\ z \end{bmatrix} \in M$, et M est convexe.

(b) M.q. $0 \notin \mathrm{Int}(N-M)$: sinon, $\exists \epsilon > 0$ t.q. $\begin{bmatrix} \epsilon \\ 0 \end{bmatrix} \in N - M$, ce qui veut dire que $\exists x \in S \; \exists r, z, \underline{r}, \underline{z}$ avec $\underline{r} \leq f(\underline{x})$, $\underline{z} \geq 0$, $-g(x) - z \geq 0$, $f(x) \leq r$ et avec $\underline{r} - r = \epsilon$ et $\underline{z} - z = 0$. Par conséquent,

$$f(x) + \epsilon = f(x) + \underline{r} - r \leq \underline{r} \leq f(\underline{x}), \quad -g(x) = -g(x) - z + \underline{z} \geq 0,$$

c'est-à-dire, x est réalisable pour (P), avec $f(x) < f(\underline{x})$, en contradiction avec l'optimalité de $\underline{x}$.

(c) Appliquons le théorème 2.2.2 de séparation faible séparant le point $0 \in \mathbb{R}^{p+1}$ et le convexe $N - M$: sachant que $\begin{bmatrix} f(x) \\ -g(x) \end{bmatrix} \in M$ pour tout $x \in S$ alors $\exists \phi = [\gamma, \lambda] \in \mathbb{R}^{1\times(p+1)}$, $\phi \neq 0$, t.q.

$$(*) \quad \forall x \in S \, \forall r \leq f(\underline{x}) \, \forall z \geq 0 : \quad \phi \begin{bmatrix} r - f(x) \\ g(x) + z \end{bmatrix} = \gamma(r - f(x)) + \lambda(z + g(x)) \geq \phi 0 = 0.$$

(d) Montrons que $\lambda \geq 0$: il suffit de remplacer dans $(*)$ la quantité $z \geq 0$ par tz avec $t \in [0, +\infty)$ et $z \geq 0$, et de faire tendre $t \to \infty$, ce qui donne que $\lambda z \geq 0$ pour tout $z \geq 0$, et alors $\lambda \geq 0$.

(e) Montrons que $\gamma \leq 0$: il suffit de faire tendre $r \to -\infty$ dans $(*)$.

(f) Montrons que $\gamma \neq 0$: sinon, $\gamma = 0$ et alors $\lambda \neq 0$ par (c) ce qui implique qu'il existe un $y \in \mathbb{R}^p$ t.q. $\lambda y > 0$. Avec $\widetilde{x}$ comme dans la condition (2.1) de Slater, posons $z = 0$ et $x = \widetilde{x}$ dans $(*)$, alors $\lambda g(\widetilde{x}) \geq 0$. Pour $\epsilon > 0$ suffisament petit, (2.1) implique que $g(\widetilde{x}) + \epsilon y \leq 0$, mais $\lambda(g(\widetilde{x}) + \epsilon y) = \lambda g(\widetilde{x}) + \epsilon \lambda y \geq \epsilon \lambda y > 0$, en contradiction avec la partie (d).

(g) Posons $\underline{\lambda} := -\lambda/\gamma$, alors $\underline{\lambda} \in K^*$ par (d)–(f). En posant $r = f(\underline{x})$ dans $(*)$, nous concluons que $f(\underline{x}) - f(x) \leq \underline{\lambda}(z + g(x))$ pour tout $x \in S$ et $z \geq 0$. En posant $z = -g(\underline{x})$ nous concluons que $L(x, \underline{\lambda}) \geq L(\underline{x}, \underline{\lambda})$ pour tout $x \in S$ et, en posant $z = 0$ et $x = \underline{x}$, que $\underline{\lambda} g(\underline{x}) \geq 0$.

(h) M.q. $\underline{\lambda} g(\underline{x}) = 0$: comme $\underline{x}$ est réalisable pour (P) par hypothèse et $\underline{\lambda} \in K^*$, nous obtenons $\underline{\lambda} g(\underline{x}) \leq 0$, et pour conclure il suffit de combiner avec la dernière inégalité de la partie (g).

(j) Concluons que $(\underline{x}, \underline{\lambda})$ est un point-col : il suffit d'appliquer (h) et d'observer que, pour tout $\lambda \in K^*$

$$L(\underline{x}, \underline{\lambda}) - L(x, \underline{\lambda}) = (\underline{\lambda} - \lambda) g(\underline{x}) = -\lambda g(\underline{x}) \geq 0.$$

□

Une combinaison de 2.3.9 avec 2.3.8(a) donne une preuve du théorème 2.1.1 de Karush, Kuhn et Tucker.

2.3.10. Retour à l'exemple 2.3.6.
Dans cet exemple, il existe une solution optimale $\underline{x} = 0$ de (P), avec valeur optimale $f(\underline{x}) = 1$. Aussi, 1 est une valeur optimale du problème (D), mais il n'existe pas une solution optimale de (D), et donc, par le lemme 2.3.8(b), nous n'avons pas un point-col. Ceci n'est pas en contradiction avec le théorème 2.3.9 car, pour cet exemple, la condition de Slater faisant partie des hypothèses du théorème 2.1.1 n'était pas satisfaite. Cherchons à comprendre à quel endroit la preuve du théorème 2.3.9 cesse d'être valable. Comme dans cet exemple

$d = p = 1, m = 0$, les ensembles M et N sont des sous-ensembles convexes du $\mathbb{R}^2$, que l'on peut tracer. Introduisons la courbe paramétrée

$$\sigma : \mathbb{R} \mapsto \mathbb{R}^2, \quad \sigma(x) = \begin{bmatrix} f(x) \\ -g(x) \end{bmatrix} = \begin{cases} \begin{bmatrix} (x+1)^2 \\ 0 \end{bmatrix} & si\ x \leq 0 \\ \begin{bmatrix} (x+1)^2 \\ -x^2 \end{bmatrix} & si\ x > 0 \end{cases}$$

alors on vérifie que

$$M = \sigma(\mathbb{R}) - U, \quad N = \sigma(0) + U,$$

avec l'orthant $U = (-\infty, 0] \times [0, +\infty)$. Un petit dessin montre que $N - M = \sigma(0) - \sigma(\mathbb{R}) + U \subset \mathbb{R} \times [0, \infty)$ est un ensemble fermé, que 0 appartient au bord de $N - M$, et que le bord de $N - M$ dans un voisinage de l'origine est donné par une partie de $\sigma(0) - \sigma(\mathbb{R})$. Comme $\sigma'(0) = [1, 0]^T$ existe, tout plan séparant l'origine de $N - M$ est forcement un plan d'appui en 0, avec une normale $\phi \in \mathbb{R}^{1\times 2}$ orthogonale à $\sigma'(0)$, et donc $\gamma = 0$, en contradiction avec la partie (f) de la preuve.

2.3.11. Exercice : *Nous souhaitons montrer que le saut de dualité entre les deux programmes linéaires dans le 2.3.4(b) vaut zéro. On suppose que le programme (P) admet une solution optimale $\underline{x}$, et que $\beta < f\underline{x}$.*

(a) En se servant du Lemme de Farkas 2.2.4, montrer qu'il existe un $\gamma \in \mathbb{R}$ et un $\mu \in \mathbb{R}^{1\times m}$ de sorte que

$$-\mu A + \gamma f \geq 0, \quad -\mu a + \gamma\beta < 0.$$

(b) En multipliant par $\underline{x}$, vérifier que $\gamma > 0$, et que l'on peut choisir $\gamma = 1$.

(c) Conclure.

2.3.12. Exercice :

Considérons le problème quadratique

$$(P)\,\{f(x) : \forall j = 1, ..., m : g_j(x) \leq 0\}, \quad f(x) = \frac{1}{2}x^T A_0 x + b_0^T x, \quad g_j(x) = \frac{1}{2}x^T A_j x + b_j^T x + c_j,$$

avec A_0 sdp d'ordre n, $A_1, ..., A_m$ ssdp d'ordre n, $b_0, ..., b_m \in \mathbb{R}^n$, et $c_1, ..., c_m \in \mathbb{R}$. Notre but est de démontrer que si $\underline{\lambda}$ est une solution optimale du problème dual (D) alors $x(\underline{\lambda})$ (qui minimise $x \mapsto L(x, \underline{\lambda})$ sur $\mathbb{R}^n$) est l'unique solution optimale de (P), avec un saut de dualité qui vaut 0.

(a) Écrire le problème dual (D) et vérifier que

$$\forall \lambda \geq 0 : \quad w(\lambda) = L(x(\lambda), \lambda) = c(\lambda) - \frac{1}{2}b(\lambda)^T A(\lambda)^{-1} b(\lambda), \quad x(\lambda) = -A(\lambda)^{-1} b(\lambda),$$

avec $A(\lambda) = A_0 + \lambda_1 A_1 + ... + \lambda_m A_m$, $b(\lambda) = b_0 + \lambda_1 b_1 + ... + \lambda_m b_m$, et $c(\lambda) = \lambda_1 c_1 + ... + \lambda_m c_m$.

(b) Pour M, N des matrices d'ordre n, M inversible, et $t \in \mathbb{R}$, vérifier que

$$\frac{d}{dt}(M + tN)(0) = N, \quad \frac{d}{dt}(M + tN)^{-1}(0) = -M^{-1} N M^{-1}.$$

En déduire que, dans un voisinage de l'orthant positif, w admet des dérivéees partielles, avec valeur

$$\frac{\partial w}{\partial \lambda_j}(\lambda) = g_j(x(\lambda)).$$

(N.B. : généralement, l'objectif w du problème (D) n'est pas différentiable).

(c) Vérifier que $\underline{\lambda}$ est une solution optimale de notre (D) ssi

$$\forall j = 1, ..., m : \frac{\partial w}{\partial \lambda_j}(\underline{\lambda}) \leq 0, \quad \underline{\lambda}_j \frac{\partial w}{\partial \lambda_j}(\underline{\lambda}) = 0.$$

Conclure.

Chapitre 3

Algorithmes d'optimisation sans contraintes

Dans la suite de ce cours on s'intéressera à comment approcher numériquement la valeur optimale (et éventuellement une solution optimale) d'un problème d'optimisation convexe. On commencera par traiter le cas d'un programme sans contraintes

$$(P): \quad \inf\{f(x) : x \in \mathbb{R}^d\}.$$

Pour pouvoir analyser la convergence de notre algorithme, il sera utile de supposer les propriétés suivantes

$(H1)$ f est convexe de classe $\mathcal{C}^2(\mathbb{R}^d)$;
$(H2)$ l'ensemble de niveau $C(x^{(0)}) = \{x \in \mathbb{R}^d : f(x) \leq f(x^{(0)})\}$ est compact ;
$(H3)$ f est elliptique : $\exists m > 0$ t.q. $\forall x \in C(x^{(0)})\ \forall v \in \mathbb{R}^d : v^T \nabla^2 f(x) v \geq m v^T v$.

Rappelons que les hypothèses $(H1), (H2)$ assurent l'existence d'une solution optimale, et $(H3)$ la convexité stricte de f et donc l'unicité de cette solution optimale, notée par $\underline{x}$.

Les hypothèses $(H1), (H2)$ assurent aussi que $x \mapsto \|\nabla^2 f(x)\|$ est continue, et alors

$$\exists M > 0 \text{ t.q. } \forall x \in C(x^{(0)})\ \forall v \in \mathbb{R}^d : v^T \nabla^2 f(x) v \leq M v^T v. \tag{3.1}$$

Rappelons du 1.5.2(c) que $\underline{x}$ est l'unique solution du système $\nabla f(x) = 0$ à d inconnues et d équations non linéaires. Plus généralement, le résultat suivant montre que $\|\nabla f(x)\|$ "petit" signifie que $f(x)$ est "proche" de la valeur optimale $f(\underline{x})$.

3.0.1. Exercice
En s'inspirant de la preuve du 3.1.2, montrer que, sous les hypothèses $(H1), (H2), (H3)$,

$$\forall x \in C(x^{(0)}) : \quad \frac{\|\nabla f(x)\|^2}{2m} \geq f(x) - f(\underline{x}) \geq \frac{\|\nabla f(x)\|^2}{2M}.$$

3.0.2. Exercice : *Soit f elliptique dans $\mathbb{R}^d$*

$$\exists \widetilde{m} \text{ t.q. } \forall v \in \mathbb{R}^x \forall x \in \mathbb{R}^d : \quad v^T \nabla^2 f(x) v \geq \widetilde{m} \|v\|^2.$$

En déduire que $C(y)$ est compact pour tout $y \in \mathbb{R}^d$.

3.1 Algorithmes de descente

Nous cherchons à construire une suite $((x^{(j)}))_j \subset \mathbb{R}^d$ que l'on souhaite faire converger vers une solution optimale de (CP), par minimisation successive d'une fonction à une variable réelle. A l'étape j on dispose de $x^{(j)}$ et on cherche à construire une direction $d^{(j)}$ ainsi qu'un nouvel itéré $x^{(j+1)} = x^{(j)} + t_j d^{(j)}$ avec $t_j \in [0, +\infty)$ déduit par minimisation

$$t_j = \arg\min\{q_j(t) : t \in [0, +\infty)\}, \quad q_j(t) := f(x^{(j)} + td^{(j)}).$$

Pour s'assurer que les valeurs $f(x^{(j)})$ décroissent, la direction $d^{(j)}$ est construite de sorte que la dérivée directionnelle vérifie $f'(x^{(j)}; d^{(j)}) = \nabla f(x^{(j)})d^{(j)} = q_j'(0) \leq 0$, et $q_j'(0) = 0$ ssi $x^{(j)}$ est une solution optimale.

3.1.1. Lemme : monotonie des valeurs
Si $q_{j-1}'(0) < 0$ *alors* $q_{j-1}(0) = f(x^{(j-1)}) > q_j(0) = f(x^{(j)})$, *en particulier* $x^{(j)} \in C(x^{(0)})$.

Le lemme 3.1.1 nous dit que la suite de valeurs $(f(x^{(j)}))_j$ est décroissante, et bornée car incluse dans le compact $C(x^{(0)})$ par $(H1)$, par conséquent elle converge. Cependant, pour relier la limite avec la valeur optimale de (CP), il faudra savoir plus sur les directions $d^{(j)}$.

Si $\nabla f(x^{(j)}) \neq 0$, parmi toutes les directions $d^{(j)}$ de longueur fixe $\|\nabla f(x^{(j)})\|$ c'est la transposée du gradient $d^{(j)} = -\nabla f(x^{(j)})^T$ qui minimise la dérivée directionnelle $f'(x^{(j)}; d^{(j)})$. Cette direction donne lieu à la méthode de la plus forte descente, aussi appelée méthode de Cauchy.

3.1.2. Méthode de la plus forte descente=Méthode de Cauchy

Posons $d^{(j)} = -\nabla f(x^{(j)})^T$ *pour tout* j, *alors* $x^{(j)}$ *est une solution optimale de* (CP) *ssi* $q_j'(0) = 0$. *Si on exclut ce cas pour tout* j, *alors*

(a) Sous les hypothèses $(H1)$ *et* $(H2)$ *:* $\nabla f(x^{(j)}) \to 0$ *pour* $j \to \infty$.

(b) Sous les hypothèses $(H1)$, $(H2)$ *et* $(H3)$ *:* $(x^{(j)})_j$ *converge vers l'unique solution* $\underline{x}$ *de* (CP).

(c) Pour tout $j \geq 0$

$$\Big(f(x^{(j+1)}) - f(\underline{x})\Big) \leq \Big(1 - \frac{m}{M}\Big)\Big(f(x^{(j)}) - f(\underline{x})\Big).$$

Structure de la preuve :

1. on calcule $q_j'(t), q_j'(0), q_j''(t)$;
2. par TAF pour q_j' on montre que $t_j \geq 1/M$;
3. par Taylor pour q_j en $1/M$ on montre que $\frac{|q_j'(0)|}{2M} \leq f(x^{(j)}) - f(x^{(j+1)})$;
4. par une somme télescopique on déduit que $q_j'(0) \to 0$ et alors $\nabla f(x^{(j)}) \to 0$ et $x^{(j)} \to \underline{x}$;
5. en minorant f par une forme quadratique, on montre que $\frac{|q_j'(0)|}{2m} \geq f(x^{(j)}) - f(\underline{x})$;
6. ...et on conclut.

Démonstration. En utilisant le 1.4.3, nous obtenons les derivées

$$\begin{aligned}
q_j'(t) &= -\nabla f(x^{(j)})\nabla f(x^{(j)} - t\nabla f(x^{(j)})^T),\\
q_j'(0) &= -\|\nabla f(x^{(j)})\|^2,\\
q_j''(t) &= -\nabla f(x^{(j)})\nabla^2 f(x^{(j)} - t\nabla f(x^{(j)})^T)\nabla f(x^{(j)})^T.
\end{aligned}$$

Montrons dans un premier temps que $t_j \geq 1/M$. Par TAF pour q_j', il existe un $\eta \in [0,1]$ de sorte que $0 = q_j'(t_j) = q_j'(0) + t_j q_j''(\eta t_j) \leq q_j'(0) + t_j M|q_j'(0)|$ en utilisant (3.1) et le fait que $x^{(j)}, x^{(j)} + t_j d^{(j)} \in C(x^{(0)})$ et que donc par convexité $x^{(j)} + \eta t_j d^{(j)} \in C(x^{(0)})$. Comme $q_j'(0) < 0$, ceci implique que $t_j \geq 1/M$.
Nous en déduisons que $x^{(j)} + (1/M)d^{(j)} \in C(x^{(0)})$ et alors par Taylor il existe $\eta' \in [0,1]$ de sorte que

$$q_j(t_j) \leq q_j(\frac{1}{M}) = q_j(0) + q_j'(0)\frac{1}{M} + \frac{q_j''(\eta'\frac{1}{M})}{2}\frac{1}{M^2} \leq q_j(0) - \frac{|q_j'(0)|}{2M},$$

où dans la dernière étape on a de nouveau appliqué (3.1). Ceci implique que

$$\frac{\|\nabla f(x^{(j)})\|^2}{2M} \leq f(x^{(j)}) - f(x^{(j+1)}) \leq f(x^{(j)}) - f(\underline{x}). \tag{3.2}$$

D'après le lemme 3.1.1, la suite des valeurs $(f(x^{(j)}))_j$ est décroissante et bornée, et donc convergeante avec limite F, ce qui avec (3.3) implique que

$$\sum_{j=0}^{k-1} \frac{\|\nabla f(x^{(j)})\|^2}{2M} \leq \sum_{j=0}^{k-1} (f(x^{(j)}) - f(x^{(j+1)}) \leq f(x^{(0)}) - F < \infty.$$

Ceci implique que $\nabla f(x^{(j)}) \to 0$ pour $j \to \infty$ ce qui démontre la partie (a).
D'après le lemme 3.1.1, la suite $(x^{(j)})_j$ reste dans le compact $C(x^{(0)})$, et tout point d'accumulation $\underline{x}$ vérifie forcément $\nabla f(\underline{x}) = 0$. La convexité stricte découlant de $(H3)$ implique alors que $(x^{(j)})_j$ admet la limite $\underline{x}$ qui est l'unique solution optimale de (CP), ce qui démontre la partie (b).
Pour une preuve de (c), rappelons que pour tout $x, y \in C(x^{(0)})$ nous avons par $(H3)$

$$\begin{aligned} f(y) &\geq f(x) + \nabla f(x)(y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(x + \eta''(y-x))(y-x) \\ &\geq f(x) + \nabla f(x)(y-x) + \frac{m}{2}\|y-x\|^2 =: h(y). \end{aligned}$$

Notons que le minimum de h pour $y \in \mathbb{R}^d$ est atteint pour $\widetilde{y} = x - \frac{1}{m}\nabla f(x)^T$ (mais on ne sait pas si ce $\widetilde{y}$ appartient à $C(x^{(0)})$). Par conséquent,

$$f(\underline{x}) = \inf_{y \in C(x^{(0)})} f(y) \geq \inf_{y \in C(x^{(0)})} h(y) \geq \inf_{y \in \mathbb{R}^d} h(y) = h(\widetilde{y}) = f(x) - \frac{\|\nabla f(x)\|^2}{2m},$$

et alors pour $x = x^{(j)}$

$$\frac{\|\nabla f(x^{(j)})\|^2}{2m} \leq f(x^{(j)}) - f(\underline{x}).$$

On déduit de cette dernière inégalité et de (3.2) que

$$0 \leq f(x^{(j+1)}) - f(\underline{x}) \leq f(x^{(j)}) - f(\underline{x}) - \frac{\|\nabla f(x^{(j)})\|^2}{2M} \leq \Big(1 - \frac{m}{M}\Big)\Big(f(x^{(j)}) - f(\underline{x})\Big).$$

□

On déduit alors la convergence $\left(f(x^{(j)})-f(\underline{x})\right) \leq \left(1-\frac{m}{M}\right)^j \left(f(x^{(0)})-f(\underline{x})\right) \to 0$ avec un taux géométrique $(1-m/M)$ mais proche de 1 car généralement $m \ll M$ (sauf si le Hessien $\nabla^2 f(x)$ est proche de l'identité). Cette convergence lente s'explique par un comportement Zigzag: deux directions consécutives sont orthogonales car $(d^{(j)})^T d^{(j+1)} = q_j'(t_j) = 0$.

En dimension $d = 2$, on peut s'imaginer $f(x)$ comme la hauteur d'une montagne (convexe) au point x dont on cherche à rejoindre la vallée $\underline{x}$. A l'étape j, si $x^{(j)}$ n'est pas encore une solution optimale de (CP), on se tourne et poursuit son trajet sur la demi-droite partant de $x^{(j)}$ dans la direction de la plus forte descente $d^{(j)}$ (qui est orthogonale à la courbe de niveau $\{x : f(x) = f(x^{(j)})\}$). On s'arrête au point le plus bas $x^{(j+1)}$ sur cette demi-droite (la démi-droite est tangentz à la courbe de niveau $\{x : f(x) = f(x^{(j+1)})\}$). Par deux petits dessins pour $f(x) = \|x\|^2$ et $f(x) = x_1^2 + 100x_2^2$ on se rend facilement compte que le comportement Zigzag peut nuire à la vitesse de convergence. [1]

La moralité de ces observations est que, pour trouver une bonne direction $y \in \mathbb{R}^d$ partant de $x \in \mathbb{R}^d$, il vaut mieux de ne pas linéariser f mais plutôt l'approcher par une forme quadratique

$$\begin{aligned} f(x+y) &\approx f(x) + \nabla f(x) y + \frac{1}{2} y^T W^{-1} y \\ &= f(x) - \frac{1}{2} \nabla f(x) W \nabla f(x)^T + \frac{1}{2} (y + W \nabla f(x)^T)^T W^{-1} (y + W \nabla f(x)^T) \end{aligned}$$

1. Une façon de remédier à ce comportement Zigzag est de choisir comme direction $d^{(j)}$ une combinaison linéaire appropriée de $d^{(j-1)}$ et $\nabla f(x^{(j)})^T$ (ou alors de $d^{(0)}$,...,$d^{(j-1)}$ et $\nabla f(x^{(j)})^T$), donnant lieu à des algorithmes CG(2) et CG non-linéaires. Ces algorithmes ne seront pas discutés dans ce cours, pour des formes quadratiques voir le cours d'ANAC2.

où W est sdp, et plus précisement une approximation de l'inverse du Hessien $\nabla^2 f(x)$. Dans ce cas, le second membre est minimum pour $y \in \mathbb{R}^d$ ssi $y = -W\nabla f(x)^T$. Rappelons qu'une matrice W est sdp ssi elle admet une décomposition de Cholesky $W = VV^T$ avec V inversible.

3.1.3. Méthode de quasi-Newton

Posons $d^{(j)} = -W_j \nabla f(x^{(j)})^T$ *pour tout* j, *avec* $W_j = V_j V_j^T$ *sdp de sorte que*

$$(H3'): \quad \exists m, M > 0 \, t.q. \, \forall j \geq 0 \forall x \in C(x^{(0)}) \forall y \in \mathbb{R}^d : \quad m \leq \frac{(V_j y)^T \nabla^2 f(x)(V_j y)}{\|y\|^2} \leq M.$$

Alors $x^{(j)}$ *est une solution optimale de* (CP) *ssi* $q_j'(0) = 0$. *Si on exclut ce cas pour tout* j, *alors, sous les hypothèses* $(H1)$, $(H2)$ *et* $(H3')$:

(a) $\nabla f(x^{(j)}) \to 0$ *pour* $j \to \infty$, *et* $(x^{(j)})_j$ *converge vers l'unique solution* $\underline{x}$ *de* (CP).

(b) Pour tout $j \geq 0$

$$\Big(f(x^{(j+1)}) - f(\underline{x})\Big) \leq \Big(1 - \frac{m}{M}\Big)\Big(f(x^{(j)}) - f(\underline{x})\Big).$$

Démonstration. En utilisant le 1.4.3, nous obtenons les derivées

$$\begin{aligned}
q_j'(t) &= -\nabla f(x^{(j)}) W_j \nabla f(x^{(j)} - tW_j \nabla f(x^{(j)})^T), \\
q_j'(0) &= -\nabla f(x^{(j)}) W_j \nabla f(x^{(j)})^T = -\|\nabla f(x^{(j)}) V_j\|^2, \\
q_j''(t) &= -\nabla f(x^{(j)}) W_j \nabla^2 f(x^{(j)} - tW_j \nabla f(x^{(j)})^T) W_j \nabla f(x^{(j)})^T.
\end{aligned}$$

Montrons dans un premier temps que $t_j \geq 1/M$. Par TAF pour q_j', il existe un $\eta \in [0,1]$ de sorte que $0 = q_j'(t_j) = q_j'(0) + t_j q_j''(\eta t_j) \leq q_j'(0) + t_j M |q_j'(0)|$, où dans

l'inégalité on a utilisé $(H3')$ et le fait que $x^{(j)}, x^{(j)} + t_j d^{(j)} \in C(x^{(0)})$ et donc par convexité $x^{(j)} + \eta t_j d^{(j)} \in C(x^{(0)})$. Comme $q_j'(0) < 0$, ceci implique que $t_j \geq 1/M$. Nous en déduisons que $x^{(j)} + (1/M)d^{(j)} \in C(x^{(0)})$, et alors, par Taylor, il existe $\eta' \in [0, 1]$ de sorte que

$$q_j(t_j) \leq q_j(\frac{1}{M}) = q_j(0) + q_j'(0)\frac{1}{M} + \frac{q_j''(\eta' \frac{1}{M})}{2}\frac{1}{M^2} \leq q_j(0) - \frac{|q_j'(0)|}{2M},$$

où dans la dernière étape on a de nouveau appliqué $(H3')$. Ceci implique que

$$\frac{|q_j'(0)|}{2M} \leq f(x^{(j)}) - f(x^{(j+1)}) \leq f(x^{(j)}) - f(\underline{x}). \tag{3.3}$$

D'après le lemme 3.1.1, la suite des valeurs $(f(x^{(j)}))_j$ est décroissante et bornée, et donc convergeante avec limite F, ce qui avec (3.3) implique que

$$\sum_{j=0}^{k-1} \frac{|q_j'(0)|}{2M} \leq \sum_{j=0}^{k-1} (f(x^{(j)}) - f(x^{(j+1)}) \leq f(x^{(0)}) - F < \infty.$$

Ceci implique que $q_j'(0) \to 0$ pour $j \to \infty$. Par hypothèses $(H1), (H2)$, le Hessien $\nabla^2 f$ est borné en norme sur le compact $C(x^{(0)})$ par une constante $c > 0$, et alors

$$\|\nabla f(x^{(j)})\|^2 \leq \frac{c}{m}\|V_j \nabla f(x^{(j)})^T\|^2 = \frac{c}{m}\,|q_j(0)| \to 0$$

par $(H3')$. Finalement, d'après le lemme 3.1.1, la suite $(x^{(j)})_j$ reste dans le compact $C(x^{(0)})$, et tout point d'accumulation $\underline{x}$ vérifie forcément $\nabla f(\underline{x}) = 0$. La convexité stricte découlant de $(H3')$ implique alors que $(x^{(j)})_j$ admet la limite $\underline{x}$ qui est l'unique

solution optimale de (CP), ce qui démontre la partie (a).
Pour une preuve de (b), rappelons que pour tout $x, y \in C(x^{(0)})$ nous avons par $(H3')$

$$\begin{aligned} f(y) &\geq f(x) + \nabla f(x)(y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(x+\eta''(y-x))(y-x) \\ &\geq f(x) + \nabla f(x)(y-x) + \frac{m}{2}\|V_j^{-1}(y-x)\|^2 =: h(y). \end{aligned}$$

Notons que le minimum de h pour $y \in \mathbb{R}^d$ est atteint pour $\widetilde{y} = x - \frac{1}{m} V_j V_j^T \nabla f(x)^T$ (mais on ne sait pas si ce $\widetilde{y}$ appartient à $C(x^{(0)})$). Par conséquent,

$$f(\underline{x}) = \inf_{y \in C(x^{(0)})} f(y) \geq \inf_{y \in C(x^{(0)})} h(y) \geq \inf_{y \in \mathbb{R}^d} h(y) = h(\widetilde{y}) = f(x) - \frac{\|\nabla f(x) V_j\|^2}{2m},$$

et alors pour $x = x^{(j)}$

$$\frac{|q_j'(0)|}{2m} \leq f(x^{(j)}) - f(\underline{x}).$$

On déduit de cette dernière inégalité et de (3.3) que

$$0 \leq f(x^{(j+1)}) - f(\underline{x}) \leq f(x^{(j)}) - f(\underline{x}) - \frac{|q_j'(0)|}{2M} \leq \Big(1 - \frac{m}{M}\Big)\Big(f(x^{(j)}) - f(\underline{x})\Big).$$

□

Notons que la méthode 3.1.2 de Cauchy est aussi une méthode de quasi-Newton avec $W_j = V_j = I$ pour tout j, ici l'hypothèse $(H3')$ se réduit à $(H3)$ et (3.1). D'ailleurs, la preuve du 3.1.2 est obtenue de celle du 3.1.3 en posant $W_j = V_j = I$. Le 3.1.3 et sa preuve nous permet de donner l'algorithme avec une condition d'arrêt et d'affirmer que l'algorithme s'arrête.

3.1.4. Algorithme de quasi-Newton

En entrée : $x^{(0)} \in \mathbb{R}^d$ *avec* $(H1)$, *tolérance* $\epsilon > 0$, *suite de matrices* W_j *sdp.*
Algo : *Pour* $j = 0, 1, ...$
poser $q_j(t) = f(x^{(j)} - tW_j \nabla f(x^{(j)})^T)$
calculer $q_j'(0) = -\nabla f(x^{(j)}) W_j \nabla f(x^{(j)})^T$
arrêt si $|q_j'(0)| < \epsilon$
minimisation : trouver $t_j = \arg\min_{t \geq 0} q_j(t)$
poser $x^{(j+1)} = x^{(j)} - t_j W_j \nabla f(x^{(j)})^T$
En sortie : $x^{(j)}$ *avec* $f(x^{(j)}) \leq \min_{x \in \mathbb{R}^d} f(x) + \frac{\epsilon}{2m}$.

Le cas particulier le plus important inclus dans 3.1.3 est $W_j = \nabla^2 f(x^{(j)})^{-1}$ donnant lieu à la formule

$$x^{(j+1)} = x^{(j)} - t_j \nabla^2 f(x^{(j)})^{-1} \nabla f(x^{(j)})^T, \quad t_j = \arg \min_{t \in [0, +\infty)} q_j(t).$$

Cette formule avec $t_j = 1$ pour tout j est connue comme la méthode de Newton pour résoudre le système non linéaire $\nabla f(x)^T = 0$, mais l'exemple évoqué dans l'exercice 3.1.9 montre que le choix de notre t_j est essentiel : en prenant $t_j = 1$ pour tout j on peut observer divergence pour $x^{(0)}$ loin de $\underline{x} = 0$.

3.1.5. Remarques sur la convergence pour la direction de Newton $W_j = \nabla^2 f(x^{(j)})^{-1}$

En supposant que $f \in \mathcal{C}^2$ *avec Hessien* $\nabla^2 f(x)$ *sdp pour tout* $x \in \mathbb{R}^d$, *la quantité*

$$M_j = \sup_{x, x' \in C(x^{(j)})} \max_{y \in \mathbb{R}^d} \frac{y^T \nabla^2 f(x) y}{y^T \nabla^2 f(x') y}$$

est finie par conpacité de $C(x^{(j)})$. Par conséquent, $(H3')$ est valable pour $W_j = \nabla^2 f(x^{(j)})^{-1}$ en prenant $m = 1/M_0$ et $M = M_0$. Comme $f(x^{(j)}) \to f(\underline{x})$, les ensembles $C(x^{(j)})$ sont des voisinages de plus en plus petits de $\underline{x}$, et donc $M_j \to 0$ pour $j \to \infty$. Une légère modification de la preuve du 3.1.3 donne

$$\Big(f(x^{(j+1)}) - f(\underline{x})\Big) \leq \Big(1 - \frac{1}{M_j^2}\Big)\Big(f(x^{(j)}) - f(\underline{x})\Big),$$

c'est-à-dire, la convergence est plus rapide que géométrique. En effet, si $f \in \mathcal{C}^3$, alors on peut montrer qu'il existe un $L > 0$ de sorte que

$$\Big(f(x^{(j+1)}) - f(\underline{x})\Big) \leq L\Big(f(x^{(j)}) - f(\underline{x})\Big)^2,$$

on obtient alors convergence quadratique : si $x^{(j)}$ est déjà si proche de $\underline{x}$ de sorte que $\epsilon_j := L(f(x^{(j)}) - f(\underline{x}))$ est petit, alors $\epsilon_{j+1} \leq \epsilon_j^2$ est bien plus petit (on double des chiffres significatifs de l'erreur après la virgule).

Mentionnons sans étude de convergence deux autres méthodes de quasi-Newton.

3.1.6. La direction de Newton simplifiée

Comme pour des larges dimensions il peut être assez coûteux d'évaluer l'inverse du Hessien à chaque itération, on peut s'imaginer avec $p > 0$ un entier une variante de la forme :

$$d^{(j)} = -W_j \nabla f(x^{(j)}), \quad W_j = \begin{cases} \nabla^2 f(x^{(j)})^{-1} & \text{si } j \text{ est un multiple de } p, \\ W_{j-1} & \text{sinon.} \end{cases}$$

On peut montrer que l'on obtient un comportement de convergence similaire mais ralenti par rapport à 3.1.5.

3.1.7. La méthode BFGS

Ici la matrice W_j est calculée récursivement par

$$W_0 = I, \quad W_{j+1} = W_j + [s, W_j y] \begin{bmatrix} \frac{1}{y^T s} + \frac{y^T W_j y}{(y^T s)^2} & \frac{-1}{y^T s} \\ \frac{-1}{y^T s} & 0 \end{bmatrix} [s, W_j y]^T,$$

avec $s = x^{(j+1)} - x^{(j)}$, $y = \nabla f(x^{(j+1)})^T - \nabla f(x^{(j)})^T$. Une preuve de convergence n'est pas triviale, on renvoie le lecteur sur [M]. Ici les matrices W_j sont spd et $W_{j+1} y = s$ (la propriété essentielle de BFGS car elle fait le lien entre W_j et l'inverse du Hessien $\nabla^2 f$), mais généralement l'hypothèse $(H3')$ n'est pas valable.

Dans l'exercice suivant on aborde des choix alternatifs pour notre pas t_j.

3.1.8. Exercice

Peut-on assurer convergence si on choisit dans (3.1.2) *ou* (3.1.3) *le pas t_j (non unique) de sorte que*

$$q_j(t_j) \geq \frac{1}{2}(q_j(0) + \min_{t \in [0, +\infty)} q_j(t)? \tag{3.4}$$

Et si on limite le pas en se donnant $\beta \geq 1/M$ et

$$t_j = \arg \min_{t \in [0, \beta]} q_j(t)$$

pour tout j ?

3.1.9. Exercice

Considérons la fonction univariée $f(x) = \log(\cosh(x))$. *Vérifier que le minimum de* f *sur* $\mathbb{R}$ *est atteint en* $\underline{x} = 0$, *et que les hypothèses* $(H1), (H2), (H3), (H3')$ *pour* $W_j = 1/f''(x^{(j)})$ *sont valables. Vérifier également que la méthode de Newton (où* $t_j = 1$ *pour tout* j*) ne converge pas forcément.*

3.1.10. Exercice : Conditionnement des ensembles de niveau

On définit la largeur d'un convexe $C \subset \mathbb{R}^n$ *dans la direction* q, $\|q\| = 1$ *par*

$$W(C, q) = \sup_{z \in C} q^T z - \inf_{z \in C} q^T z.$$

La largeur minimale et la largeur maximale de C *sont définies par*

$$w_{\min} = \inf_{\|q\|=1} W(C, q), \quad w_{\max} = \sup_{\|q\|} W(C, q),$$

et le conditionnement du convexe C *est donné par*

$$cond(C) = \frac{W_{\max}^2}{W_{\min}^2}.$$

Il mesure l'excentricité de l'ensemble (petit si l'ensemble est presque sphérique, grand si l'ensemble est plus large dans certaines directions que d'autres).

(a) *Soit* $\mathcal{E} = \{x : \quad (x - x_0)^T A^{-1} x - x_0) \leq 1\}$ *avec* A *sdp. Calculer son conditionnement.*

(b) *On définit les sous-ensembles de niveau de* f *fonction elliptique par*

$$C_\alpha = \{x : \quad f(x) \leq \alpha\} \text{ avec } f(\underline{x}) \leq \alpha < f(x^{(0)}).$$

Le conditionnement de ces ensembles est relié à la vitesse de convergence des méthodes itératives de descente. Nous allons donner une borne supérieure de ce conditionnement.

(b1) Comparons C à des boules. Montrer que C_α vérifie $B_{\min} \subset C_\alpha \subset B_{\max}$, avec

$$B_{\min} = B\left(\overline{x}, \left(\frac{2(\alpha - f(\underline{x}))}{M}\right)^{1/2}\right), \quad B_{\max} = B\left((\overline{x}, \left(\frac{2(\alpha - f(\underline{x}))}{m}\right)^{1/2}\right)$$

(b2) Montrer que alors $cond(C_\alpha) \leq \frac{M}{m}$.

3.1.11. Exercice : direction de la plus forte descente dans la norme ℓ^1

On considère une méthode itérative où la direction (normalisée) à chaque étape est donnée par

$$d_n(x) = arg\min\left\{\nabla f(x)^T v, \quad \| v \|_1 \leq 1\right\}$$

et on prend comme direction $d(x) = \| \nabla f(x) \|_\infty d_n(x)$.

1. *Déterminer $d(x)$.*
2. *On considère la méthode itérative*

$$x^{(k+1)} = x^{(k)} + t^{(k)} d(x^{(k)}) \text{ avec } t^{(k)} = arg\min_{t \geq 0} f(x^{(k)} + t d(x^{(k)})).$$

Montrer la convergence de la méthode sous les conditions habituelles pour f.

3. *Expliciter cet algorithme appliqué au problème suivant*

$$\min \sum_{i,j=1}^{n} M_{i,j}^2 d_i^2/d_j^2 = \| \, DMD^{-1} \, \|_F,$$

avec D matrice diagonale et $M = (M_{i,j}) \in \mathbb{R}^{n\times n}$ Suggestion : considérer le changement de variables $x_i = 2\log d_i$.

3.1.12. Exercice : Critère d'arrêt pour la méthode de Newton

Soit

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

l'approximation quadratique de f au voisinage de x. On pose $d_N(x) = -\nabla^2 f(x)^{-1} \nabla f(x)^T$ la direction de Newton et $\lambda^2(x) = \nabla f(x) \nabla^2 f(x)^{-1} \nabla f(x)^T$. En calculant $f(x) - \inf_y \tilde{f}(y)$, montrer qu'un critère d'arrêt possible pour la méthode de Newton est $\lambda^2(x)/2 < \epsilon$.

3.1.13. Exercice : Méthode de Fletcher-Reeves

On considère l'algorithme suivant :

Initialisations : $x^{(0)}$ donné, $d_0 = -\nabla f(x^{(0)})$;
Etape k

choisir λ_k minimisant $f(x^{(k)} + \lambda d_k)$
poser $x^{(k+1)} = x^{(k)} + \lambda_k d_k$
définir $d_{k+1} = -\nabla f(x^{(k+1)}) + \beta_k d_k$, avec $\beta_k = \|\nabla f(x^{(k+1)})\|^2 / \|\nabla f(x^{(k)})\|^2$
test d'arrêt : si vérifié alors FIN.

Vérifier que si f est quadratique c'est la méthode du gradient conjugué. Il s'agit donc d'une extension du gradient conjugué à des fonctions quelconques.

3.1.14. Exercice : Méthodes de quasi-Newton

Les méthodes de quasi-Newton sont définies par une formule itérative du type

$$x^{(k+1)} = x^{(k)} - \lambda_k H_k \nabla f(x^{(k)}) \text{ avec} \tag{3.5}$$

- λ_k *choisi de faqn à minimiser* $g(\lambda) = f(x^{(k)} + \lambda d_k)$ *dans la direction* $d_k = -H_k \nabla f(x^{(k)})$;
- H_k *une suite de matrices qui approchent l'inverse du Hessien.*

On va considérer le choix particulier suivant pour ces matrices :

$$\begin{cases} H_0 \text{ matrice symétrique quelconque} \\ H_{k+1} = H_k + \Delta_k \end{cases}$$

avec Δ_k *une correction de rang 1,* $\Delta_k = \alpha_k u_k u_k^T$ *choisie de sorte que*

$$H_{k+1}[\nabla f(x^{(k+1)}) - \nabla f(x^{(k)})] = x^{(k+1)} - x^{(k)}.$$

(a) Montrer que $\forall k$ H_k *est symétrique.*

(b) Obtenir la formule explicite pour le calcul de H_{k+1} *:*

$$H_{k+1} = H_k + \frac{(\delta_k - H_k\gamma_k)(\delta_k - H_k\gamma_k)^T}{\gamma_k^T(\delta_k - H_k\gamma_k)} \tag{3.6}$$

avec $\delta_k = x^{(k+1)} - x^{(k)}$ *et* $\gamma_k = \nabla f(x^{(k+1)}) - \nabla f(x^{(k)})$.

(c) *Soit* $H_k = B_k B_k^T$, *et considérons le changement de variables* $x = \phi(\widetilde{x}) = B_k\widetilde{x}$, $\widetilde{f}(\widetilde{x}) = f(x) = f(B_k\widetilde{x})$. *Vérifier que si* $x^{(k)} = \phi(\widetilde{x}^{(k)})$ *alors* $x^{(k+1)} = \phi(\widetilde{x}^{(k+1)})$, *avec* $\widetilde{x}^{(k+1)}$ *obtenu par une itération de la méthode de Cauchy pour* $\widetilde{f}$ *partant de* $\widetilde{x}^{(k)}$.

(d) *Soit* f *une fonction quadratique avec Hessien* A *défini positif. On considère une suite de points* $x^{(0)}, x^{(1)} = x^{(0)} + \delta_0, \cdots, x^{(n)} = x^{(n-1)} + \delta_{n-1}$ *avec* $\{\delta_i\}_{i=0}^{n-1}$ *des directions linéairement indépendantes. Montrer que la suite des matrices* H_k *obtenues par (3.6) converge en au plus* n *étapes vers la matrice* A^{-1}.

3.2 La recherche linéaire

Les algorithmes abordés dans le chapitre précédent nécessitent à chaque itération de minimiser (de manière exacte) la fonction d'une variable réelle $q_j(t) = f(x^{(j)} + td^{(j)})$ sur $[0, +\infty)$ sachant que $q_j'(0) < 0$. Ceci revient à chercher un zéro de la fonction q_j'.

On pourrait s'imaginer d'appliquer une méthode de dichotomie pour approcher un zéro de q_j', il reste alors à spécifier une condition d'arrêt pour assurer par exemple (3.4). Ceci nécessite l'évaluation répétée de q_j'. Nous discutons ici une autre approche : la recherche linéaire de Goldstein nécessite seulement d'évaluer $q_j'(0)$ et quelques valeurs de q_j pour chaque j, ce qui simplifie grandement la recherche du pas t_j. Une variante dite d'Armijo et une variante dite de Wolfe seront abordées en TD. Dans la recherche linéaire de Goldstein on se fixe deux scalaires $0 < \gamma_1 < \gamma_2 < 1$ (par exemple $\gamma_1 = 0.3, \gamma_2 = 0.7$). On dira que le scalaire t vérifie les conditions de

Goldstein pour q_j si $t > 0$, et

$$q_j(0) + t\gamma_2 q_j'(0) \leq q_j(t) \leq q_j(0) + t\gamma_1 q_j'(0).$$

Comme $q_j'(0) < 0$, un petit dessin montre que l'inégalité à gauche (et à droite) assure que t est assez grand (et assez petit, respectivement). Par convexité de q_j, l'ensemble des pas vérifiant la condition de Goldstein est un intervalle compact. Comme la méthode de Cauchy est une méthode de quasi-Newton avec $W_j = I$, il suffit d'analyser la convergence de 3.1.3 pour ce nouveau choix du pas.

3.2.1. Méthode de quasi-Newton avec recherche linéaire de Goldstein

Sous les hypothèses $(H1)$, $(H2)$ *et* $(H3')$, *considérons la suite définie récursivement par*

$$x^{(j+1)} = x^{(j)} - t_j W_j \nabla f(x^{(j)})^T, =$$

avec $t_j > 0$ *vérifiant les conditions de Goldstein pour* $q_j(t) = f(x^{(j)} - tW_j \nabla f(x^{(j)})^T)$. *Alors* $x^{(j)}$ *est une solution optimale de* (CP) *ssi* $q_j'(0) = 0$. *Si on exclut ce cas pour tout* j *alors, pour tout* $j \geq 0$,

$$\Big(f(x^{(j+1)}) - f(\underline{x})\Big) \leq \Big(1 - 4\gamma_1(1-\gamma_2)\frac{m}{M}\Big)\Big(f(x^{(j)}) - f(\underline{x})\Big).$$

Démonstration. En utilisant la deuxième condition de Goldstein en $t = t_j$

$$f(x^{(j+1)}) = q_j(t_j) \leq q_j(0) + \gamma_1 t_j q_j'(0) \leq q_j(0) = f(x^{(j)})$$

pour tout j et alors $[x^{(j)}, x^{(j+1)}] \subset C(x^{(0)})$ par convexité d'un ensemble de niveau. Comme dans la preuve du 3.1.3, un développement de Taylor d'ordre 2 nous donne

un $\eta \in [0,1]$ avec $q_j(t) = q_j(0) + tq_j'(0) + q_j''(\eta t)t^2/2$, ce qui avec $(H3')$ donne le majorant

$$\forall t \in [0, t_j] : \quad q_j(t) \leq h(t) := q_j(0) + tq_j'(0) + \frac{t^2}{2} M |q_j'(0)|.$$

Par stricte convexité, l'équation $q_j(0) + t\gamma_2 q_j'(0) = q_j(t)$ admet les deux solutions $t = 0$ et $t = t' > 0$, avec $t' \leq t_j$ par la première condition de Goldstein. De même, l'équation $q_j(0) + t\gamma_2 q_j'(0) = h(t)$ admet les deux solutions $t = 0$ et $t = t'' > 0$, et $t'' \leq t'$ car h majore q_j. Un petit calcul montre que $t_j \geq t'' = 2(1-\gamma_2)/M$.
En réutilisant la deuxième condition de Goldstein en $t = t_j$,

$$f(x^{(j+1)}) = q_j(t_j) \leq q_j(0) + \gamma_1 t_j q_j'(0) = f(x^{(j)}) - \gamma_1 t_j |q_j'(0)| \leq f(x^{(j)}) - \frac{2\gamma_1(1-\gamma_2)}{M} |q_j'(0)|.$$

Pour conclure comme dans la preuve du 3.1.3, il reste à établir l'inégalité

$$\frac{|q_j'(0)|}{2m} \geq f(x^{(j)}) - f(\underline{x}).$$

ce qui se fait exactement de la même manière que dans la preuve du 3.1.3. □

On renvoie le lecteur sur [BV, §9] pour des illustrations numériques. Avec la recherche linéaire de Goldstein, on s'attend alors à un taux de convergence de $1 - 4\gamma_1(1-\gamma_2)\frac{m}{M}$ qui vaut $1 - 0.36\frac{m}{M}$ si $\gamma_1 = 0.3, \gamma_2 = 0.7$. Ceci est un taux légèrement plus faible que dans 3.1.3, mais l'avantage est que le pas t_j est plus facile à trouver.

La recherche d'un pas t_j approprié est encore plus simple dans la recherche linéaire d'Armijo, voir l'exercice suivant, où on peut s'imaginer une procédure de backtracking : en initialisant avec le pas de l'itération précédente, on doit ajuster le pas t_j en multipliant avec ou divisant par un paramètre $\rho > 1$.

3.2.2. Exercice : *algorithme de quasi-Newton avec back-tracking d'Armijo*
Sous les hypothèses $(H1), (H2)$ *et* $(H3')$, *soient* $\gamma \in]0,1[$ *et* $\rho > 1$. *On exclut le cas* $q_j'(0) = 0$ *pour tout* j. *Considérons la suite définie récursivement par*

$$x^{(j+1)} = x^{(j)} - t_j W_j \nabla f(x^{(j)})^T, =$$

avec $t = t_j > 0$ *vérifiant les conditions d'Armijo pour* $q_j(t) = f(x^{(j)} - tW_j \nabla f(x^{(j)})^T)$

$$q_j(t) < q_j(0) + \gamma t q_j'(0) \quad et \quad q_j(\rho t) \geq q_j(0) + \gamma \rho t q_j'(0). \tag{3.7}$$

Cherchons à montrer que, pour tout $j \geq 0$,

$$\Big(f(x^{(j+1)}) - f(\underline{x})\Big) \leq \Big(1 - \frac{4\gamma(1-\gamma)}{\rho}\frac{m}{M}\Big)\Big(f(x^{(j)}) - f(\underline{x})\Big).$$

(a) *Montrer qu'il existe* $0 \leq t_j'' < t_j'$ *de sorte que* $\{t \geq 0 : q_j(t) \leq q_j(0)\} = [0, t_j']$, *et*

$$q_j(t) - q_j(0) - \gamma t q_j'(0) \begin{cases} = 0 \ \ pour\ t = 0 \ et\ t = t_j'', \\ < 0 \ \ pour\ t \in]0, t_j''[, \\ > 0 \ \ pour\ t > t_j''. \end{cases}$$

(b) *En déduire que* t *vérifie* (3.7) *ssi* $t < t_j'' \leq \rho t$ *(ceci explique la mise à jour dans l'algorithme : tant que* $t_j \geq t_j''$ *on met à jour* $t_j \longleftarrow t_j/\rho$, *et tant que* $\rho t_j < t_j''$ *on met à jour* $t_j \longleftarrow \rho t_j$*).*
(c) *Montrer que, pour* $t \in [0, t_j']$,

$$q_j(t) \leq h_j(t) := q_j(0) + t q_j'(0) + \frac{M t^2}{2} |q_j'(0)|,$$

et que la seule solution > 0 de l'équation $h_j(t) = q_j(0) + \gamma t q_j'(0)$ est donnée par

$$t_j''' = \frac{2}{M}(1-\gamma) \leq t_j'' \leq \rho t_j.$$

Conclure comme dans la preuve du 3.2.1.

En entrée : $x^{(0)} \in \mathbb{R}^d$ *avec* $(H1)$, *tolérance* $\epsilon > 0$, *suite de matrices* W_j *sdp, paramètres* $\gamma \in (0,1), \rho > 1$.

Algo : *Pour* $j = 0, 1, ...$

poser $q_j(t) = f(x^{(j)} - tW_j \nabla f(x^{(j)})^T)$

calculer $q_j'(0) = -\nabla f(x^{(j)}) W_j \nabla f(x^{(j)})^T$

arrêt si $|q_j'(0)| < \epsilon$

initialiser $t_j > 0$ *(par exemple* $t_j = t_{j-1}$ *pour* $j > 0$*)*

ajuster t_j

tant que $q_j(t_j) \geq q_j(0) + \gamma t_j q_j'(0)$: $t_j \leftarrow t_j/\rho$

tant que $q_j(\rho t_j) < q_j(0) + \gamma \rho t_j q_j'(0)$: $t_j \leftarrow \rho t_j$

poser $x^{(j+1)} = x^{(j)} - t_j W_j \nabla f(x^{(j)})^T$

En sortie : $x^{(j)}$ *avec* $f(x^{(j)}) \leq \min_{x \in \mathbb{R}^d} f(x) + \frac{\epsilon}{2m}$.

3.2.3. Exercice : Règle de Wolfe et Powell

On considère l'algorithme de plus forte pente (steepest descent) pour le calcul du minimum d'une fonction f vérifiant

- *f est bornée inférieurement ;*
- *∇f satisfait une condition de Lipschitz :* $\forall x, y \quad \| \nabla f(x) - \nabla f(y) \| \leq K \| x - y \|$.

On remplace la minimisation unidimensionnelle par la recherche linéaire de Wolfe et Powell définie par la règle suivante : fixer les constantes $m_1 \in]0,1[$ *et* $m_2 \in]m_1,1[$ *et choisir* t *vérifiant*

$$\begin{cases} q_j(t) \leq q_j(0) + m_1 t q_j'(0) \\ q_j'(t) \geq m_2 q_j'(0) \end{cases}.$$

(a) Montrer que

$$f(x^{(k+1)}) \leq f(x^{(k)}) - m_1 \|x^{(k+1)} - x^{(k)}\| \|\nabla f(x^{(k)})\|.$$

(b) En déduire que la série $\sum_{k=0}^{\infty} \|x^{(k+1)} - x^{(k)}\| \, \|\nabla f(x^{(k)})\|$ *converge.*

(c) Montrer que

$$K\|x^{(k+1)} - x^{(k)}\| \geq (1 - m_2)\|\nabla f(x^{(k)})\|.$$

En déduire que $\lim_{k\to\infty} \|\nabla f(x^{(k)})\| = 0.$

Chapitre 4

Algorithmes d'optimisation avec contraintes affines d'égalité

Dans ce chapitre nous souhaitons résoudre numériquement le problème d'optimisation convexe sous contraintes affines d'égalité

$$(P): \quad \inf\{f(x) : x \in S\}, \quad S = \{x \in \mathbb{R}^d : Ax = a\},$$

avec f convexe et de classe $\mathcal{C}^2$, $a \in \mathbb{R}^m$, et $A \in \mathbb{R}^{m\times d}$, de rang m. D'après le corollaire 1.5.2(b), $\underline{x} \in S$ est une solution optimale de (P) ssi $\exists \mu \in \mathbb{R}^{1\times m}$ t.q. $\nabla f(\underline{x}) + \mu A = 0$. Ceci nous donne un système d'équations non linéaires à $m + d$ équations et inconnues.

Nous montrerons dans un premier temps que l'on peut transformer (P) en un problème d'optimisation convexe sas contraintes dans $\mathbb{R}^{d-m}$. Une deuxième approche

dans $\mathbb{R}^{m+d}$ permettra d'exploiter le caractère creux de la matrice A.

4.1 Contraintes affines d'égalité et élimination de variables

Notre approche par élimination de variables est basée sur le résultat suivant.

4.1.1. Lemme ;
Soient $S = \{x \in \mathbb{R}^d : Ax = a\}$, *et* $x^{(0)} \in S$. *Alors il existe une matrice* $C \in \mathbb{R}^{d\times(d-m)}$ *t.q.* $S = \{x^{(0)} + C\widetilde{x} : \widetilde{x} \in R^{d-m}\}$.

Démonstration. D'après l'exercice 1.2.7, il suffit [1] de construire une matrice $C \in \mathbb{R}^{d\times(d-m)}$ avec $Im(C) = Ker(A)$. Ceci peut se faire on calculant la décomposition QR de A^T. Comme le rang de A^T est donné par le nombre de colonnes, il existe une matrice $Q \in \mathbb{R}^{d\times d}$ orthogonale et $R \in \mathbb{R}^{m\times m}$ triangulaire supérieure et inversible de sorte que

$$A^T = Q \begin{bmatrix} R \\ 0 \end{bmatrix} = Q_1 R, \quad \text{avec } Q = [Q_1, Q_2], Q_1 \in \mathbb{R}^{d\times m}, Q_2 \in \mathbb{R}^{d\times(d-m)}.$$

Par conséquent, $Im(A^T) = Im(Q_1)$, et $Ker(A) = Im(A^T)^\perp = Im(Q_1)^\perp = Im(Q_2)$. Donc on peut prendre $C = Q_2$. □

On déduit du lemme 4.1.1 que la valeur optimale de (P) coïncide avec celle du problème

$$(\widetilde{P}) : \quad \inf\{\widetilde{f}(\widetilde{x}) : \widetilde{x} \in \mathbb{R}^{d-m}\}, \quad \widetilde{f}(\widetilde{x}) = f(x^{(0)} + C\widetilde{x}),$$

1. Notons qu'une telle matrice est unique à une multiplication à droite avec une matrice inversible.

et une solution optimale $\underline{\widetilde{x}}$ de $(\widetilde{P})$ induit une solution optimale $\underline{x} = x^{(0)} + C\underline{\widetilde{x}}$ de (P). Aussi, $\widetilde{f}$ est de classe $\mathcal{C}^2$, avec

$$\nabla\widetilde{f}(\widetilde{x}) = \nabla f(x^{(0)} + C\widetilde{x})C, \quad \nabla^2\widetilde{f}(\widetilde{x}) = C^T\nabla^2 f(x^{(0)} + C\widetilde{x})C,$$

appelés également gradient réduit et Hessien réduit.

Nous allons appliquer une méthode de quasi-Newton à $(\widetilde{P})$ avec $\widetilde{x}^{(0)} = 0$. Dans la table 4.1 nous rappelons en deuxième colonne la méthode de quasi-Newton, que nous avons exprimé en troisième colonne en termes du problème de départ. Par exemple, l'hypothèse $(H3)'$ pour la direction de Newton pour $(\widetilde{P})$ s'écrit aussi en termes du Hessien réduit, notée $(H3)''$. Les hypothèses de $(H1)$, $(H2)'$ et $(H3)'$ permettant d'assurer la convergence des valeurs $f(x^{(j)}) = \widetilde{f}(\widetilde{x}^{(j)})$ vers la valeur optimale de (P), à comparer avec les 3.1.3, 3.1.5 et 3.2.1. Dans le chapitre suivant, on étudiera de plus près ce cas particulier d'une direction de Newton.

4.2 La direction de Newton pour contraintes affines d'égalité

Dans ce chapitre on revient sur notre problème $(P) : \inf\{f(x) : x \in S\}, S = \{x \in \mathbb{R}^d : Ax = b\}$, un problème d'optimisation sous contraintes affines d'égalité. Dans des nombreuses applications, on se retrouve avec une matrice $A \in \mathbb{R}^{m\times d}$ qui est grande et creuse, avec parfois $m \ll d$. Ici on ne souhaite pas construire la matrice $C \in \mathbb{R}^{d\times(d-m)}$ du chapitre précédent qui généralement est pleine. Nous allons donc chercher à obtenir la direction $d^{(j)}$ de Newton pour (P) sans connaître C, c'est-à-dire, sans exploiter les formules du tableau 4.1 ni de calculer le Hessien

problème	$(\widetilde{P})$	(P)
variable	$\widetilde{x} \in \mathbb{R}^{d-m}$	$x = x^{(0)} + C\widetilde{x} \in \mathbb{R}^d$
deplacement	$\widetilde{x}^{(j+1)} = \widetilde{x}^{(j)} + t_j \widetilde{d}^{(j)}$	$x^{(j+1)} = x^{(j)} + t_j d^{(j)}, d^{(j)} = C\widetilde{d}^{(j)}$
fonction	$\widetilde{q}_j(t) = \widetilde{f}(\widetilde{x}^{(j)} + t\widetilde{d}^{(j)})$	$= q_j(t) = f(x^{(j)} + td^{(j)})$
quasi-Newton	$\widetilde{d}^{(j)} = -\widetilde{W}_j \nabla \widetilde{f}(\widetilde{x}^{(j)})^T$	$d^{(j)} = -W_j \nabla f(x^{(j)})^T, W_j = C\widetilde{W}_j C^T$
ensemble de niveau	$(H2)$ pour $(\widetilde{P})$	$(H2)' : K = \{x \in S : f(x) \leq f(x^{(0)})\}$ compact
ellipticité	$(H3)'$ pour $(\widetilde{P})$	$(H3)'' : \sup_{x,x' \in K} \sup_{v \in \mathbb{R}^{m-d}} \frac{v^T C^T \nabla^2 f(x) C v}{v^T C^T \nabla^2 f(x') C v} < \infty.$

TABLE 4.1 – La méthode de quasi-Newton appliquée au problème réduit $(\widetilde{P})$, exprimé en dernière colonne dans le plan $x = x^{(0)} + C\widetilde{x}$.

réduit $\widetilde{W}_j^{-1} = \nabla^2 \widetilde{f}(\widetilde{x}^{(j)})$. Notons néanmoins que si $x^{(j)} \in S$ alors $x^{(j+1)} \in S$ si et seulement si $d^{(j)} \in Ker(A)$.

4.2.1. Lemme : calcul de la direction de Newton
Avec les notations précédentes, la direction de Newton $d^{(j)}$ à l'itération j est l'unique solution du problème d'optimisation

$$(P_j) : \min\{f(x^{(j)}) + \nabla f(x^{(j)})z + \frac{1}{2} z^T \nabla^2 f(x^{(j)}) z : z \in \mathbb{R}^d, Az = 0\}, \qquad (4.1)$$

et peut être obtenue en résolvant le système

$$\begin{bmatrix} \nabla^2 f(x^{(j)}) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} d^{(j)} \\ u \end{bmatrix} = \begin{bmatrix} -\nabla f(x^{(j)})^T \\ 0 \end{bmatrix}. \qquad (4.2)$$

En particulier, $x^{(j)}$ est solution optimale pour (P) si et seulement si $d^{(j)} = 0$, et dans le cas contraire $q_j'(0) < 0$.

Démonstration. Avec la matrice C du chapitre précédent

$$Ker(A) = \{x \in \mathbb{R}^d : Ax = 0\} = \{C\widetilde{x} : \widetilde{x} \in \mathbb{R}^{d-m}\}$$

et insérant cette substitution $x = C\widetilde{x}$ dans (4.1), nous obtenons le problème équivalent

$$\min\{\widetilde{f}(\widetilde{x}^{(j)}) + \nabla\widetilde{f}(\widetilde{x}^{(j)})\widetilde{z} + \frac{1}{2}\widetilde{z}^T\nabla^2\widetilde{f}(\widetilde{x}^{(j)})\widetilde{z} : \widetilde{z} \in \mathbb{R}^{d-m}\}.$$

Avec $\nabla^2 f(x^{(j)})$, aussi le Hessien réduit $\nabla\widetilde{f}(\widetilde{x}^{(j)})$ est s.d.p.. Une unique solution optimale de ce dernier problème existe, et est donné par la direction $\widetilde{d}^{(j)}$ de quasi-Newton pour $\widetilde{W}_j = \nabla^2\widetilde{f}(\widetilde{x}^{(j)})^{-1}$. Par conséquent, l'unique solution optimale de (4.1) est donnée par $d^{(j)} = C\widetilde{d}^{(j)}$.
D'après le corollaire 1.5.2(b), $d^{(j)}$ est solution optimale de (4.1) si et seulement si il existe un $\mu \in \mathbb{R}^{1\times m}$ avec $(d^{(j)})^T\nabla^2 f(x^{(j)}) + \nabla f(x^{(j)}) = \mu A$, et $Ad^{(j)} = 0$. En passant à la transposée de la première équation, nous concluons que $d^{(j)}$ et $u = -\mu^T$ donnent bien une solution du système (4.2). Pour montrer l'unicité, soit $d \in \mathbb{R}^d$ et $u \in \mathbb{R}^m$ tels que

$$\begin{bmatrix} \nabla^2 f(x^{(j)}) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} d \\ u \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Alors $Ad = 0$, impliquant que $\exists y$ t.q. $d = Cy$. En multipliant la première équation par C^T, nous concluons que $C^T\nabla^2 f(x^{(j)})Cy = -C^TA^Tu = 0$ et alors $y = 0$ car le Hessien réduit est s.d.p. et donc inversible. Par conséquent, $A^Tu = 0$, mais comme

les colonnes de A^T sont libres par hypothèse sur le rang de A, il en suit que $d = 0$ et $u = 0$, d'où l'unicité d'une solution de (4.2).
Finalement, d'après le corollaire 1.5.2(b), $x^{(j)} \in S$ est solution optimale de (P) ssi $\exists u$ t.q. $A^T u = -\nabla f(x^{(j)})^T$ ssi la quantité $d^{(j)}$ dans (4.2) vaut 0. Si on exclut ce cas, alors par (4.2)

$$q_j'(0) = \nabla f(x^{(j)})d^{(j)} = -(d^{(j)})^T \nabla^2 f(x^{(j)})d^{(j)} = -(\widetilde{d}^{(j)})^T \nabla^2 \widetilde{f}(\widetilde{x}^{(j)})\widetilde{d}^{(j)} < 0$$

car le Hessien réduit est s.d.p.. □

Notons que résoudre efficacement un système (dit KKT) avec matrice de coefficients

$$\begin{bmatrix} H & A^T \\ A & 0 \end{bmatrix}, \quad H \in \mathbb{R}^{d\times d} \text{ s.d.p., et } A \in \mathbb{R}^{m\times d} \text{ de rang } m$$

par des méthodes directes ou itératives en creux est un sujet de recherche de grande actualité, en particulier dans le cas où H est de plus une matrice diagonale (ce qui dans notre contexte correspond à un objectif à d variables séparées). On rencontre ces systèmes aussi dans la discrétisation d'une EDP de Stokes.

Nous résumons ces considérations dans l'algorithme suivant.

4.2.2. Algorithme de Newton pour contraintes affines d'égalité

En entrée : $x^{(0)} \in \mathbb{R}^d$ *avec* $Ax^{(0)} = b$, *tolérance* $\epsilon > 0$
Algo : *Pour* $j = 0, 1, ...$
calculer $d^{(j)}$ *par le 4.2.1*
poser $q_j(t) = f(x^{(j)} + td^{(j)})$
calculer $q_j'(0) = \nabla f(x^{(j)})d^{(j)}$

arrêt si $|q_j'(0)| < \epsilon$
minimisation : trouver $t_j = \arg\min_{t\geq 0} q_j(t)$ *(ou recherche linéaire)*
poser $x^{(j+1)} = x^{(j)} + t_j d^{(j)}$

En sortie : $x^{(j)}$ *avec* $f(x^{(j)})$ *proche de* $\min_{x\in\mathbb{R}^d} f(x)$.

Dans la littérature on trouve aussi un Lagrangien modifié $\widetilde{L}$ associé à notre problème (P) avec contraintes d'égalités, défini par

$$\mathbb{R}^d \times \mathbb{R}^m \ni (x,u) \mapsto \widetilde{L}(x,p) = f(x) + (Ax-b)^T u.$$

On vérifie aisément que $\underline{x} \in R^d$ est solution optimale de (P) ssi $\exists u$ t.q. $\nabla\widetilde{L}(\underline{x},u) = 0$ (ici on prend le gradient par rapport aux $m+d$ variables données par les composantes de x et u). Il existe également des variantes de Newton nommées Newton primal-dual. Ici on utilise aussi la direction de Newton pour résoudre le système non-linéaire $\nabla\widetilde{L}(\underline{x},c) = 0$, mais les itérées $x^{(j)}$ ne sont plus forcement réalisable pour (P), voir [BV] et l'exercice 4.2.4.

4.2.3. Exercice :
Avec $W_j \in \mathbb{R}^{d\times d}$ *s.d.p., la direction de quasi-Newton* $d^{(j)}$ *à l'itération* j *est une solution optimale du problème d'optimisation*

$$(P_j) : \min\{f(x^{(j)}) + \nabla f(x^{(j)})z + \frac{1}{2}z^T W_j^{-1} f(x^{(j)})z : z \in \mathbb{R}^d, Az = 0\}.$$

(a) Montrer que $d^{(j)}$ *est unique, et peut être obtenu en résolvant le système*

$$\begin{bmatrix} W_j^{-1} & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} d^{(j)} \\ u \end{bmatrix} = \begin{bmatrix} -\nabla f(x^{(j)})^T \\ 0 \end{bmatrix}. \qquad (4.3)$$

(b) *Montrer que* AW_jA^T *est inversible. Posons* $\Pi_j = I - A^T(AW_jA^T)^{-1}AW_j$. *En multipliant le premier bloc d'équations par* AW_j, *conclure que* $u = (AW_jA^T)^{-1}AW_j(-\nabla f(x^{(j)})^T)$, *et alors*

$$d^{(j)} = W_j\Pi_j(-\nabla f(x^{(j)})^T) = \Pi_j^T W_j(-\nabla f(x^{(j)})^T).$$

En déduire que $d^{(j)}$ *peut être obtenu par des produits matrices-vecteurs et la résolution d'un système d'ordre* m.

(c) *Considérons finalement le cas particulier* $W_j = I$ *de la méthode de la plus forte descente, et la matrice* C *du lemme4.1.1. Vérifier que* $(A^T, C)^T(A^T, C)$ *est inversible et diagonal par blocs. En déduire que* $I - A^T(AA^T)^{-1}A = C(C^TC)^{-1}C^T$ *et alors*

$$d^{(j)} = C(C^TC)^{-1}C^T(-\nabla f(x^{(j)})).$$

Comparer avec la direction steepest descent obtenue dans la table4.1.

4.2.4. Exercice : Newton primal-dual pour contraintes affines d'égalités

On considère le Lagrangien modifié $\widetilde{L}(\begin{bmatrix} x \\ u \end{bmatrix}) = f(x) + (Ax - b)^T u$ *pour* $x \in \mathbb{R}^d$ *(la variable primale) et* $u \in \mathbb{R}^m$ *(la variable duale, ici un vecteur colonne). On considère l'algorithme*

Pour $x^{(0)} \in \mathbb{R}^d, u^{(0)} \in \mathbb{R}^m$, *pas forcément* $x^{(0)} \in S$
Pour $j = 0, 1, ...$
calculer direction de Newton $\nabla^2\widetilde{L}(\begin{bmatrix} x^{(j)} \\ u^{(j)} \end{bmatrix})\begin{bmatrix} \Delta x^{(j)} \\ \Delta u^{(j)} \end{bmatrix} = -\nabla\widetilde{L}(\begin{bmatrix} x^{(j)} \\ u^{(j)} \end{bmatrix})^T$
avec $t_j > 0$, *mettre à jour* $\begin{bmatrix} x^{(j+1)} \\ u^{(j+1)} \end{bmatrix} = \begin{bmatrix} x^{(j)} \\ u^{(j)} \end{bmatrix} + t_j\begin{bmatrix} \Delta x^{(j)} \\ \Delta u^{(j)} \end{bmatrix}$

(a) *Calculer* $\nabla \widetilde{L}$ *et* $\nabla^2 \widetilde{L}$, *et montrer que pour tout* $v, v' \in \mathbb{R}^d$

$$\nabla^2 \widetilde{L}(\begin{bmatrix} x^{(j)} \\ v' \end{bmatrix}) \begin{bmatrix} \Delta x^{(j)} \\ \Delta u^{(j)} + u^{(j)} - v \end{bmatrix} = -\nabla \widetilde{L}(\begin{bmatrix} x^{(j)} \\ v \end{bmatrix})^T.$$

En déduire l'équivalence que $x^{(j)}$ *est solution optimale de* (P) *ssi* $\exists v \in \mathbb{R}^p$ *avec* $\nabla \widetilde{L}(\begin{bmatrix} x^{(j)} \\ v \end{bmatrix}) = 0$ *ssi* $\Delta x^{(j)} = 0$.

(b) *Montrer que* $Ax^{(j+1)} - b = (1 - t_j)(Ax^{(j)} - b)$. *En déduire que si* $t_j = 1$ *alors* $x^{(\ell)}$ *est réalisable pour* (P) *pour tout* $\ell > j$.

(c) *Soit* $x^{(j)}$ *réalisable pour* (P). *Montrer que que l'on obtient les mêmes vecteurs* $d^{(j)} = \Delta x^{(j)}$ *et* $u = u^{(j)} + \Delta u^{(j)}$ *que dans l'algorithme 4.2.2.*

(d) *Pour* $t \geq 0$, *posons* $q_j(t) = f(x^{(j)} + t\Delta x^{(j)})$. *Montrer que* $q_j'(0) > 0$ *est possible (et alors le choix de* t *ne se fait pas à l'aide de cette fonction univariée).*

(e) *Pour* $t > 0$, *posons*

$$\widetilde{q}_j(t) = \|\nabla \widetilde{L}(\begin{bmatrix} x^{(j)} \\ u^{(j)} \end{bmatrix} + t \begin{bmatrix} \Delta x^{(j)} \\ \Delta u^{(j)} \end{bmatrix})\|^2.$$

Vérifier que $\widetilde{q}_j(t) \geq 0$, *avec* $q_j(0) = 0$ *impliquant que* $x^{(j)}$ *est solution optimale de* (P). *Aussi, vérifier que* $\widetilde{q}_j'(0) = -2q_j(0)$.

On déduit de la dernière partie que t_j *peut être obtenu par minimisation de* $\widetilde{q}_j$ *ou alors par recherche linéaire avec* $\widetilde{q}_j$. *Ici on s'arrête si* $|\widetilde{q}_j(0)|$ *est assez petit. En imposant par exemple que* $\nabla^2 \widetilde{L}(\begin{bmatrix} x \\ u \end{bmatrix})^{-1}$ *borné uniformément, on montre convergence comme dans le chapitre 3, voir par exemple [BV, Chapitre 10.3.3].*

Chapitre 5

Divers algorithmes d'optimisation sous contraintes

Le but de ce chapitre est d'étudier divers algorithmes pour la résolution numérique du problème général d'optimisation convexe

$$\inf\{f(x) : x \in C\}, \quad C = \{x \in S : g(x) \leq 0\}, \quad S = \{x \in \mathbb{R}^d : Ax = b\}, \quad (5.1)$$

où comme avant $g = (g_1, ..., g_p)^T$, f, g_k sont des fonctions convexes de classe $\mathcal{C}^1$, $A \in \mathbb{R}^{m \times d}$ de rang m, $b \in \mathbb{R}^d$. D'autres hypothèses seront précisées ultérieurement.

Nous commençerons à étudier deux méthodes où on se ramène à une suite de problèmes paramétrés par un scalaire, en rapportant les contraintes "compliquées" $g(x) \leq 0$ dans l'objectif, grâce à deux techniques différentes. Dans le chapitre 5.1, on pénalisera l'utilisation des itérés dans $S \setminus C$, ce qui nous donnera une suite d'itérés

qui ne sont pas réalisables pour (P), mais qui convergent vers une solution optimale de (P).

5.1 Méthode de pénalités extérieures

Pour résoudre le problème (P) énoncé au (5.1), l'idée de la méthode de pénalités extérieures est de construire et résoudre une suite de problèmes paramétrés d'optimisation sous contraintes affines d'égalité (ou carrément sans contraintes dans le cas $m = 0$) de sorte que l'utilisation des points dans $S \setminus C$ soit de plus en plus pénalisée. Considérons la fonction

$$H(x) = \sum_{k=1}^{p} \max\{0, g_k(x)\}^2.$$

alors H est une fonction [1] convexe de classe $\mathcal{C}^1$ et $H(x) \geq 0$ pour tout $x \in S$, avec $H(x) = 0$ ssi $x \in C$. Pour $\beta \in]0, +\infty[$, on considère le problème paramétré

$$(P)_\beta : \min\{f_\beta(x) : x \in S\}, \qquad f_\beta(x) = f(x) + \beta H(x),$$

et on note par $x(\beta)$ une solution optimale de $(P)_\beta$. L'idée de base est que si $\beta \to +\infty$ alors le terme pénalisant $\beta H(x)$ devient de plus en plus important, et alors $x(\beta)$ devrait s'approcher de C (là où $H(x) = 0$) et éventuellement converger vers une solution optimale de (P). Les deux résultats théoriques suivants confirment cette idée.

1. Autres fonctions ayant ces mêmes propriétés sont imaginables, notamment parfois 2 est remplacé par un exposant > 2, voir l'exercice 5.1.4.

5.1.1. Théorème :
Supposons que C est non vide, et que $f(x) \to \infty$ pour $||x|| \to \infty$. Alors, pour toute suite $(\beta_j)_j$ strictement croissante avec limite $+\infty$,

(a) la suite $(x(\beta_j))_j$ reste bornée ;

(b) tout point d'accumulation $\widetilde{x}$ de $(x(\beta_j))_j$ est solution optimale de (P) (notamment $\widetilde{x} \in C$) ;

(c) $(f_{\beta_j}(x(\beta_j)))_j$ est croissante, et converge vers la valeur optimale de (P) ;

(d) $\beta_j H(x(\beta_j)) \to 0$ pour $j \to \infty$;

(e) si de plus f est strictement convexe alors $(x(\beta_j))_j$ converge vers l'unique solution optimale $\underline{x}$ de (P).

Démonstration. Soit $x^{(0)} \in C$, et $\beta > 0$. Par hypothèse, $f(z) \to \infty$ et alors $f_\beta(x) \to \infty$ pour $\|x\| \to \infty$. Alors d'après le théorème 1.5.4(b), (c) et sa preuve, les ensembles de niveau $\{x \in S : f(x) \leq f(x^{(0)})\}$ et $\{x \in S : f_\beta(x) \leq f_\beta(x^{(0)})\}$ sont compacts, impliquant que (P) admet une solution optimale $\underline{x}$, et (P_β) une solution optimale $x(\beta)$, avec

$$x(\beta) \in \{x \in S : f_\beta(x) \leq f_\beta(x^{(0)}) = f(x^{(0)})\} \subset \{x \in S : f(x) \leq f(x^{(0)})\},$$

montrant la propriété (a).
L'inégalité $f_{\beta_j}(x) \leq f_{\beta_{j+1}}(x)$ pour tout $x \in S$ implique que

$$f_{\beta_j}(x(\beta_j)) = \min_{x \in S} f_{\beta_j}(x) \leq \min_{x \in S} f_{\beta_{j+1}}(x) = f_{\beta_{j+1}}(x(\beta_{j+1})) \leq f_{\beta_{j+1}}(\underline{x}) = f(\underline{x}),$$

et donc la suite des valeurs $f_{\beta_j}(x(\beta_j))_j$ est croissante et bornée supérieurement par $f(\underline{x})$, donc convergeante avec limite $L \leq f(\underline{x})$. Montrons par absurde que $L = f(\underline{x})$,

c'est-à-dire, la propriété (c). Sinon, par (a), il existe une sous-suite de paramètres aussi nommée $(\beta_j)_j$ de sorte que $x(\beta_j) \to \widetilde{x}$, et donc $f(x(\beta_j)) \to f(\widetilde{x})$. Il en suit que $\beta_j H(x(\beta_j))$ admet une limite finie $L - f(\widetilde{x})$, et donc forcément $H(x(\beta_j)) \to 0$. Ceci implique par continuité que $H(\widetilde{x}) = 0$, ou alors $\widetilde{x} \in C$ par construction de H. En rappelant que $\underline{x}$ est une solution optimale de (P), nous concluons que $f(\underline{x}) > L = f(\widetilde{x}) \geq f(\underline{x})$, une contradiction. Donc la propriété (c) est valable, et par un argument similaire on démontre (b) et (d). Finalement, (e) découle de (b) et de l'unicité d'une solution optimale de (P).[2] □

En pratique on ne résout pas le problème paramétré $(P)_{\beta_j}$ pour un nombre infini de paramètres, mais seulement jusqu'au rang où $\beta_j H(x(\beta_j)) \leq \epsilon$ et ϵ/β_j "petit" pour une tolérance $\epsilon > 0$ donnée. Nous affirmons que dans ce cas l'argument $x(\beta_j)$ permet d'atteindre la valeur minimale de f à $p\epsilon$ près dans un convexe fermé

$$C' := \{x \in S : \forall k, g_k(x) \leq \sqrt{\epsilon/\beta_j}\},$$

"légèrement" plus grand que C. Pour le voir, notons d'abord que par construction de C' nous avons $0 \leq \beta_j H(x) \leq p\epsilon$ pour tout $x \in C'$. Donc $f(x) \leq f_{\beta_j}(x) \leq f(x) + p\epsilon$ pour tout $x \in C'$. Comme $x(\beta_j) \in C'$ par hypothèse sur β_j, nous déduisons que

$$\min_{x \in C'} f(x) \leq f(x(\beta_j)) \leq f_{\beta_j}(x(\beta_j)) = \min_{x \in S} f_{\beta_j}(x) \leq \min_{x \in C'} f_{\beta_j}(x) \leq \min_{x \in C'} f(x) + p\epsilon,$$

ce qu'il fallait démontrer. D'où la question du taux de convergence de la suite $(\beta_j H(x(\beta_j))$, ici sous des hypothèses plus fortes.

---

2. f strictement convexe implique aussi que f_β est strictement convexe et donc $x(\beta)$ est unique.

5.1.2. Corollaire :
Sous les hypothèses du 5.1.1 et 5.1.1(e), supposons de plus que les gradients des contraintes actives en $\overline{x}$ et les lignes de A sont linéairement indépendants. Alors $\beta_j H(x(\beta_j)) = \mathcal{O}(1/\beta_j)_{j\to\infty}$.

Démonstration. Par KKT appliqué à $(P)_{\beta_j}$, il existe $\lambda_j \in \mathbb{R}^{1\times p}$ et $\mu_j \in \mathbb{R}^{1\times m}$ de sorte que

$$\nabla f(x(\beta_j)) + \lambda_j \nabla g(x(\beta_j)) = \mu_j A, \quad \lambda_{j,k} = 2\beta_j \max\{0, g_k(x(\beta_j))\}. \tag{5.2}$$

Comme $\|\lambda_j\|^2 = 4\beta_j^2 H(x(\beta_j))$, il suffit de montrer par absurde que la suite $(\lambda_j)_j$ est bornée. Supposons le contraire, alors $\|(\lambda_j, \mu_j)\| \to \infty$ pour $j \to \infty$. Posons

$$(\widetilde{\lambda}_j, \widetilde{\mu}_j) := \frac{\lambda_j, \mu_j}{\|(\lambda_j, \mu_j)\|},$$

des vecteurs dans la sphère d'unité du $\mathbb{R}^{p+m}$, un compact. En passant à des sous-suites, on peut alors supposer que

$$(\widetilde{\lambda}_j, \widetilde{\mu}_j) \to (\underline{\lambda}, \underline{\mu}) \neq 0$$

pour $j \to \infty$. Notons par $I = \{k : g_k(\underline{x}) = 0\}$ l'ensemble des indices des contraintes actives en $\underline{x}$. D'après le 5.1.1(e), $g_k(x(\beta_j)) \to g_k(\underline{x})$ pour $j \to \infty$. Comme la limite est < 0 pour $k \notin I$, nous concluons que $\lambda_{j,k} = 0$ et donc $\widetilde{\lambda}_{j,k} = 0$ pour tout $k \notin I$ et j assez grand, et alors $\underline{\lambda}_k = 0$ pour $k \notin I$. En divisant (5.2) par $\|(\lambda_j, \mu_j)\|$, nous obtenons pour $j \to \infty$ la limite

$$0 = \lim_{j\to\infty} \Big(\frac{\nabla f(x(\beta_j))}{\|(\lambda_j, \mu_j)\|} + \sum_{k\in I} \widetilde{\lambda}_{j,k} \nabla g_k(x(\beta_j)) - \widetilde{\mu}_j A \Big) = 0 + \sum_{k\in I} \underline{\lambda}_k \nabla g_k(\underline{x}) - \underline{\mu} A,$$

une combinaison linéaire non triviale entre les lignes de A et les gradients en $\underline{x}$ des contraintes actives en $\underline{x}$, une contradiction. [3] □

En pratique on ne cherche pas forcément la solution optimale $x(\beta_j)$ du problème paramétré $(P)_{\beta_j}$ mais plutôt une approximation $x^{(j)}$ de $x(\beta_j)$. On initialise en se donnant $\beta_1 > 0$ et un $x^{(0)} \in S$. A l'étape j, $x^{(j)}$ est obtenu en faisant quelques n_j itérations d'une des méthodes vue au §3 (pour $m = 0$) ou §4 appliquées à $(P)_{\beta_j}$ partant de $x^{(j-1)}$, et ensuite on choisit $\beta_{j+1} > \beta_j$ (pas trop large car $(P)_{\beta_{j+1}}$ devient "mal conditionné"). Ici la convergence est plus délicate (en fonction du choix de $\beta_{j+1} - \beta_j$), on renvoie le lecteur intéressé sur la littérature spécialisée.

5.1.3. Exercice :
Pour $(P) : \inf\{x^2 : x \in \mathbb{R}, 1 - x \leq 0\}$, *vérifier que le problème auxiliaire* $(P)_\beta$ *admet une et une seule solution* $x(\beta) = \frac{\beta}{\beta+1}$. *Vérifier les propriétés énoncées dans le théorème 5.1.1, et montrer que le taux du corollaire 5.1.2 est atteint.*

En partant de $x^{(0)} = 0$, *soit* $x^{(j)}$ *obtenu de* $x^{(j-1)}$ *par une itération de steepest descent appliqué a* $(P)_{\beta_j}$. *Vérifier que* $x^{(j)} = x(\beta_j)$.

5.1.4. Exercice :
Soit $\ell > 2$. *Vérifier que* $H(x) = \sum_k \max\{g_k(x), 0\}^\ell$ *est une autre fonction pénalité, de classe* $\mathcal{C}^2$. *Peut-on adapter les résultats de ce chapitre ?*

3. En fait, notre hypothèse implique unicité des vecteurs λ et μ dans KKT appliqué à (P), et donc $\lambda_j \to \lambda$, en passant à la limite dans (5.2).

5.2 Méthodes de points intérieurs

Le désavantage de la méthode du §5.1 est que tous les iterés sont non réalisables (sauf peut être le dernier). Néanmoins, on obtient des bornes inférieures pour la valeur optimale de (P). On expose maintenant une méthode alternative qui pénalise des contraintes actives, et ainsi construit une suite d'itérés réalisables pour (P) sans aucune contrainte active. Pour simplifier, on suppose que $g_1, ..., g_p \in \mathcal{C}^2$, que l'ensemble C dans (5.1) est compact et vérifie la condition de Slater, et que l'objectif f et strictement convexe de sorte que (P) admet une solution optimale unique $\underline{x}$. Néanmoins, ces conditions peuvent être relâchées.

Considérons une barrière logarithmique définie sur un ensemble $C' \subset C$ comme suit

$$B(x) = -\sum_{k=1}^{p} \log(-g_k(x)), \quad C' = \{x \in S : g_k(x) < 0 \text{ pour tout } k = 1, ..., p\},$$

ainsi que pour $\alpha \in]0, \infty[$ le problème paramétré

$$(P)_\alpha : \inf\{F_\alpha(x) : x \in C'\}, \quad F_\alpha(x) = f(x) + \alpha B(x).$$

Dans le résultat technique suivant on montre que $(P)_\alpha$ est un problème d'optimisation convexe sur un ensemble C' n'étant pas forcément fermé, et que pourtant on peut assurer l'existence et unicité d'une solution optimale $x(\alpha) \in C'$.

5.2.1. Lemme :
Pour tout $\alpha \in]0, \infty[$, *notre problème* $(P)_\alpha$ *est un problème d'optimisation*

convexe, et admet une solution optimale unique $x(\alpha) \in C'$, caractérisée par l'existence d'un $\mu \in \mathbb{R}^{1\times m}$ tel que

$$\nabla f(x(\alpha)) + \lambda(\alpha)\nabla g(x(\alpha)) = \mu A, \quad \lambda(\alpha) = -\Big(\frac{\alpha}{g_1(x(\alpha))}, ..., \frac{\alpha}{g_p(x(\alpha))}\Big) \geq 0.$$

Démonstration. On laissera le soin au lecteur de vérifier que C' est convexe. Pour montrer que $x \mapsto B(x)$ est convexe sur C', on note que

$$\nabla B(x) = \sum_{k=1}^{p}\Big(\frac{\nabla g_k(x)}{(-g_k(x))}\Big), \quad \nabla^2 B(x) = \sum_{k=1}^{p}\Big(\frac{\nabla^2 g_k(x)}{(-g_k(x))} + \frac{\nabla g_k(x)^T \nabla g_k(x)}{g_k(x)^2}\Big).$$

Donc le Hessien de B est ssdp, et alors $x \mapsto B(x)$ est convexe sur C', et $x \mapsto F_\alpha(x)$ est strictement convexe sur C'. Soit $\widetilde{x}$ comme dans la condition de Slater, alors $\widetilde{x} \in C'$, et on obtient l'existence d'une solution optimale de $(P)_\alpha$ en montrant que l'ensemble de niveau

$$C'' := \{x \in C' : F_\alpha(x) \leq F_\alpha(\widetilde{x})\}$$

est compact. Comme $C'' \subset C$ et C est compact, il reste à montrer que C'' est fermé. Soient alors $x^{(j)} \in C''$ avec $x^{(j)} \to x$ pour $j \to \infty$. Comme F_α est continue sur C', il suffit de montrer que $x \in C'$. La fonction g_k étant continue sur le compact C, il existe un γ de sorte que $g_k(x^{(j)}) \geq \gamma$ pour tout $j > 0$ et tout $k \in \{1, ..., p\}$, et alors

$$-\alpha \log(-g_k(x^{(j)})) = F_\alpha(x^{(j)}) - f(x^{(j)}) + \alpha \sum_{\ell \neq k} \log(-g_\ell(x^{(j)})) \leq F_\alpha(\widetilde{x}) - f(\underline{x}) + \alpha p \log(-\gamma).$$

Donc $g_k(x) = \lim_{j\to\infty} g_k(x^{(j)}) < 0$ et $x \in C'$, permettant d'affirmer qu'une solution optimale $x(\alpha)$ de $(P)_\alpha$ existe. L'unicité d'une solution optimale découle de la

convexité stricte de F_α.
Finalement, en ajoutant des contraintes vérifiés en $x(\alpha)$ au problème auxilliaire $(P)_\alpha$, nous obtenons la même solution optimale. En conséquence, $x(\alpha)$ est l'unique solution optimale du nouveau problème

$$\min\{F_\alpha(x) : x \in S, \forall k = 1, ..., p : g_k(x) \leq g_k(x(\alpha))/2\},$$

un problème qui vérifie la condition de Slater pour $x = x(\alpha)$, et aucune contrainte d'inégalité est active en $x = x(\alpha)$. Donc d'après le théorème KKT, $x(\alpha)$ est uniquement caractérisé par l'existence d'un $\mu \in \mathbb{R}^{1\times m}$ de sorte que $\nabla F_\alpha(x(\alpha)) = \mu A$. En observant que $\nabla F_\alpha(x(\alpha)) = \nabla f(x(\alpha)) + \alpha \nabla B(x(\alpha)) = \nabla f(x(\alpha)) + \lambda(\alpha)\nabla g(x(\alpha))$, notre lemme en découle. □

Dans la littérature, l'ensemble $\{(\alpha, x(\alpha)) : \alpha \in]0, +\infty[\}$ est appelé le "central path". D'après le lemme 5.2.1, $x = x(\alpha)$ et $\lambda = \lambda(\alpha)$ vérifient le système

$$x \in S, \quad \nabla f + \lambda \nabla g(x) = \mu A, \quad \forall k = 1, ..., p : g_k(x) \leq 0, \quad \lambda_k g_k(x) = -\alpha$$

pour un $\mu \in \mathbb{R}^{1\times m}$, c'est-à-dire, un système KKT avec une relation de complémentarité modifiée. Formellement on aurait envie de faire tendre $\alpha \to 0+$, bien que l'on sache pas encore si $x(\alpha)$ et $\lambda(\alpha)$ admettent une limite pour $\alpha \to 0+$. Pour un lien plus précis entre $(P)_\alpha$ et (P), on fera appel aux propriétés du Lagrangien $L(x, \lambda) = f(x) + \lambda g(x)$.

5.2.2. Théorème des points intérieurs :
Nous avons pour tout $\alpha \in]0, +\infty[$ *que*

$$f(x(\alpha)) - p\alpha \leq f(\underline{x}) \leq f(x(\alpha)),$$

autrement dit, $x(\alpha)$ minimise f sur C à $p\alpha$ près. En particulier, $f(x(\alpha)) \to f(\underline{x})$ et $x(\alpha) \to \underline{x}$ pour $\alpha \to 0$.

Démonstration. La caractérisation donnée dans le lemme 5.2.1 nous donne un $\mu \in \mathbb{R}^{1\times m}$ de sorte que

$$\nabla F_\alpha(x(\alpha)) = \nabla_x L(x(\alpha), \lambda(\alpha)) = \mu A.$$

Par la théorie du Lagrangien, nous concluons que

$$w(\lambda(\alpha)) = L(x(\alpha), \lambda(\alpha)) = \min_{x\in S} L(x, \lambda(\alpha)) \leq f(\underline{x}).$$

Comme $L(x(\alpha), \lambda(\alpha)) = f(x(\alpha)) + \lambda(\alpha)g(x(\alpha)) = f(x(\alpha)) - p\epsilon$, nous concluons que $f(x(\alpha)) - p\epsilon \leq f(\underline{x})$, l'inégalité $f(\underline{x}) \leq f(x(\alpha))$ vient de la définition de $\underline{x}$ et du fait que $x(\alpha) \in C$. La relation $f(x(\alpha)) \to f(\underline{x})$ pour $\alpha \to 0$ en découle directement.
Pour en déduire la relation $x(\alpha) \to \underline{x}$ pour $\alpha \to 0$, on raisonne par absurde. Supposons alors qu'il existe $\alpha_j \to 0$ avec $x(\alpha_j) \not\to \underline{x}$ pour $j \to \infty$. Par compacité de C, on peut extraire une sous-suite convergeante $x(\alpha_j) \to x \neq \underline{x}$, mais $f(x(\alpha_j)) \to f(\underline{x})$ par la première partie et donc $f(\underline{x}) = f(x)$ par continuité de f, ce qui est en contradiction avec l'unicité d'une solution optimale de (P). □

En posant $F_\alpha(x) = +\infty$ pour $x \in \mathbb{R}^d \setminus C'$, nous obtenons une fonction $F_\alpha : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ continue, et notre problème auxiliaire $(P)_\alpha$ peut être reformulé comme

$$\min\{F_\alpha(x) : x \in S, \quad F_\alpha(x) \leq F_\alpha(\widetilde{x})\},$$

avec $\widetilde{x} \in C'$ quelconque. Autrement dit, tant que l'on reste dans l'ensemble de niveau de paramètre $F_\alpha(\widetilde{x})$, on doit résoudre un problème auxiliaire sans contraintes (si $m = 0$) ou avec contraintes affines d'égalité, par une des méthodes vues dans §3 (pour $m = 0$) ou §4.

Comme dans le chapitre 5.1, en pratique on ne cherche pas forcément la solution optimale $x(\alpha_j)$ du problème paramétré $(P)_{\alpha_j}$ mais plutôt une approximation $x^{(j)}$ de $x(\alpha_j)$. On initialise en se donnant $\alpha_1 > 0$ et un $x^{(0)} \in S$ vérifiant la condition de Slater, c'est-à-dire, $x^{(0)} \in C' \cap S$. A l'étape j, $x^{(j)}$ est obtenu en faisant quelques n_j itérations d'une des méthodes vue au §3 (pour $m = 0$) ou §4 appliquées à $(P)_{\alpha_j}$ partant de $x^{(j-1)}$. Ensuite on choisit le nouveau paramètre $\alpha_{j+1} < \alpha_j$.

Il existe des études de complexité pour des sous-classes de problèmes (comme par exemple la classe des programmes linéaires de la forme $\inf\{hx : x \geq 0, Ax = b\}$) qui montrent que, même pour le choix $n_j = 1$ on obtient pour $j = J$ une erreur $\|x^{(J)} - \underline{x}\|$ inférieure à la précision machine pour un J qui ne dépend pas des données. Par exemple, un choix approprié de α_{j+1} en fonction de $x^{(j)}$ nous amène à la fameuse méthode de Karmarkar, une des premières méthodes de complexité polynômiale pour résoudre des programmes linéaires. Dans le cas général, prendre une suite géométrique $\alpha_j = \kappa^j \alpha_0$ pour un paramètre $\kappa \in]0, 1[$ et $n_j = 5$ est suggéré dans la littérature.

5.2.3. Exercice :
Considérons le programme linéaire

$$(P) : \min\{hx : Cx \leq c\}, \quad h \in \mathbb{R}^{1\times d}, \quad C \in \mathbb{R}^{p\times d}, \quad c \in \mathbb{R}^p.$$

(a) *Écrire le problème paramétré $(P)_\alpha$ avec objectif $F_\alpha(x) = f(x) + \alpha B(x)$, et la solution optimale $x(\alpha)$. Montrer que F_α est convexe. Donner une condition*

suffisante sur C de sorte que F_α est strictement convexe.
(b) *Montrer que h est un multiple scalaire de $\nabla B(x(\alpha))$, autrement dit, l'hyperplan $\{x \in \mathbb{R}^d : hx = hx(\alpha)\}$ est tangent à la courbe de niveau $\{x \in \mathbb{R}^d : B(x) = B(x(\alpha))\}$.*

5.2.4. Exercice :
Considérons le programme linéaire

$$(P) : \min\{hx : x \geq 0, Ax = b\}, \quad h \in \mathbb{R}^{1\times d}, \quad A \in \mathbb{R}^{m\times d}, \quad b \in \mathbb{R}^m.$$

(a) *Écrire le problème paramétré $(P)_\alpha$ avec objectif $F_\alpha(x) = f(x) + \alpha B(x)$, et la solution optimale $x(\alpha)$. Montrer que F_α est strictement convexe.*
(b) *Notons $X_j = \operatorname{diag}(x^{(j)}), A_j = AX_j$, et $e = (1, ..., 1)^T$. Montrer que une itération de Newton pour $(P)_\alpha$ sous contraintes affines d'égalité est donné par*

$$x^{(j+1)} = x^{(j)} + t_j d^{(j)} = X_j\Big(e + t_j(I - A_j^T(A_jA_j^T)^{-1}A_j)(e - \frac{1}{\alpha}X_jh^T)\Big).$$

5.2.5. Exercice :
Pour démarrer la méthode du point intérieur pour (P), on a besoin d'un point x' de sorte que $Ax' = b$ et $g_k(x') < 0$ pour $k = 1, ..., p$.
(a) *Soit x'' avec $Ax'' = b$. Comment démarrer une méthode de point intérieur pour résoudre*

$$\min\{s : Ax = b, \ \textit{pour } k = 1, ..., p : \quad g_k(x) \leq s\}$$

de valeur optimale s^* ?
(b) *Conclure en fonction du signe de s^*.*

5.3 Le gradient à pas fixe avec projection sur un convexe

Le but de ce chapitre est de résoudre le problème

$$(CP) \quad \min\{f(x) : x \in C\}$$

où on supposera que f est de classe $\mathcal{C}^1$ (plus d'autres conditions sur le gradient), par contre, la forme précise de convexe fermé $C \subset \mathbb{R}^d$ n'est pas imposée, il faudra juste savoir calculer pour tout $y \in \mathbb{R}^d$ la projection $\Pi(y)$ sur C, c'est-à-dire, l'unique élément le plus proche de y dans C, voir les 1.3.5 et 1.5.5. Pour un pas fixe ρ à spécifier plus tard, on envisage l'algorithme

$$x^{(0)} \in C, \quad \text{et pour } j = 1, 2, \ldots : \quad x^{(j+1)} = \Pi(x^{(j)} - \rho \nabla f(x^{(j)})^T). \tag{5.3}$$

Notons que dans le cas $C = \mathbb{R}^d$, $\Pi(y) = y$ pour tout $y \in \mathbb{R}^d$, et donc (5.3) se réduit à l'algorithme du steepest descent à pas fixe.

L'étude de convergence va être différente à celle du chapitre 3 : pour une solution optimale $\underline{x}$, on cherchera pas à estimer la différence des valeurs $f(x^{(j)}) - f(\underline{x})$, mais directement la différence des arguments $\|\underline{x} - \underline{x}^{(j)}\|$, à l'aide du théorème du point fixe dans $\mathbb{R}^d$.

5.3.1. Théorème du point fixe.
Soit $F \subset \mathbb{R}^d$ un fermé, $h : F \mapsto F$, et supposons qu'il existe $\kappa < 1$ de sorte que

$$\forall x, y \in F : \quad \|h(x) - h(y)\| \leq \kappa \|x - y\| \tag{5.4}$$

(h est dit une contraction sur F de rapport κ). Alors il existe un et un seul $\underline{x} \in F$ vérifiant $\underline{x} = h(\underline{x})$ (dit point fixe de h). De plus, soit $x^{(j)}$ obtenu par l'itération de Picard $x^{(0)} \in F$, et pour $j = 0, 1, 2, \ldots$: $x^{(j+1)} = h(x^{(j)})$. Alors $(x^{(j)})_j$ admet la limite $\underline{x}$, avec estimations d'erreur

$$\|\underline{x} - x^{(j)}\| \leq \kappa \|\underline{x} - x^{(j-1)}\| \leq \frac{\kappa}{1-\kappa} \|x^{(j)} - x^{(j-1)}\| \leq \frac{\kappa^j}{1-\kappa} \|x^{(1)} - x^{(0)}\|. \quad (5.5)$$

Démonstration. L'unicité du point fixe découle directement de (5.4). Pour montrer l'existence d'un point fixe, étudions la convergence de la suite de Picard. Par récurrence sur j en utilisant le fait que $h(F) \subset F$, on montre que $x^{(j)} \in F$, et que donc la suite de Picard est bien définie. De nouveau par récurrence on montre que $\|x^{(j+1)} - x^{(j)}\| \leq \kappa^j \|x^{(1)} - x^{(0)}\|$, et que, pour $m > j \geq 0$,

$$\|x^{(m)} - x^{(j)}\| \leq \sum_{k=0}^{m-n-1} \|x^{(j+k+1)} - x^{j+k}\| \leq \frac{\|x^{(j+1)} - x^{(j)}\|}{1-\kappa} \leq \frac{\kappa^j}{1-\kappa} \|x^{(1)} - x^{(0)}\|.$$

On déduit de la dernière inégalité que la suite de Picard est une suite de Cauchy, et donc admet une limite $\underline{x}$, et $\underline{x} \in F$ par fermeture de F. La relation (5.4) implique que h est continue en $\underline{x}$ et alors

$$\underline{x} = \lim_{j \to \infty} x^{(j+1)} = \lim_{j \to \infty} h(x^{(j)}) = h(\underline{x}),$$

et donc $\underline{x}$ est un point fixe. Les estimations d'erreur sont obtenus en faisant tendre m vers $+\infty$. □

Notre algorithme (5.3) est alors l'itération de Picard pour la fonction $h(x) = \Pi(x - \rho \nabla f(x)^T)$, ce qui nous fournira les estimations d'erreur. Il faudra alors analyser de plus près cette fonction h.

5.3.2. Lemme
Soit $\Pi(y)$ la projection de $y \in \mathbb{R}^d$ sur le convexe fermé C, alors

$$\forall x, y \in \mathbb{R}^d : \quad \|\Pi(y) - \Pi(x)\| \leq \|y - x\|.$$

Démonstration. Rappelons du 1.5.5 la caractérisation $(\Pi(y) - y)^T(z - \Pi(y)) \geq 0$ pour tout $z \in C$. Alors

$$\begin{aligned}
&\|\Pi(y) - \Pi(x)\|^2 \\
&= (\Pi(y) - y)^T(\Pi(y) - \Pi(x)) + (y - x)^T(\Pi(y) - \Pi(x)) + (x - \Pi(x))^T(\Pi(y) - \Pi(x)) \\
&\leq (y - x)^T(\Pi(y) - \Pi(x)),
\end{aligned}$$

et le résultat en découle en appliquant Cauchy-Schwarz. □

5.3.3. Lemme
Avec la fonction $h(x) = \Pi(x - \rho\nabla f(x)^T)$, $\underline{x} \in \mathbb{R}^d$ est un point fixe de h sur C si et seulement si $\underline{x}$ est une solution optimale de (CP).

Démonstration. Posons $y = \underline{x} - \rho\nabla f(\underline{x})^T$. Alors d'après le théorème 1.5.1 on a les équivalences

$$\begin{aligned}
&\underline{x} = \Pi(y) \quad \Longleftrightarrow \quad \underline{x} = \arg\min\{\|y - z\|^2 : z \in C\} \quad \Longleftrightarrow \quad \forall z \in C : (\underline{x} - y)^T(z - \underline{x}) \geq 0 \\
&\Longleftrightarrow \quad \forall z \in C : \nabla f(\underline{x})(z - \underline{x}) \geq 0 \quad \Longleftrightarrow \quad \underline{x} = \arg\min\{f(z) : z \in C\}
\end{aligned}$$

□

Ce travail préliminaire nous permet de donner un théorème de convergence.

5.3.4. Théorème de convergence pour l'algorithme du gradient à pas fixe avec projection sur un convexe.

Soient f de classe $\mathcal{C}^1$ et $\alpha, \beta \in \mathbb{R}$ de sorte que, pour tout $x, y \in C$

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|, \tag{5.6}$$

$$\Big(\nabla f(x) - \nabla f(y)\Big)(x - y) \geq \alpha \|x - y\|^2, \tag{5.7}$$

et soit[4]

$$\rho \in]0, \frac{2\alpha}{\beta^2}[, \quad \kappa := \sqrt{1 - 2\rho\alpha + \rho^2\beta^2} \in [0, 1[.$$

Alors (CP) admet une unique solution optimale $\underline{x}$.
Finalement, soit la suite $(x^{(j)})_j$ calculée par l'algorithme (5.3). Alors on a stationnarité $x^{(j+1)} = x^{(j)}$ ssi $x^{(j)} = \underline{x}$, et sinon $(x^{(j)})_j$ converge vers $\underline{x}$, avec estimations d'erreur (5.5).

Démonstration. D'après le théorème du point fixe 5.3.1 et le lemme 5.3.2 il suffit de démontrer que $h : C \mapsto C$ définie par $h(x) = \Pi(x - \rho\nabla f(x)^T$ est une contraction de rapport κ sur C. On obtient pour $x, y \in C$

$$\begin{aligned}
\|h(x) - h(y)\|^2 &\leq \|x - y - \rho(\nabla f(x)^T - \nabla f(y)^T)\|^2 \qquad \text{(par le lemme 5.3.2)} \\
&= \|x - y\|^2 - 2\rho(\nabla f(x) - \nabla f(y))(x - y) + \rho^2\|\nabla f(x) - \nabla f(y)\|^2 \\
&\leq \|x - y\|^2(1 - 2\alpha\rho + \beta^2\rho^2) \qquad \text{(par l'hypothèse sur } f) \\
&= \kappa^2 \|x - y\|^2.
\end{aligned}$$

□

4. Le lecteur intéressé vérifiera que κ prend la valeur minimale $\sqrt{1 - \alpha^2/\beta^2}$ pour le choix $\rho = \alpha/\beta^2$.

Pour une tolérance $\epsilon > 0$, les estimations (5.5) nous donnent aussi la condition d'arrêt $\|x^{(j+1)} - x^{(j)}\| \leq \epsilon$ car dans ce cas $\|x^{(j+1)} - \underline{x}\| \leq \epsilon\kappa/(1-\kappa)$.

5.3.5. Exercice

Dans le cas ou f est une forme quadratique avec Hessien H s.d.p., vérifier que les hypothèses du théorème 5.3.4 sont valables avec $\beta = \|H\|$ et $\alpha = 1/\|H^{-1}\|$, et que le meilleur taux de convergence est donné par $\rho = \frac{\alpha}{\beta^2}$ et $\kappa = \sqrt{1 - 1/cond^2(H)}$. Comparer avec le taux obtenu au chapitre 3.

5.3.6. Exercice

Soit f comme dans le théorème 5.3.4 et, de plus, on suppose que les inégalités (5.6) et (5.7) soient valables pour $x, y \in \mathbb{R}^d$.

(a) En observant que $q(1)-q(0)-q'(0) = \int_0^1 (q'(t)-q'(0))dt$ pour toute fonction $q \in \mathcal{C}^1([0,1])$, montrer que

$$\forall x, y \in \mathbb{R}^d : \quad f(y) - f(x) - \nabla f(x)(y-x) \begin{cases} \leq \beta\|y-x\|^2/2, \\ \geq \alpha\|y-x\|^2/2. \end{cases}$$

(b) En déduire que f est strictement convexe sur $\mathbb{R}^d$, et que $f(x) \to \infty$ si $\|x\| \to \infty$, et qu'il existe une solution optimale unique $\underline{x}$ de (CP).

(c) Soit de plus $f \in \mathcal{C}^2$. En posant $y = x + tv$, vérifier que les hypothèses $(H3)$ et (3.1) du chapitre 3 sont valables.

On rappelle que l'algorithme (5.3) nécessite d'évaluer la projection $\Pi(y)$ pour certains $y \in \mathbb{R}^d$, ce qui est assez facile pour les ensembles convexes et fermés C

discutés dans le 1.5.5, mais il n'existe que rarement de formules explicites pour le cas général

$$C = \bigcap_{k=1}^{p} \{x \in \mathbb{R}^d : g_k(x) \leq 0\}$$

(sans contraintes d'égalité) avec g_k convexe, ou même pour le cas d'un polyèdre C. Ici, en utilisant la théorie de la dualité et du Lagrangien $L(x, \lambda) = f(x) + \lambda g(x)$ exposée dans le chapitre 2, on peut se ramener à un problème dual (D) avec comme seule contrainte que $\lambda \geq 0$. L'application au problème dual (D) de l'algorithme du gradient à pas fixe avec projection sur un convexe (ici l'orthant positif) donne lieu à l'algorithme d'Uzawa que l'on va spécifier dans le reste de ce sous-chapitre.

Le premier ingrédient du problème dual est de trouver pour un vecteur ligne $\lambda \in \mathbb{R}^{1\times p}$ avec $\lambda \geq 0$ fixé un minimiseur de la fonction $x \mapsto L(x, \lambda)$ sur $\mathbb{R}^d$. Par l'exercice 5.3.6(b), cette fonction est strictement convexe et admet un et un seul minimiseur sur $\mathbb{R}^d$, noté par $x(\lambda)$. On sait aussi que $x(\lambda)$ est l'unique solution x du système $\nabla_x L(x, \lambda) = \nabla f(x) + \lambda \nabla g(x)$. Nous arrivons alors au problème dual

$$\max\{w(\lambda) : \lambda \in \mathbb{R}^{1\times p}, \lambda \geq 0\}, \quad w(\lambda) = L(x(\lambda), \lambda).$$

Rappelons que $\lambda^T \mapsto -w(\lambda)$ est convexe sur l'orthant positif, mais peut ne pas être différentiable. Pour combler à cette difficulté, on supposera ici que l'application $\lambda^T \mapsto x(\lambda)$ est différentiable (ce qui nécessite d'ajouter des hypothèses sur f et g, voir le DS1 et l'exercice2.3.12). Le lecteur intéressé pourrait vérifier que dans ce cas aussi $\lambda^T \mapsto w(\lambda)$ est différentiable sur l'orthant positif, avec dérivée

$$\frac{\partial w}{\partial \lambda_\ell}(\lambda) = g_\ell(x(\lambda)).$$

L'algorithme d'Uzawa prend alors la forme suivante.

5.3.7. Algorithme d'Uzawa

En entrée : $\lambda^{(0)} \geq 0$, *tolérance* $\epsilon > 0$, *pas fixe* $\rho > 0$

Algo : *Pour* $j = 0, 1, ...$

chercher minimiseur $x^{(j)}$ *de* $x \mapsto L(x, \lambda^{(j)})$ *sur* $\mathbb{R}^d$, *calculer* $g(x^{(j)})$

pour $\ell = 1, ..., p$ *calculer* $\lambda_\ell^{(j+1)} = \max\{0, \lambda_\ell^{(j)} + \rho g_\ell(x^{(j)})\}$

arrêt si $\|\lambda^{(j+1)} - \lambda^{(j)}\| < \epsilon$

En sortie : $\lambda^{(j)}$ *approximation de* $\underline{\lambda}$ *sol. opt. de* (D) *avec* $\|\lambda^{(j)} - \underline{\lambda}\| \leq \epsilon/(1-\kappa)$,

$x^{(j)}$ *approximation de* $\underline{x}$ *sol. opt. de* (CP) *avec* $\|x^{(j)} - \underline{x}\| \leq \frac{\epsilon\beta\|B^\dagger\|}{(1-\kappa)^{3/2}}$.

5.3.8. Théorème de convergence pour Uzawa sur polyèdre C

Considérons $g(x) = Bx - b$ *avec* $B \in \mathbb{R}^{p\times d}$ *de rang* p, *de sorte qu'il existe une inverse à droite* $B^\dagger = B^*(BB^*)^{-1}$. *Soit* f *comme dans le théorème 5.3.4 et, de plus, on suppose que les inégalités* (5.6) *et* (5.7) *soient valables pour* $x, y \in \mathbb{R}^d$. *Soit finalement*

$$\rho \in]0, \frac{2\alpha}{\|B\|^2}[, \quad \kappa := \sqrt{1 - \frac{2\rho\alpha - \rho^2\|B\|^2}{\beta^2\|B^\dagger\|^2}} \in [0, 1[.$$

Alors (D) *admet une seule solution optimale* $\underline{\lambda}$, *et nous avons les estimations d'erreur*

$$\|\underline{\lambda} - \lambda^{(j)}\| \leq \kappa\|\underline{\lambda} - \lambda^{(j-1)}\| \leq \frac{\kappa}{1-\kappa}\|\lambda^{(j)} - \lambda^{(j-1)}\| \leq \frac{\kappa^j}{1-\kappa}\|\lambda^{(1)} - \lambda^{(0)}\|$$

et

$$\|\lambda^{(j)} - \underline{\lambda}\| \leq \frac{\|\lambda^{(j+1)} - \lambda^{(j)}\|}{1-\kappa}, \quad \|x^{(j)} - \underline{x}\| \leq \frac{\beta\|B^\dagger\|\,\|\lambda^{(j+1)} - \lambda^{(j)}\|}{(1-\kappa)^{3/2}}$$

Démonstration. Avec Π la projection sur l'orthant positif $M \subset \mathbb{R}^{1\times p}$, montrons que $h : M \mapsto M$ défini par $h(\lambda) = \Pi(\lambda + \rho g(x(\lambda))^T)$ est une contraction de rapport κ. En se servant du lemme 5.3.2, nous obtenons pour tout $\lambda, \mu \in M$

$$\begin{aligned}
&\|h(\lambda) - h(\mu)\|^2 \leq \|(\lambda-\mu) + \rho(x(\lambda) - x(\mu))^T B^T\|^2 \\
&= \|\lambda-\mu\|^2 + \rho^2\|B(x(\lambda) - x(\mu))\|^2 + 2\rho(\lambda-\mu)B(x(\lambda) - x(\mu)) \\
&\leq \|\lambda-\mu\|^2 + \rho^2\|B\|^2\|x(\lambda) - x(\mu)\|^2 - 2\rho(\nabla f(x(\lambda)) - \nabla f(x(\mu)))(x(\lambda) - x(\mu))
\end{aligned}$$

où dans la dernière inégalité nous avons utilisé le fait que $\nabla f(x(\mu)) = -\mu B$ et $\nabla f(x(\lambda)) = -\lambda B$. En utilisant l'hypothèse sur f, nous arrivons à

$$\|h(\lambda) - h(\mu)\|^2 \leq \|\lambda-\mu\|^2 - [-\rho^2\|B\|^2 + 2\alpha\rho]\,\|x(\lambda) - x(\mu)\|^2 \tag{5.8}$$

où on remarque que le terme entre crochets est > 0 par hypothèse sur ρ. Il reste alors de minorer

$$\|x(\lambda) - x(\mu)\| \geq \frac{1}{\beta}\|\nabla f(x(\lambda)) - \nabla f(x(\mu))\| = \frac{\|(\lambda-\mu)B\|}{\beta} \geq \frac{\|\lambda-\mu\|}{\beta\|B^\dagger\|}.$$

En combinant les deux inégalités, nous obtenons bien une contraction de rapport κ. Par conséquent, il existe un seul point fixe $\underline{\lambda} = h(\underline{\lambda})$, et on montre comme dans le Lemme 5.3.3 que $\underline{\lambda}$ est alors la seule solution optimale du problème dual (D). Les estimations d'erreur pour $\lambda^{(j)}$ découlent alors du théorème du point fixe. De plus, par la forme particulière du projecteur, nous obtenons pour $\underline{x} = x(\underline{\lambda})$ que

$$\underline{\lambda} = \Pi(\underline{\lambda} + \rho g(\underline{x})^T) \geq \underline{\lambda} + \rho g(\underline{x})^T$$

impliquant que $\underline{x}$ est réalisable pour (CP), et $\underline{\lambda} \geq 0$, avec $\underline{\lambda}_k > 0$ impliquant que $g_k(\underline{x}) = 0$. Donc $(\underline{x}, \underline{\lambda})$ est un point-côl de notre Lagrangien, et alors $\underline{x} = x(\underline{\lambda})$ est (l'unique) solution optimale de notre problème (CP). De l'équation (5.8) avec $\lambda = \lambda^{(j)}$, $x(\lambda) = x^{(j)}$ et $\mu = \underline{\lambda}$, $x(\mu) = \underline{x}$ nous déduisons que

$$\|\lambda^{(j)} - \underline{\lambda}\|^2 \geq [-\rho^2\|B\|^2 + 2\alpha\rho] \, \|x^{(j)} - \underline{x}\|^2 = \beta^2\|B^\dagger\|^2(1-\kappa^2) \, \|x^{(j)} - \underline{x}\|^2$$

ce qui implique la dernière inégalité (conditions d'arrêt). □

5.4 La méthode d'Active set

Dans ce dernier sous-chapitre on souhaite donner un algorithme fini pour la minimisation d'une forme quadratique strictement convexe sur un polyèdre de la forme $\{x \in \mathbb{R}^d : Bx \leq b\}$. Deux raisons expliquent pourquoi un tel cas particulier est de grande importance

— un tel problème doit être résolu si on veut projeter sur un polyèdre sous la forme indiquée ci-dessus ;

— dans la procédure du sequential quadratic programming (SQP), on construit une suite approchant un minimiseur de f sur $\bigcap_{k=1}^p \{x \in \mathbb{R}^d : g_k(x) \leq 0\}$ comme suit : étant donné $x^{(j)}$, on remplace l'objectif par son développement de Taylor d'ordre 2, et les contraintes par leur linéarisé (=leur développement

de Taylor d'ordre 1), donnant lieu à l'itération [5]

$$x^{(j+1)} = \arg\min\{f(x^{(j)}) + \nabla f x^{(j)}(x - x^{(j)}) + \frac{1}{2}(x - x^{(j)})^T \nabla^2 f(x^{(j)})(x - x^{(j)}) : \\ \forall k = 1, ..., p : g_k(x^{(j)}) + \nabla g_k(x^{(j)})(x - x^{(j)}) \leq 0\}.$$

Notre but est alors de résoudre le problème

$$(P) : \quad \min\{f(x) : Bx \leq b\}, \quad f(x) = \frac{1}{2}x^T Hx + h^T x,$$

avec $H \in \mathbb{R}^{d\times d}$ s.d.p., $B \in \mathbb{R}^{p\times d}$ de rang p, $b \in \mathbb{R}^p$, $h \in \mathbb{R}^d$. Rappelons que ce problème admet une solution unique $\underline{x}$ par stricte convexité de f et par le fait que les ensembles de niveau de f sont compacts. ' Posons $J = \{1, ..., p\}$, et pour un sous-ensemble $I \subset J$ on notera $B_I \in \mathbb{R}^{|I|\times d}$ la sous-matrice obtenue en extrayant les lignes à indice $j \in I$ de B, et par b_I le sous-vecteur en extrayant les lignes à indice $j \in I$ de b.

Au lieu de résoudre directement (P), nous allons résoudre pour une suite de différents ensembles I les problèmes avec contraintes affines d'égalité

$$(P)_I : \quad \min\{f(x) : B_I x = b_I\}, \quad \text{avec solution optimale } x(I),$$

où l'existence et unicité de $x(I)$ se démontre comme pour (P). On voit que le théorème KKT pour $(P)_I$ se réduit à un système d'équations linéaires (comparer avec le lemme 4.2.1 dans le cas $b_I = 0$) : $x(I)$ est solution optimale du problème

5. Dans ce document on ne va pas se retarder de montrer convergence pour SQP, ce qui nécessite des résultats supplémentaires sur nos fonctions $f, g_1, ..., g_p$.

$(P)_I$ ssi il existe $\mu(I) \in \mathbb{R}^{1\times|I|}$ dit multiplicateur de Lagrange de sorte que

$$\begin{bmatrix} H & B_I^T \\ B_I & 0 \end{bmatrix} \begin{bmatrix} x(I) \\ \mu(I)^T \end{bmatrix} = \begin{bmatrix} -h \\ b_I \end{bmatrix}. \tag{5.9}$$

Rappelons que la matrice dans (5.9) est inversible car les lignes de B et donc celles de B_I sont libres. Pour un $x \in \mathbb{R}^d$ réalisable pour (P), on notera par $E(x)$ l'ensemble des indices des contraintes actives : $E(x) = \{j \in J : g_j(x) = 0\}$, $g(x) = Bx - b$. Le lien entre (P) et $(P)_I$ est donné par le lemme suivant qui spécifie la condition d'arrêt de notre algorithme.

5.4.1. Lemme.
Soit x réalisable pour (P), et $I \subset E(x)$. Si $f(x) \leq f(x(I))$ alors $x = x(I)$, la solution optimale de $(P)_I$.
Si de plus $\mu(I) \geq 0$ alors $x = x(I) = \underline{x}$, la solution solution optimale de (P).

Démonstration. Par définition de $E(x)$, $B_I x = b_I$, autrement dit, x est réalisable pour $(P)_I$, et donc $f(x) \geq f(x(I))$ par définition de $x(I)$. Si alors $f(x) \leq f(x(I))$ alors x coincide avec l'unique solution optimale $x(I)$ de $(P)_I$. Pour démontrer le résultat de la deuxième phrase, soit $\lambda \in \mathbb{R}^{1\times p}$ défini par $\lambda_j = \mu(I)_j$ pour $j \in I$, et $\lambda_j = 0$ sinon. Alors, en utilisant (5.9),

$$\lambda \geq 0, \quad \nabla f(x) = x^T H + h^T = -\mu(I) B_I = -\lambda B,$$

aussi, x est réalisable pour (P), et finalement, $\lambda_j g_j(x) = 0 g_j(x) = 0$ pour $j \notin I$ et $\lambda_j g_j(x) = \lambda_j 0 = 0$ pour $j \in I$, c'est-à-dire, nous obtenons une solution (x, λ) au système (KKT) pour (P), et $x = \underline{x}$ est l'unique solution optimale de (P). □

5.4.2. Algorithme "Active set" pour minimiser une forme quadratique sur un polyèdre

En entrée : $x^{(0)} \in \mathbb{R}^d$ *réalisable pour* (P), $I^{(0)} = E(x^{(0)})$.
Algo : *Pour* $j = 0, 1, ...$
On dispose de $x^{(j)} \in \mathbb{R}^d$ *réalisable pour* (P), $I^{(j)} \subset E(x^{(j)})$.
calculer $y = x(I^{(j)})$ *et* $\mu(I^{(j)})$ *par* (5.9)
si $f(x^{(j)}) \leq f(y)$
si (cas (1a)) $\mu(I^{(j)}) \geq 0$ *alors STOP*
sinon (cas (1b)) chercher $k \in I^{(j)}$ *avec* $\mu(I^{(j)})_k < 0$
poser $x^{(j+1)} = x^{(j)}$, $I^{(j+1)} = E(x^{(j)}) \setminus \{k\}$.
sinon (cas (2))
calculer $\alpha_j = \max\{\alpha \in [0,1] : (1-\alpha)x^{(j)} + \alpha y$ *réalisable pour* $(P$
poser $x^{(j+1)} = (1-\alpha_j)x^{(j)} + \alpha_j y$, $I^{(j+1)} = E(x^{(j+1)})$.
En sortie : $x^{(j)}$ *solution optimale de* (P).

5.4.3. Théorème : Finitude.

L'algorithme 5.4.2 est fini.

Démonstration. Notons plus explicitement $y^{(j)} = x(I^{(j)})$. Si dans le cas (2) $\alpha_j = 1$ (ssi $y^{(j)}$ est réalisable pour (P)) on parlera du sous-cas (2a), et sinon du sous-cas (2b). Supposons par absurde que l'algorithme ne s'arrête pas.
(i) Montrons que la suite $(f(x^{(j)}))_j$ est décroissante. Dans le cas (2) nous obtenons par convexité de f que $f(x^{(j+1)}) \leq (1-\alpha_j)f(x^{(j)}) + \alpha_j f(y^{(j)})$ et $f(y^{(j)}) < f(x^{(j)})$ par construction. Donc $f(x^{(j+1)}) \leq f(x^{(j)})$ ce qui trivialement est aussi vrai pour le cas (1b).

(ii) Montrons que si le cas (1b) est vrai pour l'indice j alors le cas (2) est valable pour l'indice $j+1$. Rappelons que $f(x^{(j)}) \leq f(y^{(j)})$ et $I^{(j)} \subset E(x^{(j)})$ impliquent que $x^{(j)} = y^{(j)}$ par le lemme 5.4.1, et alors $x^{(j+1)} = y^{(j)}$. Par absurde, supposons que $f(x^{(j+1)}) \leq f(y^{(j+1)}$ et alors $f(x^{(j)}) \leq f(y^{(j+1)})$. Comme par construction $I^{(j+1)} \subset E(x^{(j)})$, nous déduisons du lemme 5.4.1 pour $I^{(j+1)})$ que $y^{(j+1)} = x^{(j)} = y^{(j)}$, et donc

$$\mu(I^{(j+1)})B_{I^{(j+1)}} = \mu(I^{(j)})B_{I^{(j)}}$$

par (5.9). Par construction $k \in I^{(j)} \setminus I^{(j+1)}$ et $\mu(I^{(j)})_k \neq 0$, cette dernière relation implique que la $k$ième ligne de B est une combinaison linéaire des autres lignes de B, une contradiction avec l'hypothèse sur B. Donc $f(x^{(j+1)}) > f(y^{(j+1)}$, c'est-à-dire, nous avons le cas (2) pour l'indice $j+1$.
(iii) Montrons que si (1b) est vrai pour l'indice j alors $f(x^{(j+2)}) < f(x^{(j)})$. Dans la partie (ii) nous avons montré que le cas (2) est valable pour l'indice $j+1$, et $I^{(j+1)} = E(x^{(j+1)}) \setminus \{k\}$, ainsi que $f(x^{(j+1)}) = f(x^{(j)}) > f(y^{(j+1)})$ par construction. Revenons sur la preuve de (i). Supposons un instant que $\alpha_{j+1} > 0$ ce qui sera montré ci-dessous. Dans ce cas, $x^{(j+2)} = (1-\alpha_{j+1})x^{(j+1)} + \alpha_{j+1}y^{(j+1)}$ pour un $\alpha_{j+1} \in]0,1]$, et par convexité de f

$$\begin{aligned} x^{(j+2)} &\leq (1-\alpha_{j+1})f(x^{(j+1)}) + \alpha_{j+1}f(y^{(j+1)}) \\ &< (1-\alpha_{j+1})f(x^{(j+1)}) + \alpha_{j+1}f(x^{(j+1)}) = f(x^{(j+1)}) = f(x^{(j)}), \end{aligned}$$

ce qu'il fallait démontrer. Par absurde, supposons que $\alpha_{j+1} = 0$. Dans ce cas, par construction il existe un indice ℓ avec $g_\ell(x^{(j+1)}) = 0$ et $g_\ell(y^{(j+1)}) > 0$, c'est-à-dire, forcément $\ell \in E(x^{(j+1)}) \setminus I^{(j+1)}$, et alors $\ell = k \in I^{(j)}$. Posons $\mu_k = -\mu(I^{(j)})_k$, $\mu_\ell = \mu(I^{(j)})_\ell$ pour $\ell \in I^{(j)} \setminus \{k\}$, et $\mu_\ell = 0$ pour tout autre indice dans $E(x^{(j)})$.

Alors en comparant avec (5.9) on vérifie que $(y(I^{(j)}), \mu)$ est solution du système (KKT) pour le problème convexe

$$(P') : \min\{f(x) : g_k(x) \geq 0, \forall i \in I^{(j+1)} : g_i(x) = 0\}.$$

et alors $x^{(j+1)} = y^{(j)} = x(I^{(j)})$ est solution optimale de (P'), et $y^{(j+1)}$ réalisable pour (P'), en contradiction avec l'hypothèse que $f(x^{(j+1)}) > f(y^{(j+1)})$.
(iv) Montrons que si le cas (2a) est vrai pour l'indice j alors le cas (1) est valable pour l'indice $j+1$. Par construction nous avons que $x^{(j+1)} = y^{(j)}$ réalisable pour (P), et $I^{(j)} \subset E(y^{(j)}) = I^{(j+1)}$. Donc $y^{(j)}$ est réalisable pour le problème $(P)_{I^{(j+1)}}$, ce qui implique que $f(y^{(j)}) \geq f(y^{(j+1)})$ et alors $f(x^{(j+1)}) \geq f(y^{(j+1)})$, le cas (1) pour l'indice $j+1$.
(v) Montrons que si le cas (2b) est valable pour les indices successives $j, j+1, ..., j'$ alors $j' - j \leq p$. En effet, si $\ell \in \{j+1, ..., j'\}$ alors $I^{(\ell)} = E(x^{(\ell)})$ (car le cas (2b) était valable pour l'indice $\ell - 1$). Pour tout $k \in I^{(\ell)} = E(x^{(\ell)})$ nous avons par construction que $g_k(x^{(\ell)}) = 0 = g_k(y^{(\ell)}) = g_k(x^{(\ell+1)})$, mais par définition de α_ℓ on doit avoir un indice $k \notin I^{(\ell)}$ avec $g_k(x^{(\ell+1)}) = (1 - \alpha_\ell) g_k(x^{(\ell)}) + \alpha_\ell g_k(y^{(\ell)}) = 0$. Donc $I^{(\ell+1)} = E(x^{(\ell+1)}) \supset I^{(\ell)} \cup \{k\} \supsetneq I^{(\ell)}$, autrement dit, $I^{(\ell+1)}$ contient au moins un élément de plus que $I^{(\ell)}$, mais ne peut pas contenir plus que p éléments. Donc $j' - j \leq p$.
Il découle du (ii), (iv) et (v) que si l'algorithme ne se termine pas alors le cas (1b) doit être vrai pour un nombre infini d'indices, disons pour $j \in \Lambda$. Si $j_1 < j_2$ sont des éléments de Λ alors on a le cas (2) pour l'indice $j_1 + 1$ par (ii) et alors $j_2 \geq j_1 + 2$. En utilisant $j_1 \in \Lambda$, le (iii), le (i), et finalement $j_2 \in \Lambda$ nous concluons que

$$f(x(I^{(j_1)})) = f(x^{(j_1)}) > f(x^{(j_1+2)}) \geq f(x^{(j_2)}) = f(x(I^{(j_2)})),$$

ce qui implique que $I^{(j_1)} \neq I^{(j_2)}$. Donc les listes $I^{(j)}$ pour $j \in \Lambda$ devraient être distincts, mais on peut seulement construire un nombre fini de sous-ensembles distincts de $\{1, ..., p\}$, une contradiction. Par conséquent, l'algorithme se termine. □

5.4.4. Remarque.
Presque le même algorithme (et la même preuve de finitude) s'applique si on veut résoudre un problème de la forme

$$\min\{f(x) : \forall k = 1, ..., m : (Bx - b)_k = 0, \forall k = m + 1, ..., p : (Bx - b)_k \leq 0\}.$$

Il suffit de remplacer le test $\mu(I^{(j)}) \geq 0$ *par* $\mu(I^{(j)})_\ell \geq 0$ *pour* $\ell = m + 1, ..., p$, *et de choisir sur la ligne suivante* $k > m$. *L'algorithme aura alors la propriété que* $\{1, ..., m\} \subset I^{(j)}$ *pour tout* j, *c'est-à-dire, les contraintes d'égalité seront bien actives pour* $x(I^{(j)})$.

Bibliographie

[BV] S. Boyd, L. Vandenberghe, Convex optimization, Cambridge (2004), gratuitement disponible à l'adresse http://www.ee.ucla.edu/∼vandenbe/cvxbook.html. On trouve les cours Convex Optimisation 1 (et 2) basés sur ce livre et donnés par S. Boyd à l'Université de Stanford sur youtube à l'adresse https://www.youtube.com/watch?v=McLq1hEq3UY&list=PL3940DI

[H] Raphaèle Herbin, Cours d'Analyse numérique, Licence de mathématiques de l'Université Aix Marseille, mai 2020 gratuitement disponible à l'adresse https://www.i2m.univ-amu.fr/perso/raphaele.herbin/PUBLI/anamat.pdf. On trouve des exercices corrigés dans les chapitres 3.1.3, 3.2.3, 3.3.5 (assez longue et un peu plus ambitieux), 3.4.5, et 3.5.3 (assez longue et plus ambitieux).

[HU] J.-B. Hiriat-Uruty, Optimisation et analyse convexe, EDP Sciences (2009), contient exercices corrigés.

[M] Michel Minoux, Programmation mathématique : théorie et algorithmes, Dunod, 1989.