COLLECTION ENSEIGNEMENT SUP //// Mathématiques

L3M1

Optimisation et analyse convexe

EXERCICES CORRIGÉS

Jean-Baptiste Hiriart-Urruty

EDP SCIENCES

This page intentionally left blank

OPTIMISATION ET ANALYSE CONVEXE

Exercices et problèmes corrigés, avec rappels de cours

Jean-Baptiste Hiriart-Urruty

Collection dirigée par Daniel Guin

EDP SCIENCES

17, avenue du Hoggar
Parc d'activités de Courtabœuf, BP 112
91944 Les Ulis Cedex A, France

Illustration de couverture : un corps convexe d'épaisseur presque constante et son ombre ; reproduit avec la gracieuse permission de Christof Weber (université de Zurich).

Imprimé en France

ISBN : 978-2-7598-0373-6

TABLE DES MATIÈRES

INTRODUCTION

« *Good modern science implies good variational problems* »
M.S. Berger (1983)

Le recueil d'exercices et problèmes corrigés que nous proposons ici concerne les domaines des Mathématiques répertoriées sous les vocables d'**Optimisation** et **Analyse convexe**. L'Optimisation est traitée dans ses aspects suivants : la clé de voûte que constituent les conditions d'optimalité (chapitres II et III) ; le rôle (incontournable) de la dualisation de problèmes (chapitre IV) ; le monde particulier (et toujours en haut de l'affiche depuis ses débuts) de l'Optimisation linéaire (chapitre V). L'Analyse convexe (moderne) n'est pas traitée en tant que telle mais par l'utilisation qu'on peut en avoir en Optimisation ; il s'agit en fait d'une initiation à la manipulation de concepts et de résultats concernant essentiellement : la projection sur un convexe fermé (au chapitre VI), le calcul sous-différentiel et de transformées de Legendre-Fenchel (chapitre VII). L'Analyse linéaire et bilinéaire (ou, plutôt, l'Analyse matricielle) ainsi que le Calcul différentiel interviennent de manière harmonieuse en Optimisation et Analyse convexe : un chapitre de revision des bases leur est consacré (chapitre I). Près de 160 exercices et problèmes sont corrigés, parfois commentés et situés dans un contexte d'utilisation ou de développement historique, gradués dans leur difficulté par un, deux ou trois $*$:

$*$ Exercices plutôt faciles (applications immédiates d'un résultat du Cours, vérification d'un savoir-faire de base, etc.) ;

$**$ Exercices que le lecteur-étudiant doit pouvoir aborder après une bonne compréhension et assimilation du Cours. De difficulté moyenne, ce sont de loin les plus nombreux ;

$***$ Exercices plus difficiles, soit à cause de certains calculs à mener à bien, soit simplement en raison d'un degré de maturité plus grand que leur résolution requiert.

Comme tous les exercices de mathématiques, ceux présentés ici ne seront profitables au lecteur-étudiant que si celui-ci les travaille, un crayon à la main, sans

regarder la correction dans un premier temps. Qu'il garde à l'esprit ce proverbe chinois :

« J'entends et j'oublie, (cours oral)

je vois et je retiens, (étude du cours)

je fais et je comprends » . (exercices)

Le *cadre de travail* choisi est volontairement simple (celui des espaces de dimension finie), et nous avons voulu insister sur les *idées* et *mécanismes de base* davantage que sur les généralisations possibles ou les techniques particulières à tel ou tel contexte. Les problèmes dits *variationnels* requièrent dans leur traitement une intervention plus grande de la Topologie et de l'Analyse fonctionnelle, à commencer par le cadre – fondamental – des espaces de Hilbert ; ils seront abordés dans un prochain recueil.

Les *connaissances mathématiques* pour tirer profit des exercices et problèmes du recueil présent sont maintenues minimales, celles normalement acquises après une formation scientifique à Bac + 2 ou Bac + 3 (suivant les cas).

Chaque chapitre débute par des rappels de résultats essentiels, ce qui ne doit pas empêcher le lecteur-étudiant d'aller consulter les références indiquées à la fin du livre. L'approche retenue est celle d'une progression en spirale plutôt que linéaire au sens strict : ainsi, par exemple, la fonction $A \in \mathcal{M}_n(\mathbb{R}) \longmapsto \ln(\det A)$ est d'abord considérée pour un calcul de différentielles, puis pour sa convexité, puis plus tard en raison de son rôle comme fonction-barrière dans des problèmes d'optimisation matricielle.

Pour ce qui est de l'*enseignement*, les aspects de l'Optimisation et Analyse convexe traités en exercices ici trouvent leur place dans les formations de niveau deuxième cycle universitaire (modules généralistes ou professionnalisés) et dans la formation mathématique des ingénieurs, sur une durée d'un semestre environ ; la connaissance de ces aspects est un préalable à des formations plus en aval, en optimisation numérique par exemple.

La plupart des exercices et problèmes proposés, sinon tous, ont été posés en séances d'exercices ou examens à l'Université Paul Sabatier de Toulouse.

Je voudrais remercier les anciens étudiants ou jeunes collègues qui ont bien voulu relire une première version de ce document et y relever une multitude de petites fautes (il en reste sûrement...), parmi eux : D. Mallard, M. Torki, Y. Lucet, C. Imbert et J. Benoist. Enfin je ne voudrais pas oublier A. Andrei pour la part primordiale qui a été la sienne dans la saisie informatique de l'ouvrage.

Toulouse, 1989–1997
J.-B. Hiriart-Urruty

Depuis sa publication il y a dix ans (en mars 1998), cet ouvrage a subi les vicissitudes d'un document de formation destiné à un public (d'étudiants en sciences) en nette diminution. Il a été traduit en russe par des collègues de Kiev (Ukraine) en 2004, mais la version française originelle n'est plus disponible depuis 2006. Ainsi, pour répondre à une demande de collègues et étudiants, un nouveau tirage a été envisagé. Je remercie les éditions EDP Sciences, notamment mon collègue D. Guin (directeur de la collection Enseignement Sup – Mathématiques), d'avoir accueilli ce projet. Aude Rondepierre a donné un coup de main pour reprendre les fichiers informatiques anciens ; qu'elle soit remerciée de sa bonne volonté et efficacité.

Toulouse, printemps 2009
J.-B. Hiriart-Urruty

This page intentionally left blank

ABRÉVIATIONS ET NOTATIONS

$:=$: égal par définition.

cf. : *confer*, signifie « se reporter à ».

i.e. : *id est*, signifie « c'est-à-dire ».

ln : notation normalisée pour le logarithme népérien.

$\mathbb{R}_*^+$, $\mathbb{R}_+^*$ ou $]0,+\infty[$: ensemble des réels strictement positifs.

u^+ : partie positive du réel u.

$x = (x_1, \ldots, x_n)$ ou $x = (\xi_1, \ldots, \xi_n)$: notation générique pour un vecteur de $\mathbb{R}^n$.

u^+ signifie $(u_1^+, \ldots, u_n^+)$ lorsque $u = (u_1, \ldots, u_n) \in \mathbb{R}^n$.

Lorque u et v sont deux vecteurs de $\mathbb{R}^n$, $u \leqslant v$ signifie « $u_i \leqslant v_i$ pour tout $i = 1, \ldots, n$ ».

$\{u_k\}$ ou (u_k) : notations utilisées pour les suites indexées par des entiers naturels.

Pour une fonction f différentiable en x (resp. deux fois différentiable en x), $Df(x)$ désigne la différentielle (première) de f en x (resp. $D^2f(x)$ désigne la différentielle seconde de f en x). Si la variable est réelle (et notée t), on utilise la notation $f'(t)$ (resp. $f''(t)$) pour la dérivée de f en t (resp. la dérivée seconde de f en t) [ce sont des éléments de l'espace d'arrivée et non des applications linéaires].

Pour une fonction numérique f définie sur un ouvert $\mathcal{O}$ de $\mathbb{R}^n$, différentiable en $x \in \mathcal{O}$ (resp. deux fois différentiable en $x \in \mathcal{O}$), $\nabla f(x)$ (resp. $\nabla^2 f(x)$) désigne le (vecteur) *gradient* de f en x (resp. la matrice *hessienne* de f en x).

Lorsqu'elle existe, la dérivée directionnelle de f en x dans la direction d est notée $f'(x, d)$.

Pour une fonction vectorielle $f : \mathcal{O} \subset \mathbb{R}^n \to \mathbb{R}^m$ différentiable en $x \in \mathcal{O}$, $Jf(x)$ désigne la matrice *jacobienne* de f en x (matrice à m lignes et n colonnes).

$\mathcal{M}_{m,n}(\mathbb{R})$: ensemble des matrices (m, n) (m lignes et n colonnes) à coefficients réels ; $\mathcal{M}_n(\mathbb{R})$ est une abréviation de $\mathcal{M}_{n,n}(\mathbb{R})$.

$[a_{ij}]$: matrice de terme général a_{ij} (à la i-ème ligne et j-ème colonne).

diag $(\lambda_1, \ldots, \lambda_n)$: matrice diagonale dont les éléments diagonaux sont $\lambda_1, \ldots, \lambda_n$.

I_n (ou I quand il n'y a pas d'ambiguïté) : matrice-unité de $\mathcal{M}_n(\mathbb{R})$, *i.e.* diag (1,...,1).

$A^\top$ ou ${}^tA$: *transposée* de $A \in \mathcal{M}_{m,n}(\mathbb{R})$ [les deux notations sont d'un usage très courant ; par contre A^t est à proscrire car génératrice de confusions].

Lorsque A est inversible, $A^{-\top}$ désigne l'inverse de $A^\top$ (ou, ce qui revient au même, la transposée de A^{-1}).

tr A : trace de $A \in \mathcal{M}_n(\mathbb{R})$.

dét A : déterminant de $A \in \mathcal{M}_n(\mathbb{R})$.

cof A : matrice des cofacteurs de $A \in \mathcal{M}_n(\mathbb{R})$, *i.e.* celle dont le terme (i, j) est $(-1)^{i+j}$ dét A_{ij}, où A_{ij} est obtenue à partir de A en enlevant la i-ème ligne et la j-ème colonne.

$\mathcal{S}_n(\mathbb{R})$: ensemble des matrices de $\mathcal{M}_n(\mathbb{R})$ qui sont symétriques.

$\oplus$ symbolise la somme directe de sous-espaces vectoriels.

vect$\{v_1, \ldots, v_k\}$: sous-espace vectoriel engendré par les vecteurs $v_1, \ldots, v_k$.

Sauf indication contraire, $\mathbb{R}^n$ est muni de sa base canonique ; ainsi *à* $A \in \mathcal{M}_{m,n}(\mathbb{R})$ *est canoniquement associée une application linéaire de* $\mathbb{R}^n$ *dans* $\mathbb{R}^m$, d'où les notations Ker A, Im A, etc.

L'isomorphisme canonique de $\mathbb{R}^n$ sur $\mathcal{M}_{n,1}(\mathbb{R})$ est celui qui à $x = (x_1, \ldots, x_n)$ associe la matrice unicolonne $X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$; des expressions comme AX (ou Ax) ne devraient pas arrêter l'étudiant-lecteur. Si, par exemple, u et v sont deux vecteurs de $\mathbb{R}^n$, $uv^\top$ *est une matrice carrée de taille* n dont le terme général est $u_i v_j$, alors que $u^\top v$ est la matrice-scalaire (ou scalaire) $\sum_{i=1}^{n} u_i v_j$.

$\langle \cdot, \cdot \rangle$, $\ll \cdot, \cdot \gg$, $(\cdot \mid \cdot)$: notations utilisées pour les produits scalaires (dans des espaces euclidiens). Sauf indication contraire, $\langle \cdot, \cdot \rangle$ désigne dans $\mathbb{R}^n$ le produit scalaire usuel (celui qui à $x = (\xi_1, \ldots, \xi_n)$ et $y = (\eta_1, \ldots, \eta_n)$ associe $\langle x, y \rangle := \sum_{i=1}^{n} \xi_i \eta_i$, soit encore $x^\top y$ (*cf.* supra)). Bien des problèmes d'optimisation se posent dans des espaces de matrices : si $X := \mathcal{M}_{m,n}(\mathbb{R})$, le produit scalaire standard sur X est défini par $\ll M, N \gg := \text{tr}\,(M^\top N)$.

Soit $A \in \mathcal{M}_n(\mathbb{R})$, u et v des vecteurs de $\mathbb{R}^n$: $u^\top Av$ est un(e) (matrice-) scalaire égal(e) à (sa transposée) $uA^\top v$; un mécanisme plus commode d'utilisation et engendrant moins de fautes est d'écrire $\langle u, Av\rangle = \left\langle A^\top u, v\right\rangle$.

Si l est une application linéaire d'un espace euclidien $(E, \ll \cdot, \cdot \gg)$ dans un autre espace euclidien $(F, \langle\cdot,\cdot\rangle)$, l'adjointe l^* de l est l'application linéaire de F dans E définie par

$$\ll l^*(y), x \gg = \langle y, l(x)\rangle \text{ pour tout } (x, y) \in E \times F.$$

Si l'on représente l'ensemble des formes linéaires sur E par E *via* le produit scalaire $\ll \cdot, \cdot \gg$ (idem pour F), prendre l'adjointe de l ou sa transposée revient au même. Lorsque E et F sont munies de bases orthonormales (comme c'est le cas des espaces euclidiens $(\mathbb{R}^n, \langle\cdot,\cdot\rangle)$ munies de leurs bases canoniques), la matrice représentant l'adjointe de l (= sa transposée) dans ces bases est la transposée de la matrice représentant l.

$\|\cdot\|$: norme dans un espace vectoriel normé X. Si X est muni d'un produit scalaire $\langle\cdot,\cdot\rangle$ (exemples typiques : $\mathbb{R}^n$, $\mathcal{M}_{m,n}(\mathbb{R})$), et en l'absence d'autres précisions, $\|\cdot\|$ désignera la norme dérivée du produit scalaire (*i.e.* $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$). Lorsqu'interviennent à la fois des normes de vecteurs et de matrices, on évite les confusions en utilisant $\|\cdot\|$ pour les vecteurs et $|||\cdot|||$ pour les matrices).

int S ou $\overset{\circ}{S}$: intérieur de S ; $\overline{S}$: adhérence de S ;
fr S : frontière de S.
$\overline{B}(x, r)$: boule fermée de centre x et de rayon r.

Δ_n : *simplexe-unité* de $\mathbb{R}^n$, c'est-à-dire l'ensemble des $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$ tels que $\alpha_1 + \ldots + \alpha_n = 1$ et $\alpha_i \geqslant 0$ pour tout i (coefficients qui servent dans la prise de combinaisons convexes).

$A \in \mathcal{S}_n(\mathbb{R})$ est dite (symétrique) *semi-définie positive* lorsque $\langle Ax, x\rangle \geqslant 0$ pour tout $x \in \mathbb{R}^n$ [cette appellation est préférable à « positive » qui peut signifier, dans certains cas, « à coefficients positifs »]. On désignera par $\mathcal{P}_n(\mathbb{R})$ l'ensemble des matrices semi-définies positives de taille n.

$A \in \mathcal{S}_n(\mathbb{R})$ est dite (symétrique) *définie positive* lorsque $\langle Ax, x\rangle > 0$ pour tout $x \neq 0$ de $\mathbb{R}^n$ (ce qui revient à : A semi-définie positive et inversible). On désignera par $\overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ l'ensemble des matrices définies positives de taille n.

Le caractère de semi-définie ou définie positivité ne sera considéré dans ce recueil que pour des matrices symétriques.

This page intentionally left blank

I

RÉVISION DE BASES : CALCUL DIFFÉRENTIEL, ALGÈBRE LINÉAIRE ET BILINÉAIRE

Rappels

I.1. Algèbre linéaire et bilinéaire

Si $l : \mathbb{R}^n \to \mathbb{R}$ est linéaire, il existe un unique $v \in \mathbb{R}^n$ tel que $l(x) = \langle v, x\rangle$ pour tout $x \in \mathbb{R}^n$; les fonctions affines à valeurs réelles sont de la forme $x \longmapsto \langle v, x\rangle + c$, où $v \in \mathbb{R}^n$ et $c \in \mathbb{R}$. Même chose si l est une forme linéaire sur un espace euclidien $(E, \langle\cdot,\cdot\rangle)$: l s'écrit de manière unique sous la forme $\langle v, \cdot\rangle$, où $v \in E$.

Si $q : \mathbb{R}^n \to \mathbb{R}$ est une forme quadratique, il existe un unique $Q \in \mathcal{S}_n(\mathbb{R})$ tel que $q(x) = \dfrac{1}{2}\langle Qx, x\rangle$ pour tout $x \in \mathbb{R}^n$ (le coefficient $1/2$ est mis là pour simplifier les calculs).

Rappel du mécanisme de transposition : $\langle Ax, y\rangle = \langle x, A^\top y\rangle$ (produits scalaires dans $\mathbb{R}^m$ et $\mathbb{R}^n$ respectivement).

Si H est un sous-espace vectoriel de $\mathbb{R}^n$, $H^\perp$ désigne son sous-espace orthogonal. Rappel : si A est linéaire, $(\mathrm{Ker}A)^\perp = \mathrm{Im}(A^\top)$.

Un résultat-clé de l'Algèbre linéaire et bilinéaire : si $S \in \mathcal{S}_n(\mathbb{R})$, toutes les valeurs propres λ_i de S sont réelles et il existe une matrice orthogonale U telle que $U^\top SU = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$; ainsi il existe des vecteurs propres x_i unitaires (x_i associé à λ_i) tels que $S = \displaystyle\sum_{i=1}^{n} \lambda_i x_i x_i^\top$ (ceci est appelé une *décomposition spectrale* de S).

Les formulations variationnelles suivantes de la plus grande et la plus petite valeurs propres de $S \in \mathcal{S}_n(\mathbb{R})$ sont essentielles :

$$\lambda_{max}(S) = \max_{x \neq 0} \frac{\langle Sx, x\rangle}{\| x \|^2}, \quad \lambda_{min}(S) = \min_{x \neq 0} \frac{\langle Sx, x\rangle}{\| x \|^2}.$$

$\mathcal{S}_n(\mathbb{R})$ est muni de l'ordre (partiel) suivant : $A \succeq B$ dans $\mathcal{S}_n(\mathbb{R})$ lorsque $A - B$ est semi-définie positive ($A - B \in \mathcal{P}_n(\mathbb{R})$).

Si $X \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R})$, la racine carrée positive de X (notée $X^{1/2}$) est l'unique $S \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ telle que $S^2 = X$. L'application $X \longmapsto X^{1/2}$ a des propriétés similaires à celles de la fonction racine carrée sur $\mathbb{R}_*^+$; prudence tout de même... Mêmes remarques pour l'application $X \longmapsto X^{-1}$.

Si K est un cône convexe fermé d'un espace euclidien $(E, \langle \cdot, \cdot \rangle)$, le cône polaire K° de K est le cône convexe fermé de E défini comme suit :

$$K^\circ := \{s \in E \mid \ \langle s, x\rangle \leqslant 0 \text{ pour tout } \ x \in K\}.$$

I.2. Calcul différentiel

$f : \mathcal{O} \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ définie sur un ouvert $\mathcal{O}$ de $\mathbb{R}^n$ est dite *différentiable* en $\overline{x} \in \mathcal{O}$ s'il existe $l : \mathbb{R}^n \longrightarrow \mathbb{R}$ linéaire telle que :

$$f(\overline{x} + h) = f(\overline{x}) + l(h) + \| h \| \varepsilon(h), \text{ avec } \ \varepsilon(h) \to 0 \text{ quand } h \underset{\neq}{\longrightarrow} 0. \qquad (1.1)$$

l est représentée (dans $(\mathbb{R}^n, \langle \cdot, \cdot \rangle)$ euclidien) par un unique vecteur, appelé *gradient* de f en $\overline{x}$, et noté $\nabla f(\overline{x})$ (ça se lit « nabla f de $\overline{x}$ »). Désignant par $\partial_j f(\overline{x})$ la *dérivée partielle* en $\overline{x}$ de f par rapport à la j^{e} variable, $\nabla f(\overline{x})$ est le vecteur de $\mathbb{R}^n$ de composantes $\partial_1 f(\overline{x}), ..., \partial_n f(\overline{x})$. Une manière équivalente d'exprimer (1.1) est :

$$\lim_{t \to 0^+} \frac{f(\overline{x} + td) - f(\overline{x})}{t} = \langle \nabla f(\overline{x}), d \rangle \qquad (1.2)$$

et la convergence est uniforme par rapport à d lorsque celui-ci reste dans un ensemble borné de $\mathbb{R}^n$ (la sphère-unité par exemple).

Plus généralement, $F : \mathcal{O} \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$ qui à $x \in \mathcal{O}$ associe $F(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{pmatrix}$ est dite différentiable en $\overline{x} \in \mathcal{O}$ si chacune des fonctions-composantes $f_1, ..., f_m$ l'est. On appelle alors matrice *jacobienne* de F en $\overline{x}$, et on note $JF(\overline{x})$ la matrice de $\mathcal{M}_{m,n}(\mathbb{R})$ dont les lignes sont $[\nabla f_1(\overline{x})]^\top, ..., [\nabla f_m(\overline{x})]^\top$, c'est-à-dire que le terme (i, j) de $JF(\overline{x})$ est $\partial_j f_i(\overline{x})$.

Revenant à une fonction numérique $f : \mathcal{O} \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ différentiable sur $\mathcal{O}$, on dit que f est deux fois différentiable en $\overline{x} \in \mathcal{O}$ lorsque $\nabla f : \mathcal{O} \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ est différentiable en $\overline{x}$. La matrice jacobienne de ∇f en $\overline{x}$ est appelée matrice *hessienne* de f en $\overline{x}$ et notée $\nabla^2 f(\overline{x})$; il s'agit d'une matrice symétrique de taille n dont le terme (i,j) est $\partial_i(\partial_j f)(\overline{x}) = \partial^2_{ij} f\,(\overline{x})$ (dérivée partielle d'ordre 2 en $\overline{x}$ de f par rapport à la i^{e} et j^{e} variables).

Rappel : dans le cas où f est deux fois différentiable en $\overline{x} \in \mathcal{O}$, on a le développement de Taylor-Young d'ordre deux suivant :

$$f\,(\overline{x}+h) = f\,(\overline{x}) + \langle \nabla f\,(\overline{x})\,, h\rangle + \frac{1}{2}\langle \nabla^2 f\,(\overline{x})\,h, h\rangle + \parallel h \parallel^2 \varepsilon\,(h)\,, \tag{1.3}$$

avec $\varepsilon\,(h) \to 0$ quand $h \underset{\neq}{\longrightarrow} 0$.

Enfin deux ensembles de résultats de Calcul différentiel sont essentiels en Optimisation : le théorème de la fonction implicite et le théorème d'inversion locale ; les développements de Taylor sous leurs formes diverses. À revoir si nécessaire.

I.3. Fonctions convexes

Soit C un convexe de $\mathbb{R}^n$; $f : C \to \mathbb{R}$ est dite *convexe* sur C si pour tout $(x, x') \in C \times C$ et tout $\alpha \in]0,1[$ on a :

$$f\left(\alpha x + (1-\alpha)\,x'\right) \leqslant \alpha f\,(x) + (1-\alpha)\,f(x'). \tag{1.4}$$

f est dite *strictement convexe* sur C quand l'inégalité (1.4) est stricte dès que $x \neq x'$. Une propriété encore plus forte est comme suit : f est dite *fortement convexe* sur C, de module de forte convexité $c > 0$, lorsque

$$f\left(\alpha x + (1-\alpha)\,x'\right) \leqslant \alpha f\,(x) + (1-\alpha)\,f(x') - \frac{1}{2} c\,\alpha\,(1-\alpha) \parallel x' - x \parallel^2 \tag{1.5}$$

pour tout $(x, x') \in C \times C$ et tout $\alpha \in]0,1[$.

Rappelons deux résultats essentiels :

Théorème. *Soit f différentiable sur un ouvert $\mathcal{O}$ de $\mathbb{R}^n$ et C un convexe de $\mathcal{O}$. Alors :*

(i) f est convexe sur C si et seulement si

$$f\,(x) \geqslant f\,(\overline{x}) + \langle \nabla f\,(\overline{x})\,, x - \overline{x}\rangle \text{ pour tout } (x, \overline{x}) \in C \times C\ ; \tag{1.6}$$

(*ii*) *f est strictement convexe sur C si et seulement si l'inégalité* (1.6) *est stricte dès que* $\overline{x} \neq x$;

(*iii*) *f est fortement convexe sur C de module c si et seulement si*

$$f(x) \geqslant f(\overline{x}) + \langle \nabla f(\overline{x}), x - \overline{x} \rangle + \frac{1}{2} c \parallel x - \overline{x} \parallel^2, \text{ pour tout } (x, \overline{x}) \in C \times C. \quad (1.7)$$

Théorème. *Soit f deux fois différentiable sur un ouvert convexe $\mathcal{O}$. Alors :*

(i) f est convexe sur $\mathcal{O}$ si et seulement si $\nabla^2 f(x)$ est semi-définie positive pour tout $x \in \mathcal{O}$;

(ii) Si $\nabla^2 f(x)$ est définie positive pour tout $x \in \mathcal{O}$, alors f est strictement convexe sur $\mathcal{O}$;

(iii) f est fortement convexe sur $\mathcal{O}$ de module c si et seulement si la plus petite valeur propre de $\nabla^2 f(x)$ est minorée sur $\mathcal{O}$ par c , soit encore :

$$\langle \nabla^2 f(x)\, d, d \rangle \geqslant c \parallel d \parallel^2 \text{ pour tout } x \in \mathcal{O} \text{ et tout } d \in \mathbb{R}^n.$$

Références. Parmi les exercices de [15] figurent des applications simples à l'Analyse numérique matricielle et l'Optimisation. Outre les références suggérées en pp. 305 307 de [15] (et rééditées à présent), signalons [16], ouvrage complet et très diffusé, dans lequel l'aspect matriciel (celui qui prévaut dans les applications) est privilégié.

Pour les fonctions convexes différentiables, on pourra consulter [12], Chapitre IV, Section 4, par exemple.

***Exercice I.1.** Soit $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ continûment différentiable.

1°) Soit x_0 tel que $\nabla f(x_0) \neq 0$. Que représente $\nabla f(x_0)$ pour la surface de niveau $S := \{x \in \mathbb{R}^n \mid f(x) = f(x_0)\}$?

2°) Rappeler l'équation de l'hyperplan affine (de $\mathbb{R}^{n+1}$) tangent au graphe de f en $(x_0, f(x_0))$.

Donner à l'aide de $\nabla f(x_0)$ un vecteur normal à cet hyperplan.

3°) On suppose qu'il existe $L > 0$ tel que

$$\parallel \nabla f(x) - \nabla f(x') \parallel \leqslant L \parallel x - x' \parallel \text{ pour tout } (x, x') \in \mathbb{R}^n \times \mathbb{R}^n.$$

Montrer qu'alors

$$|f(x + d) - f(x) - \langle \nabla f(x), d \rangle| \leqslant \frac{L}{2} \parallel d \parallel^2 \text{ pour tout } (x, d) \in \mathbb{R}^n \times \mathbb{R}^n.$$

Solution : 1°) $\nabla f(x_0)$ est un vecteur normal à S en x_0 ; le sous-espace tangent à S en x_0 est orthogonal à $\nabla f(x_0)$.

2°) Soit $\mathrm{gr} f := \{(x, f(x)) \mid x \in \mathbb{R}^n\}$ le graphe de f (partie de $\mathbb{R}^{n+1}$ donc). Le sous-espace affine tangent à $\mathrm{gr} f$ en $(x_0, f(x_0))$ a pour équation

$$y = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle.$$

Un vecteur normal à cet hyperplan est fourni par $(\nabla f(x_0), -1) \in \mathbb{R}^n \times \mathbb{R}$.

On peut voir aussi $\mathrm{gr} f$ comme la surface de niveau 0 de la fonction

$$g : (x, r) \in \mathbb{R}^n \times \mathbb{R} \longmapsto g(x, r) := f(x) - r.$$

Comme $\nabla g\,(x_0, f(x_0)) = (\nabla f(x_0), -1)$ n'est jamais nul, le vecteur $(\nabla f(x_0), -1)$ est normal à $\mathrm{gr} f$ en $(x_0, f(x_0))$.

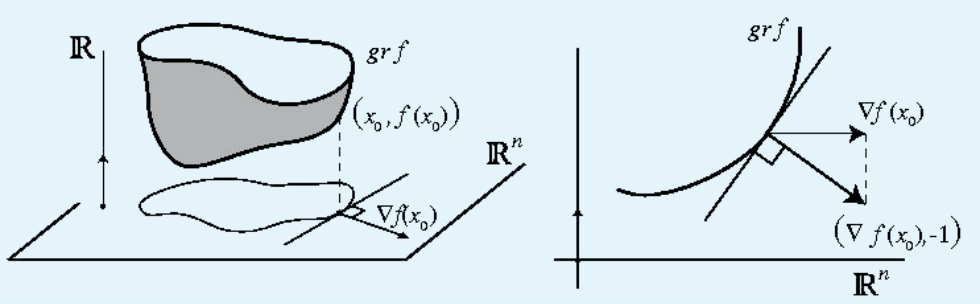


Figure 1.

3°) f étant continûment différentiable, on a grâce à la formule de Taylor à l'ordre zéro avec reste sous forme d'intégrale

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

Sachant que $\langle \nabla f(x), d \rangle = \displaystyle\int_0^1 \langle \nabla f(x), d \rangle dt$, on transforme l'expression précédente en

$$f(x + d) - f(x) - \langle \nabla f(x), d \rangle = \int_0^1 \langle \nabla f(x + td) - \nabla f(x), d \rangle dt.$$

Il s'ensuit grâce à l'hypothèse faite sur ∇f :

$$|\langle \nabla f\,(x+td) - \nabla f(x), d\rangle| \leqslant \| \nabla f\,(x+td) - \nabla f(x) \| \cdot \| d \| \\ \leqslant tL \| d \|^2,$$

d'où :

$$|f\,(x+d) - f(x) - \langle \nabla f(x), d\rangle| \leqslant \int_0^1 (tL \| d \|^2) dt = \frac{L}{2} \| d \|^2 .$$

Commentaire :

– Prolongement de l'exercice. On suppose à présent que f est deux fois différentiable sur $\mathbb{R}^n$. Comment traduire alors en termes de $\nabla^2 f$ l'hypothèse dans la 3$^{\text{e}}$ question ? Proposer ensuite une démonstration de l'inégalité de la 3$^{\text{e}}$ question qui s'appuie sur une autre formule de Taylor.

– Il importe de bien assimiler ce que représentent géométriquement les objets $\nabla f(x)$ et $\nabla^2 f(x)$ vis-à-vis des surfaces de niveau et du graphe de f.

***Exercice I.2.**

Situations :	À déterminer :
1°) $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, $x \longmapsto f(x) := \langle c, x\rangle + \gamma$.	$\nabla f(x), \nabla^2 f(x)$ en tout point x.
2°) $F : \mathbb{R}^n \longrightarrow \mathbb{R}^m$, $x \longmapsto F(x) := Lx + b$, où $L \in \mathcal{M}_{m,n}(\mathbb{R})$.	$JF(x)$.
3°) $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, $x \longmapsto f(x) := \frac{1}{2}\langle Ax, x\rangle + \langle d, x\rangle + \delta$, où $A \in \mathcal{M}_n(\mathbb{R})$.	$\nabla f(x), \nabla^2 f(x)$.
4°) $g : \mathbb{R}^n \longrightarrow \mathbb{R}$, $x \longmapsto g(x) := \sum_{i=1}^{m} [r_i(x)]^2$, où les $r_i : \mathbb{R}^n \longrightarrow \mathbb{R}$ sont deux fois différentiables.	$\nabla g(x), \nabla^2 g(x)$.

Commentaire : Ne pas se lancer tête baissée dans le calcul de dérivées partielles !

Pour les fonctions de base que sont les fonctions affines, quadratiques, sommes de carrés, etc., il faut savoir déterminer $\nabla f(x), \nabla^2 f(x)$ rapidement et sans erreur, « mécaniquement » presque.

Solution :

1°) $\nabla f(x) = c$, $\nabla^2 f(x) = 0$ pour tout $x \in \mathbb{R}^n$.

2°) $JF(x) = L$ (L est la partie linéaire de la fonction affine F).

3°) $\nabla f(x) = \frac{1}{2}\left(A + A^\top\right) x + d$, $\nabla f(x) = Ax + d$ si A est symétrique.
$\nabla^2 f(x) = \frac{1}{2}\left(A + A^\top\right)$, $\nabla^2 f(x) = A$ si A est symétrique.

4°) $\nabla g(x) = 2\sum_{i=1}^{m} r_i(x)\nabla r_i(x)$,

$$\nabla^2 g(x) = 2\sum_{i=1}^{m}\left\{\nabla r_i(x)\left[\nabla r_i(x)\right]^\top + r_i(x)\nabla^2 r_i(x)\right\}.$$

Ainsi, si $(r_1(x), ..., r_m(x))$ est « proche de 0 dans $\mathbb{R}^m$ », on peut considérer que $2\sum_{i=1}^{m}\nabla r_i(x)\left[\nabla r_i(x)\right]^\top$ est une « bonne » approximation de $\nabla^2 g(x)$; approximation qui, de plus, ne fait appel qu'aux gradients des fonctions r_i.

***Exercice I.3.**

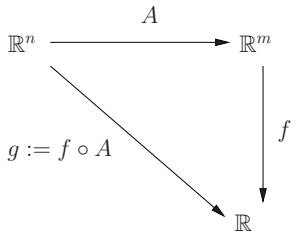
Situations :

1°) $g : \mathbb{R}^n \longrightarrow \mathbb{R}$

$$x \longmapsto g(x) := \sum_{i=1}^{m}\left[f_i^+(x)\right]^2,$$

où les $f_i : \mathbb{R}^n \longrightarrow \mathbb{R}$ sont deux fois différentiables et a^+ désigne $\max(0, a)$.

2°)



où $A : u \longmapsto Au = A_0 u + b$, avec $A_0 \in \mathcal{M}_{m,n}(\mathbb{R})$, et f deux fois différentiable.

Application : Soit : $\mathbb{R}^n \longrightarrow \mathbb{R}$ deux fois différentiable, $x_0 \in \mathbb{R}^n$, $d \in \mathbb{R}^n$. On pose $\varphi(t) := f(x_0 + td)$.

Questions :

g est-elle différentiable deux fois différentiable ?

Déterminer $\nabla g, \nabla^2 g$ en fonction des éléments correspondants ∇f et $\nabla^2 f$.

Calculer $\varphi'(t)$ (resp. $\varphi''(t)$) en fonction de $\nabla f(x_0 + td)$ (resp. $\nabla^2 f(x_0 + td)$).

Solution : 1°) La fonction $u \in \mathbb{R} \longmapsto (u^+)^2$ est dérivable une fois, mais pas deux fois (d'accord ?). Par suite, g est différentiable sur $\mathbb{R}^n$ et

$$\nabla g(x) = 2\sum_{i=1}^{m} f_i^+(x) \cdot \nabla f_i(x) \text{ pour tout } x \in \mathbb{R}^n.$$

Mais, en général, g n'est pas deux fois différentiable en les points x où l'un des f_i s'annule.

Lorsque les f_i sont de classe $\mathcal{C}^2$, la fonction g est dans une classe intermédiaire entre $\mathcal{C}^2(\mathbb{R}^n)$ et $\mathcal{C}^1(\mathbb{R}^n)$, celle des fonctions différentiables dont le gradient est localement lipschitzien.

2°) $\nabla g(u) = A_0^\top \nabla f(Au)$; $\nabla^2 g(u) = A_0^\top \nabla^2 f(Au) A_0$.

Application : $A : \mathbb{R} \longrightarrow \mathbb{R}^n$ définie par $At := x_0 + td$. Comme $A_0 : t \longmapsto A_0 t = td$, nous avons $A_0^\top : \mathbb{R}^n \longrightarrow \mathbb{R}$ qui à $x \in \mathbb{R}^n$ associe $\langle x, d\rangle$. Par conséquent, les éléments *de* $\mathbb{R}$ que sont les dérivées première et seconde (usuelles) de φ sont données par :

$$\varphi'(t) = \langle \nabla f(x_0 + td), d\rangle, \qquad \varphi''(t) = \langle \nabla^2 f(x_0 + td)d, d\rangle.$$

***Exercice I.4.** Soit $\mathcal{M}_n(\mathbb{R})$ structuré en espace euclidien grâce au produit scalaire $\ll A, B \gg := \text{ tr } \left(A^\top B\right)$; soit Ω l'ouvert de $\mathcal{M}_n(\mathbb{R})$ constitué des matrices inversibles. On considère les applications suivantes :

1°) tr : $A \in \mathcal{M}_n(\mathbb{R}) \longmapsto \text{tr}A \in \mathbb{R}$
2°) dét : $A \in \mathcal{M}_n(\mathbb{R}) \longmapsto \text{dét}A \in \mathbb{R}$
3°) $g : A \in \Omega \longmapsto g(A) := \ln|\text{dét}A| \in \mathbb{R}$
4°) $f_{-1} : A \in \Omega \longmapsto f_{-1}(A) := A^{-1} \in \Omega$
5°) $p \in \mathbb{N}^*$, $f_p : A \in \mathcal{M}_n(\mathbb{R}) \longmapsto f_p(A) := A^p \in \mathcal{M}_n(\mathbb{R})$

$$f_{-p} : A \in \Omega \longmapsto f_{-p}(A) := A^{-p} \in \Omega.$$

Indiquer rapidement pourquoi ces applications sont différentiables. Déterminer en tout point de leurs domaines de définition le gradient pour les trois premières et la différentielle pour les dernières.

Commentaire : Si $l : \mathcal{M}_n(\mathbb{R}) \longrightarrow \mathbb{R}$ est linéaire (l est une forme linéaire sur $\mathcal{M}_n(\mathbb{R})$ donc), l est continue et il existe un unique $M \in \mathcal{M}_n(\mathbb{R})$ tel que

$$l(H) = \ll M, H \gg \text{ pour tout } H \in \mathcal{M}_n(\mathbb{R}).$$

Lorsque l est la différentielle $Df(x_0)$ d'une fonction $f : \Omega \subset \mathcal{M}_n(\mathbb{R}) \longrightarrow \mathbb{R}$ différentiable en $x_0 \in \Omega$, son représentant M dans $\mathcal{M}_n(\mathbb{R})$ est le gradient de f en x_0 (et est noté $\nabla f(x_0)$).

Bien maîtriser le calcul différentiel sur les fonctions de matrices est essentiel en Analyse matricielle, Optimisation et (surtout) Statistique.

Solution : 1°) $A \longmapsto \mathrm{tr}A$ est linéaire (continue) ; donc sa différentielle en tout $A \in \mathcal{M}_n(\mathbb{R})$ est elle-même :

$$D\ \mathrm{tr}\,(A) : H \longmapsto \mathrm{tr}H = \ll I_n, H \gg .$$

Par conséquent

$$\nabla\ \mathrm{tr}\,(A) = I_n \text{ pour tout } A \in \mathcal{M}_n(\mathbb{R}).$$

2°) L'application *dét* peut être vue comme une application multilinéaire (continue) de $\mathcal{M}_n(\mathbb{R}) \equiv \mathbb{R}^n \times \cdots \times \mathbb{R}^n$ dans $\mathbb{R}$. Elle est donc différentiable en tout $A = \left[a^1, ..., a^n\right]$ avec

$$D\ \text{dét}\ (A) : H = \left[h^1, ..., h^n\right] \longmapsto \sum_{j=1}^{n} \text{dét}\ \left[a^1, ..., a^{j-1}, h^j, a^{j+1}, ..., a^n\right].$$

En développant $\text{dét}\left[a^1, ..., a^{j-1}, h^j, a^{j+1}, ..., a^n\right]$ par rapport à la j^{e} colonne, on obtient :

$$\text{dét}\left[a^1, ..., a^{j-1}, h^j, a^{j+1}, ..., a^n\right] = \sum_{i=1}^{n} h_i^j\,(\mathrm{cof}A)_{ij}\,,$$

où $cof\ A$ désigne la matrice des cofacteurs de A. Il s'ensuit :

$$[D\ \text{dét}(A)]\,(H) = \sum_{i,j=1}^{n} h_i^j\,(\mathrm{cof}A)_{ij} = \ll \mathrm{cof}A, H \gg .$$

En définitive,

$$\nabla\text{dét}(A) = \mathrm{cof}A \text{ en tout } A \in \mathcal{M}_n(\mathbb{R}).$$

3°) Par application du théorème de composition des applications différentiables, g est différentiable et

$$Dg(A) : H \longmapsto \frac{\ll \mathrm{cof}A, H \gg}{\text{dét}\ A} = \ll \left(A^{-1}\right)^{\top}, H \gg .$$

Donc

$$\nabla g(A) = \left(A^{-1}\right)^{\top} \text{ en tout } A \in \Omega.$$

4°) L'application $A \longmapsto \text{cof}\, A$ est différentiable car chacune de ses applications-composantes (qui sont des déterminants) est différentiable ; il s'ensuit que l'application $f_{-1} : A \longmapsto A^{-1} = \dfrac{(\text{cof}A)^{\top}}{\det A}$ est différentiable. Le calcul de la différentielle de f_{-1} en A se fait rapidement à partir de la relation

$$f_1(A) f_{-1}(A) = I_n \text{ pour tout } A \in \Omega,$$

que l'on différentie. On trouve :

$$Df_{-1}(A) : H \longmapsto [D\ f_{-1}(A)]\, H = -A^{-1} H A^{-1}.$$

5°) L'application qui à $(A_1, \ldots, A_p) \in \mathcal{M}_n(\mathbb{R}) \times \ldots \times \mathcal{M}_n(\mathbb{R})$ associe le produit $A_1 A_2 \ldots A_p$ est multilinéaire (continue) ; elle est donc différentiable et sa différentielle en $(A_1, \ldots, A_p)$ est :

$$(H_1, \ldots, H_p) \longmapsto \sum_{i=1}^{p} A_1 \ldots A_{i-1} H_i A_{i+1} \ldots A_p.$$

Il s'ensuit alors :

$$Df_p(A) : H \longmapsto [Df_p(A)]\,(H) = \sum_{i=1}^{p} A^{i-1} H A^{p-i} \text{ pour tout } A \in \mathcal{M}_n(\mathbb{R}) \ ;$$

$$Df_{-p}(A) : H \longmapsto [Df_{-p}(A)]\,(H) = -\sum_{i=1}^{p} A^{-i} H A^{i-1-p} \text{ pour tout } A \in \Omega.$$

***Exercice I.5.** Soit $A \in \mathcal{S}_n(\mathbb{R})$ et $f : x \in \mathbb{R}^n \setminus \{0\} \longmapsto f(x) := \dfrac{\langle Ax, x\rangle}{\| x \|^2}$. Calculer $\nabla f(x)$ en $x \neq 0$ et le comparer à la projection orthogonale de Ax sur le sous-espace vectoriel H_x orthogonal à x.

Solution : On a aisément $\nabla f(x) = \dfrac{2}{\| x \|^2} [Ax - f(x)x]$.

La décomposition $Ax = \left[Ax - \dfrac{\langle Ax, x\rangle}{\| x \|^2} x\right] + \dfrac{\langle Ax, x\rangle}{\| x \|^2} x$ est la décomposition orthogonale de Ax suivant H_x et $(H_x)^{\perp} = \mathbb{R}x$. Donc, $\nabla f(x)$ est au coefficient multiplicatif $\dfrac{2}{\| x \|^2}$ près la projection orthogonale de Ax sur H_x.

En $\overline{x} \in \mathbb{R}^n \setminus \{0\}$ minimisant (ou maximisant) f sur $\mathbb{R}^n \setminus \{0\}$, on a nécessairement $\nabla f(\overline{x}) = 0$, soit $A\overline{x} = f(\overline{x})\,\overline{x}$; de tels $\overline{x}$ sont donc des vecteurs propres associés à la valeur propre $f(\overline{x})$ (à suivre dans l'Exercice III.4).

***Exercice I.6.** Soit $\mathcal{M}_{m,n}(\mathbb{R})$ structuré en espace euclidien grâce au produit scalaire $\ll M, N \gg := \mathrm{tr}\big(M^\top N\big)$ et soit $B \in \mathcal{S}_n(\mathbb{R})$.

On définit $f, g : \mathcal{M}_{m,n}(\mathbb{R}) \longrightarrow \mathbb{R}$ par

$$f(A) := \mathrm{tr}\left(ABA^\top\right) \text{ et } g(A) := \text{ dét } \left(ABA^\top\right) \text{ respectivement.}$$

Déterminer les différentielles (premières) et les gradients de f et g en tout $A \in \mathcal{M}_{m,n}(\mathbb{R})$.

Solution : f peut être vue comme $f_2 \circ f_1$, où

$$f_1 : A \longmapsto f_1(A) := ABA^\top \text{ et } f_2 : C \longmapsto f_2(C) := \mathrm{tr}\, C.$$

Puisque $f_1(A+H) = f_1(A) + HBA^\top + ABH^\top + HBH^\top$, il vient que f_1 est différentiable en A avec $Df_1(A) : H \to [Df_1(A)](H) = HBA^\top + ABH^\top$. Quant à f_2 qui est linéaire (continue), $Df_2(C) = f_2$. Il s'ensuit :

$$\begin{aligned} Df(A) : H \longmapsto [Df(A)](H) &= \mathrm{tr}\left(HBA^\top + ABH^\top\right) \\ &= 2\ \mathrm{tr}(ABH^\top) \begin{pmatrix} \text{car } \mathrm{tr}(HBA^\top) = \mathrm{tr}(ABH^\top) \\ \text{en raison de la symétrie de B} \end{pmatrix} \\ &= 2 \ll AB, H \gg . \end{aligned}$$

D'où $\nabla f(A) = 2AB$.

Si $f_3 : C \in \mathcal{S}_m(\mathbb{R}) \longmapsto f_3(C) := \text{dét}\, C$, on a $g = f_3 \circ f_1$ par définition même de g. L'application f_3 est différentiable et

$$Df_3(C) : K \to [Df_3(C)](K) = \mathrm{tr}\left((\mathrm{cof}\ C)\,K\right) \ (cf. \text{ Exercice 1.4})$$

Par suite :

$$\begin{aligned} Dg(A) : H \longmapsto [Dg(A)](H) &= 2\ \mathrm{tr}\left(\mathrm{cof}\left(ABA^\top\right) ABH^\top\right) \\ &= 2 \ll \mathrm{cof}\left(ABA^\top\right) AB, H \gg \end{aligned}$$

et $\nabla g(A) = 2\ \mathrm{cof}\big(ABA^\top\big)\, AB$.

En particulier, $\nabla g(A) \cdot A^\top = 2\ \mathrm{cof}\left(ABA^\top\right) ABA^\top = 2\text{dét}\big(ABA^\top\big) \cdot I_m$.

***Exercice I.7.** Soit $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ convexe et différentiable sur $\mathbb{R}^n$. Montrer l'équivalence des trois propriétés suivantes (où L est une constante > 0) :

$(i) \| \nabla f(x) - \nabla f(x') \| \leqslant L \| x - x' \|$ pour tout x,x' (∇f est L-lipschitzienne sur $\mathbb{R}^n$).

$(ii) f(x') - f(x) - \langle \nabla f(x), x' - x \rangle \geqslant \frac{1}{2L} \| \nabla f(x) - \nabla f(x') \|^2$ pour tout x, x'.

$(iii) \langle \nabla f(x) - \nabla f(x'), x - x' \rangle \geqslant \frac{1}{L} \| \nabla f(x) - \nabla f(x') \|^2$ pour tout x, x'.

Indication. Pour $[(i) \Rightarrow (ii)]$ on pourra considérer :

$$f_x : y \longmapsto f(y) - f(x) - \langle \nabla f(x), y - x \rangle.$$

Solution : $[(i) \Rightarrow (ii)]$. Soit $(x, x') \in \mathbb{R}^n \times \mathbb{R}^n$ et montrons que l'inégalité (ii) est vérifiée pour ce couple de points. Considérons

$$f_x : \quad y \in \mathbb{R}^n \longmapsto f_x(y) := f(y) - f(x) - \langle \nabla f(x), y - x \rangle.$$

Il est clair que f_x est convexe avec ∇f_x L-lipschitzienne sur $\mathbb{R}^n$ (puisque $\nabla f_x = \nabla f - \nabla f(x)$). Par conséquent, pour tout y $\in \mathbb{R}^n$,

$$\begin{aligned} f_x(y) &= f_x(x') + \int_0^1 \langle \nabla f_x(x' + t(y - x')), y - x' \rangle dt \\ &= f_x(x') + \langle \nabla f_x(x'), y - x' \rangle + \int_0^1 \langle \nabla f_x(x' + t(y - x')) - \nabla f_x(x'), y - x' \rangle dt \\ &\leqslant f_x(x') + \langle \nabla f_x(x'), y - x' \rangle + \int_0^1 tL \| y - x' \|^2 \, dt \\ &\leqslant f_x(x') + \langle \nabla f_x(x'), y - x' \rangle + \frac{L}{2} \| y - x' \|^2 . \end{aligned} \tag{1.8}$$

Le point x minimise f_x sur $\mathbb{R}^n$. Par suite

$$0 = f_x(x) \leqslant f_x\left(x' - \frac{\nabla f_x(x')}{L} \right).$$

D'après l'inégalité (1.8) dans laquelle on fait $y = x' - \dfrac{\nabla f_x(x')}{L}$

$$(0 \leqslant) \; f_x\left(x' - \frac{\nabla f_x(x')}{L} \right) \leqslant f_x(x') - \frac{1}{L} \langle \nabla f_x(x'), \nabla f_x(x') \rangle + \frac{L}{2} \left\| \frac{\nabla f_x(x')}{L} \right\|^2 ,$$

soit

$$0 \leqslant f_x(x') - \frac{1}{2L} \parallel \nabla f_x(x') \parallel^2$$

ce qui n'est autre que l'inégalité (ii).

$[(ii) \Rightarrow (iii)]$. On écrit (ii) pour (x, x') et (x', x) successivement, et on additionne membre à membre les inégalités pour obtenir (iii).

$[(iii) \Rightarrow (i)]$. Résulte de l'inégalité de Cauchy-Schwarz.

***Exercice I.8.** Soit $A \in \mathcal{M}_n(\mathbb{R})$ inversible, $V \in \mathcal{M}_{m,n}(\mathbb{R})$ et $U \in \mathcal{M}_{n,m}(\mathbb{R})$. On suppose que $I_m - VA^{-1}U$ est inversible.

1°) Montrer que $A - UV$ est alors inversible, avec

$$(A - UV)^{-1} = A^{-1} + A^{-1}U\left(I_m - VA^{-1}U\right)^{-1}VA^{-1}.$$

2°) *Application 1.* Soit $u, v \in \mathbb{R}^n$. Quelle condition assurerait que $A + uv^\top$ est inversible ? Quelle est alors l'inverse de $A + uv^\top$?

3°) *Application 2.* Rappeler pourquoi $A + \varepsilon I_n$ est inversible pour $|\,\varepsilon\,|$ suffisamment petit ; donner alors une expression de $(A + \varepsilon I_n)^{-1}$.

4°) *Application 3.* On remplace dans A la k^{e} colonne $\begin{pmatrix} a_{1k} \\ \vdots \\ a_{nk} \end{pmatrix}$ par $\begin{pmatrix} \tilde{a}_{1k} \\ \vdots \\ \tilde{a}_{nk} \end{pmatrix}$, et on appelle $\tilde{A}$ la nouvelle matrice ainsi obtenue. Exprimer $\tilde{A}^{-1}$ sous la forme $E_k A^{-1}$, où E_k est une matrice que l'on déterminera ; en déduire l'expression des termes $\tilde{a}_{ij}^{(-1)}$ de $\tilde{A}^{-1}$ en fonction de ceux $a_{kl}^{(-1)}$ de A^{-1}.

5°) *Application 4.* On suppose ici que A est de plus symétrique. Soit u et v dans $\mathbb{R}^n$, α et β des réels ; on se propose de déterminer l'inverse de $A + \alpha uu^\top + \beta vv^\top$.

On pose $d := (1 + \alpha \langle A^{-1}u, u\rangle)(1 + \beta \langle A^{-1}v, v\rangle) - \alpha\beta(\langle A^{-1}u, v\rangle)^2$. Montrer que $A + \alpha uu^\top + \beta vv^\top$ est inversible dès lors que $d \neq 0$, avec

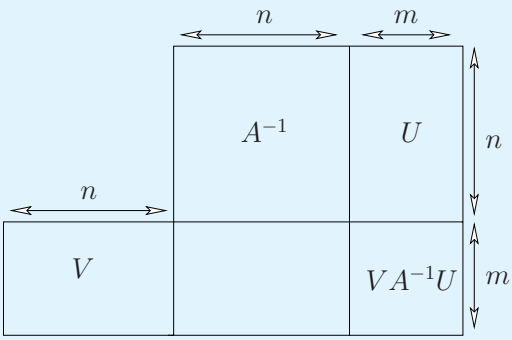
$$\begin{aligned}\left(A + \alpha uu^\top + \beta vv^\top\right)^{-1} = A^{-1} - \frac{1}{d}\Big\{&\alpha(1 + \beta \langle A^{-1}v, v\rangle)A^{-1}uu^\top A^{-1} \\ &+ \beta(1 + \alpha \langle A^{-1}u, u\rangle)A^{-1}vv^\top A^{-1} \\ &-\alpha\beta \langle A^{-1}u, v\rangle \left(A^{-1}vu^\top A^{-1} + A^{-1}uv^\top A^{-1}\right)\Big\}.\end{aligned}$$

Indication. Pour la 4e question, on utilisera le résultat de la 2e question avec

$$u = \begin{pmatrix} a_{1k} - \tilde{a}_{1k} \\ \vdots \\ \\ \vdots \\ \\ \vdots \\ a_{nk} - \tilde{a}_{nk} \end{pmatrix} \quad et \quad v = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow k^{\text{e}} \text{ ligne.}$$

Pour la 5e question, on utilisera le résultat de la 1re question avec $U := [\alpha u \quad \beta v] \in M_{n,2}(\mathbb{R})$ *et* $V := \begin{bmatrix} -u^\top \\ -v^\top \end{bmatrix} \in \mathcal{M}_{2,n}(\mathbb{R})$.

Solution : $UV \in \mathcal{M}_n(\mathbb{R})$, $VA^{-1}U \in \mathcal{M}_m(\mathbb{R})$ par construction.



1°) Puisqu'on propose une expression E de $(A - UV)^{-1}$, il suffit de vérifier que $(A - UV)\,E = I_n$ (ou bien $E\,(A - UV) = I_n$).

On a :

$$\begin{aligned}
(A - UV)\,E &= (A - UV)\left[A^{-1} + A^{-1}U\left(I_m - VA^{-1}U\right)^{-1}VA^{-1}\right] \\
&= AA^{-1} + AA^{-1}U\left(I_m - VA^{-1}U\right)^{-1}VA^{-1} - UVA^{-1} \\
&\quad - UVA^{-1}U\left(I_m - VA^{-1}U\right)^{-1}VA^{-1} \\
&= I_n - UVA^{-1} + U\left(I_m - VA^{-1}U\right)\left(I_m - VA^{-1}U\right)^{-1}VA^{-1} \\
&= I_n - UVA^{-1} + UVA^{-1} = I_n.
\end{aligned}$$

2°) On se place dans le cas où $m = 1$: $-u$ et v sont dans $\mathbb{R}^n \equiv \mathcal{M}_{n,1}(\mathbb{R})$, et la condition « $I_m - VA^{-1}U$ est inversible » revient à « $1 + \langle v, A^{-1}u\rangle \neq 0$ ».

Donc, si $\langle A^{-1}u, v\rangle \neq -1$, la matrice $A + uv^\top$ est inversible avec

$$\left(A + uv^\top\right)^{-1} = A^{-1} - \frac{1}{1 + \langle A^{-1}u, v\rangle} A^{-1}uv^\top A^{-1}. \tag{1.9}$$

Cas particuliers :

$\star$ $A = I_n$: Si $\langle u, v\rangle \neq -1$, $I_n + uv^\top$ est inversible avec

$$\left(I_n + uv^\top\right)^{-1} = I_n - \frac{1}{1 + \langle u, v\rangle} uv^\top. \tag{1.10}$$

$\star$ $A = I_n$, $u = v$: $I_n + uu^\top$ est inversible avec

$$\left(I_n + uu^\top\right)^{-1} = I_n - \frac{1}{1 + \| u \|^2} uu^\top. \tag{1.11}$$

La formule (1.9) donne l'inverse d'une matrice perturbée par une matrice de rang 1 ; on rappelle en liaison avec (1.10) le résultat suivant sur les déterminants :

$$\text{dét}\left(I_n + uv^\top\right) = 1 + \langle u, v\rangle \quad \text{(à revoir si nécessaire)}.$$

3°) L'ensemble Ω des matrices $n \times n$ inversibles est un ouvert de $\mathcal{M}_n(\mathbb{R})$. Puisque $A \in \Omega$, il existe donc $\overline{\varepsilon} > 0$ tel que $A + \varepsilon I_n \in \Omega$ pour $| \varepsilon | \leqslant \overline{\varepsilon}$.

Appliquons le résultat de la 1re question avec $U = \varepsilon I_n$ et $V = -I_n$:

$$(A + \varepsilon I_n)^{-1} = A^{-1} - \varepsilon A^{-1}\left(I_n + \varepsilon A^{-1}\right)^{-1} A^{-1}.$$

4°) En prenant $u = \begin{pmatrix} a_{1k} - \tilde{a}_{1k} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ a_{nk} - \tilde{a}_{nk} \end{pmatrix}$ et $v = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow k^{\text{e}}$ ligne,

on constate que $\tilde{A} = A - uv^\top$ (c'est fait pour !). Appliquons le résultat de la 2e question : si $\left(A^{-1}u\right)_k \neq 1$, la matrice $\tilde{A}$ est inversible et

$$\tilde{A}^{-1} = A^{-1} + \frac{1}{1-\left(A^{-1}u\right)_k} A^{-1} \begin{array}{c} \text{k}^{\text{e}} \text{ colonne} \\ \downarrow \\ \begin{bmatrix} u_1 \\ 0 \ \vdots \ 0 \\ u_n \end{bmatrix} \end{array} A^{-1}$$

$$= \left[\underbrace{I_n + \frac{1}{1-\left(A^{-1}u\right)_k} \begin{bmatrix} \left(A^{-1}u\right)_1 \\ 0 \quad \vdots \quad 0 \\ \left(A^{-1}u\right)_n \end{bmatrix}}_{E_k} \right] A^{-1}.$$

Explicitons $\tilde{A}^{-1} = \left[\tilde{a}_{ij}^{(-1)}\right]$: E_k est de la forme $\begin{array}{c} \text{k}^{\text{e}} \text{ colonne} \\ \downarrow \\ \begin{bmatrix} 1 & 0 & 0 & \vdots & & \\ 0 & 1 & 0 & \vdots & 0 & \\ 0 & 0 & 1 & \vdots & & \\ & & & \vdots & & \\ & & & \vdots & 1 & 0 & 0 \\ & 0 & & \vdots & 0 & 1 & 0 \\ & & & \vdots & 0 & 0 & 1 \end{bmatrix} \end{array}$,

et $\tilde{A}^{-1} = E_k A^{-1}$; d'où :

- $\tilde{a}_{k,j}^{(-1)} = a_{k,j}^{(-1)} \left(\frac{1}{1-\left(A^{-1}u\right)_k} \right) = a_{k,j}^{(-1)} \frac{1}{\sum\limits_{l=1}^{n} a_{kl}^{(-1)} \cdot \tilde{a}_{lk}}$;
- si $i \neq k$, $\tilde{a}_{i,j}^{(-1)} = a_{ij}^{(-1)} - \frac{\sum\limits_{l=1}^{n} a_{il}^{(-1)} \cdot \tilde{a}_{lk}}{\sum\limits_{l=1}^{n} a_{kl}^{(-1)} \cdot \tilde{a}_{lk}} a_{kj}^{(-1)} = a_{ij}^{(-1)} - \left[\sum_{l=1}^{n} a_{il}^{(-1)} \cdot \tilde{a}_{lk} \right] \tilde{a}_{kj}^{(-1)}$.

5°) Avec $U := [\alpha u \ \beta v]$ et $V := \begin{bmatrix} -u^\top \\ -v^\top \end{bmatrix}$, on a

$$I_2 - VA^{-1}U = \begin{bmatrix} 1+\alpha\langle A^{-1}u,u\rangle & \beta\langle A^{-1}u,v\rangle \\ \alpha\langle A^{-1}u,v\rangle & 1+\beta\langle A^{-1}v,v\rangle \end{bmatrix} \quad (\in \mathcal{M}_2(\mathbb{R})).$$

Cette matrice est inversible si et seulement si

$$\begin{aligned} d :&= \det\left(I_2 - VA^{-1}U\right) \\ &= (1+\alpha\langle A^{-1}u,u\rangle)(1+\beta\langle A^{-1}v,v\rangle) - \alpha\beta(\langle A^{-1}u,v\rangle)^2 \neq 0. \end{aligned}$$

Dans ce cas

$$\left[I_2 - VA^{-1}U\right]^{-1} = \frac{1}{d}\begin{bmatrix} 1+\beta\langle A^{-1}v,v\rangle & -\beta\langle A^{-1}u,v\rangle \\ -\alpha\langle A^{-1}u,v\rangle & 1+\alpha\langle A^{-1}u,u\rangle \end{bmatrix}.$$

L'application de la formule démontrée à la 1re question conduit alors – après quelques calculs, certes – à la formule annoncée.

****Exercice I.9.** *Inégalité de Kantorovitch*

Soit A symétrique définie positive. Montrer que pour tout $x \in \mathbb{R}^n$

$$\| x \|^4 \leqslant \langle Ax, x\rangle \cdot \langle A^{-1}x, x\rangle \leqslant \frac{1}{4}\left(\sqrt{\frac{\lambda_1}{\lambda_n}} + \sqrt{\frac{\lambda_n}{\lambda_1}}\right)^2 \| x \|^4, \qquad (1.12)$$

où λ_1 et λ_n désignent respectivement la plus grande et la plus petite valeur propre de A.

Indication. Il y a intérêt à diagonaliser A (et donc A^{-1}). On utilisera ensuite les propriétés de convexité de la fonction $x > 0 \longmapsto 1/x$.

Solution : Par homogénéité, il suffit de démontrer l'inégalité (1.12) pour $\| x \| = 1$.

Rangeons les valeurs propres de A par ordre décroissant : $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n$, et considérons une matrice orthogonale P diagonalisant A :

$$A = {}^tP\Delta P, \quad \text{avec} \quad \Delta := \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n).$$

$A^{-1} = ({}^tP\Delta P)^{-1} = {}^tP\Delta^{-1}P$, ou $\Delta^{-1} = \mathrm{diag}(1/\lambda_1, 1/\lambda_2, \ldots, 1/\lambda_n)$;
$\langle Ax, x\rangle = \langle {}^tP\Delta Px, x\rangle = \langle \Delta(Px), Px\rangle$;
$\langle A^{-1}x, x\rangle = \langle \Delta^{-1}(Px), Px\rangle$.

La transformation $P : x \in S := \{x \mid \ \| x \| = 1\} \longmapsto y = Px \in S$ est une bijection de la sphère-unité euclidienne sur elle-même. Opérons alors le changement de variables $y = Px$. Il faut donc démontrer que pour tout vecteur unitaire y de $\mathbb{R}^n$,

$$1 \leqslant \langle \triangle y, y\rangle \cdot \langle \triangle^{-1} y, y\rangle \leqslant \frac{1}{4}\left(\sqrt{\frac{\lambda_1}{\lambda_n}} + \sqrt{\frac{\lambda_n}{\lambda_1}}\right)^2 . \tag{1.13}$$

Considérons pour cela le graphe (et ce qui est au-dessus, appelé l'épigraphe) de la fonction $\theta : x > 0 \longmapsto 1/x$. (Il est supposé que $\lambda_n < \lambda_1$, sinon il n'y a rien à démontrer.)

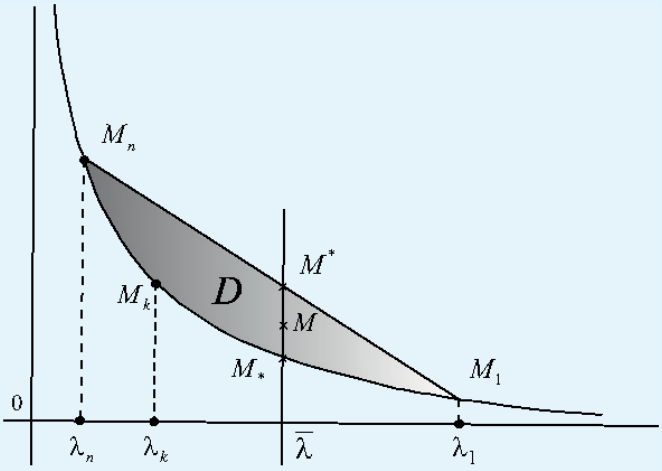


Figure 2.

Étant donné $y = (y_1, ..., y_n) \in \mathbb{R}^n$ de norme 1, posons $\alpha_i := y_i^2$. Ainsi $\langle \triangle y, y\rangle = \sum_{i=1}^{n} y_i^2 \lambda_i$ (resp. $\langle \triangle^{-1} y, y\rangle = \sum_{i=1}^{n} y_i^2 \frac{1}{\lambda_i}$) est une combinaison convexe des λ_i (resp. des $1/\lambda_i$).

Pour $k = 1, 2, ..., n$, soit M_k le point du graphe de θ de coordonnées λ_k et $1/\lambda_k$.

Le point M, barycentre des points M_k affectés des coefficients α_i, est dans le domaine convexe $\mathcal{D}$ délimité par le graphe de θ et la droite M_1M_n (et même dans le polygone convexe de sommets M_1, M_2,..., M_n). Deux points *de même abscisse que* M (abscisse notée $\overline{\lambda}$) jouent un rôle particulier :

– le premier, M_*, est le point correspondant du graphe de θ;

– le second, M^*, est le point correspondant sur la droite M_1M_n.

Ainsi :

M est d'ordonnée $\displaystyle\sum_{i=1}^{n} \alpha_i \frac{1}{\lambda_i}$; M_* est d'ordonnée $\displaystyle\frac{1}{\sum_{i=1}^{n} \alpha_i \lambda_i} = \frac{1}{\overline{\lambda}}$;

M^* est d'ordonnée $\displaystyle y(\overline{\lambda}) = \frac{1}{\lambda_1} + \frac{1}{\lambda_n} - \frac{\overline{\lambda}}{\lambda_1 \lambda_n}$ (par un calcul algébrique facile).

Écrire que l'ordonnée de M est supérieure à celle de M_* conduit à

$$\sum_{i=1}^{n} \alpha_i \frac{1}{\lambda_i} \geqslant \frac{1}{\overline{\lambda}} \text{ , soit } \langle \triangle^{-1} y, y\rangle \cdot \langle \triangle y, y\rangle \geqslant 1.$$

De même, écrivons que l'ordonnée de M^* est supérieure à celle de M :

$$\sum_{i=1}^{n} \alpha_i \frac{1}{\lambda_i} \leqslant y(\overline{\lambda}),$$

d'où

$$\left(\sum_{i=1}^{n} \alpha_i \lambda_i\right)\left(\sum_{i=1}^{n} \alpha_i \frac{1}{\lambda_i}\right) \leqslant \overline{\lambda}\ y\left(\overline{\lambda}\right) = \frac{\overline{\lambda}\left(\lambda_1 + \lambda_n - \overline{\lambda}\right)}{\lambda_1 \lambda_n}.$$

Comme la fonction $u \longmapsto \dfrac{u\left(\lambda_1 + \lambda_n - u\right)}{\lambda_1 \lambda_n}$ atteint son maximum pour $u = \dfrac{\lambda_1 + \lambda_n}{2}$, l'expression $\overline{\lambda}\ y\left(\overline{\lambda}\right)$ est majorée par $\left(\dfrac{\lambda_1 + \lambda_n}{2}\right)^2 \dfrac{1}{\lambda_1 \lambda_n}$, soit encore

$$\frac{1}{4}\left(\frac{\lambda_1}{\lambda_n} + \frac{\lambda_n}{\lambda_1} + 2\right) = \frac{1}{4}\left(\sqrt{\frac{\lambda_1}{\lambda_n}} + \sqrt{\frac{\lambda_n}{\lambda_1}}\right)^2.$$

La deuxième inégalité dans (1.13) est donc démontrée.

***Exercice I.10.** *Géométrie de l'ensemble des matrices symétriques semi-définies positives*

On part de $\mathcal{S}_n(\mathbb{R})$, $n \geqslant 2$, structuré en espace euclidien à l'aide du produit scalaire fondamental $(A, B) \longmapsto \ll A, B \gg := \mathrm{tr}(AB)$. On désigne par $\mathcal{P}_n(\mathbb{R})$ (resp. $\overset{\circ}{\mathcal{P}}_n(\mathbb{R})$) l'ensemble de $A \in \mathcal{S}_n(\mathbb{R})$ qui sont semi-définies positives (resp. définies positives).

1°) Montrer que $\mathcal{P}_n(\mathbb{R})$ est un cône convexe fermé de $\mathcal{S}_n(\mathbb{R})$ et que l'intérieur de $\mathcal{P}_n(\mathbb{R})$ est exactement $\overset{\circ}{\mathcal{P}}_n(\mathbb{R})$.

2°) Montrer que le cône polaire $[\mathcal{P}_n(\mathbb{R})]^\circ$ de $\mathcal{P}_n(\mathbb{R})$ n'est autre que $-\mathcal{P}_n(\mathbb{R})$.

3°) On suppose ici que $n = 2$.

a) Rappeler quelles conditions sur les réels a, b, et c sont nécessaires et suffisantes pour que $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$ soit semi-définie positive (resp. définie positive).

b) On définit $\varphi : \mathcal{S}_2(\mathbb{R}) \longrightarrow \mathbb{R}^3$ par

$$A = \begin{bmatrix} a & b \\ b & c \end{bmatrix} \longmapsto \varphi(A) := (a, b, c).$$

Dans $\mathbb{R}^3$ rapporté à un repère orthonormé $(O; \vec{i}, \vec{j}, \vec{k})$, représenter $\varphi\left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\right)$, $\varphi\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$, $\varphi\left(\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\right)$ et $\varphi\left(\mathcal{P}_2(\mathbb{R})\right)$.

c) Montrer que toutes les matrices non nulles $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$ se trouvant sur la frontière de $\mathcal{P}_2(\mathbb{R})$ sont de forme $xx^\top$, où x est un élément non nul de $\mathbb{R}^2$. En déduire une description des $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$ se trouvant sur cette frontière à l'aide de conditions sur les coefficients a, b, c de A.

Indication. Parmi les matrices de $\mathcal{P}_n(\mathbb{R})$, il y a celles de la forme $A = xx^\top$, avec $x \in \mathbb{R}^n \setminus \{0\}$. Pour ces matrices A et un $B \in \mathcal{S}_n(\mathbb{R})$, on note que $\ll A, B \gg = \langle Bx, x \rangle$.

De plus, si $S \in \mathcal{S}_n(\mathbb{R})$, une décomposition spectrale de S est $S = \sum_{i=1}^{n} \lambda_i x_i x_i^\top$, où les λ_i sont les valeurs propres de A, $x_i \in \mathbb{R}^n$ est un vecteur propre unitaire associé à λ_i.

Solution : 1°) Soit $A, B \in \mathcal{P}_n(\mathbb{R})$ et $\alpha \in [0,1]$. On a :

$$\forall x \in \mathbb{R}^n, \quad \langle(\alpha A + (1-\alpha)B)x, x\rangle = \alpha\langle Ax, x\rangle + (1-\alpha)\langle Bx, x\rangle \geqslant 0,$$

ce qui implique $\alpha A + (1-\alpha)B \in \mathcal{P}_n(\mathbb{R})$.

De même, il est immédiat de constater que $\alpha A \in \mathcal{P}_n(\mathbb{R})$ lorsque $A \in \mathcal{P}_n(\mathbb{R})$ et $\alpha > 0$. Donc $\mathcal{P}_n(\mathbb{R})$ est bien un cône convexe de $\mathcal{S}_n(\mathbb{R})$.

Soit $\{A_k\}$ une suite d'éléments de $\mathcal{P}_n(\mathbb{R})$ convergeant vers A. Outre le fait – clair – que $A \in \mathcal{S}_n(\mathbb{R})$, l'inégalité

$$\langle A_k x, x\rangle \geqslant 0 \quad \text{pour tout} \quad x \in \mathbb{R}^n,$$

induit par passage à la limite sur k : $\langle Ax, x\rangle \geqslant 0$ pour tout $x \in \mathbb{R}^n$. Par conséquent $A \in \mathcal{P}_n(\mathbb{R})$. Et $\mathcal{P}_n(\mathbb{R})$ est bien fermé dans $\mathcal{S}_n(\mathbb{R})$.

Soit $A \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ et $\lambda_n > 0$ la plus petite valeur propre de A. Rappelons à cet égard l'inégalité suivante (que l'on reverra dans l'Exercice 3.4) : $\langle Ax, x\rangle \geqslant \lambda_n \parallel x \parallel^2$ pour tout $x \in \mathbb{R}^n$. Soit à présent $M \in \mathcal{S}_n(\mathbb{R})$. Puisque

$$\parallel M \parallel := \sqrt{\ll M, M \gg} = \left(\sum_{i=1}^n \mu_i^2\right)^{1/2} \qquad (\mu_1 \geqslant \ldots \geqslant \mu_n, \text{ valeurs propres de } M),$$

il suffit de prendre $\parallel M \parallel \leqslant \lambda_n$ pour être sûr d'avoir

$$\langle(A+M)x, x\rangle \geqslant (\lambda_n + \mu_n) \parallel x \parallel^2 \geqslant 0 \quad \text{pour tout} \quad x \in \mathbb{R}^n,$$

soit $A + M \in \mathcal{P}_n(\mathbb{R})$. Donc A est bien à l'intérieur de $\mathcal{P}_n(\mathbb{R})$.

Réciproquement, soit A à l'intérieur de $\mathcal{P}_n(\mathbb{R})$. Il existe alors $\varepsilon > 0$ assez petit tel que $A - \varepsilon I_n \in \mathcal{P}_n(\mathbb{R})$. En conséquence, l'inégalité

$$\langle(A - \varepsilon I_n)x, x\rangle \geqslant 0 \quad \text{pour tout } x \in \mathbb{R}^n$$

induit

$$\langle Ax, x\rangle \geqslant \varepsilon \parallel x \parallel^2 \quad \text{pour tout } x \in \mathbb{R}^n,$$

soit $A \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R})$.

Le résultat de cette 1re question explique la notation $\overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ utilisée pour l'ensemble des matrices symétriques définies positives.

La frontière de $\mathcal{P}_n(\mathbb{R})$ est donc constituée des matrices semi-définies positives qui sont singulières ; parmi celles-là figurent les matrices de rang 1, c'est-à-dire du type $xx^\top$ avec $x \neq 0$.

2°) Soit $B \in \mathcal{S}_n(\mathbb{R})$ dans le cône polaire de $\mathcal{P}_n(\mathbb{R})$. Puisque $\ll B, A \gg \leqslant 0$ pour tout $A \in \mathcal{P}_n(\mathbb{R})$, en particulier $\ll B, xx^\top \gg = \langle Bx, x\rangle \leqslant 0$ pour tout $x \in \mathbb{R}^n$; donc B est semi-définie négative.

Réciproquement, soit B semi-définie négative et $A \in \mathcal{P}_n(\mathbb{R})$. En décomposant A sous la forme $\sum_{i=1}^{n} \lambda_i x_i x_i^\top$ (avec $\lambda_1 \geqslant \cdots \geqslant \lambda_n \geqslant 0$, valeurs propres de A), on a

$$\ll B, A \gg = \ll B, \sum_{i=1}^{n} \lambda_i x_i x_i^\top \gg = \sum_{i=1}^{n} \lambda_i \langle Bx_i, x_i \rangle \leqslant 0.$$

D'où $B \in [\ \mathcal{P}_n(\mathbb{R})]^\circ$.

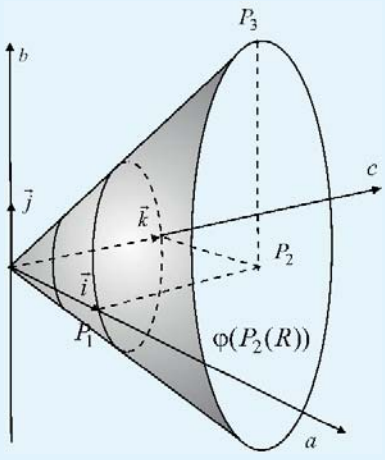
Ce résultat $[\mathcal{P}_n(\mathbb{R})]^\circ = -\mathcal{P}_n(\mathbb{R})$ apparaît ainsi comme la généralisation de la relation de polarité $(\mathbb{R}^n_+)^\circ = -\mathbb{R}^n_+$ dans l'espace euclidien standard $(\mathbb{R}^n, \langle \cdot, \cdot \rangle)$.

3°) a)

$$\left(A = \begin{bmatrix} a & b \\ b & c \end{bmatrix} \in \overset{\circ}{\mathcal{P}}_2(\mathbb{R}) \right) \Leftrightarrow (a > 0 \ \text{ et } \ ac - b^2 > 0)\ ;$$

$$\left(A = \begin{bmatrix} a & b \\ b & c \end{bmatrix} \in \mathcal{P}_2(\mathbb{R}) \right) \Leftrightarrow \begin{pmatrix} a \geqslant 0,\ c \geqslant 0 \\ \text{et} \ \ ac - b^2 \geqslant 0 \end{pmatrix}.$$

b)



$$P_1 = \varphi\left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \right)$$
$$P_2 = \varphi\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$
$$P_3 = \varphi\left(\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \right)$$

FIGURE 3.

Remarquons que le produit scalaire $\ll A, B \gg$ n'est pas le produit scalaire usuel de $\varphi(A)$ et $\varphi(B)$ dans $\mathbb{R}^3$; en effet

$$\ll \begin{bmatrix} a & b \\ b & c \end{bmatrix}, \begin{bmatrix} a' & b' \\ b' & c' \end{bmatrix} \gg = aa' + 2bb' + cc',$$

tandis que

$$\left\langle \varphi\left(\begin{bmatrix} a & b \\ b & c \end{bmatrix}\right), \varphi\left(\begin{bmatrix} a' & b' \\ b' & c' \end{bmatrix}\right)\right\rangle = aa' + bb' + cc'.$$

c) Les matrices de la frontière de $\mathcal{P}_2(\mathbb{R})$ sont les matrices semi-définies positives singulières ; mise à part la matrice nulle, il s'agit donc de matrices de rang 1, *i.e.* du type

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \cdot [x_1 \ x_2] = \begin{bmatrix} x_1^2 & x_1x_2 \\ x_1x_2 & x_2^2 \end{bmatrix}$$

où $x = (x_1, x_2)$ est un élément non nul de $\mathbb{R}^2$.

En définitive, la frontière de $\mathcal{P}_2(\mathbb{R})$ peut être décrite comme

$$\left\{ \begin{bmatrix} a & \pm\sqrt{ac} \\ \pm\sqrt{ac} & c \end{bmatrix}, \quad a \geqslant 0 \quad \text{et} \quad c \geqslant 0 \right\}.$$

Remarque : On aurait pu utiliser

$$\psi : A := \begin{bmatrix} a & b \\ b & c \end{bmatrix} \in \mathcal{S}_2(\mathbb{R}) \longmapsto \psi(A) := (a, \sqrt{2}\, b, c) \in \mathbb{R}^3$$

qui a l'avantage de conserver les produits scalaires usuels dans $\mathcal{S}_2(\mathbb{R})$ et $\mathbb{R}^3$, ou mieux, $\Psi : A = \begin{bmatrix} a & b \\ b & c \end{bmatrix} \longmapsto \Psi(A) := \frac{1}{\sqrt{2}}(2b, c-a, c+a)$ qui permet de représenter $\Psi(\mathcal{P}_2(\mathbb{R}))$ par le cône de $\mathbb{R}^3$ d'équation $w \geqslant \sqrt{u^2+v^2}$ et les éléments de $\Psi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \cdot [x_1 \ x_2]\right) = \frac{1}{\sqrt{2}} \begin{pmatrix} 2x_1x_2 \\ x_2^2 - x_1^2 \\ x_2^2 + x_1^2 \end{pmatrix}$ par la frontière de ce cône, d'équation : $w^2 = u^2 + v^2, \ w \geqslant 0$.

***Exercice I.11.** Soit $A \in \mathcal{M}_n(\mathbb{R})$ symétrique et inversible, soit H un sous-espace vectoriel de $\mathbb{R}^n$. Montrer qu'une condition nécessaire et suffisante pour que A soit définie positive est :

$$(\mathcal{C}) \begin{cases} \langle Ax, x\rangle > 0 \ \text{ pour tout } x \in H \setminus \{0\} \\ \text{et} \\ \langle A^{-1}x, x\rangle > 0 \ \text{ pour tout } x \in H^\perp \setminus \{0\}. \end{cases}$$

Solution : La nécessité de $(\mathcal{C})$ est immédiate puisque A^{-1} est (symétrique) définie positive dès que A l'est.

La définie positivité de A sous la condition $(\mathcal{C})$ sera démontrée en deux étapes.

– Montrons tout d'abord que $\mathbb{R}^n = H \oplus A^{-1}(H^\perp)$.

On a $\dim H + \dim A^{-1}(H^\perp) = \dim H + \dim H^\perp = n$. Il suffit donc de démontrer que $H \cap A^{-1}(H^\perp) = \{0\}$. Prenons pour cela $x \in H \cap A^{-1}(H^\perp)$. Comme $Ax \in H^\perp$, $\langle Ax, y\rangle = 0$ pour tout $y \in H$. Ainsi : $x \in H$ et $\langle Ax, x\rangle = 0$. Or, d'après le premier volet de la condition $(\mathcal{C})$, ceci n'est possible que si $x = 0$.

– Montrons à présent que

$$\langle A(x+y), x+y\rangle \geqslant 0 \text{ pour tout } x \in H \text{ et tout } y \in A^{-1}(H^\perp),$$

ce qui assurera la semi-définie positivité de A.

Soit donc $x \in H$ et $y \in A^{-1}(H^\perp)$. Alors

$$\langle A(x+y), x+y\rangle = \langle Ax, x\rangle + 2\langle x, Ay\rangle + \langle Ay, y\rangle,$$

avec :

$$\langle Ax, x\rangle \geqslant 0 \text{ puisque } x \in H,$$
$$\langle x, Ay\rangle = 0 \text{ puisque } x \in H \text{ et } Ay \in H^\perp,$$

et

$$\langle Ay, y\rangle = \langle x', A^{-1}x'\rangle \text{ puisque } x' := Ay \in H^\perp.$$

En définitive, $\langle A(x+y), x+y\rangle \geqslant 0$.

Commentaire : Il y a en fait une caractérisation de « A définie positive » plus générale que $(\mathcal{C})$, avec un cône convexe fermé K au lieu d'un sous-espace vectoriel H : dans $(\mathcal{C})$ on remplace H par K et $H^\perp$ par le cône polaire K°. Mais la démonstration est plus difficile. On retiendra néanmoins la belle symétrie du résultat puisque $(A^{-1})^{-1} = A$ et $(K^\circ)^\circ = K$.

***Exercice I.12.** Soit $A = [a_{ij}] \in \mathcal{S}_n(\mathbb{R})$. Rassemblez toutes les caractérisations que vous connaissez de :

« A est définie positive », et de « A est semi-définie positive ».

Solution : Dans $\mathcal{S}_n(\mathbb{R})$ structuré en espace euclidien grâce au produit scalaire $\ll A, B \gg := \mathrm{tr}(AB)$, on considère :

$$\overset{\circ}{\mathcal{P}}_n(\mathbb{R}) := \text{ensemble des matrices définies positives,}$$
$$\mathcal{P}_n(\mathbb{R}) := \text{ensemble des matrices semi-définies positives.}$$

Nous serons aussi amenés à considérer la forme quadratique q_A sur $\mathbb{R}^n$ associée à A, à savoir : $x \in \mathbb{R}^n \longmapsto q_A(x) = \langle Ax, x\rangle$.

* Propriétés de $\mathcal{P}_n(\mathbb{R})$ et $\overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ (*cf.* Exercice I.10) :

– $\mathcal{P}_n(\mathbb{R})$ est un cône convexe fermé de $\mathcal{S}_n(\mathbb{R})$, dont l'intérieur est $\overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ précisément ($\overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ est donc un cône convexe ouvert de $\mathcal{S}_n(\mathbb{R})$).

– Le cône polaire de $\mathcal{P}_n(\mathbb{R})$ est exactement $-\mathcal{P}_n(\mathbb{R})$ (le cône des matrices semi-définies négatives), c'est-à-dire :

$$(\ll B, A \gg \leqslant 0 \text{ pour tout } A \in \mathcal{P}_n(\mathbb{R})) \Leftrightarrow (-B \in \mathcal{P}_n(\mathbb{R}))\,.$$

Cela se voit facilement en utilisant, pour $S \in \mathcal{S}_n(\mathbb{R})$, une décomposition spectrale de S :

$$S = \sum_{i=1}^{n} \lambda_i x_i x_i^\top$$

(où les λ_i sont les valeurs propres de S, $x_i \in \mathbb{R}^n$ est un vecteur propre unitaire associé à λ_i) et en observant que $\ll S, yy^\top \gg = \langle Sy, y\rangle$.

Sur la frontière du cône $\mathcal{P}_n(\mathbb{R})$ se trouvent donc toutes les matrices semi-définies positives qui sont singulières ; il y a

– la matrice nulle (c'est la pointe du cône) ;

– les matrices de rang 1, *i.e.* celles de la forme $xx^\top$ avec $x \neq 0$ (ce sont les génératrices extrémales du cône $\mathcal{P}_n(\mathbb{R})$) ;

– les matrices de rang 2, 3, ..., $n-1$.

Lorsque $n = 2$, on peut visualiser $\mathcal{P}_2(\mathbb{R})$ dans $\mathcal{S}_2(\mathbb{R}) \equiv \mathbb{R}^3$ (*cf.* Exercice I.10) mais ceci est très particulier dans la mesure où la frontière de $\mathcal{S}_2(\mathbb{R})$, privée de l'origine, est « lisse » (n'y figurent que des génératrices extrémales).

* *Caractérisations diverses de « $A \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ » :*

– Toutes les valeurs propres λ_i de A sont strictement positives.

– $\det A_k > 0$ pour tout $k = 1, \ldots, n$, où $A_k = [a_{ij}]_{\substack{1 \leqslant i \leqslant k \\ 1 \leqslant j \leqslant k}}$.

– Il existe B inversible telle que $A = BB^\top$ (auquel cas $\langle Ax, x\rangle = \| Bx \|^2$).

Il y a différentes factorisations possibles de A sous cette forme :

- la factorisation de Cholesky, où B est triangulaire inférieure avec éléments diagonaux strictement positifs ;
- en prenant B symétrique définie positive (B est dans ce cas la racine carrée positive de A, notée $A^{1/2}$).

– Les invariants principaux $i_k(A)$ de A, $k = 1, \dots, n$, sont strictement positifs.

Si P_A est le polynôme caractéristique de A, on a :

$$P_A = (-X)^n + i_1(A)(-X)^{n-1} + \cdots + i_k(A)(-X)^{n-k} + \cdots + i_n(A).$$

Rappelons que

$$i_k(A) = \sum_{1 \leqslant i_1 < \cdots < i_k \leqslant n} \lambda_{i_1} \lambda_{i_2} \dots \lambda_{i_k};$$

par exemple $i_1(A) = \mathrm{tr}(A)$ et $i_n(A) = \det A$.

– Une caractérisation lorsque A est inversible. Soit K un cône convexe fermé de $\mathbb{R}^n$ et K° son cône polaire ; alors

$$\left(A \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R}) \right) \Leftrightarrow \begin{pmatrix} \langle Ax, x \rangle > 0 \ \text{ pour tout } \ x \in K \setminus \{0\} \ \text{ et} \\ \langle A^{-1}x, x \rangle > 0 \ \text{ pour tout } \ x \in K^\circ \setminus \{0\} \end{pmatrix}$$

(*cf.* Exercice I.11 pour une démonstration dans le cas plus simple où K est un sous-espace vectoriel de $\mathbb{R}^n$).

Il y a d'autres tests de dépistage du caractère défini positif de $A \in \mathcal{S}_n(\mathbb{R})$, mais ils sont moins utiles dans la pratique ; à titre d'exemple :

$$\left(A \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R}) \right) \Leftrightarrow (\alpha A + (1-\alpha) I_n \text{ est inversible pour tout } \alpha \in]0,1]) ;$$

$$\left(A \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R}) \right) \Leftrightarrow (\text{Il existe } B \in \mathcal{S}_n(\mathbb{R}) \text{ telle que } A = \exp B) .$$

* *Caractérisations diverses de « $A \in \mathcal{P}_n(\mathbb{R})$ » :*

– Toutes les valeurs propres de A sont positives.

– Il existe B telle que $A = BB^\top$ (alors la forme quadratique $q_A(x) = \langle Ax, x \rangle$ est minimisée (nulle en fait) sur le noyau de B).

– *Tous* les mineurs principaux d'ordre k de A, $k = 1, \dots, n$, sont positifs (et pas seulement les $\det(A_k), k = 1, \dots, n$). On rappelle qu'un mineur principal d'ordre k de A est le déterminant d'une matrice (k,k) extraite de A en retenant les lignes et colonnes de numéros $n_1, n_2, \dots, n_k$, où $1 \leqslant n_1 < n_2 < \cdots < n_k < \cdots \leqslant n$.

Considérons par exemple

$$A = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \in \mathcal{S}_3(\mathbb{R}).$$

Le calcul des mineurs principaux de A se décompose en :

– 3 déterminants de matrices (1,1) (tous égaux à 2 ici),

– 3 déterminants de matrices (2,2) (tous égaux à 3 ici),

– 1 déterminant de matrice (3,3) (ici dét $A = 0$).

Donc A est semi-définie positive.

Considérons à présent $A = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \in \mathcal{S}_2(\mathbb{R})$. Bien que dét $A_1 \geqslant 0$ et dét$A_2 \geqslant 0$, la matrice A n'est pas semi-définie positive (elle est même semi-définie négative).

* *Propriétés de la forme quadratique q_A :*

– $(A \in \mathcal{P}_n(\mathbb{R})) \quad \Leftrightarrow (q_A$ est une fonction convexe sur $\mathbb{R}^n)$

– $\left(A \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R})\right) \Leftrightarrow (q_A$ est strictement convexe sur $\mathbb{R}^n)$

$\Leftrightarrow (q_A$ est fortement convexe sur $\mathbb{R}^n)$.

***Exercice I.13.** Soit $\overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ l'ensemble des $A \in \mathcal{S}_n(\mathbb{R})$ qui sont définies positives (c'est un ouvert convexe de $\mathcal{S}_n(\mathbb{R})$). Soit

$$\begin{aligned} f : \overset{\circ}{\mathcal{P}}_n(\mathbb{R}) &\longrightarrow \mathbb{R} \\ A &\longmapsto f(A) := \ln(\text{dét } A^{-1}). \end{aligned}$$

On se propose de montrer que f est strictement convexe sur $\overset{\circ}{\mathcal{P}}_n(\mathbb{R})$. Pour cela, on prend $X_0 \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R})$, $H \in \mathcal{S}_n(\mathbb{R})$, et on considère la fonction φ de la variable réelle définie par

$$\varphi(t) := \ln(\text{dét}(X_0 + tH)^{-1}).$$

1°) Vérifier que φ est définie sur un intervalle ouvert (de $\mathbb{R}$) contenant l'origine ; on notera $I_{X_0,H}$ cet intervalle.

2°) Montrer que

$$\varphi(t) - \varphi(0) = -\sum_{i=1}^{n} \ln(1 + t\lambda_i) \text{ pour tout } t \in I_{X_0,H},$$

où les λ_i désignent les valeurs propres de $X_0^{-1/2} H X_0^{-1/2}$.

3°) Déduire de ce qui précède la stricte convexité de f.

Solution : 1°) $\overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ étant ouvert et convexe,

$$I_{X_0,H} := \{t \in \mathbb{R} \mid X_0 + tH \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R})\}$$

est également ouvert et convexe : c'est l'image inverse de $\overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ par l'application affine (continue) $t \longmapsto X_0 + tH$.

2°) De :
$$\begin{aligned} X_0 + tH &= X_0^{1/2}\left(X_0^{1/2} + tX_0^{-1/2} H\right) \\ &= X_0^{1/2}\left(I_n + tX_0^{-1/2} H X_0^{-1/2}\right) X_0^{1/2}, \end{aligned}$$

on tire :

$$(X_0 + tH)^{-1} = X_0^{-1/2}\left(I_n + tX_0^{-1/2} H X_0^{-1/2}\right)^{-1} X_0^{-1/2},$$

d'où :

$$\ln\left(\det (X_0 + tH)^{-1}\right) = \ln(\det X_0^{-1}) + \ln\left(\det\left(I_n + tX_0^{-1/2} H X_0^{-1/2}\right)^{-1}\right).$$

Si $\{\lambda_i \mid i = 1, \ldots, n\}$ est le spectre (entièrement réel) de la matrice symétrique $X_0^{-1/2} H X_0^{-1/2}$, alors $\{1 + t\lambda_i \mid i = 1, \ldots, n\}$ (resp. $\{(1 + t\lambda_i)^{-1} \mid i = 1, \ldots, n\}$) est le spectre de $I_n + tX_0^{-1/2} H X_0^{-1/2}$ (resp. de $\left(I_n + tX_0^{-1/2} H X_0^{-1/2}\right)^{-1}$). En conséquence,

$$\varphi(t) = \varphi(0) - \sum_{i=1}^{n} \ln(1 + t\lambda_i).$$

3°) φ est strictement convexe sur $I_{X_0,H}$ car $\varphi'' > 0$ sur $I_{X_0,H}$. Par suite, la fonction f dont toute « trace » φ sur une droite passant par $X_0 \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ et dirigée par $H \in \mathcal{S}_n(\mathbb{R})$ est strictement convexe, est elle-même strictement convexe.

Commentaire :

– La propriété démontrée dans l'exercice se traduit par : si A et B sont (symétriques) définies positives et différentes, et si $\alpha \in]0,1[$ alors

$$\text{dét}(\alpha A + (1-\alpha)B) > (\text{dét}A)^{\alpha}(\text{dét}B)^{1-\alpha}.$$

– Autre exemple de conséquence : $\{A \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R}) \mid \text{dét}A \geqslant 1\}$ est un ensemble convexe (ce qui ne saute pas aux yeux...).

– La propriété démontrée dans cet exercice est *essentielle* ; on y reviendra maintes et maintes fois.

***Exercice I.14.** Soit $f : \mathcal{S}_n(\mathbb{R}) \longrightarrow \mathbb{R}$
$A \longmapsto f(A) := \text{ dét } A.$

1°) Rappeler ce qu'est la différentielle de f en $A \in \mathcal{S}_n(\mathbb{R})$.

2°) Soit $g : \overset{\circ}{\mathcal{P}}_n(\mathbb{R}) \longrightarrow \mathbb{R}$
$A \longmapsto g(A) := \ln(\text{dét } A).$

Sachant que g est deux fois différentiable sur $\overset{\circ}{\mathcal{P}}_n(\mathbb{R})$, déterminer le plus économiquement possible la différentielle seconde de g en $A \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R})$.

Solution : 1°) $Df(A) : H \in \mathcal{S}_n(\mathbb{R}) \longmapsto \ll \text{cof}\, A, H \gg$; revoir à ce sujet l'Exercice 1.4 si nécessaire.

2°) $Dg(A) : H \in \mathcal{S}_n(\mathbb{R}) \longmapsto \ll A^{-1}, H \gg (= \text{tr}(A^{-1}H))$.

On sait que $D^2g(A) : \mathcal{S}_n(\mathbb{R}) \times \mathcal{S}_n(\mathbb{R}) \longrightarrow \mathbb{R}$
$(H,K) \longmapsto D^2g(A)(H,K)$
est une forme bilinéaire symétrique sur $\mathcal{S}_n(\mathbb{R})$. La voie la plus économique pour déterminer $D^2g(A)(H,K)$ est la suivante : $D^2g(A)(H,K)$ est la différentielle (première) en A de l'application $X \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R}) \longmapsto Dg(X)H$, appliquée à K. Dans le cas présent :

$$D^2g(A)(H,K) = \ll -A^{-1}KA^{-1}, H \gg = -\text{tr}(A^{-1}KA^{-1}H).$$

Il s'ensuit notamment (par un développement de Taylor-Young à l'ordre deux de g en $A \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R})$) : pour tout $H \in \mathcal{S}_n(\mathbb{R})$,

$$\ln(\text{dét}\,(A+H)) = \ln(\text{dét}\,(A)) + \text{tr}(A^{-1}H) - \text{tr}(A^{-1}HA^{-1}H) + \| H \|^2 \varepsilon(H),$$

où $\varepsilon(H) \longrightarrow 0$ quand $H \longrightarrow 0$.

***Exercice I.15.** Soit $\mathcal{O}$ un ouvert convexe de $\mathbb{R}^n$ et $f : \mathcal{O} \to \mathbb{R}_*^+$. On dit que f est *logarithmiquement convexe* (sur $\mathcal{O}$) lorsque la fonction $\ln f : \mathcal{O} \to \mathbb{R}$ est convexe.

1°) a) Montrer que si $\varphi : \mathcal{O} \to \mathbb{R}$ est convexe, il en est de même de la fonction $\exp \varphi$.

b) En déduire qu'une fonction logarithmiquement convexe est nécessairement convexe.

2°) a) On suppose ici que f est deux fois différentiable sur $\mathcal{O}$. Montrer l'équivalence des assertions suivantes :

(i) f est logarithmiquement convexe ;

(ii) $f(x)\nabla^2 f(x) \succeq \nabla f(x)\,[\nabla f(x)]^\top$ pour tout $x \in \mathcal{O}$;

(iii) La fonction $x \longmapsto \mathrm{e}^{\langle a,x\rangle} f(x)$ est convexe pour tout $a \in \mathbb{R}^n$.

b) En déduire que si f_1 et f_2 sont logarithmiquement convexes et deux fois différentiables sur $\mathcal{O}$, il en est de même de leur somme $f_1 + f_2$.

Solution : 1°) *a*) La fonction $\varphi : x \in \mathcal{O} \longmapsto \varphi(x)$ est convexe par hypothèse ; la fonction $y \in \mathbb{R} \longmapsto \exp y$ est croissante et convexe. Par suite, la fonction composée qui à $x \in \mathcal{O}$ associe $\exp \varphi(x)$ est convexe ; en effet :

$$\forall \alpha \in]0,1[\,, \quad \forall x',\ x' \in \mathcal{O}, \quad \varphi(\alpha x + (1-\alpha)x') \leqslant \alpha\varphi(x) + (1-\alpha)\varphi(x'),$$

d'où

$$\begin{aligned}(\exp\,\varphi)(\alpha x + (1-\alpha)x') &\leqslant \exp[\alpha\varphi(x) + (1-\alpha)\varphi(x')] \\ &\leqslant\ \alpha\ \exp\,\varphi(x) + (1-\alpha)\varphi(x').\end{aligned}$$

b) Appliquons le résultat précédent à $\varphi := \ln f$; il vient que $f = \exp(\ln f)$ est convexe.

2°) *a*) [$(i) \Leftrightarrow (ii)$]. On sait que la fonction (deux fois différentiable) $\varphi := \ln f$ est convexe si et seulement si :

$$\nabla^2\varphi(x) \text{ est semi-définie positive pour tout } x \in \mathcal{O}.$$

Mais

$$\nabla^2\varphi(x) = \frac{f(x)\nabla^2 f(x) - \nabla f(x)[\nabla f(x)]^\top}{f(x)^2} \quad \text{pour tout } x \in \mathcal{O}\ ;$$

l'équivalence annoncée s'ensuit.

$[(ii) \Leftrightarrow (iii)]$. Soit $g_a : x \in \mathcal{O} \longmapsto g_a(x) := \mathrm{e}^{\langle a,x\rangle} f(x)$. La fonction g_a est deux fois différentiable sur $\mathcal{O}$ avec

$\nabla^2 g_a(x) = \mathrm{e}^{\langle a,x\rangle} \left\{ f(x)aa^\top + \nabla f(x)a^\top + a[\nabla f(x)]^\top + \nabla^2 f(x) \right\}$ pour tout $x \in \mathcal{O}$.

Par conséquent, ce qu'exprime (iii) est exactement :

$f(x)(\langle a,h\rangle)^2 + 2\langle \nabla f(x),h\rangle \langle a,h\rangle + \left\langle \nabla^2 f(x)h,h\right\rangle \geqslant 0$ pour tout $x \in \mathcal{O}$, tout $a \in \mathbb{R}^n$ et tout $h \in \mathbb{R}^n$;

soit encore

$f(x)u^2 + 2\langle \nabla f(x),h\rangle u + \left\langle \nabla^2 f(x)h,h\right\rangle \geqslant 0$ pour tout $x \in \mathcal{O}$, tout $u \in \mathbb{R}$ et tout $h \in \mathbb{R}^n$.

Le trinôme du second degré mis en évidence ci-dessus est positif sur $\mathbb{R}$ si et seulement si son discriminant est négatif ; cela se traduit par

$$(\langle \nabla f(x),h\rangle)^2 - f(x)\left\langle \nabla^2 f(x)h,h\right\rangle \leqslant 0.$$

L'équivalence de (iii) et (ii) est claire à présent.

b) Si f_1 et f_2 sont logarithmiquement convexes, toutes les fonctions $x \longmapsto \mathrm{e}^{\langle a,x\rangle} f_1(x)$ et $x \longmapsto e^{\langle a,x\rangle} f_2(x)$ sont convexes (d'après $[(i) \Rightarrow (iii)]$ de la question précédente) ; par suite, toutes les fonctions $x \longmapsto \mathrm{e}^{\langle a,x\rangle}(f_1+f_2)(x)$ sont convexes, et donc $f_1 + f_2$ est logarithmiquement convexe (d'après $[(iii) \Rightarrow (i)]$ de la question précédente).

Commentaire : Prolongement de l'exercice : Démontrer l'équivalence entre (i) et (iii) de la question 2°) a) et le résultat de la question 2°) b) pour des fonctions qui ne sont pas nécessairement différentiables.

***Exercice I.16.** Soit $Q \in \mathcal{S}_p(\mathbb{R})$ et $A \in \mathcal{M}_{m,p}(\mathbb{R})$. Montrer l'équivalence des deux assertions suivantes :

(i) $\langle Qx,x\rangle > 0$ pour tout $x \in \operatorname{Ker} A \setminus \{0\}$;

(ii) Il existe $\alpha_0 \geqslant 0$ tel que $Q + \alpha_0 A^\top A$ soit définie positive.

Solution : $[(ii) \Rightarrow (i)]$. Immédiat : si $x \in \operatorname{Ker} A,\ x \neq 0$,

$$\langle Qx,x\rangle = \langle Qx,x\rangle + \alpha_0 \parallel Ax \parallel^2 = \langle (Q + \alpha_0 A^\top A)x,x\rangle > 0.$$

$[(i) \Rightarrow (ii)]$. Supposons que (ii) ne soit pas réalisée. Pour tout $n \in \mathbb{N}^*$ il existe donc un vecteur $x_n \neq 0$ tel que

$$\langle Qx_n, x_n\rangle + n \parallel Ax_n \parallel^2 \leqslant 0,$$

soit encore, en posant $u_n := x_n / \parallel x_n \parallel$,

$$(*)_n \qquad \frac{\langle Qu_n, u_n\rangle}{n} + \parallel Au_n \parallel^2 \leqslant 0.$$

On peut alors extraire de $\{u_n\}$ une sous-suite $\{u_{n_k}\}_k$ convergeant, lorsque $k \longrightarrow +\infty$, vers un vecteur u de norme 1. En passant à la limite dans l'inégalité $(*)_{n_k}$, on obtient $Au = 0$. Or, d'après (i), $\langle Qu, u\rangle > 0$, ce qui implique que $\langle Qu_{n_k}, u_{n_k}\rangle > 0$ pour k assez grand. Ceci contredit l'inégalité $(*)_{n_k}$.

Commentaire : Il est clair que si $Q + \alpha_0 A^\top A$ est définie positive pour un certain α_0, il en est de même de $Q + \alpha A^\top A$ dès que $\alpha \geqslant \alpha_0$.

***Exercice I.17.** Soit A et B dans $\mathcal{S}_n(\mathbb{R})$, A étant de plus définie positive. Montrer que :

- le spectre de AB est entièrement réel ;
- la plus grande valeur propre de AB est égal à $\sup_{u\neq 0} \dfrac{\langle Bu, u\rangle}{\langle A^{-1}u, u\rangle}$;
- AB est diagonalisable.

Indication. Considérer $C := A^{1/2}BA^{1/2}$ ou, ce qui revient au même, la réduction de la forme quadratique associée à B dans l'espace euclidien $(\mathbb{R}^n, \langle\cdot,\cdot\rangle_A)$ avec $\langle u, v\rangle_A := \langle Au, v\rangle$.

Solution : Les valeurs propres de AB sont les racines de la fonction polynôme caractéristique $P(\lambda) = \det(AB - \lambda I_n)$. Mais, en factorisant $A^{1/2}$ à gauche dans $AB - \lambda I_n$, puis en multipliant à droite par $A^{1/2}$, on constate que

$$\begin{aligned}(\det(AB - \lambda I_n) = 0) &\Leftrightarrow (\det(A^{1/2}B - \lambda A^{-1/2}) = 0)\\ &\Leftrightarrow (\det(A^{1/2}BA^{1/2} - \lambda I_n) = 0).\end{aligned}$$

Le spectre de AB est donc celui de la matrice symétrique $A^{1/2}BA^{1/2}$, entièrement constitué de réels.

De plus

$$\lambda_{\max}(AB) = \lambda_{\max}(A^{1/2}BA^{1/2}) = \sup_{x\neq 0} \frac{\langle A^{1/2}BA^{1/2}x, x\rangle}{\langle x, x\rangle}$$
$$= \sup_{u\neq 0} \frac{\langle Bu, u\rangle}{\langle A^{-1}u, u\rangle} \quad \begin{pmatrix}\text{en posant}\\ u = A^{1/2}x\end{pmatrix}.$$

Soit U une matrice orthogonale diagonalisant $A^{1/2}BA^{1/2}$:

$$U^{-1}(A^{1/2}BA^{1/2})U = \mathrm{diag}(\lambda_1, \ldots, \lambda_n).$$

Alors $P := A^{1/2}U$ diagonalise AB :

$$P^{-1}(AB)P = \mathrm{diag}(\lambda_1, \ldots, \lambda_n).$$

Remarque : Si B est semi-définie positive (resp. définie positive) il en est de même de $A^{1/2}BA^{1/2}$; le spectre de AB est alors constitué de réels $\geqslant 0$ (resp. > 0).

**** Problème I.18.** Les données sont les suivantes : N, M entiers $\geqslant 1$, A matrice symétrique définie positive de taille $N, b \in \mathbb{R}^N$, B matrice à M lignes et N colonnes ($M \leqslant N$) non nulle. On désigne par $\langle\cdot,\cdot\rangle$ le produit scalaire usuel aussi bien dans $\mathbb{R}^N$ que dans $\mathbb{R}^M$, et par $\|\cdot\|$ la norme euclidienne associée.

On considère le problème de minimisation (dans $\mathbb{R}^N$) suivant :

$$(\mathcal{P}) \begin{cases}\text{Minimiser } f(x) := \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle\\ \text{sous la contrainte } \ Bx = 0.\end{cases}$$

A – Existence, unicité, caractérisation des solutions de $(\mathcal{P})$:

1°) Quelles sont les propriétés de f relatives à la différentiabilité, la convexité, et le comportement à l'infini ?

2°) Démontrer que $(\mathcal{P})$ a une solution et une seule (que l'on notera $\overline{x}$ par la suite).

3°) a) Montrer que $\overline{x}$ est caractérisée (parmi les éléments de $\mathbb{R}^N$) par le système

$$(\mathcal{S}) \begin{cases} B\overline{x} = 0\\ A\overline{x} - b \in \mathrm{Im}(B^\top).\end{cases}$$

b) Vérifier que $f(\overline{x}) = -\frac{1}{2}\langle A\overline{x}, \overline{x}\rangle = -\frac{1}{2}\langle b, \overline{x}\rangle$:

B – Lagrangien et lagrangien augmenté :

Étant donné $r \geqslant 0$, on définit les applications $\mathcal{L}$ et $\mathcal{L}_r$ de $\mathbb{R}^N \times \mathbb{R}^M$ dans $\mathbb{R}$ par :

$$\mathcal{L}(x,\lambda) := f(x) + \langle \lambda, Bx \rangle \text{ (lagrangien usuel) ;}$$

$$\mathcal{L}_r(x,\lambda) := \mathcal{L}(x,\lambda) + \frac{r}{2} \parallel Bx \parallel^2 \text{ (lagrangien augmenté).}$$

On dit que $(\overline{x},\overline{\lambda}) \in \mathbb{R}^N \times \mathbb{R}^M$ est un point-selle (ou un col) de $\mathcal{L}$ sur $\mathbb{R}^N \times \mathbb{R}^M$ lorsque :

$$\forall (x,\lambda) \in \mathbb{R}^N \times \mathbb{R}^M, \quad \mathcal{L}(\overline{x},\lambda) \leqslant \mathcal{L}(\overline{x},\overline{\lambda}) \leqslant \mathcal{L}(x,\overline{\lambda}).$$

Définition analogue d'un point-selle de $\mathcal{L}_r$.

1°) a) Montrer que l'on a toujours :

$$\sup_{\lambda \in \mathbb{R}^M} \left(\inf_{x \in \mathbb{R}^N} \mathcal{L}(x,\lambda) \right) \leqslant \inf_{x \in \mathbb{R}^N} \left(\sup_{\lambda \in \mathbb{R}^M} \mathcal{L}(x,\lambda) \right)$$

(Cette inégalité est dans $\mathbb{R} \cup \{\pm\infty\}$, et sa démonstration ne fait pas appel à l'expression particulière de $\mathcal{L}$ comme fonction de x et λ).

b) Soit $(\overline{x},\overline{\lambda})$ un point-selle de $\mathcal{L}$. Montrer :

$$\mathcal{L}(\overline{x},\overline{\lambda}) = \max_{\lambda \in \mathbb{R}^M} \left(\inf_{x \in \mathbb{R}^N} \mathcal{L}(x,\lambda) \right) = \min_{x \in \mathbb{R}^N} \left(\sup_{\lambda \in \mathbb{R}^M} \mathcal{L}(x,\lambda) \right).$$

2°) Soit $r \geqslant 0$ et $(\overline{x},\overline{\lambda}) \in \mathbb{R}^N \times \mathbb{R}^M$.

a) Établir les équivalences suivantes :

$$\left(\mathcal{L}_r(\overline{x},\lambda) \leqslant \mathcal{L}_r(\overline{x},\overline{\lambda}) \text{ pour tout } \lambda \in \mathbb{R}^M \right) \Leftrightarrow (B\overline{x} = 0) \ ;$$

$$\left(\mathcal{L}_r(\overline{x},\overline{\lambda}) \leqslant \mathcal{L}_r(x,\overline{\lambda}) \text{ pour tout } x \in \mathbb{R}^N \right) \Leftrightarrow \left(\left(A + rB^\top B \right) \overline{x} + B^\top \overline{\lambda} = b \right).$$

b) Quelles conclusions peut-on tirer de ce qui précède concernant les liens existant entre les points-selles de $\mathcal{L}_r$ et la solution de $(\mathcal{P})$?

C – Algorithme d'Arrow-Uzawa (à pas variable) :

Les données sont : $r \geqslant 0, \lambda_0 \in \mathbb{R}^M, (\rho_n)$ une suite de réels > 0. On considère les suites (x_n) de $\mathbb{R}^N$, (λ_n) de $\mathbb{R}^M$ construites de manière récurrente comme suit :

$$\begin{cases} \forall n \in \mathbb{N},\ x_n \text{ minimise } \mathcal{L}_r(\cdot,\lambda_n) \text{ sur } \mathbb{R}^N, \\ \lambda_{n+1} = \lambda_n + \rho_n B x_n. \end{cases}$$

Soit $(\overline{x},\overline{\lambda})$ un point-selle de $\mathcal{L}_r$.

1°) Établir les relations suivantes :

$$(A + rB^\top B)(x_n - \overline{x}) + B^\top(\lambda_n - \overline{\lambda}) = 0 \ ;$$

$$\lambda_{n+1} - \overline{\lambda} = \lambda_n - \overline{\lambda} + \rho_n B(x_n - \overline{x}).$$

En déduire

$$\| \lambda_{n+1} - \overline{\lambda} \|^2 = \| \lambda_n - \overline{\lambda} \|^2 - 2 \rho_n \langle A(x_n - \overline{x}), x_n - \overline{x} \rangle + \rho_n(\rho_n - 2r) \| B(x_n - \overline{x}) \|^2 .$$

2°) On suppose désormais $\rho_n \in [\alpha_0, \alpha_1]$ pour tout n, où :

$$0 < \alpha_0 < \alpha_1 < 2\left(r + \frac{1}{\sigma_*}\right),$$

$\sigma_* :=$ la plus grande valeur propre de $A^{-1}B^\top B$.

(On admettra ici que le spectre de $A^{-1}B^\top B$ est constitué de réels $\geqslant 0$ et que $\sigma_* = \sup_{u \neq 0} \frac{\|Bu\|^2}{\langle Au, u \rangle}$; *cf.* Exercice 1.17).

a) Démontrer que la suite $\left(\| \lambda_n - \overline{\lambda} \|^2\right)_n$ est décroissante.

b) Montrer que :

$$\lim_{n \longrightarrow +\infty} B(x_n - \overline{x}) = 0 \ ; \quad \lim_{n \longrightarrow +\infty} \langle A(x_n - \overline{x}), x_n - \overline{x} \rangle = 0.$$

Déduire de ce qui précède : $\lim_{n \longrightarrow +\infty} x_n = \overline{x}$.

3°) Montrer également que la suite (λ_n) converge quand $n \longrightarrow +\infty$ vers $\lambda_\infty = \widehat{\lambda} + \lambda_{0,2}$, où $\lambda_{0,2}$ est la composante de λ_0 dans $\mathrm{Ker}(B^\top)$ provenant de la décomposition $\mathrm{Im}B \oplus \mathrm{Ker}(B^\top)$ de $\mathbb{R}^M$, et $\widehat{\lambda}$ le multiplicateur de Lagrange de norme minimale dans $\mathbb{R}^M$ pour le problème $(\mathcal{P})$ en question.

Commentaire : Ce problème, dans lequel apparaissent bien des aspects de l'Optimisation, peut être abordé sans aucune connaissance spécifique à ce domaine ; seuls sont utilisés les résultats et techniques dans les thèmes qui font l'objet de révisions dans ce chapitre.

Solution : A – 1°) f est de classe $\mathcal{C}^\infty$ sur $\mathbb{R}^N$; $\nabla f(x) = Ax - b$ et $\nabla^2 f(x) = A$ pour tout $x \in \mathbb{R}^N$.

f est convexe (et même fortement convexe) sur $\mathbb{R}^N$.

Si σ désigne la plus petite valeur propre de A ($\sigma > 0$ donc), on a :

$$\langle Ax, x\rangle \geqslant \sigma \parallel x \parallel^2, \text{ d'où } f(x) \geqslant \frac{\sigma}{2} \parallel x \parallel^2 - \parallel b \parallel \cdot \parallel x \parallel \text{ pour tout } x.$$

Ainsi $\lim\limits_{\|x\| \longrightarrow +\infty} \dfrac{f(x)}{\parallel x \parallel} = +\infty$: f est ce qu'on appelle « 1-coercive sur $\mathbb{R}^N$ ».

2°) La 1-coercivité de f est plus qu'il n'en faut pour assurer l'existence d'un minimum de f sur $\operatorname{Ker} B$. Servons-nous de la propriété : $\lim\limits_{\substack{\parallel x \parallel \longrightarrow +\infty \\ x \in \operatorname{Ker} B}} f(x) = +\infty$.

Choisissons $x_0 \in \operatorname{Ker} B$; il existe $r > \parallel x_0 \parallel$ tel que

$$(x \in \operatorname{Ker} B \text{ et } \parallel x \parallel > r) \Rightarrow (f(x) \geqslant f(x_0)).$$

De ce fait :

$$\inf_{Bx=0} f(x) = \inf_{\substack{Bx = 0 \\ \parallel x \parallel \leqslant r}} f(x).$$

L'existence d'un minimum $\overline{x}$ de f sur $\operatorname{Ker} B$ s'ensuit. L'unicité de ce minimum vient de la stricte convexité de f.

3°) $(\mathcal{S})$ $\begin{cases} B\overline{x} = 0 \\ \exists\, \overline{\lambda} \in \mathbb{R}^M \text{ tel que } A\overline{x} - b + B^\top \overline{\lambda} = 0. \end{cases}$

a) Faisons une démonstration directe du fait que $(\mathcal{S})$ *caractérise* la solution $\overline{x}$ de $(\mathcal{P})$.

– Partons de $\overline{x}$ vérifiant $(\mathcal{S})$. Soit $x \in \operatorname{Ker} B$; a-t-on

$$\frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle \geqslant \frac{1}{2}\langle A\overline{x}, \overline{x}\rangle - \langle b, \overline{x}\rangle \ ?$$

La fonction f étant convexe, on sait que $f(x) \geqslant f(\overline{x}) + \langle \bigtriangledown f(\overline{x}), x - \overline{x}\rangle$, soit ici

$$f(x) \geqslant f(\overline{x}) + \langle A\overline{x} - b, x - \overline{x}\rangle.$$

Mais $A\overline{x} - b \in \operatorname{Im}(B^\top) = (\operatorname{Ker} B)^\perp$ et $x - \overline{x} \in \operatorname{Ker} B$, donc $\langle A\overline{x} - b, x - \overline{x}\rangle = 0$.

– On suppose que $\overline{x}$ minimise f sur $\operatorname{Ker} B$. Tout d'abord, $B\overline{x} = 0$. Ensuite

$$f(\overline{x} + td) \geqslant f(\overline{x}) \text{ pour tout } t \in \mathbb{R} \text{ et } d \in \operatorname{Ker} B,$$

i.e. $\frac{1}{2}t^2\langle Ad, d\rangle + t\langle A\overline{x} - b, d\rangle \geqslant 0$ pour tout $t \in \mathbb{R}$ et $d \in \operatorname{Ker} B$.

Ceci n'est possible que si $\langle A\overline{x} - b, d\rangle = 0$. Donc $A\overline{x} - b \in (\operatorname{Ker} B)^\perp = \operatorname{Im}(B^\top)$.

Remarque : L'ensemble des $\overline{\lambda}$ vérifiant $A\overline{x} - b + B^\top\overline{\lambda} = 0$ est l'ensemble des multiplicateurs de Lagrange associé à $\overline{x}$. C'est un sous-espace affine de $\mathbb{R}^M$, de la forme $\overline{\lambda} + \mathrm{Ker}(B^\top)$. Il est réduit à $\{\overline{\lambda}\}$ lorsque B est surjective (*i.e.*, lorsque les m vecteurs-lignes de B sont linéairement indépendants), ce qu'indiquait déjà le théorème de Lagrange.

b) Puisque $A\overline{x} - b \in \mathrm{Im}(B^\top) = (\mathrm{Ker}\, B)^\perp$, on a $\langle A\overline{x} - b, \overline{x}\rangle = 0$, soit $\langle A\overline{x}, \overline{x}\rangle = \langle b, \overline{x}\rangle$.

D'où $f(\overline{x}) = -\frac{1}{2}\langle A\overline{x}, \overline{x}\rangle = -\frac{1}{2}\langle b, \overline{x}\rangle$.

B – 1°) a) Soit $(x_0, \lambda_0) \in \mathbb{R}^N \times \mathbb{R}^M$. Par définitions,

$$\inf_{x\in\mathbb{R}^N} \mathcal{L}(x, \lambda) \leqslant \mathcal{L}(x_0, \lambda_0) \leqslant \sup_{\lambda\in\mathbb{R}^M} \mathcal{L}(x, \lambda).$$

D'où :

$$\sup_{\lambda\in\mathbb{R}^M} \left(\inf_{x\in\mathbb{R}^N} \mathcal{L}(x, \lambda) \right) \leqslant \inf_{x\in\mathbb{R}^N} \left(\sup_{\lambda\in\mathbb{R}^M} \mathcal{L}(x, \lambda) \right).$$

b) Considérons maintenant un point-selle $(\overline{x}, \overline{\lambda})$ de $\mathcal{L}$ sur $\mathbb{R}^N \times \mathbb{R}^M$. On a :

$$\mathcal{L}(\overline{x}, \overline{\lambda}) = \inf_{x\in\mathbb{R}^N} \mathcal{L}(x, \overline{\lambda}) \leqslant \sup_{\lambda\in\mathbb{R}^M} \left(\inf_{x\in\mathbb{R}^N} \mathcal{L}(x, \lambda) \right);$$

$$\mathcal{L}(\overline{x}, \overline{\lambda}) = \sup_{\lambda\in\mathbb{R}^M} \mathcal{L}(\overline{x}, \lambda) \geqslant \inf_{x\in\mathbb{R}^N} \left(\sup_{\lambda\in\mathbb{R}^M} \mathcal{L}(x, \lambda) \right).$$

Par suite :

$$\underbrace{\sup_{\lambda\in\mathbb{R}^M} \left(\inf_{x\in\mathbb{R}^N} \mathcal{L}(x, \lambda) \right)}_{\text{supremum atteint pour } \lambda = \overline{\lambda}} = \underbrace{\inf_{x\in\mathbb{R}^N} \left(\sup_{\lambda\in\mathbb{R}^M} \mathcal{L}(x, \lambda) \right)}_{\text{infimum atteint pour } x = \overline{x}} = \mathcal{L}(\overline{x}, \overline{\lambda}).$$

2°) a) $\mathcal{L}(x, \lambda) = f(x) + \langle \lambda, Bx\rangle$; $\mathcal{L}_r(x, \lambda) = \mathcal{L}(x, \lambda) + \frac{r}{2}\langle B^\top Bx, x\rangle$. Alors :

– $\overline{\lambda}$ maximise $\mathcal{L}_r(\overline{x}, \cdot)$ sur $\mathbb{R}^M$ si et seulement si $\nabla_\lambda \mathcal{L}_r(\overline{x}, \overline{\lambda}) = B\overline{x} = 0$;

– $\overline{x}$ minimise la fonction quadratique convexe $\mathcal{L}_r(\cdot, \overline{\lambda})$ sur $\mathbb{R}^N$ si et seulement si

$$\nabla_x \mathcal{L}_r(\overline{x}, \overline{\lambda}) = A\overline{x} - b + B^\top\overline{\lambda} + rB^\top B\overline{x} = 0.$$

b) Pour $r \geqslant 0$ donné, un point-selle de $\mathcal{L}_r$ est exactement un couple $(\overline{x}, \overline{\lambda})$ de $\mathbb{R}^N \times \mathbb{R}^M$ vérifiant

$$B\overline{x} = 0 \quad \text{et} \quad (A + rB^\top B)\overline{x} + B^\top\overline{\lambda} = b,$$

soit encore

$$B\overline{x} = 0 \quad \text{et} \quad A\overline{x} + B^\top \overline{\lambda} = b.$$

En comparant ceci à $(\mathcal{S})$, on voit donc que la solution $\overline{x}$ de $(\mathcal{P})$ est précisément la première composante de tout point-selle de $\mathcal{L}$ (ou de $\mathcal{L}_r$, ceci pour un $r > 0$, ou quelque soit $r \geqslant 0$).

C – 1°) Puisque x_n est, par construction, un minimum de $\mathcal{L}_r(\cdot, \lambda_n)$ sur $\mathbb{R}^N$, il est caractérisé comme suit :

$$(A + rB^\top B)x_n + B^\top \lambda_n = b.$$

On a vu qu'il existe effectivement un point-selle $(\overline{x}, \overline{\lambda})$ de $\mathcal{L}_r$ sur $\mathbb{R}^N \times \mathbb{R}^M$ et que ce point-selle est caractérisé par les relations :

$$B\overline{x} = 0 \quad \text{et} \quad (A + rB^\top B)\overline{x} + B^\top \overline{\lambda} = b.$$

Par suite :

$$(A + rB^\top B)(x_n - \overline{x}) + B^\top(\lambda_n - \overline{\lambda}) = 0, \qquad (\alpha)$$

$$\lambda_{n+1} - \overline{\lambda} = \lambda_n - \overline{\lambda} + \rho_n B(x_n - \overline{x}). \qquad (\beta)$$

On tire de (β) :

$$\| \lambda_{n+1} - \overline{\lambda} \|^2 = \| \lambda_n - \overline{\lambda} \|^2 + 2\, \rho_n \langle B(x_n - \overline{x}), \lambda_n - \overline{\lambda} \rangle + \rho_n^2 \| B(x_n - \overline{x}) \|^2 .$$

Il vient de (α), en faisant le produit scalaire des deux membres avec $x_n - \overline{x}$:

$$\langle A(x_n - \overline{x}), x_n - \overline{x} \rangle + r \| B(x_n - \overline{x}) \|^2 + \langle B(x_n - \overline{x}), \lambda_n - \overline{\lambda} \rangle = 0.$$

Les deux égalités précédentes conduisent à :

$$\| \lambda_{n+1} - \overline{\lambda} \|^2 - \| \lambda_n - \overline{\lambda} \|^2 = - 2\rho_n \langle A(x_n - \overline{x}), x_n - \overline{x} \rangle + (\rho_n^2 - 2\rho_n r) \| B(x_n - \overline{x}) \|^2 . \qquad (\gamma)$$

2°) a) On doit s'assurer que le second membre de (γ) est $\leqslant 0$.
C'est le cas si $B(x_n - \overline{x}) = 0$. Sinon, il faut prouver que

$$\rho_n \| B(x_n - \overline{x}) \|^2 - 2[\langle A(x_n - \overline{x}), x_n - \overline{x} \rangle + r \| B(x_n - \overline{x}) \|^2] \leqslant 0,$$

soit

$$\rho_n \leqslant 2 \left[r + \frac{\langle A(x_n - \overline{x}), x_n - \overline{x} \rangle}{\| B(x_n - \overline{x}) \|^2} \right]. \qquad (\delta)$$

Or : $\frac{1}{\sigma_*} \leqslant \frac{\langle Au, u \rangle}{\|Bu\|^2}$ pour tout u tel que $Bu \neq 0$ et $\rho_n \leqslant \alpha_1 < 2\left(r + \frac{1}{\sigma_*}\right)$.
Donc l'inégalité (δ) a bien lieu.

b) Par choix de ρ_n, on a : $2r - \rho_n > -\frac{2}{\sigma_*}$ et $\rho_n \geqslant \alpha_0 > 0$ pour tout n. Alors

$$\| \lambda_n - \overline{\lambda} \|^2 - \| \lambda_{n+1} - \overline{\lambda} \|^2 = \rho_n[(2r - \rho_n) \| B(x_n - \overline{x}) \|^2$$

$$+ \ 2\langle A(x_n - \overline{x}), x_n - \overline{x}\rangle] \geqslant \alpha_0 \left[-\frac{2}{\sigma_*} \| B(x_n - \overline{x}) \|^2 + 2\langle A(x_n - \overline{x}), x_n - \overline{x}\rangle \right].$$

Comme $\langle A(x_n - \overline{x}), x_n - \overline{x}\rangle \geqslant \frac{1}{\sigma_*} \| B(x_n - \overline{x}) \|^2$ (voir la formulation variationnelle de σ_* et se rappeler que $\sigma_* > 0$), le second membre de l'inégalité précédente est $\geqslant 0$.

Or $\| \lambda_n - \overline{\lambda} \|^2 - \| \lambda_{n+1} - \overline{\lambda} \|^2 \longrightarrow 0$ quand $n \longrightarrow +\infty$ (puisque la suite de réels positifs $(\| \lambda_n - \overline{\lambda} \|^2)_n$ est décroissante) ; donc

$$\langle A(x_n - \overline{x}), x_n - \overline{x}\rangle - \frac{1}{\sigma_*} \| B(x_n - \overline{x}) \|^2 \longrightarrow 0 \quad \text{quand } n \longrightarrow +\infty.$$

Utilisons à présent l'inégalité stricte $\alpha_1 < 2\left(r + \frac{1}{\sigma_*}\right)$:

$$2\sigma_*\langle A(x_n - \overline{x}), x_n - \overline{x}\rangle + \sigma_*(2r - \rho_n) \| B(x_n - \overline{x}) \|^2 \qquad \begin{pmatrix} \text{suite dont on sait} \\ \text{qu'elle tend vers 0} \\ \text{quand } n \longrightarrow +\infty \end{pmatrix}$$

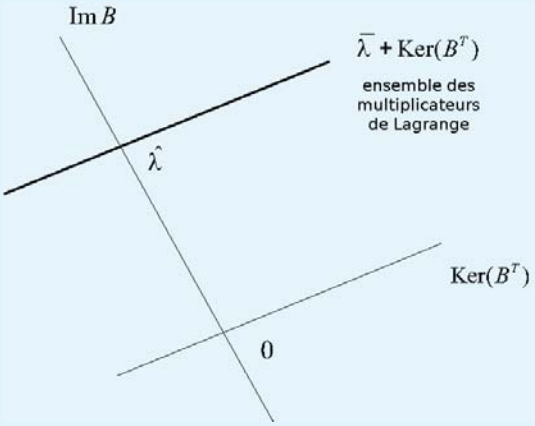
$$= \left[\underbrace{2\sigma_*\langle A(x_n - \overline{x}), x_n - \overline{x}\rangle - 2 \| B(x_n - \overline{x}) \|^2}_{\text{élément } \geqslant 0} \right]$$

$$+ \underbrace{\sigma_*}_{>0} \left[\underbrace{(2(r + \frac{1}{\sigma_*}) - \rho_n)}_{\geqslant 2(r+\frac{1}{\sigma_*}) - \alpha_1 > 0.} \| B(x_n - \overline{x}) \|^2 \right]$$

Par conséquent : $\lim\limits_{n \longrightarrow +\infty} \| B(x_n - \overline{x}) \|^2 = 0$, et $\lim_{n \longrightarrow +\infty} \langle A(x_n - \overline{x}), x_n - \overline{x}\rangle = 0$.

Comme $\langle A(x_n - \overline{x}), x_n - \overline{x}\rangle \geqslant \sigma \| x_n - \overline{x} \|^2$, avec $\sigma > 0$ plus petite valeur propre de A, il s'ensuit $\lim\limits_{n \longrightarrow +\infty} \| x_n - \overline{x} \| = 0$.

3°)



$$\mathbb{R}^M = \mathrm{Im}B \oplus \mathrm{Ker}\left(B^\top\right)$$

Tout multiplicateur de Lagrange λ (i.e., toute seconde composante de point-selle de $\mathcal{L}_r$) se décompose en $\lambda = \widehat{\lambda} + s$, où $\widehat{\lambda}$ est la projection de l'origine sur le sous-espace affine des multiplicateurs de Lagrange, et $s \in \mathrm{Ker}(B^\top)$.

Désignons par p_2 le projecteur orthogonal de $\mathbb{R}^M$ d'image $\mathrm{Ker}(B^\top)$ (et de noyau $\mathrm{Im}B$). Puisque $\lambda_{n+1} = \lambda_n + \rho B x_n$, on a : $p_2(\lambda_{n+1}) = p_2(\lambda_n)$ pour tout $n \in \mathbb{N}$, donc :

$$p_2(\lambda_n) = p_2(\lambda_0) = \lambda_{0,2} \text{ pour tout } n.$$

On sait que $\lim\limits_{n\longrightarrow+\infty} \| x_n - \overline{x} \| = 0$, et l'égalité

$$\left(A + rB^\top B\right)(x_n - \overline{x}) + B^\top(\lambda_n - \overline{\lambda}) = 0$$

vue au 1°) de C, donne : $\lim\limits_{n\longrightarrow+\infty} B^\top(\lambda_n - \overline{\lambda}) = 0$.

Or l'application $u \in \mathrm{Im}B \longmapsto \| B^\top u \|$ est une norme *sur ImB*. En écrivant

$$\lambda_n - \overline{\lambda} = (id_{\mathbb{R}^M} - p_2)(\lambda_n - \overline{\lambda}) + p_2(\lambda_n - \overline{\lambda}) \qquad \begin{pmatrix}\text{décomposition suivant}\\ \mathrm{Im}B \oplus \mathrm{Ker}(B^\top)\end{pmatrix}$$

et

$$B^\top(\lambda_n - \overline{\lambda}) = B^\top\left[(id_{\mathbb{R}^M} - p_2)(\lambda_n - \overline{\lambda})\right],$$

on a donc $\lim\limits_{n\longrightarrow+\infty} (\mathrm{id}_{\mathbb{R}^M} - p_2)(\lambda_n - \overline{\lambda}) = 0$, soit $\lim\limits_{n\longrightarrow+\infty} (id_{\mathbb{R}^M} - p_2)(\lambda_n) = \widehat{\lambda}$.

En rassemblant : $\lambda_n \longrightarrow \widehat{\lambda} + \lambda_{0,2}$ quand $n \longrightarrow +\infty$.

II

MINIMISATION SANS CONTRAINTES. CONDITIONS DE MINIMALITÉ

Rappels

– $f : \mathbb{R}^n \to \mathbb{R}$ (ou même $\mathbb{R} \cup \{+\infty\}$) est dite 0-coercive (resp. 1-coercive) sur $C \subset \mathbb{R}^n$ lorsque

$$\lim_{\substack{\| x \| \to +\infty \\ x \in C}} f(x) = +\infty \quad \text{(resp.} \quad \lim_{\substack{\| x \| \to +\infty \\ x \in C}} \frac{f(x)}{\| x \|} = +\infty\text{)}.$$

– Si C est convexe et si $f : C \to \mathbb{R}$ est strictement convexe sur C, alors il existe au plus un $\overline{x} \in C$ minimisant f sur C.

II.1. Conditions de minimalité du premier ordre

Soit $\mathcal{O}$ un ouvert de $\mathbb{R}^n$ et $f : \mathcal{O} \to \mathbb{R}$.

– Si $\overline{x} \in \mathcal{O}$ est un minimum local de f et si f est différentiable en $\overline{x}$, alors $\nabla f(\overline{x}) = 0$.

– On suppose $\mathcal{O}$ convexe et f convexe sur $\mathcal{O}$. Alors les conditions suivantes relatives à $\overline{x} \in \mathcal{O}$ sont équivalentes :

(i) $\overline{x}$ est un minimum (global) de f sur $\mathcal{O}$;

(ii) $\overline{x}$ est un minimum local de f.

Et si f est différentiable en $\overline{x}$, on a une troisième condition équivalente :

(iii) $\nabla f(\overline{x}) = 0$.

II.2. Conditions de minimalité du second ordre

Soit $\mathcal{O}$ un ouvert de $\mathbb{R}^n$ et $f : \mathcal{O} \to \mathbb{R}$.

– Si $\overline{x} \in \mathcal{O}$ est un minimum local de f et si f est deux fois différentiable en $\overline{x}$, alors

$$\nabla f(\overline{x}) = 0 \quad \text{et } \nabla^2 f(\overline{x}) \text{ est semi-définie positive.}$$

– Si $\overline{x} \in \mathcal{O}$ est un point où f est deux fois différentiable et si :

$$\nabla f(\overline{x}) = 0, \ \nabla^2 f(\overline{x}) \text{ est définie positive,}$$

alors $\overline{x}$ est un minimum local strict de f.

Références. Chapitre II de [11].

***Exercice II.1.** Soit $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ et $S \subset \mathbb{R}^n$. On suppose que :

(i) S est fermé ;

(ii) f est semi-continue inférieurement sur $\mathbb{R}^n$;

(iii) Il existe un point de S en lequel f est finie ;

(iv) $\lim\limits_{\substack{\| x \| \to +\infty \\ x \in S}} f(x) = +\infty.$

Montrer que f est bornée inférieurement sur S et qu'il existe $\overline{x} \in S$ tel que $f(\overline{x}) = \inf_{x \in S} f(x)$.

Indication. On montrera que toute suite minimisante pour f sur S est bornée, ou bien on commencera par traiter le cas où S est borné et on ramènera le cas général à ce cas grâce à l'hypothèse de 0-coercivité de f sur S (hypothèse (iv)).

Solution : On rappelle tout d'abord les différentes caractérisations de « f est semi-continue inférieurement sur $\mathbb{R}^n$ » :

– En tout point $x \in \mathbb{R}^n$, on a $\liminf_{x' \to x} f(x') \geqslant f(x)$ (inégalité dans $\mathbb{R} \cup \{+\infty\}$) ;

– $\forall\, \alpha \in \mathbb{R}, \{x \in \mathbb{R}^n \mid f(x) \leqslant \alpha\}$ est fermé ;

– $\mathrm{epi} f := \{(x, r) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) \leqslant r\}$ est fermé.

Soit $\overline{f} := \inf_S f \, (\in \mathbb{R} \cup \{-\infty\})$ et soit $\{x_k\} \subset S$ une suite minimisante pour f sur S, *i.e.*, telle que $f(x_k) \to \overline{f}$ quand $k \to +\infty$.

Montrons que $\{x_k\}$ est bornée. Si ce n'était pas le cas, il existerait une sous-suite $\{x_{k_l}\}_l$ de $\{x_k\}$ telle que $\| x_{k_l} \| \to +\infty$ quand $l \to +\infty$. Par la 0-coercivité de f sur S (hypothèse (iv)), cela impliquerait

$$\left(\overline{f} =\right) \quad \lim_{l \to +\infty} f(x_{k_l}) = +\infty,$$

ce qui est impossible.

La suite $\{x_k\}$ étant bornée, il existe une sous-suite $\{x_{k_l}\}$ qui converge vers un élément $\overline{x}$ de S. D'où, grâce à la semi-continuité inférieure de f,

$$\left(\overline{f} =\right) \quad \lim_{l \to +\infty} f(x_{k_l}) \geqslant f(\overline{x}).$$

Comme $f(\overline{x}) > -\infty$ et $f(\overline{x}) \geqslant \overline{f}$, on en déduit que la valeur $\overline{f}$ est finie et atteinte en $\overline{x}$.

Autre démonstration : Soit $x_0 \in S$ en lequel f est finie ; puisque $f(x) \to +\infty$ quand $\| x \| \to +\infty$, $x \in S$, il existe $r \ (> \| x_0 \|)$ tel que

$$f(x) \geqslant f(x_0) \text{ dès que } \ x \in S, \ \| x \| > r.$$

Donc minimiser f sur S revient à minimiser f sur $S \cap \overline{B}(0, r)$.

Commentaire : Le résultat de l'exercice est utilisé fréquemment dans le contexte suivant. Soit Ω un ouvert de $\mathbb{R}^n$ et $g : \Omega \to \mathbb{R}$ vérifiant :

$$g \text{ est continue sur } \Omega \ ; \ g(x) \to +\infty \text{ quand } x \to a \in \operatorname{fr} \Omega \ ;$$
$$\lim_{\substack{\| x \| \to +\infty \\ x \in \Omega}} g(x) = +\infty.$$

Alors, si S est un fermé tel que $S \cap \Omega \neq \emptyset$, g est bornée inférieurement sur $S \cap \Omega$ et il existe $\overline{x} \in S \cap \Omega$ tel que $g(\overline{x}) = \inf_{x \in S \cap \Omega} g(x)$. Il suffit pour voir cela de considérer $f : x \in \mathbb{R}^n \longmapsto f(x) := g(x)$ si $x \in \Omega, +\infty$ sinon.

***Exercice II.2.** Soit $\mathcal{D}_1 := \{A_1 + t\vec{v}_1 \mid t \in \mathbb{R}\}$ et $\mathcal{D}_2 := \{A_2 + t\vec{v}_2 \mid t \in \mathbb{R}\}$ deux droites affines de $\mathbb{R}^n$ définies à l'aide des points A_i et des vecteurs directeurs non nuls $\vec{v}_i$. On cherche les points $M_1 \in \mathcal{D}_1$ et $M_2 \in \mathcal{D}_2$ minimisant la distance (euclidienne) de M_1 à M_2.

1°) Formaliser le problème ci-dessus comme un problème de minimisation convexe sans contraintes.

2°) Résoudre le problème posé en précisant : les conditions d'existence et d'unicité des solutions, comment caractériser ces solutions, les propriétés particulières des solutions.

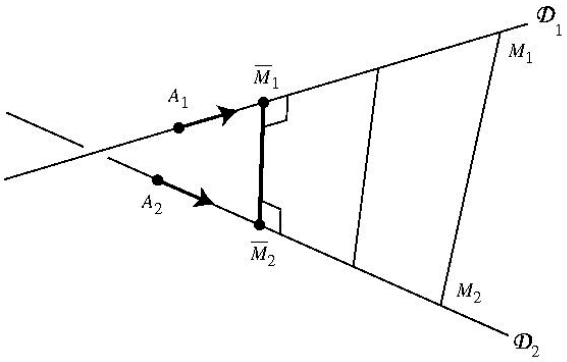


Figure 4.

Solution : 1°) Soit

$$M_1 = A_1 + t_1\vec{v}_1,\ t_1 \in \mathbb{R}$$
$$M_2 = A_2 + t_2\vec{v}_2,\ t_2 \in \mathbb{R},$$

de sorte que

$$\| \overrightarrow{M_1M_2} \|^2 = \| \overrightarrow{A_2A_1} + t_1\vec{v}_1 - t_2\vec{v}_2 \|^2 .$$

Le problème consiste à minimiser

$$f(t_1,t_2) := \| \overrightarrow{M_1M_2} \|^2, \quad (t_1,t_2) \in \mathbb{R}^2.$$

On a :

$$f(t_1,t_2) = t_1^2 \| \vec{v}_1 \|^2 + t_2^2 \| \vec{v}_2 \|^2 - 2t_1t_2\langle \vec{v}_1, \vec{v}_2 \rangle + 2t_1\langle \overrightarrow{A_2A_1}, \vec{v}_1 \rangle - 2t_2\langle \overrightarrow{A_2A_1}, \vec{v}_2 \rangle + \| \overrightarrow{A_2A_1} \|^2 .$$

f est une fonction quadratique de la variable $t = (t_1,t_2)$, convexe même car

$$\nabla^2 f(t_1,t_2) = 2 \begin{bmatrix} \| \vec{v}_1 \|^2 & -\langle \vec{v}_1, \vec{v}_2 \rangle \\ -\langle \vec{v}_1, \vec{v}_2 \rangle & \| \vec{v}_2 \|^2 \end{bmatrix}$$

est semi-définie positive pour tout $(t_1,t_2) \in \mathbb{R}^2$.

Le problème posé peut donc être formalisé en

$$(\mathcal{P}) \qquad \begin{cases} \text{Minimiser } f(t_1,t_2) \\ (t_1,t_2)\in\mathbb{R}^2. \end{cases}$$

2°) Les solutions $(\overline{t}_1,\overline{t}_2)$ de $(\mathcal{P})$ sont les solutions de $\nabla f(\overline{t}_1,\overline{t}_2)=0$, soit encore

$$(\mathcal{S}) \qquad \begin{cases} \overline{t}_1 \parallel \vec{v}_1 \parallel^2 - \overline{t}_2\langle \vec{v}_1,\vec{v}_2\rangle = -\langle \overrightarrow{A_2A_1},\vec{v}_1\rangle \\ \overline{t}_2 \parallel \vec{v}_2 \parallel^2 - \overline{t}_1\langle \vec{v}_1,\vec{v}_2\rangle = \langle \overrightarrow{A_2A_1},\vec{v}_2\rangle. \end{cases}$$

1er cas : $\vec{v}_1$ et $\vec{v}_2$ sont colinéaires, *i.e.*, $\mathcal{D}_1$ et $\mathcal{D}_2$ sont parallèles. Il y a une infinité de solutions à $(\mathcal{S})$; en posant $\vec{v}_2=\alpha\vec{v}_1$, ce sont les $(\overline{t}_1,\overline{t}_2)$ vérifiant

$$(\overline{t}_1-\alpha\overline{t}_2)\parallel \vec{v}_1\parallel^2 = -\langle \overrightarrow{A_2A_1},\vec{v}_1\rangle.$$

Les points $\overline{M}_1,\overline{M}_2$ correspondants sont

$$\overline{M}_1 = A_1+\overline{t}_1\vec{v}_1, \quad \overline{M}_2 = A_2+\left(\frac{\langle \overrightarrow{A_2A_1},\vec{v}_1\rangle}{\parallel \vec{v}_1\parallel^2}+\overline{t}_1\right)\vec{v}_1.$$

Évidemment $\overrightarrow{\overline{M}_1\overline{M}_2} = \overrightarrow{A_1A_2}+\dfrac{\langle \overrightarrow{A_2A_1},\vec{v}_1\rangle}{\parallel \vec{v}_1\parallel^2}\vec{v}_1$ est orthogonal à $\vec{v}_1$ (et à $\vec{v}_2$).

2e cas : $\vec{v}_1$ et $\vec{v}_2$ sont linéairement indépendants. Dans ce cas, $\nabla^2 f(t_1,t_2)$ est définie positive pour tout $(t_1,t_2)\in\mathbb{R}^2$ ($\parallel \vec{v}_1\parallel^2>0$ et $\det\nabla^2 f(t_1,t_2) = 4\left(\parallel \vec{v}_1\parallel^2\parallel \vec{v}_2\parallel^2-(\langle \vec{v}_1,\vec{v}_2\rangle)^2>0\right)$), et le problème $(\mathcal{P})$ a une et une seule solution $(\overline{t}_1,\overline{t}_2)$. Ce $(\overline{t}_1,\overline{t}_2)$ est l'unique solution de $(\mathcal{S})$:

$$\overline{t}_1 = \frac{-\langle \overrightarrow{A_2A_1},\vec{v}_1\rangle\parallel \vec{v}_2\parallel^2+\langle \overrightarrow{A_2A_1},\vec{v}_2\rangle\langle \vec{v}_1,\vec{v}_2\rangle}{\parallel \vec{v}_1\parallel^2\parallel \vec{v}_2\parallel^2-(\langle \vec{v}_1,\vec{v}_2\rangle)^2},$$

$$\overline{t}_2 = \frac{\langle \overrightarrow{A_2A_1},\vec{v}_2\rangle\parallel \vec{v}_1\parallel^2-\langle \overrightarrow{A_2A_1},\vec{v}_1\rangle\langle \vec{v}_1,\vec{v}_2\rangle}{\parallel \vec{v}_1\parallel^2\parallel \vec{v}_2\parallel^2-(\langle \vec{v}_1,\vec{v}_2\rangle)^2}.$$

Et on vérifie que $\overrightarrow{\overline{M}_1\overline{M}_2} = \overrightarrow{A_1A_2} + \overline{t}_2\vec{v}_2 - \overline{t}_1\vec{v}_1$ est orthogonal à $\vec{v}_1$ et $\vec{v}_2$: la droite (M_1M_2) est la « perpendiculaire commune » à $\mathcal{D}_1$ et $\mathcal{D}_2$.

La distance minimale entre deux points de $\mathcal{D}_1$ et $\mathcal{D}_2$ se déduit donc de $f\left(\overline{t}_1, \overline{t}_2\right) = \| \overrightarrow{\overline{M}_1\overline{M}_2} \|^2$.

Dans le cas particulier où $n = 3$,

$$\| \overrightarrow{\overline{M}_1\overline{M}_2} \| = \frac{\left| \left[\overrightarrow{A_1A_2}, \vec{v}_1, \vec{v}_2 \right] \right|}{\| \vec{v}_1 \wedge \vec{v}_2 \|},$$

où $\left[\overrightarrow{A_1A_2}, \vec{v}_1, \vec{v}_2\right]$ désigne le produit mixte de $\overrightarrow{A_1A_2}, \vec{v}_1, \vec{v}_2$, et $\vec{v}_1 \wedge \vec{v}_2$ le produit vectoriel de $\vec{v}_1$ et $\vec{v}_2$.

***Exercice II.3.** *Condition nécessaire de minimalité à ε près d'Ekeland*

Soit $f : \mathbb{R}^n \to \mathbb{R}$ continue et bornée inférieurement sur $\mathbb{R}^n$. Soit $\varepsilon > 0$ et u une solution à ε près du problème de minimisation de f sur $\mathbb{R}^n$, c'est-à-dire vérifiant $f(u) \leqslant \inf_{x \in \mathbb{R}^n} f(x) + \varepsilon$. Étant donné $\lambda > 0$ on considère

$$g : x \in \mathbb{R}^n \longmapsto g(x) := f(x) + \frac{\varepsilon}{\lambda} \| x - u \| .$$

1°) Montrer qu'il existe $v \in \mathbb{R}^n$ minimisant g sur $\mathbb{R}^n$.

Montrer que ce point v vérifie les conditions ci-après :

(i) $\quad f(v) \leqslant f(u)$;
(ii) $\quad \| v - u \| \leqslant \lambda$;
(iii) $\quad \forall x \in \mathbb{R}^n,\ f(v) \leqslant f(x) + \frac{\varepsilon}{\lambda} \| x - v \|$.

2°) Soit $f : \mathbb{R}^n \to \mathbb{R}$ différentiable et bornée inférieurement sur $\mathbb{R}^n$. Montrer que pour tout $\varepsilon > 0$ il existe x_ε tel que $\| \nabla f(x_\varepsilon) \| \leqslant \varepsilon$.

Solution : 1°) g est continue ; de plus, f étant bornée inférieurement, $g(x) \to +\infty$ quand $\| x \| \to +\infty$. Par conséquent, il existe un point v minimisant g sur $\mathbb{R}^n$:

$$\forall x \in \mathbb{R}^n, \quad f(v) + \frac{\varepsilon}{\lambda} \| v - u \| \leqslant f(x) + \frac{\varepsilon}{\lambda} \| x - u \| . \tag{2.1}$$

Faisons $x = u$ dans (2.1) :

$$f(v) + \frac{\varepsilon}{\lambda} \| v - u \| \leqslant f(u), \tag{2.2}$$

d'où $f(v) \leqslant f(u)$.

En notant $\overline{f} := \inf_{x\in\mathbb{R}^n} f(x)$, il vient de (2.2) :

$$\overline{f} + \frac{\varepsilon}{\lambda} \parallel v-u \parallel \leqslant \overline{f} + \varepsilon,$$

d'où $\parallel v-u \parallel \leqslant \lambda$.

Enfin l'inégalité $\parallel x-u \parallel \leqslant \parallel x-v \parallel + \parallel v-u \parallel$, introduite dans (2.1), conduit à

$$f(v) \leqslant f(x) + \frac{\varepsilon}{\lambda} \parallel x-v \parallel .$$

2°) Partant d'un minimum à ε^2 près de f et choisissant $\lambda = \varepsilon$, il existe d'après la question précédente un x_ε tel que

$$\forall x \in \mathbb{R}^n,\ f(x_\varepsilon) \leqslant f(x) + \varepsilon \parallel x - x_\varepsilon \parallel .$$

Pour $d \in \mathbb{R}^n$ et $\alpha > 0$ faisons successivement $x = x_\varepsilon + \alpha d$ et $x = x_\varepsilon - \alpha d$ dans l'inégalité précédente ; on obtient :

$$f(x_\varepsilon + \alpha d) - f(x_\varepsilon) \geqslant -\varepsilon\alpha \parallel d \parallel, \quad \text{soit } \frac{f(x_\varepsilon + \alpha d) - f(x_\varepsilon)}{\alpha} \geqslant -\varepsilon \parallel d \parallel ;$$

$$f(x_\varepsilon - \alpha d) - f(x_\varepsilon) \geqslant -\varepsilon\alpha \parallel d \parallel, \quad \text{soit } \frac{f(x_\varepsilon - \alpha d) - f(x_\varepsilon)}{\alpha} \geqslant -\varepsilon \parallel d \parallel .$$

Un passage à la limite $\alpha \to 0^+$ induit :

$$\langle\, \nabla f(x_\varepsilon), d \,\rangle \geqslant -\varepsilon \parallel d \parallel \text{ et } \langle\, \nabla f(x_\varepsilon), -d \,\rangle \geqslant -\varepsilon \parallel d \parallel,$$

soit encore $|\langle\, \nabla f(x_\varepsilon), d \,\rangle| \leqslant \varepsilon \parallel d \parallel$. Cette dernière inégalité étant vraie pour tout $d \in \mathbb{R}^n$, il s'ensuit $\parallel \nabla f(x_\varepsilon) \parallel \leqslant \varepsilon$.

***Exercice II.4.** Soit $n \geqslant 2$ et $f : \mathbb{R}^n \to \mathbb{R}$ la fonction polynomiale de degré cinq définie par

$$x = (x_1, \ldots, x_n) \in \mathbb{R}^n \longmapsto f(x) := (1+x_n)^3 \sum_{i=1}^{n-1} x_i^2 + x_n^2.$$

Montrer que $0\,(\in \mathbb{R}^n)$ est le seul point critique de f, qu'il est minimum local strict de f, mais qu'il n'est pas minimum global de f.

Solution : On a :

$$\partial_i f(x) = 2x_i(1+x_n)^3 \quad \text{pour tout } i = 1, \dots, n-1 \ ;$$
$$\partial_n f(x) = 3(1+x_n)^2 \sum_{i=1}^{n-1} x_i^2 + 2x_n.$$

Ainsi le seul point $\overline{x} \in \mathbb{R}^n$ vérifiant $\nabla f(\overline{x}) = 0$ est $\overline{x} = 0$.

Ce point $\overline{x} = 0$ est un minimum local strict de f car $\nabla^2 f(\overline{x}) = \operatorname{diag}(2, \dots, 2)$ est définie positive.

Mais $f(1, \dots, 1, x_n) = (n-1)(1+x_n)^3 + x_n^2$ est une fonction polynomiale du 3$^{\text{e}}$ degré, prenant donc toutes les valeurs réelles. Il n'y a donc pas de minimum global de f sur $\mathbb{R}^n$.

****** Exercice II.5.*** Soit $f : \mathbb{R}^n \to \mathbb{R}$ continue et $\overline{x} \in \mathbb{R}^n$. Montrer :

$$(\overline{x} \text{ est minimum global de } f) \iff \begin{pmatrix} \text{Tout } x \text{ tel que } f(x) = f(\overline{x}) \\ \text{est un minimum local de } f \end{pmatrix}.$$

Solution : L'implication « $\Rightarrow$ » étant évidente, considérons l'implication inverse. Supposons donc que tout x au même niveau que $\overline{x}$ soit un minimum local de f. Si $\overline{x}$ n'est pas un minimum global de f, c'est qu'il existe $u \in \mathbb{R}^n$ tel que $f(u) < f(\overline{x})$. Définissons $\varphi : t \in [0,1] \longmapsto \varphi(t) := f(t\overline{x} + (1-t)u)$ et posons $S_\varphi := \{t \in [0,1] \mid \varphi(t) = f(\overline{x})\}$. Ainsi S_φ est une partie non vide ($1 \in S_\varphi$) fermée (de par la continuité de f donc de φ) et minorée. Désignons par t_0 la borne inférieure de S_φ : $t_0 \in S_\varphi$ et $0 < t_0 \leqslant 1$.

Le point $t_0\overline{x} + (1-t_0)u$ étant au même niveau que $\overline{x}$, il est minimum local de f ; par conséquent t_0 est un minimum local de φ : il existe $\eta > 0$ tel que :

$$\begin{pmatrix} |t - t_0| \leqslant \eta \\ t \in [0,1] \end{pmatrix} \Rightarrow (\varphi(t) \geqslant \varphi(t_0)).$$

Comme t_0 est la borne inférieure de S_φ, on en déduit :

$$\begin{pmatrix} t_0 - \eta < t < t_0 \\ 0 < t \end{pmatrix} \Rightarrow (\varphi(t) > \varphi(t_0)).$$

Prenons un tel t, que nous appelons t_1 :

$$t_0 - \eta < t_1 < t_0, \ 0 < t_1, \ \text{et } \varphi(t_1) > \varphi(t_0).$$

Alors $\varphi(1) = \varphi(t_0) \in [\varphi(0), \varphi(t_1)]$ et, d'après le théorème des valeurs intermédiaires, il existe $t_2 \in [0, t_1]$ tel que $\varphi(t_2) = \varphi(1)$ $(= f(\overline{x}))$. Par conséquent, $t_2 \in]0, t_1] \cap S_\varphi$, et puisque $t_1 < t_0$, ceci contredit la définition de t_0 comme borne inférieure de S_φ.

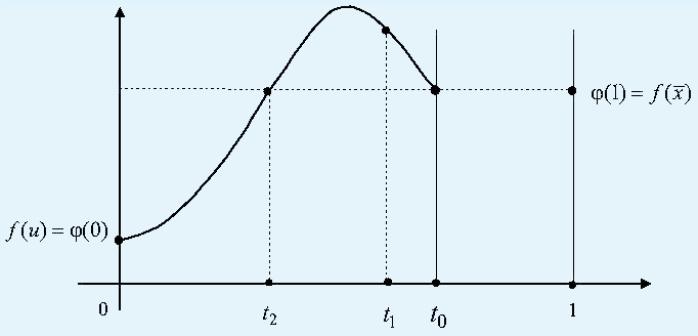


Figure 5.

Remarque : Dans l'assertion de droite de l'énoncé, on ne peut pas remplacer « minimum local de f » par « point critique de f » ; prendre l'exemple de $f : x \in \mathbb{R} \longmapsto f(x) = x^3$.

****** Exercice II.6.*** On désigne par $\overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ l'ensemble des matrices définies positives de taille n (c'est un ouvert de $\mathcal{S}_n(\mathbb{R})$). Étant donnés A et B dans $\overset{\circ}{\mathcal{P}}_n(\mathbb{R})$, on considère le problème de minimisation suivant :

$$(\mathcal{P}) \qquad \begin{cases} \text{Minimiser } f(X) := \operatorname{tr}(AX) + \operatorname{tr}\left(BX^{-1}\right) \\ X \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R}). \end{cases}$$

1°) Quelles propriétés de f, utiles pour sa minimisation (différentiabilité, convexité, coercivité) peut-on dégager ? En déduire que le problème $(\mathcal{P})$ a une et une seule solution.

2°) (i) Vérifier que la solution de $(\mathcal{P})$ est la (seule) matrice $\overline{X}$ vérifiant $\overline{X}A\overline{X} = B$.

(ii) Donner, à partir de BA (ou de AB), une procédure permettant de construire $\overline{X}$.

(iii) En déduire la valeur optimale dans $(\mathcal{P})$.

3°) On fait $n = 2$ ici. Montrer que la valeur optimale dans $(\mathcal{P})$ est $2\sqrt{\operatorname{tr}(AB) + 2\sqrt{\det(AB)}}$.

Indication. Le cas simple où $n = 1$ permet de guider la démarche et de contrôler les résultats.

Solution : $\mathcal{S}_n(\mathbb{R})$ est structuré en espace euclidien grâce au produit scalaire $\ll U, V \gg = \operatorname{tr}(UV)$.

Le problème $(\mathcal{P})$ est celui de la minimisation sur (le cône convexe ouvert) $\overset{\circ}{\mathcal{P}}_n(\mathbb{R}) =: \Omega$ de la fonction-objectif $X \longmapsto f(X) = \ll A, X \gg + \ll B, X^{-1} \gg$.

1°) L'application $X \in \Omega \longmapsto X^{-1} \in \Omega$ est bijective et de classe C^∞ ; il s'ensuit que f est C^∞ sur Ω. De l'inégalité de convexité

$$\begin{pmatrix} U, V \quad \text{dans} \quad \Omega \\ \alpha \in]0,1[\end{pmatrix} \Rightarrow \left([\alpha U + (1-\alpha) V]^{-1} \preceq \alpha U^{-1} + (1-\alpha) V^{-1} \right),$$

et de la (semi-) définie positivité de B, on déduit que $X \longmapsto \ll B, X^{-1} \gg$ est convexe sur Ω (d'accord ?).

De manière à concentrer sur une seule partie toute la contribution de A et B aux deux parties de $f(X)$, posons $C := A^{1/2} B A^{1/2}$. Comme

$$\begin{aligned} f(X) &= \operatorname{tr}\left[A^{1/2} \left(A^{1/2} X A^{1/2} \right) A^{-1/2} + C \left(A^{1/2} X A^{1/2} \right)^{-1} \right] \\ &= \operatorname{tr}\left[A^{1/2} X A^{1/2} + C \left(A^{1/2} X A^{1/2} \right)^{-1} \right] \end{aligned}$$

et que l'application $X \in \Omega \longmapsto Y := \left(A^{1/2} X A^{1/2} \right)^{-1} \in \Omega$ est visiblement une bijection de Ω sur Ω, le problème $(\mathcal{P})$ est équivalent à celui de la minimisation de

$$Y \longmapsto g(Y) := \operatorname{tr}\left(Y^{-1} + CY \right) = \ll I_n, Y^{-1} \gg + \ll C, Y \gg$$

sur Ω.

La frontière $\operatorname{fr}\Omega$ de Ω est exactement l'ensemble des matrices semi-définies positives singulières (*i.e.*, dont la plus petite valeur propre est nulle). En conséquence :

$$\begin{pmatrix} Y \in \Omega \longrightarrow Y_0 \in \operatorname{fr}\Omega \\ (\text{ou } X \in \Omega \longrightarrow X_0 \in \operatorname{fr}\Omega) \end{pmatrix} \Longrightarrow \begin{pmatrix} g(Y) \longrightarrow +\infty \\ (\text{ou } f(X) \longrightarrow +\infty) \end{pmatrix}.$$

De même, un calcul (pas trop difficile) sur les matrices définies positives conduit à élucider le comportement à l'infini :

$$\begin{pmatrix} Y \in \Omega, \quad \| Y \| \longrightarrow \infty \\ (\text{ou } X \in \Omega, \quad \| X \| \longrightarrow \infty) \end{pmatrix} \Longrightarrow \begin{pmatrix} g(Y) \longrightarrow +\infty \\ (\text{ou } f(X) \longrightarrow +\infty) \end{pmatrix}.$$

Ainsi g (ou f) agit comme ce qu'il est convenu d'appeler une « fonction-barrière » sur Ω. Il en résulte ainsi qu'il existe bien $\overline{X} \in \Omega$ minimisant f sur Ω (*cf.* Commentaire suivant l'Exercice II.1).

L'unicité de $\overline{X}$ est une conséquence de la stricte convexité de la fonction $X \in \Omega \longmapsto \ll B, X^{-1} \gg$, qui se voit mieux *via* la stricte convexité de $Y \in \Omega \longmapsto \ll I_n, Y^{-1} \gg = \mathrm{tr}(Y^{-1})$.

2°) (*i*) La différentielle $Df(X)$ de f en $X \in \Omega$ est facile à obtenir :

$$Df(X) : H \in \mathcal{S}_n(\mathbb{R}) \longmapsto \ll A - X^{-1}BX^{-1}, H \gg .$$

Le problème (essentiellement sans contraintes) $(\mathcal{P})$ étant celui de la minimisation d'une fonction convexe différentiable, $\overline{X} \in \Omega$ est solution de $(\mathcal{P})$ si et seulement si $A - \overline{X}^{-1}B\overline{X}^{-1} = 0$, soit $\overline{X}A\overline{X} = B$.

(*ii*) Étant le produit de deux matrices définies positives, BA est diagonalisable et son spectre est constitué exclusivement de réels > 0 (*cf.* Exercice 1.17). Si on diagonalise BA avec P,

$$P^{-1}(BA)P = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) \quad (\lambda_i \text{ valeurs propres de } BA),$$

la matrice $M := P\mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})P^{-1}$ vérifie $M^2 = BA$, et il vient immédiatement que $\overline{X} = M^{-1}B$ est la solution cherchée.

(*iii*) La valeur optimale dans $(\mathcal{P})$ est

$$\begin{aligned} f(\overline{X}) &= \mathrm{tr}(AM^{-1}B + B(M^{-1}B)^{-1}) \\ &= 2\ \mathrm{tr}M; \end{aligned}$$

c'est donc la somme des racines carrées des valeurs propres de BA (ou de AB).

La symétrie en A et B de la valeur optimale pouvait être notée dès le départ, en raison de la forme particulière de $(\mathcal{P})$.

3°) De l'égalité $\sqrt{\lambda_1} + \sqrt{\lambda_2} = \sqrt{\lambda_1 + \lambda_2 + 2\sqrt{\lambda_1\lambda_2}}$ écrite pour λ_1, λ_2, valeurs propres de BA, il vient

$$\sqrt{\lambda_1} + \sqrt{\lambda_2} = \mathrm{tr}M = \sqrt{\mathrm{tr}(BA) + 2\sqrt{\mathrm{dét}(BA)}},$$

d'où l'expression annoncée.

***Exercice II.7.** Soit $g : \mathbb{R}^n \longrightarrow \mathbb{R}$ une fonction convexe et différentiable sur $\mathbb{R}^n$, soit $h : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{+\infty\}$ une fonction convexe sur $\mathbb{R}^n$, finie en au moins un point. On pose $f := g + h$ et on considère le problème de la minimisation de f sur $\mathbb{R}^n$.

1°) Montrer que $\overline{x}$ minimise f sur $\mathbb{R}^n$ si et seulement si

$$< \nabla g(\overline{x}), x - \overline{x} > \ + h(x) - h(\overline{x}) \geqslant 0 \text{ pour tout } x \in \mathbb{R}^n. \tag{2.3}$$

2°) Vérifier que (2.3) est équivalente à

$$< \nabla g(x), x - \overline{x} > \ + h(x) - h(\overline{x}) \geqslant 0 \text{ pour tout } x \in \mathbb{R}^n. \tag{2.4}$$

Solution : 1°) Soit $\overline{x}$ minimisant f sur $\mathbb{R}^n$. Étant donné $x \in \mathbb{R}^n$ et $\alpha \in]0, 1[$, on a $f(\alpha x + (1-\alpha)\overline{x}) \geqslant f(\overline{x})$, ce qui se traduit par

$$g(\overline{x}) + h(\overline{x}) \leqslant g(\alpha x + (1-\alpha)\overline{x}) + h(\alpha x + (1-\alpha)\overline{x}).$$

En utilisant l'inégalité de convexité $h(\alpha x + (1-\alpha)\overline{x}) \leqslant \alpha h(x) + (1-\alpha)h(\overline{x})$ et en divisant par $\alpha > 0$, on obtient

$$0 \leqslant \frac{g(\overline{x} + \alpha(x - \overline{x})) - g(\overline{x})}{\alpha} + h(x) - h(\overline{x}).$$

En faisant tendre α vers 0, on obtient précisément (2.3).

Réciproquement, soit $\overline{x}$ vérifiant (2.3). La convexité de g induit

$$g(x) \geqslant g(\overline{x}) + \langle \nabla g(\overline{x}), x - \overline{x} \rangle,$$

ce qui, combiné avec (2.3), entraîne $g(x) + h(x) \geqslant g(\overline{x}) + h(\overline{x})$.

2°) Soit $\overline{x}$ vérifiant (2.3). La fonction g étant convexe, l'application ∇g vérifie : $\langle \nabla g(x) - \nabla g(\overline{x}), x - \overline{x} \rangle \geqslant 0$. D'où (2.4) s'ensuit.

Soit à présent $\overline{x}$ vérifiant (2.4). Considérons $x \in \mathbb{R}^n$ et $\alpha \in]0, 1[$. Il vient alors de (2.4) :

$$\alpha \langle \nabla g((1-\alpha)\overline{x} + \alpha x), x - \overline{x} \rangle + h\left((1-\alpha)\overline{x} + \alpha x\right) - h(\overline{x}) \geqslant 0,$$

ce qui, avec l'inégalité de convexité de h déjà vue plus haut, donne

$$\alpha \langle \nabla g((1-\alpha)\overline{x} + \alpha x), x - \overline{x} \rangle + \alpha[h(x) - h(\overline{x})] \geqslant 0.$$

Divisons par α et faisons tendre α vers 0 ; sachant que ∇g est continue (d'accord ?), on obtient précisément (2.3).

Remarque : Dans le cas particulier où C est un convexe fermé non vide et h la fonction indicatrice de C ($h(x) = 0$ si $x \in C$, $+\infty$ sinon), ce que dit le résultat de l'exercice est :

$$\begin{aligned}(\overline{x} \text{ minimise } g \text{ sur } C) &\Leftrightarrow (\langle \nabla g(\overline{x}), x - \overline{x}\rangle \geqslant 0 \text{ pour tout } x \in C)\\ &\quad (\text{ caractérisation classique})\\ &\Leftrightarrow (\langle \nabla g(x), x - \overline{x}\rangle \geqslant 0 \text{ pour tout } x \in C)\\ &\quad (\text{caractérisation nouvelle}).\end{aligned}$$

***Exercice II.8.** Soit $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ définie par

$$x \in \mathbb{R}^n \longmapsto f(x) := \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle + c,$$

où $A \in \mathcal{S}_n(\mathbb{R})$ est supposée définie positive, $b \in \mathbb{R}^n$ et $c \in \mathbb{R}$. On considère le problème d'optimisation suivant :

$$(\mathcal{P}) \quad \text{Minimiser } f(x),\ x \in \mathbb{R}^n.$$

1°) Rappeler pourquoi $(\mathcal{P})$ a une et une seule solution $\overline{x}$, caractérisée comme étant l'unique solution de l'équation $\nabla f(x) = 0$.

Pour approcher $\overline{x}$, on utilise l'algorithme du gradient à pas optimal, qui consiste à construire une suite (x_k) de la manière itérative suivante :

Initialisation : $x_0 \in \mathbb{R}^n$;

Définition de l'itéré x_{k+1} à partir de x_k (lorsque $\nabla f(x_k) \neq 0$) :

$$x_{k+1} := x_k + t_k d_k,$$

où $d_k := -\nabla f(x_k)$, et t_k est l'unique réel (positif) minimisant $t \longmapsto f(x_k + t d_k)$ sur $\mathbb{R}$.

Le but de l'exercice est de donner une idée de la vitesse de convergence de (x_k) vers $\overline{x}$ en fonction d'un réel associé à A appelé conditionnement de A.

2°) Vérifier les relations suivantes : Pour tout $k \in \mathbb{N}$,

$$t_k = \frac{\| d_k \|^2}{\langle Ad_k, d_k\rangle},\ d_{k+1} = d_k - t_k A d_k,\quad \langle d_{k+1}, d_k\rangle = 0,$$

$$f(x_{k+1}) - \overline{f} = [f(x_k) - \overline{f}]\left[1 - \frac{\| d_k \|^4}{\langle Ad_k, d_k\rangle\ \langle A^{-1}d_k, d_k\rangle}\right], \tag{2.5}$$

où $\overline{f}$ désigne la valeur optimale dans $(\mathcal{P})$.

3°) Soit $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$ les valeurs propres de A rangées dans l'ordre décroissant. On pose

$$c_2(A) := \frac{\lambda_1}{\lambda_n}, \text{ appelé } \textit{conditionnement de } A.$$

Démontrer les inégalités suivantes : Pour tout $k \in \mathbb{N}$,

$$f(x_k) - \overline{f} \leqslant [f(x_0) - \overline{f}] \left[\frac{c_2(A) - 1}{c_2(A) + 1}\right]^{2k}, \tag{2.6}$$

$$\| x_k - \overline{x} \| \leqslant \left[\frac{2(f(x_0) - \overline{f})}{\lambda_n}\right]^{1/2} \left[\frac{c_2(A) - 1}{c_2(A) + 1}\right]^{k}. \tag{2.7}$$

On pourra pour cela utiliser librement l'inégalité de Kantorovitch que voici :

$$\forall x \in \mathbb{R}^n, \quad \langle Ax, x\rangle \, \langle A^{-1}x, x\rangle \leqslant \frac{1}{4}\left[\sqrt{\frac{\lambda_1}{\lambda_n}} + \sqrt{\frac{\lambda_n}{\lambda_1}}\right]^2 \| x \|^4 .$$

Quels commentaires peut-on faire à partir des inégalités (2.6) et (2.7) quant à la rapidité de convergence de la méthode de gradient à pas optimal ?

Solution : 1°) f est quadratique strictement convexe, 1-coercive sur $\mathbb{R}^n$: il existe donc un et un seul $\overline{x}$ minimisant f sur $\mathbb{R}^n$. La différentiabilité de f alliée à sa convexité assurent que $\overline{x}$ est l'unique solution de l'équation $\nabla f(x) = 0$.

Si on veut préciser en fonction des données de $(\mathcal{P})$, $\overline{x} = -A^{-1}b$ et la valeur optimale est $\overline{f} = f(\overline{x}) = -\frac{1}{2}\langle A^{-1}b, b\rangle + c$.

2°) Comme $f(x_k + td_k) = f(x_k) + \frac{1}{2}t^2\langle Ad_k, d_k\rangle + t\langle Ax_k + b, d_k\rangle$, la fonction $t \in \mathbb{R} \longmapsto f(x_k + td_k)$ est minimisée sur $\mathbb{R}$ (lorsque $d_k = -\nabla f(x_k) \neq 0$) en un seul point, qui est $t_k = \dfrac{\| d_k \|^2}{\langle Ad_k, d_k\rangle}$ (> 0).

Ensuite :

$$\begin{aligned} d_{k+1} &= -(Ax_{k+1} + b) = -Ax_k - b - t_k Ad_k = d_k - t_k Ad_k, \\ \langle d_{k+1}, d_k\rangle &= \langle d_k, d_k\rangle - t_k \, \langle Ad_k, d_k\rangle \; = 0. \end{aligned}$$

En développant $f(x_{k+1}) = f(x_k + t_k d_k)$ on obtient

$$f(x_{k+1}) = f(x_k) - \frac{1}{2}\frac{\| d_k \|^4}{\langle Ad_k, d_k\rangle},$$

soit encore

$$f(x_{k+1}) - \overline{f} = [f(x_k) - \overline{f}] \left[1 - \frac{\| d_k \|^4}{2(f(x_k) - \overline{f})\ \langle Ad_k, d_k\rangle}\right]$$

(toujours sous l'hypothèse $\nabla f(x_k) \neq 0$, qui est équivalente à $f(x_k) - \overline{f} > 0$).

Un simple calcul à présent montre que

$$\begin{aligned}\langle A^{-1}d_k, d_k\rangle &= \langle A^{-1}(Ax_k + b), Ax_k + b\rangle \\ &= 2\left[\frac{1}{2}\langle Ax_k, x_k\rangle \ + \ \langle b, x_k\rangle + \frac{1}{2}\langle A^{-1}b, b\rangle\right] \\ &= 2\left[f(x_k) - \overline{f}\right] \qquad \left(\text{car } \ \frac{1}{2}\langle A^{-1}b, b\rangle \ = c - \overline{f}\right).\end{aligned}$$

D'où l'expression (2.5).

3°) De l'inégalité de Kantorovitch on déduit (toujours lorsque $d_k \neq 0$) :

$$\frac{\| d_k \|^4}{\langle Ad_k, d_k\rangle\ \langle A^{-1}d_k, d_k\rangle} \geqslant 4\left[\sqrt{\frac{\lambda_1}{\lambda_n}} + \sqrt{\frac{\lambda_n}{\lambda_1}}\right]^{-2} = 4\frac{\lambda_1/\lambda_n}{(\lambda_1/\lambda_n + 1)^2}.$$

Ainsi, d'après (2.5) :

$$\begin{aligned}\forall k \in \mathbb{N}, \quad f(x_{k+1}) - \overline{f} &\leqslant [f(x_k) - \overline{f}]\left[1 - 4\frac{c_2(A)}{(c_2(A)+1)^2}\right] \\ &\leqslant [f(x_k) - \overline{f}]\left[\frac{c_2(A) - 1}{c_2(A) + 1}\right]^2.\end{aligned}$$

L'inégalité (2.6) s'ensuit alors facilement.

Par ailleurs on a :

$$\begin{aligned}f(x_k) - \overline{f} &= \frac{1}{2}\langle Ax_k, x_k\rangle + \langle b, x_k\rangle + c - \overline{f} \\ &= \frac{1}{2}\langle A(x_k - \overline{x}), x_k - \overline{x}\rangle \qquad \begin{pmatrix}\text{car } \ A\overline{x} = -b \\ \text{et } c - \overline{f} = \frac{1}{2}\langle A^{-1}b, b\rangle\end{pmatrix} \\ &\geqslant \frac{1}{2}\lambda_n \| x_k - \overline{x} \|^2 \ ;\end{aligned}$$

d'où (2.7).

D'après (2.6) et (2.7), plus $c_2(A)$ est proche de 1, plus la méthode (du gradient à pas optimal) converge rapidement. Le cas limite serait celui où $c_2(A) = 1$, ce qui suppose $f(x) = \lambda \| x - \overline{x} \|^2$ pour un certain $\lambda > 0$, auquel cas on atteint $\overline{x}$ dès la 1re itération x_1.

Par contre, lorsque $c_2(A)$ est grand, c'est-à-dire lorsque les valeurs propres extrêmes sont très différentes, la méthode est (dans le pire des cas) très lente.

Lorsque $n = 2$, les courbes de niveau de f sont alors des ellipses très effilées, et on observe facilement la convergence lente en zigzags de (x_k) vers $\overline{x}$.

Pour être sûr d'avoir $\frac{f(x_k)-\overline{f}}{f(x_0)-\overline{f}} \leqslant \varepsilon$, il faudrait

$$k \geqslant \frac{\ln \varepsilon}{2 \ln \left(\frac{c_2(A)-1}{c_2(A)+1}\right)} \sim \frac{c_2(A)}{4} \ln\left(\frac{1}{\varepsilon}\right) \quad (\text{quand } c_2(A) \longrightarrow +\infty).$$

***Exercice II.9.** Soit $\mathcal{O}$ un ouvert de $\mathbb{R}^n$ et $f : \mathcal{O} \longrightarrow \mathbb{R}$ quatre fois différentiable en $\overline{x} \in \mathcal{O}$. On suppose que $\overline{x}$ est un minimum local de f, ce qui implique :

$$\nabla f(\overline{x}) = 0, \ \nabla^2 f(\overline{x}) \text{ est semi-définie positive.}$$

On suppose également que $\nabla^2 f(\overline{x})$ n'est pas nulle et on pose $H := \operatorname{Ker} \nabla^2 f(\overline{x})$.

1°) Montrer que l'on a nécessairement :

(a) $\nabla^3 f(\overline{x})(u,u,u) = 0$ pour tout $u \in H$;

(b) $\nabla^4 f(\overline{x})(u,u,u,u) \cdot \langle \nabla^2 f(\overline{x})v, v\rangle - 3\left[\nabla^3 f(\overline{x})(u,u,v)\right]^2 \geqslant 0$ pour tout $(u,v) \in H \times H^\perp$.

2°) On prend l'exemple d'une fonction f de deux variables, où $\overline{x} = (0,0)$ est un point critique de f et où $\nabla^2 f(\overline{x}) = \begin{bmatrix} 0 & 0 \\ 0 & \lambda \end{bmatrix}$ avec $\lambda = \dfrac{\partial^2 f}{\partial x_2^2}(\overline{x}) > 0$.

Quelles formes prennent les conditions (a) et (b) dans ce cas ?

Application. Soit $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$

$$(x_1, x_2) \longmapsto f(x_1, x_2) := 3x_1^4 - 4x_1^2 x_2 + x_2^2.$$

Au vu de ce qui a été établi, décider si $\overline{x} = (0,0)$ est un minimum local de f ou non.

Commentaire : Le cas où $\nabla f(\overline{x}) = 0$ et $\nabla^2 f(\overline{x})$ est semi-définie positive est ce qu'on peut appeler « cas d'incertitude », car les conditions de minimalité du second ordre ne permettent pas de décider. Les conditions (a) et (b) considérées ici sont des *conditions nécessaires de minimalité d'ordre trois et quatre.* Comme bien entendu, on s'attend à des conditions suffisantes de minimalité locale en remplaçant l'inégalité au sens large de (b) par une inégalité stricte (pour tous les vecteurs non nuls u et v).

En règle générale, les conditions de minimalité d'ordre supérieur (à deux) sont difficiles à utiliser car, contrairement à $\nabla^2 f(\overline{x})$, il n'y a pas de « théorie spectrale » simple qui va avec les formes 2p-linéaires $\nabla^{2p} f(\overline{x})$.

Indication. $\nabla^3 f(\overline{x})$ (resp. $\nabla^4 f(\overline{x})$) désigne la forme trilinéaire (resp. la forme quadrilinéaire) différentielle d'ordre trois (resp. d'ordre quatre) de f en $\overline{x}$. Exprimer f en $\overline{x}+tu+t^2v$ à l'aide du développement de Taylor-Young à l'ordre quatre de f en $\overline{x}$.

Solution : 1°) Étant donnés $u \in H$, $v \in \mathbb{R}^n$ et $t \neq 0$, exprimons f en $\overline{x}+tu+t^2v = \overline{x}+t(u+tv)$ à l'aide du développement de Taylor-Young à l'ordre quatre de f en $\overline{x}$. Sachant que $\nabla f(\overline{x}) = 0$ et $\nabla^2 f(\overline{x})u = 0$, on obtient :

$$f(\overline{x}+tu+t^2v) = f(\overline{x}) + \tfrac{t^3}{6}\nabla^3 f(\overline{x})(u,u,u) + \tfrac{t^4}{24}\left[12\langle\nabla^2 f(\overline{x})v,v\rangle + 12\nabla^3 f(\overline{x})(u,u,v) + \nabla^4 f(\overline{x})(u,u,u,u)\right] + t^4\varepsilon(t), \tag{2.8}$$

où $\varepsilon(t) \longrightarrow 0$ quand $t \longrightarrow 0$.

Puisque $f(\overline{x}+tu+t^2v) \geqslant f(\overline{x})$ pour t suffisamment petit, il vient du développement ci-dessus

$$\left[f(\overline{x}+tu+t^2v) - f(\overline{x})\right] / \frac{t^3}{6} \underset{\to 0}{\longrightarrow} \nabla^3 f(\overline{x})(u,u,u) \geqslant 0.$$

La même inégalité étant valable en $-u$ $(\in H)$,

$$0 \leqslant \nabla^3 f(\overline{x})(-u,-u,-u) = -\nabla^3 f(\overline{x})(u,u,u),$$

on en déduit (a).

Ensuite, en procédant de la même manière pour le bloc facteur de $t^4/24$ dans (2.8), on arrive à :

$$12\langle\nabla^2 f(\overline{x})v,v\rangle + 12\nabla^3 f(\overline{x})(u,u,v) + \nabla^4 f(\overline{x})(u,u,u,u) \geqslant 0 \text{ pour tout } v \in \mathbb{R}^n.$$

Écrivons cette inégalité en $\alpha v, \alpha \in \mathbb{R}$; il vient :

$$\begin{cases} 12\alpha^2\langle\nabla^2 f(\overline{x})v,v\rangle + 12\alpha\nabla^3 f(\overline{x})(u,u,v) + \nabla^4 f(\overline{x})(u,u,u,u) \geqslant 0 \\ \text{pour tout } \alpha \in \mathbb{R},\ v \in \mathbb{R}^n. \end{cases} \tag{2.9}$$

Choisissons $v \neq 0$ dans $H^\perp$; alors la positivité du trinôme du second degré dans (2.9) se traduit en disant que son discriminant est négatif, d'où (b).

2°) Les conditions (a) et (b) prennent ici les formes suivantes :

$$\frac{\partial^3 f}{\partial x_1^3}(\overline{x}) = 0 \quad \text{et} \quad \frac{\partial^2 f}{\partial x_2^2}(\overline{x}) \cdot \frac{\partial^4 f}{\partial x_1^4}(\overline{x}) - 3\left(\frac{\partial^3 f}{\partial x_1^2 \partial x_2}(\overline{x})\right)^2 \geqslant 0.$$

Dans l'exemple proposé, la deuxième condition ci-dessus n'est pas vérifiée, et donc $(0,0)$ ne saurait être un minimum local de f.

***Exercice II.10.** Soit $\mathcal{O}$ un ouvert convexe de $\mathbb{R}^n$ et $f : \mathcal{O} \to \mathbb{R}$ deux fois différentiable sur $\mathcal{O}$. On suppose que ∇f et $\nabla^2 f$ sont lipschitziennes sur $\mathcal{O}$ de constante de Lipschitz L, c'est-à-dire :

$$\| \nabla f(x) - \nabla f(x') \| \leqslant L \| x - x' \| \text{ et } |||\nabla^2 f(x) - \nabla^2 f(x')||| \leqslant L \| x - x' \|$$

pour tout x, x' dans $\mathcal{O}$, où $\| \cdot \|$ désigne la norme euclidienne usuelle sur $\mathbb{R}^n$ et $||| \cdot |||$ désigne la norme sur $\mathcal{M}_n(\mathbb{R})$ déduite du produit scalaire $\ll \cdot, \cdot \gg$ $\left(|||A|||^2 = \ll A, A \gg = \operatorname{tr}\left(A^T A\right)\right)$.

1°) Étant donné $x_k \in \mathcal{O}$, on pose

$$x \in \mathcal{O} \longmapsto \theta^1_{x_k}(x) := f(x_k) + \langle \nabla f(x_k), x - x_k \rangle,$$
$$x \in \mathcal{O} \longmapsto \theta^2_{x_k}(x) := f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2}\langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle .$$

$\theta^1_{x_k}$ (resp. $\theta^2_{x_k}$) est ainsi l'approximation de f fournie par l'information du premier ordre (resp. du deuxième ordre) au point x_k.

Montrer que pour tout $x \in \mathcal{O}$:

$$\left|f(x) - \theta^1_{x_k}(x)\right| \leqslant \frac{L}{2} \| x - x_k \|^2, \ \left|f(x) - \theta^2_{x_k}(x)\right| \leqslant \frac{L}{6} \| x - x_k \|^3 .$$

2°) Pour avoir une approximation du gradient de f en $x_k \in \mathcal{O}$ à l'aide d'évaluations de f, on utilise les vecteurs suivants :

$\triangle f(x_k) :=$ vecteur de composantes

$$\frac{f(x_k + h_j e_j) - f(x_k)}{h_j} \quad \begin{pmatrix}\text{différences finies}\\ \text{en avant}\end{pmatrix},$$

$\overline{\triangle} f(x_k) :=$ vecteur de composantes

$$\frac{f(x_k + h_j e_j) - f(x_k - h_j e_j)}{2h_j} \quad \begin{pmatrix}\text{différences finies}\\ \text{centrées}\end{pmatrix},$$

où $\mathrm{e}_1, \ldots, \mathrm{e}_n$ sont les vecteurs de la base canonique de $\mathbb{R}^n$ et $h_1, \ldots, h_n$ des réels strictement positifs.

Montrer que pour tout $j = 1, \ldots, n$:

$$\left| [\triangle f(x_k)]_j - \partial_j f(x_k) \right| \leqslant \frac{L}{2} h_j,$$
$$\left| [\overline{\triangle} f(x_k)]_j - \partial_j f(x_k) \right| \leqslant \frac{L}{6} h_j^2.$$

3°) On suppose ici qu'on a accès aux évaluations de ∇f et on propose une approximation de $\nabla^2 f(x_k)$ de la forme suivante :

$$\tilde{\nabla}^2 f(x_k) := \left[\frac{[\nabla f(x_k + h_j e_j)]_i - [\nabla f(x_k - h_j e_j)]_i}{2h_j} \right]_{1 \leqslant i,j \leqslant n}.$$

Cette approximation est obtenue par différences finies centrées à partir du gradient de f, le terme (i,j) de $\tilde{\nabla}^2 f(x_k)$ est une approximation de $\partial^2_{ij} f(x_k)$.

Montrer qu'il existe une et une seule matrice symétrique la plus proche de $\tilde{\nabla}^2 f(x_k)$ au sens de la distance associée à $|||\cdot|||$, et déterminer cette matrice en fonction de $\tilde{\nabla}^2 f(x_k)$.

Solution : 1°) Grâce aux développements de Taylor avec restes sous formes d'intégrales on a :

$$f(x) = f(x_k) + \int_0^1 \langle\, \nabla f(x_k + t(x - x_k)), x - x_k \,\rangle \, dt, \tag{2.10}$$

$$f(x) = f(x_k) + \langle\, \nabla f(x_k), x - x_k \,\rangle + \tag{2.11}$$
$$\int_0^1 (1-t) \langle\, \nabla^2 f(x_k + t(x - x_k))(x - x_k), x - x_k \,\rangle \, dt.$$

Il vient alors de (2.10)

$$f(x) - \theta^1_{x_k}(x) = \int_0^1 \langle\, \nabla f(x_k + t(x - x_k)) - \nabla f(x_k), x - x_k \,\rangle \, dt,$$

d'où, grâce à la propriété de Lipschitz de ∇f,

$$\left| f(x) - \theta^1_{x_k}(x) \right| \leqslant \int_0^1 Lt \parallel x - x_k \parallel^2 dt = \frac{L}{2} \parallel x - x_k \parallel^2.$$

D'une manière similaire, on déduit de (2.11) :

$$\begin{aligned} & f(x)-\theta_{x_k}^2(x) \\ &= \int_0^1 \left\langle \left[(1-t)\,\nabla^2 f\left(x_k+t\left(x-x_k\right)\right)-\frac{1}{2}\nabla^2 f(x_k)\right]\left(x-x_k\right), x-x_k \right\rangle\, dt \\ &= \int_0^1 \left\langle\, (1-t)\left[\nabla^2 f\left(x_k+t\left(x-x_k\right)\right)-\nabla^2 f(x_k)\right]\left(x-x_k\right), x-x_k \,\right\rangle\, dt, \end{aligned}$$

d'où, toujours grâce à la propriété de Lipschitz de $\nabla^2 f$, et à l'inégalité $\| Au \| \leqslant |||A|||\cdot \| u \|$ encore valable pour la norme $|||\cdot|||$ (d'accord ?) :

$$\left|f(x)-\theta_{x_k}^2(x)\right| \leqslant \int_0^1 Lt\,(1-t)\,\| x-x_k \|^3\, dt = \frac{L}{6}\,\| x-x_k \|^3 .$$

2°) On a : $\partial_j f\left(x_k\right) = \langle\, \nabla f\left(x_k\right), e_j \rangle$, d'où

$$\begin{aligned} \left[\triangle f\left(x_k\right)\right]_j - \partial_j f\left(x_k\right) &= \frac{1}{h_j}\left(f\left(x_k+h_j e_j\right)-f\left(x_k\right)-\langle\, \nabla f\left(x_k\right), h_j e_j\rangle\right) \\ &= \frac{1}{h_j}\left(f\left(x_k+h_j e_j\right)-\theta_{x_k}^1\left(x_k+h_j e_j\right)\right). \end{aligned}$$

Il résulte alors de la 1re question :

$$\left|\left[\triangle f(x_k)\right]_j - \partial_j f\left(x_k\right)\right| \leqslant \frac{1}{h_j}\frac{L}{2}h_j^2 = \frac{L}{2}h_j.$$

De manière similaire, on a pour les différences finies centrées :

$$\begin{aligned} \left[\overline{\triangle} f(x_k)\right]_j - \partial_j f\left(x_k\right) &= \frac{1}{2h_j}\left(f\left(x_k+h_j e_j\right)-f\left(x_k-h_j e_j\right)-2\left\langle \nabla f\left(x_k\right), h_j e_j\right\rangle\right) \\ &= \frac{1}{2h_j}\left\{\left(f\left(x_k+h_j e_j\right)-\theta_{x_k}^2\left(x_k+h_j e_j\right)\right)\right. \\ &\quad \left.-\left(f\left(x_k-h_j e_j\right)-\theta_{x_k}^2\left(x_k-h_j e_j\right)\right)\right\}. \end{aligned}$$

Par suite

$$\left|\left[\overline{\triangle} f(x_k)\right]_j - \partial_j f\left(x_k\right)\right| \leqslant \frac{1}{2h_j}\left(2\frac{L}{6}h_j^3\right) = \frac{L}{6}h_j^2.$$

3°) $\tilde{\nabla}^2 f(x_k)$ n'étant pas symétrique alors que $\nabla^2 f(x_k)$ l'est toujours, on cherche $S \in \mathcal{S}_n(\mathbb{R})$ la plus proche de $\tilde{\nabla}^2 f(x_k)\,(\in \mathcal{M}_n(\mathbb{R}))$.

$\mathcal{M}_n(\mathbb{R})$ est structuré en espace euclidien grâce au produit scalaire $\ll A, B \gg = \operatorname{tr}\left(A^\top B\right)$, et la distance considérée est celle associée à ce produit scalaire :

$|||A-B||| = (\ll A-B, A-B \gg)^{1/2}$. L'ensemble $\mathcal{S}_n(\mathbb{R})$ étant un sous-espace vectoriel de $\mathcal{M}_n(\mathbb{R})$, le problème posé est donc celui de *la projection*

orthogonale de $\tilde{\nabla}^2 f(x_k)$ *sur* $\mathcal{S}_n(\mathbb{R})$. L'élément $\overline{S} \in \mathcal{S}_n(\mathbb{R})$ à distance minimale de $A = \tilde{\nabla}^2 f(x_k)$ existe, est unique, et caractérisé par la propriété suivante :

$$\begin{cases} \overline{S} \in \mathcal{S}_n(\mathbb{R}) \\ \text{et} \quad \ll A - \overline{S}, S \gg = 0 \ \text{ pour tout } S \in \mathcal{S}_n(\mathbb{R}). \end{cases}$$

L'élément $\overline{S} := \frac{1}{2}\left(A + A^\top\right)$ est la solution cherchée car

$$\begin{aligned} \ll A - \overline{S}, S \gg &= \frac{1}{2}\left[\ll A, S \gg - \ll A^\top, S \gg\right] \\ &= 0 \text{ pour tout } S \in \mathcal{S}_n(\mathbb{R}). \end{aligned}$$

**** Exercice II.11.*** Une situation d'application importante où la fonction-objectif à minimiser est de la forme $x \longmapsto \sum_{i=1}^{m} [r_i(x)]^2$ (« *moindres carrés* ») est l'ajustement de données expérimentales. D'une manière habituelle les choses se présentent comme suit : on dispose de m données $d_1, d_2, \ldots, d_m$ correspondant aux valeurs $t_1, t_2, \ldots, t_m$ d'une variable t ; l'objectif est d'accorder « au mieux » aux données (t_1, d_1), $(t_2, d_2), \ldots,$ (t_m, d_m), une fonction-modèle $t \longmapsto \varphi(t, x)$ qui contient des paramètres ajustables $x_1, x_2, \ldots, x_n$.

En général la forme de la fonction-modèle dépend du contexte d'origine des données. On appelle *résidus* les différences entre la valeur de la fonction-modèle $\varphi(t_i, x)$ et la valeur d_i correspondant à t_i, c'est-à-dire $r_i(x) = \varphi(t_i, x) - d_i$ $(i = 1, \ldots, m)$. Le critère mesurant l'écart (dans $\mathbb{R}^m$) entre $(\varphi(t_1, x), \ldots, \varphi(t_m, x))$ et $(r_1, \ldots, r_m)$ fait l'objet d'un choix : c'est souvent le carré de la distance euclidienne,

$$x \longmapsto f(x) = \sum_{i=1}^{m} [\varphi(t_i, x) - d_i]^2 = \sum_{i=1}^{m} [r_i(x)]^2 .$$

Exemple. Un utilisateur a obtenu les données expérimentales suivantes pour un problème de nature physique dans lequel le comportement des valeurs observées est gouverné par une fonction inconnue de la variable $t \in [0, 1]$.

i	1	2	3	4	5	6
t_i	0,2	0,4	0,6	0,8	0,9	0,95
d_i	0,7123	1,754	4.852	22,27	94,91	388,2

Selon l'utilisateur, le phénomène physique sous-jacent suggère que la fonction $t \longmapsto d(t)$ doit avoir les propriétés suivantes :

– un comportement en t^{x_1} au voisinage de 0, pour un certain $x_1 > 0$;

– un comportement en $\frac{1}{(1-t)^{x_2}}$ au voisinage de 1, pour un certain $x_2 > 0$.

La fonction-modèle suggérée est

$$t \longmapsto \varphi(t,x) := x_3 t^{x_1} \frac{1}{(1-t)^{x_2}},$$

où $x_1 > 0,\ x_2 > 0,\ x_3 > 0$.

La fonction-objectif à minimiser est donc

$$x = (x_1, x_2, x_3) \longmapsto f(x) = \sum_{i=1}^{6} \left[\frac{x_3 t^{x_1}}{(1-t)^{x_2}} - d_i \right]^2 .$$

Questions :

– Mettre en œuvre votre méthode de minimisation favorite pour obtenir un jeu de paramètres $(\overline{x}_1, \overline{x}_2, \overline{x}_3)$ acceptable (point de départ suggéré : (1,1,1)).

– Déterminer la valeur $f(\overline{x})$ de la somme des carrés des résidus.

– Tracer sur un même graphique la fonction $t \longmapsto \varphi(t, \overline{x})$ et les données $(t_i, d_i),\ i = 1, \ldots, 6$.

Solution : $\overline{x}_1 = 0,5532$; $\overline{x}_2 = 1,9897$; $\overline{x}_3 = 1,0298$; $f(\overline{x}) = 0,0205$.

Commentaire : Les problèmes d'optimisation du type « moindres carrés » sont ceux les plus fréquemment rencontrés dans les applications.

III

MINIMISATION AVEC CONTRAINTES. CONDITIONS DE MINIMALITÉ

Rappels

III.1. Conditions de minimalité du premier ordre

Soit $f : \mathcal{O} \subset \mathbb{R}^n \to I\!R$ différentiable sur un ouvert $\mathcal{O}$ de $\mathbb{R}^n$, soit S un ensemble-contrainte décrit par m égalités et p inégalités

$$S := \{x \in \mathbb{R}^n \mid h_1(x) = 0, \ldots, h_m(x) = 0,\ g_1(x) \leqslant 0, \ldots, g_p(x) \leqslant 0\}$$

où les $m + p$ fonctions $h_i, g_j : \mathbb{R}^n \to \mathbb{R}$ sont supposées continûment différentiables sur $\mathbb{R}^n$.

***Théorème* (F. John).** *Si $\overline{x} \in \mathcal{O} \cap S$ est un minimum local de f sur S, alors il existe $\overline{\lambda}_1, \ldots, \overline{\lambda}_m$, $\overline{\mu}_0, \overline{\mu}_1, \ldots, \overline{\mu}_p$ non tous nuls tels que :*

(i) $\overline{\mu}_0 \nabla f(\overline{x}) + \sum_{i=1}^{m} \overline{\lambda}_i \nabla h_i(\overline{x}) + \sum_{j=1}^{p} \overline{\mu}_j \nabla g_j(\overline{x}) = 0$;

(ii) $\overline{\mu}_j \geqslant 0$ *pour tout* $j = 0, 1, \ldots, p$;

(iii) $\overline{\mu}_j = 0$ *si* $g_j(\overline{x}) < 0$.

On écarte l'éventualité $\overline{\mu}_0 = 0$ (peu informative sur $\overline{x}$) en faisant une hypothèse de *qualification des contraintes* en $\overline{x}$, c'est-à-dire une hypothèse sur les

fonctions-contraintes h_i et g_j représentant S. On considère par exemple la condition $(QC)_{\overline{x}}$ de Mangasarian-Fromovitz suivante :

$$(QC)_{\overline{x}} \begin{cases} \text{Il existe } d_0 \text{ tel que } \langle\, \nabla h_i\,(\overline{x})\,, d_0\,\rangle = 0 \ \text{ pour tout } i = 1, \ldots, m \\ \text{et } \langle\, \nabla g_j\,(\overline{x})\,, d_0\,\rangle < 0 \text{ pour tout } j \text{ vérifiant } g_j\,(\overline{x}) = 0\ ; \\ \text{les vecteurs } \nabla h_1\,(\overline{x})\,, \ldots, \nabla h_m\,(\overline{x}) \ \text{ sont linéairement indépendants.} \end{cases}$$

***Théorème* (Karush-Kuhn-Tucker).** *Si $\overline{x} \in \mathcal{O} \cap S$ est un minimum local de f sur S, et si $(QC)_{\overline{x}}$ a lieu, il existe alors $\overline{\lambda} = \left(\overline{\lambda}_1, \ldots, \overline{\lambda}_m\right) \in \mathbb{R}^m$ et $\overline{\mu} = (\overline{\mu}_1, \ldots, \overline{\mu}_p) \in \mathbb{R}^p$ tels que :*

(i) $\quad \nabla f\,(\overline{x}) + \sum\limits_{i=1}^{m} \overline{\lambda}_i \nabla h_i\,(\overline{x}) + \sum\limits_{j=1}^{p} \overline{\mu}_j \nabla g_j\,(\overline{x}) = 0\ ;$

(ii) $\quad \overline{\mu}_j \geqslant 0 \ \ \textit{pour tout} \ \ j = 1, \ldots, p\ ;$

(iii) $\ \overline{\mu}_j = 0 \ \ \textit{si}\ g_j\,(\overline{x}) < 0.$

Les $\overline{\lambda}_i$ et $\overline{\mu}_j$ mis en évidence dans ce théorème de Karush-Kuhn-Tucker (KKT en abrégé) sont appelés *multiplicateurs de Lagrange* (ou de Lagrange-KKT).

Une hypothèse de qualification des contraintes en $\overline{x}$, plus forte que $(QC)_{\overline{x}}$, est la suivante :

$(QC)'_{\overline{x}}$ Les vecteurs $\nabla h_1\,(\overline{x})\,, \ldots, \nabla h_m\,(\overline{x})\,,\ \nabla g_j\,(\overline{x})\,,\ j \in J\,(\overline{x})$, sont linéairement indépendants, où $J\,(\overline{x}) := \{j \mid g_j\,(\overline{x}) = 0\}$ est l'ensemble des indices des contraintes (de type inégalité) *actives* (ou *saturées*) en $\overline{x}$.

Si les fonctions h_i et g_j sont *affines*, on peut se dispenser de toute hypothèse de qualification des contraintes et déduire quand même les conditions nécessaires de minimalité de KKT.

Les conditions (i) - (ii) - (iii) du théorème de KKT (appelées conditions de KKT) deviennent suffisantes en présence de convexité dans les données. Considérons à cet effet le contexte suivant :

$$(\mathcal{C}) \begin{cases} f \text{ est convexe sur l'ouvert convexe } \mathcal{O}\,; \\ \text{les fonctions } h_i : \mathbb{R}^n \to \mathbb{R} \text{ sont affines}\,; \\ \text{les fonctions } g_j : \mathbb{R}^n \to \mathbb{R} \text{ sont convexes.} \end{cases}$$

Théorème. *Dans la situation décrite par $(\mathcal{C})$ ci-dessus, les conditions de KKT sont suffisantes pour que $\overline{x}$ soit un minimum de f sur $\mathcal{O} \cap S$.*

Dans le contexte décrit par $(\mathcal{C})$, il y a une condition d'énoncé simple assurant $(QC)_{\overline{x}}$ en tout point de S ; c'est la condition dite de Slater que voici : en notant

$h_i : x \longmapsto h_i(x) = \langle a_i, x\rangle - b_i$ les fonctions-contraintes du type égalité, et $Ax = b$ la conjonction des m contraintes du type égalité $h_i(x) = 0,\ i = 1, \ldots, m$,

$$(\mathcal{S}) \begin{cases} \text{Les vecteurs } a_i \text{ sont linéairement indépendants} \\ (i.e.\ A \text{ est surjective}) ; \\ \text{Il existe } x_0 \text{ tel que } Ax_0 = b \text{ et } g_j(x_0) < 0 \text{ pour tout } \ j = 1, \ldots, p. \end{cases}$$

III.2. Cône tangent, cône normal à un ensemble

Soit $\overline{x} \in \overline{S}$. On dit que $d \in \mathbb{R}^n$ est une *direction tangente à* S en $\overline{x}$ lorsqu'il existe une suite $\{x_k\} \subset S$ convergeant vers $\overline{x}$ et une suite $\{t_k\} \subset \mathbb{R}^+_*$ convergeant vers 0 telles que $(x_k - \overline{x})/t_k \to d$ quand $k \to +\infty$.

Autre manière de dire la même chose : d est tangente à S en $\overline{x}$ si et seulement si : il existe $\{d_k\} \to d$, il existe $\{t_k\} \subset \mathbb{R}^+_* \to 0$, telles que $\overline{x} + t_k d_k \in S$ pour tout k.

L'ensemble de telles directions est appelé *cône tangent à* S *en* $\overline{x}$, et noté $T(S, \overline{x})$. Il s'agit bien d'un cône et même d'un cône fermé (de sommet 0).

De plus $T(S, \overline{x}) = T\left(\overline{S}, \overline{x}\right)$, ce qui fait qu'en général nous ne considérerons ce concept de tangence que pour des S fermés.

Théorème. *Sous l'hypothèse* $(QC)_{\overline{x}}$, *on a*
$T(S, \overline{x}) = \{d \mid \langle\ \nabla h_i(\overline{x}), d\ \rangle = 0$ *pour* $i = 1, \ldots, m$ *et*

$$\langle\ \nabla g_j(\overline{x}), d\rangle \leqslant 0 \quad \textit{pour} \quad j \in J(\overline{x})\}.$$

Cette relation reste vraie en l'absence de $(QC)_{\overline{x}}$ lorsque les fonctions h_i et g_j sont affines.

Considérons le problème de la minimisation de f sur un ensemble-contrainte S qui n'est pas nécessairement représenté par des égalités ou inégalités. La condition nécessaire de minimalité locale s'énonce comme suit.

Théorème. *Si* $\overline{x} \in S$ *est un minimum local de* f *sur* S *et si* f *est différentiable en* $\overline{x}$, *alors*

$$-\nabla f(\overline{x}) \in [T(S, \overline{x})]^\circ,$$

où $[T(S, \overline{x})]^\circ$ *désigne le cône polaire de* $T(S, \overline{x})$, *c'est-à-dire*

$$[T(S, \overline{x})]^\circ := \{s \mid \langle s, d\rangle \leqslant 0 \textit{ pour tout } d \in T(S, \overline{x})\}.$$

III.3. Prise en compte de la convexité

– Si S est un convexe fermé, $T(S,\overline{x})$ est exactement l'adhérence du cône engendré par $S-\overline{x}$:

$$\begin{aligned} T(S,\overline{x}) &= \text{ adhérence de } \{d \mid d=\alpha(c-\overline{x}) \text{ avec } c\in S \text{ et } \alpha\geqslant 0\} \\ &= \overline{\mathbb{R}^+(S-\overline{x})}. \end{aligned}$$

$T(S,\overline{x})$ est alors un cône convexe fermé.

– Une direction $s\in\mathbb{R}^n$ est dite *normale à S en $\overline{x}$* lorsque

$$\langle s, c-\overline{x}\rangle \leqslant 0 \text{ pour tout } c\in S.$$

L'ensemble des directions normales à S en $\overline{x}$ est appelé *cône normal à S en $\overline{x}$*, et noté $N(S,\overline{x})$. Lorsque S est un convexe fermé, $N(S,\overline{x})$ n'est autre que le cône polaire de $T(S,\overline{x})$:

$$N(S,\overline{x}) = [T(S,\overline{x})]^\circ \quad (\text{et } [N(S,\overline{x})]^\circ = T(S,\overline{x})).$$

Théorème. *Soit $f:\mathbb{R}^n\to\mathbb{R}$ convexe et différentiable, soit S convexe fermé. Alors les minima de f sur S sont exactement les $\overline{x}\in S$ pour lesquels*

$$-\nabla f(\overline{x}) \in N(S,\overline{x}).$$

Par exemple, la projection $\overline{x}$ de x sur le convexe fermé S est l'élément de S minimisant l'application $u\longmapsto \| u-x\|$ (ou $1/2\,\| u-x\|^2$) sur S ; cet élément $\overline{x}$ est unique et caractérisé par la relation suivante :

$$\langle x-\overline{x}, c-\overline{x}\rangle \leqslant 0 \quad \text{pour tout} \quad c\in S.$$

III.4. Conditions de minimalité du second ordre

Soit $f:\mathcal{O}\subset\mathbb{R}^n$ deux fois différentiable sur un ouvert $\mathcal{O}$ de $\mathbb{R}^n$, soit S un ensemble-contrainte défini par m égalités $h_i(x)=0$ où les fonctions $h_i:\mathbb{R}^n\to\mathbb{R}$ sont supposées deux fois différentiables. On désigne par $\mathcal{L}$ le lagrangien usuel associé au problème de minimisation de f sur S, *i.e.*

$$(x,\lambda)\in\mathcal{O}\times\mathbb{R}^m \longmapsto \mathcal{L}(x,\lambda) := f(x)+\sum_{i=1}^m \lambda_i h_i(x).$$

Théorème. *Si $\overline{x} \in \mathcal{O} \cap S$ est un minimum local de f sur S et si les vecteurs $\nabla h_1(\overline{x}), \ldots, \nabla h_m(\overline{x})$ sont linéairement indépendants, il existe alors $\overline{\lambda} = (\overline{\lambda}_1, \ldots, \overline{\lambda}_m) \in \mathbb{R}^m$ (unique) tel que :*

(i) $\nabla_x \mathcal{L}(\overline{x}, \overline{\lambda}) = 0$

et

(ii) $\langle \nabla^2_{xx} \mathcal{L}(\overline{x}, \overline{\lambda}) d, d \rangle \geqslant 0$ *pour toute direction d dans le sous-espace tangent* $T(S, \overline{x}) = \{d \in \mathbb{R}^n \mid \langle \nabla h_i(\overline{x}), d \rangle = 0$ *pour tout* $i = 1, \ldots, m\}$.

Théorème. *Soit $\overline{x} \in \mathcal{O} \cap S$. Supposons les $\nabla h_1(\overline{x}), \ldots, \nabla h_m(\overline{x})$ linéairement indépendants, et supposons qu'il existe $\overline{\lambda} \in \mathbb{R}^m$ tel que :*

(i) $\nabla_x \mathcal{L}(\overline{x}, \overline{\lambda}) = 0$

et

(ii) $\langle \nabla^2_{xx} \mathcal{L}(\overline{x}, \overline{\lambda}) d, d \rangle > 0$ *pour toute direction non nulle d dans le sous-espace tangent* $T(S, \overline{x}) = \{d \in \mathbb{R}^n \mid \langle \nabla h_i(\overline{x}), d \rangle = 0$ *pour tout* $i = 1, \ldots, m\}$.

Alors $\overline{x}$ est un minimum local strict de f sur S.

Références. Chapitre III de [11].

***Exercice III.1.** On considère le problème de la minimisation d'une fonction différentiable $f : \mathbb{R}^2 \to \mathbb{R}$ sous la contrainte $h(\xi_1, \xi_2) := \xi_1^3 - \xi_2^2 = 0$.

Quels sont les points vérifiant les conditions de minimalité du 1er ordre ?

Résoudre le problème avec $f : (\xi_1, \xi_2) \longmapsto f(\xi_1, \xi_2) := \xi_1 + \xi_2^2$.

Solution : La fonction h définissant la contrainte (du type égalité) est continûment différentiable, mais $\nabla h(\overline{x}) = 0$ en $\overline{x} = (0,0)$. Par conséquent, les points $\overline{x} = (\overline{\xi}_1, \overline{\xi}_2)$ vérifiant les conditions de minimalité du 1er ordre sont ceux vérifiant :

$$\begin{cases} h(\overline{x}) = 0, & \text{(conditions de} \\ \exists\, (\overline{\lambda}_0, \overline{\lambda}_1) \neq (0,0) \text{ tel que } \overline{\lambda}_0 \nabla f(\overline{x}) + \overline{\lambda}_1 \nabla h(\overline{x}) = 0 & \text{F. John).} \end{cases}$$

Cela conduit aux candidats suivants :

- $\overline{x} \neq (0,0)$ vérifiant $h(\overline{x}) = 0$ et $\nabla f(\overline{x}) = \overline{\lambda} \nabla h(\overline{x})$ pour un certain $\overline{\lambda} \in \mathbb{R}$ (conditions usuelles de Lagrange) ;
- $\overline{x} = (0,0)$ (à considérer dans tous les cas de figure).

Dans l'exemple considéré, avec $f(\xi_1, \xi_2) := \xi_1 + \xi_2^2$, il n'y a pas de point vérifiant des conditions de minimalité de type Lagrange ; seul $\overline{x} = (0,0)$ est à retenir. En effet, $\overline{x} = (0,0)$ est le minimum de $f(\xi_1, \xi_2)$ sous la contrainte $h(\xi_1, \xi_2) = 0$.

***Exercice III.2.** Soit $s \neq 0$ fixé dans $\mathbb{R}^n$, $r \in \mathbb{R}$, et H_r l'hyperplan de $\mathbb{R}^n$ d'équation $\langle s, x\rangle = r$.

Étant donné $v \in \mathbb{R}^n$ on se propose de déterminer la projection orthogonale $\overline{v}_r$ de v sur H_r.

1°) Déterminer $\overline{v}_r$ à partir de l'observation suivante : $\overline{v}_r$ est la solution du problème de minimisation convexe suivant :

$$(\mathcal{P}_r)\begin{cases}\text{Minimiser } \frac{1}{2} \parallel x - v \parallel^2 \\ x \in H_r.\end{cases}$$

2°) Posons $d(r) := \frac{1}{2} \parallel \overline{v}_r - v \parallel^2$. Vérifier que d est une fonction dérivable de r et que sa dérivée $d'(r)$ est, au signe près, le multiplicateur de Lagrange dans le problème $(\mathcal{P}_r)$.

Solution : 1°) $\overline{v}_r$ est la solution de $(\mathcal{P}_r)$ si et seulement si :

$$\begin{cases}\langle s, \overline{v}_r\rangle = r \\ \exists\ \overline{\lambda}_r \in \mathbb{R} \ \text{ tel que } \ \overline{v}_r - v + \overline{\lambda}_r s = 0.\end{cases} \qquad \text{(d'accord ?)} \qquad (3.1)$$

$\overline{\lambda}_r$ est le multiplicateur de Lagrange associé à la seule contrainte de $(\mathcal{P}_r)$, à savoir la contrainte-égalité $\langle s, x\rangle - r = 0$.

Il vient de la 2e relation de (3.1)

$$\langle \overline{v}_r - v, s\rangle + \overline{\lambda}_r \parallel s \parallel^2 = 0$$

d'où, avec la 1re relation de (3.1) , $\overline{\lambda}_r = \frac{\langle s,v\rangle - r}{\|s\|^2}$.

Par suite

$$\overline{v}_r = v - \frac{\langle s, v\rangle - r}{\parallel s \parallel^2} s.$$

$v - \overline{v}_r$ est dirigé par s, ce qui est... normal.

La distance de v à H_r vaut donc $\frac{|\langle s,v\rangle - r|}{\|s\|}$.

2°) La fonction $r \in \mathbb{R} \longmapsto d(r) := \frac{1}{2} \parallel \overline{v}_r - v \parallel^2 = \frac{1}{2}\frac{(\langle s,v\rangle - r)^2}{\|s\|^2}$ est dérivable sur $\mathbb{R}$ et

$$d'(r) = -\frac{\langle s, v\rangle - r}{\parallel s \parallel^2} = -\overline{\lambda}_r.$$

Ainsi, la « sensibilité au 1er ordre » de la valeur minimale $d(r)$ aux variations de r (second membre dans l'équation de H_r) est donnée, au signe près, par le multiplicateur de Lagrange associé à la contrainte dans $(\mathcal{P}_r)$.

***Exercice III.3.** Soient $n_1, \dots, n_k$ k entiers naturels de somme $N > 0$ et $f : \mathbb{R}^k \to \mathbb{R}$ définie par :

$$f(p_1, \dots, p_k) := \prod_{i=1}^{k} p_i^{n_i}.$$

Maximiser f sur le simplexe-unité Λ_k de $\mathbb{R}^k$

$$\left(\Lambda_k := \left\{ p = (p_1, \dots, p_k) \in \mathbb{R}^k \mid p_i \geqslant 0 \text{ pour tout } i, \text{ et} \sum_{i=1}^{k} p_i = 1 \right\} \right).$$

Solution : f est continue, Λ_k est compact (d'accord ?) ; il existe donc au moins un $\overline{p} \in \Lambda_k$ maximisant f sur Λ_k.

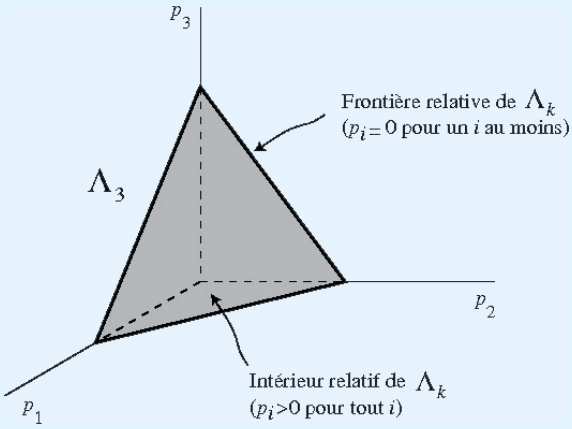


Figure 6.

Comme f est nulle sur la frontière relative de Λ_k, les $\overline{p}$ se trouvent nécessairement dans l'intérieur relatif de Λ_k, *i.e.* $\overline{p}_i > 0$ pour tout $i = 1, \dots, k$ (« relatif » signifie dans le sous-espace affine engendré par Λ_k, à savoir l'hyperplan affine d'équation $\sum_{i=1}^{k} p_i = 1$).

Donc, si on définit $g : (\mathbb{R}_*^+)^k \to \mathbb{R}$ par $g(p_1, \dots, p_k) := \ln f(p_1, \dots, p_k)$ et $h : \mathbb{R}^k \to \mathbb{R}$ par $h(p_1, \dots, p_k) := \sum_{i=1}^{k} p_i - 1$, il est clair que $\overline{p}$ maximise $g(p)$ sous la contrainte $h(p) = 0$. Puisque $\nabla h(\overline{p}) \neq 0$ nécessairement, il existe $\overline{\lambda} \in \mathbb{R}$

unique tel que $\nabla g(\overline{p}) = \overline{\lambda}\nabla h(\overline{p})$, soit :

$$\frac{n_i}{\overline{p}_i} = \overline{\lambda} \quad \text{pour tout } i = 1, \ldots, k.$$

Comme $\sum_{i=1}^{k} \overline{p}_i = 1$, on en déduit $\overline{\lambda} = N$ et, par suite,

$$\overline{p}_i = \frac{n_i}{N} \quad \text{pour tout } i = 1, \ldots, k. \tag{3.2}$$

Sachant que le problème de maximisation de départ a une solution et qu'on n'a détecté qu'un seul point vérifiant les conditions nécessaires d'optimalité, le point en question est *la* solution du problème.

Remarque : On aurait pu faire la même démarche avec f sans recourir à g ; mais « casser » un produit en une somme en prenant le logarithme peut simplifier les calculs.

Commentaire : Dans les problèmes d'estimation en Statistique, on est amené à maximiser des fonctions appelées *fonctions de vraisemblance*, et cela est source de bien des problèmes d'optimisation. L'exemple traité dans cet exercice en est une illustration.

***Exercice III.4.** Soit $A \in \mathcal{S}_n(\mathbb{R})$, λ_1 (resp. λ_n) la plus grande valeur propre (resp. la plus petite valeur propre) de A. Montrer que :

$$\lambda_1 = \max_{\|x\|=1} \langle Ax, x\rangle \quad (\text{resp. } \lambda_n = \min_{\|x\|=1} \langle Ax, x\rangle).$$

À quelle condition a-t-on $\lambda_1 = \max_{\|x\|\leqslant 1} \langle Ax, x\rangle$ (resp. $\lambda_n = \min_{\|x\|\leqslant 1} \langle Ax, x\rangle$) ?

On demande des démonstrations basées sur des techniques et résultats d'Optimisation, sans recourir à des résultats d'Algèbre bilinéaire (comme une décomposition spectrale de A).

Solution : – Soient $f : x \in \mathbb{R}^n \longmapsto f(x) := \langle Ax, x\rangle$ et $S := \{x \in \mathbb{R}^n \mid \| x \| = 1\}$. Il est clair que S peut être représenté comme une contrainte du type égalité, $h(x) = 0$, avec $h(x) := \| x \|^2 - 1$.

La fonction-objectif f étant continue et S compact, il existe bien $\overline{x} \in S$ tel que $f(\overline{x}) = \max_{x\in S} f(x)$. Comme h est continûment différentiable et $\nabla h(\overline{x}) = 2\overline{x} \neq 0$, il existe $\overline{\lambda} \in \mathbb{R}$ (unique) tel que $\nabla f(\overline{x}) = \overline{\lambda}\nabla h(\overline{x})$, soit

$A\overline{x} = \overline{\lambda}\overline{x}$. Donc $\overline{x}$ est un vecteur propre (unitaire) de A associé à la valeur propre (exhibée) $\overline{\lambda}$; la valeur correspondante $f(\overline{x})$ est $\langle A\overline{x}, \overline{x}\rangle = \overline{\lambda}$. Reste à montrer que $\overline{\lambda}$ est la plus grande des valeurs propres de A. Soit μ une valeur propre de A et x un vecteur propre unitaire associé : $Ax = \mu x$, $\| x \| = 1$. Il s'ensuit :

$$\langle Ax, x\rangle = \mu \leqslant \max_{\|x\|=1} \langle Ax, x\rangle = \overline{\lambda}.$$

– Posons $\alpha := \max_{\|x\|\leqslant 1} \langle Ax, x\rangle$. Il va de soi que $\lambda_1 \leqslant \alpha$ et $\alpha \geqslant 0$.

Si $\alpha = \lambda_1$, on a $\lambda_1 \geqslant 0$ évidemment. Montrons la réciproque, à savoir :

$(\lambda_1 \geqslant 0) \Rightarrow (\langle Ax, x\rangle \leqslant \lambda_1$ pour tout x tel que $\| x \| \leqslant 1)$.

Si $x = 0$, l'inégalité voulue est bien vérifiée ; si $x \neq 0$, $\| x \| \leqslant 1$, on note que

$$\langle Ax, x\rangle = \| x \|^2 \left\langle A\frac{x}{\| x \|}, \frac{x}{\| x \|}\right\rangle \leqslant \lambda_1.$$

Donc λ_1 *est le maximum de la forme quadratique* $\langle Ax, x\rangle$ *sur la boule unité fermée (euclidienne) de* $\mathbb{R}^n$ *si, et seulement si, la plus grande valeur propre de* A *est* $\geqslant 0$.

Par un raisonnement similaire à celui fait pour λ_1, nous avons $\lambda_n = \min_{\|x\|=1} \langle Ax, x\rangle$.

Posons $\beta := \min_{\|x\|\leqslant 1} \langle Ax, x\rangle$. Il est clair que $\beta \leqslant \lambda_n$ et $\beta \leqslant 0$.

Si $\beta = \lambda_n$, on a $\lambda_n \leqslant 0$ bien sûr. Réciproquement, si $\lambda_n \leqslant 0$,

$$\langle Ax, x\rangle \geqslant \lambda_n \| x \|^2 \geqslant \lambda_n \text{ pour tout } x \text{ tel que } \| x \| \leqslant 1.$$

Donc : $\lambda_n = \min_{\|x\|\leqslant 1} \langle Ax, x\rangle$ si, et seulement si, l'une des valeurs propres de A est $\leqslant 0$ (ce qui équivaut à $\lambda_n \leqslant 0$).

***Exercice III.5.** Soit $A \in \mathcal{S}_n(\mathbb{R})$. À quelle condition les deux problèmes suivants sont-il équivalents

$$(\mathcal{P}) \quad \begin{cases} \text{Min } \langle Ax, x\rangle \\ \| x \| = 1 \end{cases} \qquad (\hat{\mathcal{P}}) \quad \begin{cases} \text{Min } \langle Ax, x\rangle \\ \| x \| \leqslant 1 \end{cases}$$

(au sens où les valeurs optimales et les solutions sont les mêmes dans $(\mathcal{P})$ et $(\hat{\mathcal{P}})$) ?

Solution : Les valeurs optimales dans $(\mathcal{P})$ et $(\hat{\mathcal{P}})$ sont les mêmes si, et seulement si, la plus petite valeur propre λ_n de A est $\leqslant 0$ (voir exercice précédent). Mais on demande plus ici : on veut que les solutions soient les mêmes dans $(\mathcal{P})$ et $(\hat{\mathcal{P}})$ ou, ce qui revient au même, on veut qu'aucune solution $\overline{x}$ de $(\hat{\mathcal{P}})$ ne soit intérieure à la boule unité (euclidienne) de $\mathbb{R}^n$ ($\| \overline{x} \| < 1$). Or une solution $\overline{x}$ de $(\hat{\mathcal{P}})$ est intérieure à la boule unité (euclidienne) de $\mathbb{R}^n$ si, et seulement si, A est semi-définie positive (d'accord ?).

Donc, $(\mathcal{P})$ et $(\hat{\mathcal{P}})$ sont équivalents si, et seulement si, $\lambda_n < 0$.

***Exercice III.6.** Étant donnés $A \in \mathcal{S}_n(\mathbb{R})$ et $b \in \mathbb{R}^n$, on considère le problème d'optimisation suivant :

$$(\mathcal{P}) \quad \begin{cases} \text{Min } f(x) := \frac{1}{2} \langle Ax, x \rangle + \langle b, x \rangle \\ \| x \| = 1. \end{cases}$$

1°) On suppose $b = 0$. Rappeler alors ce que vaut $\overline{f} := \inf \{ f(x) : \| x \| = 1 \}$ et quels sont les $\overline{x}$ de norme 1 pour lesquels $f(\overline{x}) = \overline{f}$.

2°) Soient λ_1 la plus grande valeur propre de A et p un réel strictement inférieur à $-\lambda_1$. On pose

$$A_p := A + pI_n \text{ et } f_p : x \in \mathbb{R}^n \longmapsto f_p(x) := \frac{1}{2} \langle A_p x, x \rangle + \langle b, x \rangle .$$

(a) Indiquer pourquoi f_p est strictement concave.

(b) On considère le problème d'optimisation suivant :

$$\left(\tilde{\mathcal{P}}_p\right) \quad \begin{cases} \text{Min } f_p(x) \\ \| x \| \leqslant 1. \end{cases}$$

Montrer que

$$\inf \{ f_p(x) : \| x \| \leqslant 1 \} = \inf \{ f(x) : \| x \| = 1 \} + \frac{1}{2} p$$

et que les solutions de $(\mathcal{P})$ et $\left(\tilde{\mathcal{P}}_p\right)$ sont les mêmes.

Solution : 1°) $\overline{f} = \frac{1}{2}\lambda_n$, où λ_n désigne la plus petite valeur propre de A. De plus

$$\left(\| \overline{x} \| = 1,\ f(\overline{x}) = \overline{f}\right) \Leftrightarrow \left(\| \overline{x} \| = 1 \text{ et } A\overline{x} = \lambda_n \overline{x}\right).$$

Les solutions de $(\mathcal{P})$ sont, dans ce cas, les vecteurs propres unitaires associés à λ_n.

2°) (a) On a $\nabla^2 f_p(x) = A_p$ pour tout $x \in \mathbb{R}^n$, et par choix de p

$$\begin{aligned}\langle A_p x, x\rangle &= \langle Ax, x\rangle + p \| x \|^2 \leqslant (\lambda_1 + p) \| x \|^2 \\ &< 0 \text{ si } x \neq 0.\end{aligned}$$

Donc f_p est une fonction strictement concave (et même fortement concave sur $\mathbb{R}^n$).

(b) Notons que $\inf\{f_p(x) : \| x \| \leqslant 1\}$ est nécessairement atteint en des points $\overline{x}$ tels que $\| \overline{x} \| = 1$. En effet, si cette borne inférieure était atteinte en un point $\overline{x}$ intérieur à l'ensemble-contrainte de $\left(\tilde{\mathcal{P}}_p\right)$ on aurait : $\nabla f_p(\overline{x}) = 0$ et $\nabla^2 f_p(\overline{x})$ semi-définie positive, ce qui est exclu.

Il s'ensuit

$$\begin{aligned}\inf\{f_p(x) : \| x \| \leqslant 1\} &= \inf\{f_p(x) : \| x \| = 1\} \\ &= \inf\{f(x) : \| x \| = 1\} + \frac{p}{2} \\ \text{(puisque } f_p(x) &= f(x) + \frac{p}{2} \text{ lorsque } \| x \| = 1).\end{aligned}$$

D'une manière claire :

$$\begin{aligned}\overline{x} \text{ est solution de } \left(\tilde{\mathcal{P}}_p\right) &\Leftrightarrow f_p(\overline{x}) = \inf\{f_p(x) : \|x\| \leqslant 1\} \text{ et } \|\overline{x}\| = 1 \ ; \\ &\Leftrightarrow f_p(\overline{x}) = \inf\{f_p(x) : \|x\| = 1\} \text{ et } \|\overline{x}\| = 1 \ ; \\ &\Leftrightarrow f(\overline{x}) + \frac{p}{2} = \inf\{f(x) : \|x\| = 1\} + \frac{p}{2} \text{ et } \|\overline{x}\| = 1 \ ; \\ &\Leftrightarrow \overline{x} \text{ est solution de } (\mathcal{P}).\end{aligned}$$

Les solutions de $(\mathcal{P})$ et $\left(\tilde{\mathcal{P}}_p\right)$ sont bien les mêmes.

Remarque : L'ensemble-contrainte dans $\left(\tilde{\mathcal{P}}_p\right)$ est un convexe compact simple (la boule unité fermée pour la norme euclidienne) et la fonction à minimiser est strictement concave. Ce n'est pas pour autant un problème simple, et la stricte

concavité de la fonction objectif f_p n'implique nullement qu'il n'y a qu'une seule solution (contrairement à la *maximisation* de f_p sur le même ensemble-contrainte).

***Exercice III.7.** Soit $\lambda_n : \mathcal{S}_n(\mathbb{R}) \to \mathbb{R}$

$$A \longmapsto \lambda_n(A) := \text{ plus petite valeur propre de } A.$$

1°) Vérifier que λ_n est une fonction concave.

2°) Représenter $\mathcal{P}_n(\mathbb{R})$, ensemble des matrices semi-définies positives de taille n, sous la forme $\{A \in \mathcal{S}_n(\mathbb{R}) \mid g(A) \leqslant 0\}$, où g est une fonction convexe que l'on proposera.

Solution : 1°) On a (*cf.* Exercice III.4 par exemple) :
$\forall A \in \mathcal{S}_n(\mathbb{R})$, $\lambda_n(A) = \inf_{\|x\|=1} \langle Ax, x\rangle$.
Si $\ll \cdot, \cdot \gg$ symbolise le produit scalaire usuel dans $\mathcal{S}_n(\mathbb{R})$ (*i.e.* $\ll A, B \gg = \mathrm{tr}(AB)$), $\langle Ax, x\rangle$ n'est autre que $\ll A, xx^T \gg$. Ainsi

$$\lambda_n(A) = \inf_{\|x\|=1} \ll A, xx^T \gg .$$

D'où la fonction λ_n est l'infimum de la famille de formes linéaires $A \longmapsto \ll A, xx^T \gg$, indexée par x vérifiant $\| x \| = 1$. Il s'ensuit que λ_n est une fonction concave positivement homogène (*i.e.*, $\lambda_n(\alpha A) = \alpha \lambda_n(A)$ pour tout $\alpha \geqslant 0$ et $A \in \mathcal{S}_n(\mathbb{R})$).

2°) On a : $(A \in \mathcal{P}_n(\mathbb{R})) \Leftrightarrow (\lambda_n(A) \geqslant 0)$.
Par suite : $\mathcal{P}_n(\mathbb{R}) = \{A \in \mathcal{S}_n(\mathbb{R}) \mid g(A) \leqslant 0\}$, où $g := -\lambda_n$.
Il s'agit d'une « bonne » représentation de $\mathcal{P}_n(\mathbb{R})$ sous forme de contrainte de type inégalité, car il existe $A_0 \in \mathcal{S}_n(\mathbb{R})$ tel que $g(A_0) < 0$ (par exemple $A_0 = I_n$) : l'hypothèse de qualification des contraintes de Slater est satisfaite. On retrouve alors que

$$\text{int } \mathcal{P}_n(\mathbb{R}) = \{A \in \mathcal{S}_n(\mathbb{R}) \mid g(A) < 0\}$$
(ensemble des matrices définies positives de taille n).

Notons toutefois que g n'est pas nécessairement différentiable.

***Exercice III.8.** Étant donné $a_1, \dots, a_n$ des réels différents de zéro, on considère l'ellipsoïde plein (ou convexe compact elliptique) de $\mathbb{R}^n$ défini comme suit :

$$\mathcal{E} := \left\{ u = (u_1, \dots, u_n) \in \mathbb{R}^n \mid \sum_{i=1}^{n} \left(\frac{u_i}{a_i} \right)^2 \leqslant 1 \right\}.$$

Soient $x \notin \mathcal{E}$ fixé et $\overline{x}$ sa projection sur $\mathcal{E}$. Montrer que

$$\overline{x}_i = \frac{a_i^2 x_i}{a_i^2 + \overline{\mu}} \quad \text{pour tout } i = 1, \dots, n,$$

où $\overline{\mu} > 0$ est la seule solution de l'équation (en μ) $\displaystyle\sum_{i=1}^{n} \frac{a_i^2 x_i^2}{(a_i^2 + \mu)^2} = 1$.

Solution : $\overline{x} = (\overline{x}_1, \dots, \overline{x}_n)$ est solution du problème de minimisation suivant :

$$(\mathcal{P}) \quad \begin{cases} \text{Minimiser } f(u) := \| x - u \|^2 \\ \text{sous la contrainte } \displaystyle\sum_{i=1}^{n} \left(\frac{u_i}{a_i} \right)^2 \leqslant 1. \end{cases}$$

Il est clair ici que la (seule) contrainte inégalité est active en $\overline{x}$: $\displaystyle\sum_{i=1}^{n} \left(\frac{\overline{x}_i}{a_i} \right)^2 = 1$ (d'accord ?). La condition de minimalité caractérisant $\overline{x}$ est : il existe un multiplicateur $\overline{\mu} \geqslant 0$ tel que

$$\overline{x}_i - x_i + \overline{\mu} \frac{\overline{x}_i}{a_i^2} = 0 \quad \text{pour tout} \quad i = 1, \dots, n.$$

Il s'ensuit que $\overline{\mu} > 0$ (car $x \neq \overline{x}$) et $\overline{x}_i = \frac{a_i^2 x_i}{a_i^2 + \overline{\mu}}$ pour tout $i = 1, \dots, n$.

Comment trouver ce $\overline{\mu}$? On écrit tout simplement la condition $\displaystyle\sum_{i=1}^{n} \left(\frac{\overline{x}_i}{a_i} \right)^2 = 1$.

La fonction $\mu \longmapsto d(\mu) := \displaystyle\sum_{i=1}^{n} \frac{a_i^2 x_i^2}{(a_i^2 + \mu)^2}$ est continue et strictement décroissante sur $\mathbb{R}^+$; elle décroît de $d(0) = \displaystyle\sum_{i=1}^{n} \frac{x_i^2}{a_i^2} > 1$ (car $x \notin \mathcal{E}$) à $0 \left(= \lim_{\mu \to +\infty} d(\mu) \right)$.

L'unique $\overline{\mu} > 0$ vérifiant $d(\overline{\mu}) = 1$ est le multiplicateur recherché.

***Exercice III.9.** Dans le processus de minimisation d'une fonction f deux fois différentiable sur $\mathbb{R}^n$, la « direction de Newton » à partir d'un point x_k où $\nabla f(x_k) \neq 0$ et $\nabla^2 f(x_k)$ est définie positive est obtenue

– soit en minimisant $d \longmapsto \langle \nabla f(x_k), d\rangle + \frac{1}{2}\left\langle \nabla^2 f(x_k)\, d, d\right\rangle$ sur $\mathbb{R}^n$

– soit en minimisant $d \longmapsto \langle \nabla f(x_k), d\rangle$ sous la contrainte $\left\langle \nabla^2 f(x_k)\, d, d\right\rangle \leqslant 1$.

Montrer que les directions obtenues comme solutions de ces deux problèmes de minimisation sont les mêmes à une constante positive multiplicative près.

Solution : Le premier problème (sans contraintes) est celui de la minimisation d'une fonction quadratique strictement convexe sur $\mathbb{R}^n$: sa solution est celle de l'équation

$$\nabla f(x_k) + \nabla^2 f(x_k)\, d = 0, \text{ soit } d = -\left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k).$$

Dans le deuxième problème (avec contrainte) il s'agit de minimiser une forme linéaire non nulle sur le convexe compact (elliptique) décrit par l'inégalité $\left\langle \nabla^2 f(x_k)\, d, d\right\rangle \leqslant 1$.

La solution $\overline{d}$ est donc nécessairement sur la frontière de l'ensemble-contrainte (frontière d'équation $\left\langle \nabla^2 f(x_k)\, d, d\right\rangle = 1$) et elle est caractérisée par l'existence de $\overline{\mu} \geqslant 0$ (le multiplicateur) tel que

$$\nabla f(x_k) + 2\overline{\mu} \nabla^2 f(x_k)\, \overline{d} = 0.$$

Comme $\overline{\mu} \neq 0$ (car $\nabla f(x_k) \neq 0$), il s'ensuit $\overline{d} = -\dfrac{\left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)}{2\overline{\mu}}$.

***Exercice III.10.** *Minimisation sur le simplexe-unité de $\mathbb{R}^n$*

1°) Soit $\Lambda_3 := \{(x_1, x_2, x_3) \in \mathbb{R}^3 \mid x_1 + x_2 + x_3 = 1,\ x_1 \geqslant 0,\ x_2 \geqslant 0,\ x_3 \geqslant 0\}$.

Dessiner Λ_3 et déterminer le cône tangent et le cône normal à Λ_3 en $(1/3, 1/3, 1/3)$, $(0, 1/2, 1/2)$ et $(0, 0, 1)$ respectivement. (On ne demande pas une démonstration ou un calcul détaillé mais une réponse au vu des définitions et de la représentation graphique de Λ_3.)

2°) Soit Λ_n le simplexe-unité de $\mathbb{R}^n$, c'est-à-dire

$$\Lambda_n := \left\{ x = (x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_i \geqslant 0 \text{ pour tout } i \text{ et } \sum_{i=1}^{n} x_i = 1 \right\}.$$

Pour $\overline{x} \in \Lambda_n$ on note $I(\overline{x})$ l'ensemble (éventuellement vide) des i tels que $\overline{x}_i = 0$.

Montrer que

– le cône tangent à Λ_n en $\overline{x}$ est

$$T_{\Lambda_n}(\overline{x}) = \{d = (d_1, \ldots, d_n) \in \mathbb{R}^n \mid d_i \geqslant 0 \text{ pour tout } i \in I(\overline{x}), \text{ et } \sum_{i=1}^{n} d_i = 0\} ;$$

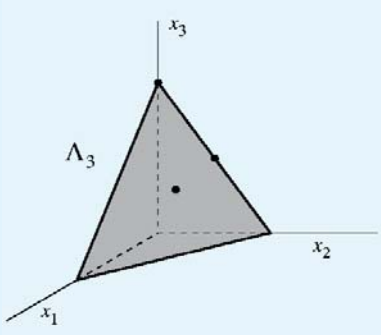
– le cône normal à Λ_n en $\overline{x}$ est

$$N_{\Lambda_n}(\overline{x}) = \{(\alpha_0, \ldots, \alpha_0) \mid \alpha_0 \in \mathbb{R}\} + \{(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n \mid \alpha_i = 0 \text{ pour tout } i \notin I(\overline{x}), \ \alpha_i \leqslant 0 \text{ pour tout } i \in I(\overline{x})\}.$$

3°) Soit $f : \mathbb{R}^n \to \mathbb{R}$ convexe et différentiable. Montrer qu'une condition nécessaire et suffisante pour que $\overline{x} \in \Lambda_n$ minimise f sur Λ_n est :

$$\partial_i f(\overline{x}) = \text{constante } \overline{c} \text{ pour tout } i \notin I(\overline{x})$$
$$\partial_i f(\overline{x}) \geqslant \overline{c} \text{ pour tout } i \in I(\overline{x}).$$

Solution :



1°) $T_{\Lambda_3}(1/3, 1/3, 1/3)$ est le plan d'équation $d_1 + d_2 + d_3 = 0$;
$T_{\Lambda_3}(0, 1/2, 1/2)$ est le demi-plan d'équation : $d_1 + d_2 + d_3 = 0$ et $d_1 \geqslant 0$;
$T_{\Lambda_3}(0, 0, 1)$ est le cône d'équation : $d_1 + d_2 + d_3 = 0$ $d_1 \geqslant 0$ et $d_2 \geqslant 0$.

Figure 7.

$T_{\Lambda_3}(\overline{x})$ est dans chaque cas $\mathbb{R}^+(\Lambda_3 - \overline{x})$.

$N_{\Lambda_3}(1/3, 1/3, 1/3)$ est la droite dirigée par le vecteur $(1, 1, 1)$;

$N_{\Lambda_3}(0, 1/2, 1/2)$ est le demi-plan constitué des vecteurs de la forme $(\alpha_0 + \alpha_1, \alpha_0, \alpha_0)$, où $\alpha_0 \in \mathbb{R}$ et $\alpha_1 \leqslant 0$.

$N_{\Lambda_3}(0, 0, 1)$ est le cône constitué des vecteurs de la forme $(\alpha_0 + \alpha_1, \alpha_0 + \alpha_2, \alpha_0)$, où $\alpha_0 \in \mathbb{R}$, $\alpha_1 \leqslant 0$ et $\alpha_2 \leqslant 0$.

2°) Il y a différentes manières d'arriver aux expressions annoncées de $T_{\Lambda_n}(\overline{x})$ et $N_{\Lambda_n}(\overline{x})$:

– à partir des définitions : le caractère polyédral du convexe (fermé) Λ_n facilite grandement les choses

$$T_{\Lambda_n}(\overline{x}) = \{\alpha(x-\overline{x}) \mid \alpha \geqslant 0 \text{ et } x \in \Lambda_n\}$$
$$N_{\Lambda_n}(\overline{x}) = [T_{\Lambda_n}(\overline{x})]^\circ \ ;$$

– à partir de la représentation de Λ_n comme conjonction d'un nombre fini d'inégalités affines $\langle a_j, x\rangle \leqslant b_j$; auquel cas :

$$T_{\Lambda_n}(\overline{x}) = \{d \in \mathbb{R}^n \mid \langle a_j, d\rangle \leqslant 0 \ \text{ pour tout } \ j \in J(\overline{x})\},$$
$$N_{\Lambda_n}(\overline{x}) = \left\{ \sum_{j\in J(\overline{x})} \alpha_j a_j \mid \alpha_j \geqslant 0 \ \text{ pour tout } \ j \in J(\overline{x}) \right\},$$

où $J(\overline{x})$ désigne $\{j \mid \langle a_j, \overline{x}\rangle = b_j\}$.

Dans le cas présent, si l'on pose $a_0 = (1,\ldots,1)$, $a_1 = -(1,0,\ldots,0),\ldots,$ $a_n = -(0,\ldots,0,1)$, Λ_n est représenté comme conjonction de $n+2$ inégalités affines

$$\langle a_0, x\rangle \leqslant 1,\ \langle -a_0, x\rangle \leqslant -1,\ \langle a_1, x\rangle \leqslant 0,\ldots,\ \langle a_n, x\rangle \leqslant 0.$$

Les deux premières contraintes-inégalités sont toujours actives en $\overline{x} \in \Lambda_n$, et certaines parmi les n autres peuvent l'être (de aucune à presque toutes $(n-1)$). En conséquence :

$$T_{\Lambda_n}(\overline{x}) = \left\{ d = (d_1,\ldots,d_n) \in \mathbb{R}^n \mid \sum_{i=1}^n d_i \leqslant 0,\ -\sum_{i=1}^n d_i \leqslant 0, \right.$$
$$\left. -d_i \leqslant 0 \ \text{ pour tout } \ i \in I(\overline{x}) \right\} ;$$
$$N_{\Lambda_n}(\overline{x}) = \mathbb{R}a_0 - \left\{ \sum_{i=1}^n \alpha_i' (0,\ldots,1,0,\ldots,0) \mid \alpha_i' \geqslant 0 \text{ si } \overline{x}_i = 0 \right.$$
$$\left. \text{et } \alpha_i' = 0 \text{ si } \overline{x}_i > 0 \right\}.$$

3°) Une condition nécessaire et suffisante pour que $\overline{x} \in \Lambda_n$ minimise la fonction (convexe et différentiable) f sur Λ_n est

$$-\nabla f(\overline{x}) = (-\partial_1 f(\overline{x}), \ldots, -\partial_n f(\overline{x})) \in N_{\Lambda_n}(\overline{x}).$$

Le résultat escompté découle de l'expression de $N_{\Lambda_n}(\overline{x})$ trouvée à la question précédente.

***Exercice III.11.** Soit S un ensemble non vide de $\mathbb{R}^n$; on désigne par S^c le complémentaire de S, et par fr$S$ la frontière de S. Montrer que si $\overline{x} \in \text{fr}S$,

$$T(S,\overline{x}) \cup T(S^c,\overline{x}) = \mathbb{R}^n, \quad T(frS,\overline{x}) = T(S,\overline{x}) \cap T(S^c,\overline{x}).$$

Solution : – Soit d quelconque dans $\mathbb{R}^n$ et posons $x_k := \overline{x} + \frac{1}{k}d, \ k \in \mathbb{N}^*$. Si $\{x_k\}$ possède une sous-suite contenue dans S, il est alors clair que $d \in T(S,\overline{x})$. Dans le cas contraire, il existe k_0 tel que $x_k \in S^c$ pour tout $k \geqslant k_0$; il s'ensuit que $d \in T(S^c,\overline{x})$.

En résumé, $d \in T(S,\overline{x}) \cup T(S^c,\overline{x})$.

– Puisque $\text{fr}S = \overline{S} \cap \overline{S^c}$, l'inclusion $T(\text{fr}S,\overline{x}) \subset T(S,\overline{x}) \cap T(S^c,\overline{x})$ est acquise.

Réciproquement, considérons $d \in T(S,\overline{x}) \cap T(S^c,\overline{x})$:

$\exists \{d_k\} \to d, \ \exists \{t_k\} \subset \mathbb{R}_*^+$ convergeant vers 0, telles que

$$\overline{x} + t_k d_k \in S \text{ pour tout } k ;$$

$\exists \left\{d_k'\right\} \to d', \ \exists \left\{t_k'\right\} \subset \mathbb{R}_*^+$ convergeant vers 0, telles que

$$\overline{x} + t_k' d_k' \in S^c \text{ pour tout } k.$$

Puisque $\overline{x} + t_k d_k \in S$ et $\overline{x} + t_k' d_k' \in S^c$, il existe $\alpha_k \in]0,1[$ tel que $\alpha_k(\overline{x} + t_k d_k) + (1-\alpha_k)(\overline{x} + t_k' d_k') \in \text{fr}S$. Posons

$$\tau_k := \alpha_k t_k + (1-\alpha_k) t_k', \ \ \delta_k := \frac{t_k \alpha_k}{\tau_k} d_k + \frac{(1-\alpha_k) t_k'}{\tau_k} d_k'.$$

Alors : $\{\tau_k\} \subset \mathbb{R}_*^+$ converge vers 0 ;

$\{\delta_k\}$ converge vers d (car chaque δ_k est une combinaison convexe de d_k et d_k', d'où $\| \delta_k - d \| \leqslant \max(\| d_k - d \|, \ \| d_k' - d \|)$) ; $\overline{x} + \tau_k \delta_k \in \text{fr}S$ pour tout k. Donc, $d \in T(\text{fr}S,\overline{x})$.

Remarque : La première formule n'offre pas un grand intérêt puisque, de manière générale, $T(A \cup B, \overline{x}) = T(A, \overline{x}) \cup T(B, \overline{x})$; la deuxième est plus inattendue puisque seule l'inclusion $T(A \cap B, \overline{x}) \subset T(A, \overline{x}) \cap T(B, \overline{x})$ est valide en général.

***Exercice III.12.** Soit S un fermé non vide de $\mathbb{R}^n$ et $\overline{x} \in S$. On désigne par $T(S, \overline{x})$ le cône tangent à S en $\overline{x}$.

1°) Montrer que $p \in [T(S, \overline{x})]^\circ$ si et seulement si :

$$\begin{cases} \forall\ \varepsilon > 0,\ \exists\ \delta > 0 \text{ tel que} \\ (x \in S \text{ et } \| x - \overline{x} \| \leqslant \delta) \Rightarrow (\langle p, x - \overline{x} \rangle \leqslant \varepsilon \| x - \overline{x} \|) . \end{cases} \tag{3.3}$$

2°) Soit f une fonction numérique telle que, pour tout $d \in \mathbb{R}^n$, le quotient différentiel suivant

$$\frac{f\left(\overline{x} + td'\right) - f(\overline{x})}{t}$$

admette une limite (finie) quant $t \to 0^+$et $d' \to d$. Cette limite est notée $f'(\overline{x}, d)$ (et appelée *dérivée directionnelle tangentielle* de f en $\overline{x}$ dans la direction d).

Montrer que la condition

$$f'(\overline{x}, d) > 0 \text{ pour tout } \ 0 \neq d \in T(S, \overline{x}) \tag{3.4}$$

entraîne que $\overline{x}$ est un minimum local strict de f sur S.

Que deviendrait (3.4) si $S = \mathbb{R}^n$ et f était différentiable en $\overline{x}$?

3°) Soit f une fonction numérique différentiable en $\overline{x}$. Vérifier l'équivalence des énoncés suivants :

(i) $\ -\nabla f(\overline{x}) \in [T(S, \overline{x})]^\circ$;

(ii) $\ \forall \varepsilon > 0,\ \langle \nabla f(\overline{x}), d \rangle > -\varepsilon \| d \|$ pour tout $0 \neq d \in T(S, \overline{x})$.

On suppose que $\overline{x}$ vérifie l'une des conditions équivalentes ci-dessus. Montrer que pour tout $\varepsilon > 0$, $\overline{x}$ est un minimum local strict de $f + \varepsilon \| \cdot - \overline{x} \|$ sur S.

Solution : 1°) Soit p vérifiant la propriété (3.3), et soit $d \in T(S, \overline{x})$; montrons que $\langle p, d \rangle \leqslant 0$.

Puisque $d \in T(S, \overline{x})$, il existe $\{t_k\} \subset \mathbb{R}_*^+$ tendant vers 0, $\{d_k\}$ tendant vers d, tels que $x_k := \overline{x} + t_k d_k \in S$ pour tout k. Donc $\| x_k - \overline{x} \| \leqslant \delta$ pour tout k assez grand (disons $k \geqslant k_0$). Par suite :

$$\forall \varepsilon > 0,\ \exists\ k_0 \text{ tel que } \langle p, d_k \rangle \leqslant \varepsilon \| d_k \| \ \text{ pour tout } k \geqslant k_0.$$

Donc, à la limite, $\langle p, d \rangle \leqslant 0$.

Réciproquement, soit $p \in [T(S,\overline{x})]^\circ$ et supposons que (3.3) n'ait pas lieu. Il existe donc $\overline{\varepsilon} > 0$, une suite $\{x_k\} \subset S$ convergeant vers $\overline{x}$, tels que

$$\langle p, x_k - \overline{x}\rangle > \overline{\varepsilon} \parallel x_k - \overline{x} \parallel \text{ pour tout } k. \tag{3.5}$$

Nécessairement $x_k \neq \overline{x}$ pour tout k. Posons $d_k := \frac{x_k - \overline{x}}{\|x_k - \overline{x}\|}$. La suite $\{d_k\}$ est dans la sphère-unité de $\mathbb{R}^n$; prenons-en une sous-suite convergente : $d_{k_l} \to d$ quand $l \to +\infty$. Par construction, $d \in T(S,\overline{x})$. Il vient de (3.5) :

$$\langle p, d\rangle \geqslant \overline{\varepsilon} > 0$$

ce qui entre en contradiction avec le fait que $p \in [T(S,\overline{x})]^\circ$.

2°) Supposons qu'il existe $\{x_k\} \subset S,\ x_k \neq \overline{x},\ x_k \to \overline{x}$ telle que $f(x_k) \leqslant f(\overline{x})$ pour tout k. Posons $d_k := \frac{x_k - \overline{x}}{\|x_k - \overline{x}\|}$, de sorte que $x_k = \overline{x} + \| x_k - \overline{x} \| d_k$. Quitte à prendre une sous-suite, on peut supposer que $d_k \to d$ quand $k \to +\infty$.

Évidemment $\| d \| = 1$ et $d \in T(S,\overline{x})$. Posant $t_k := \| x_k - \overline{x} \|$

$$0 \geqslant \frac{f(x_k) - f(\overline{x})}{t_k} = \frac{f(\overline{x} + t_k d_k) - f(\overline{x})}{t_k} \underset{k \to +\infty}{\longrightarrow} f'(\overline{x}, d) > 0.$$

D'où contradiction.

Si f était différentiable en $\overline{x}$, $f'(\overline{x}, d) = \langle \nabla f(\overline{x}), d\rangle$ pour tout $d \in \mathbb{R}^n$, et la condition (3.4) deviendrait alors :

$$\langle \nabla f(\overline{x}), d\rangle > 0 \text{ pour tout } d \in \mathbb{R}^n \setminus \{0\}$$

ce qui est impossible à réaliser !

3°) (i) dit : $\langle \nabla f(\overline{x}), d\rangle \geqslant 0$ pour tout $d \in T(S,\overline{x})$.

(ii) dit : $\forall \varepsilon > 0,\ \langle \nabla f(\overline{x}), d\rangle > -\varepsilon \| d \|$ pour tout $0 \neq d \in T(S,\overline{x})$.

Il est clair que $(i) \Leftrightarrow (ii)$.

Soit $g : x \longmapsto g(x) := f(x) + \varepsilon \| x - \overline{x} \|$. Il est facile de vérifier que $g'(\overline{x}, d) = \langle \nabla f(\overline{x}), d\rangle + \varepsilon \| d \|$.

D'après le résultat de la 2e question, appliqué à g, on obtient que $\overline{x}$ est un minimum local strict de g sur S.

En résumé :

– Si f n'est pas différentiable en le point $\overline{x}$ considéré, on peut espérer avoir une *condition suffisante de minimalité du 1er ordre* pour f (basée sur $f'(\overline{x},\cdot)$).

– En minimisation sans contraintes, une *condition nécessaire de minimalité du 1er ordre* comme $\nabla f(\overline{x}) = 0$ ne donne une information (du type d'être un minimum local) que sur la fonction *perturbée* $f + \varepsilon \parallel \cdot - \overline{x} \parallel$.

***Exercice III.13.** Soit $\mathcal{M}_n(\mathbb{R})$ structuré en espace euclidien à l'aide du produit scalaire

$$(A,B) \longmapsto \ll A,B \gg = \operatorname{tr}\left(A^\top B\right);$$

soit $\mathcal{N}$ une norme sur $\mathcal{M}_n(\mathbb{R})$ pour laquelle on pose :

$$\mathcal{S} := \{M \in \mathcal{M}_n(\mathbb{R}) : \mathcal{N}(M) = 1\} \quad \text{(sphère-unité)},$$
$$\mathcal{B} := \{M \in \mathcal{M}_n(\mathbb{R}) : \mathcal{N}(M) \leqslant 1\} \quad \text{(boule-unité fermée)}.$$

On définit
$\mathcal{N}_* : A \in \mathcal{M}_n(\mathbb{R}) \longmapsto \mathcal{N}_*(A) := \max\{\ll A,M \gg : M \in \mathcal{S}\}$.
$\mathcal{N}_*$ est une norme sur $\mathcal{M}_n(\mathbb{R})$ (ce que l'on ne demande pas de démontrer).
On considère la fonction $f : \mathcal{M}_n(\mathbb{R}) \to \mathbb{R}$ qui à X associe
$f(X) :=$ dét X.

1°) Montrer que le maximum de f sur $\mathcal{S}$ est également le maximum de f sur $\mathcal{B}$.

2°) Soit $\overline{X}$ un point de $\mathcal{S}$ où f atteint son maximum. Montrer que $\overline{X}$ est inversible et que

$$\mathcal{N}_*\left[\left(\overline{X}^{-1}\right)^\top\right] \geqslant n.$$

3°) a) Quelle est la différentielle de f en $X \in \mathcal{M}_n(\mathbb{R})$ (forme linéaire sur $\mathcal{M}_n(\mathbb{R})$) ? le gradient de f en X (élément de $\mathcal{M}_n(\mathbb{R})$) ? On demande simplement de rappeler les résultats.

b) En écrivant la condition nécessaire de maximalité en $\overline{X}$, montrer :

$$\ll \left(\overline{X}^{-1}\right)^\top, M - \overline{X} \gg \leqslant 0 \text{ pour tout } M \in \mathcal{B}.$$

En déduire que $\mathcal{N}_*\left[\left(\overline{X}^{-1}\right)^\top\right] = n$.

Solution : L'application $M \longmapsto \ll A, M \gg$ est une forme linéaire sur $\mathcal{M}_n(\mathbb{R})$, donc continue (pourquoi ?). La sphère-unité $\mathcal{S}$ étant compacte, il existe bien $\overline{M} \in \mathcal{S}$ telle que

$$\mathcal{N}_*(A) = \ll A, \overline{M} \gg = \sup\{\ll A, M \gg : M \in \mathcal{S}\}.$$

Si $M \in \mathcal{S}$, il en est de même de $-M$, d'où $\mathcal{N}_*(A) \in \mathbb{R}^+$.

Les propriétés suivantes sont immédiates :

$$\mathcal{N}_*(0) = 0 ;$$

$$\forall A \in \mathcal{M}_n(\mathbb{R}), \forall \lambda \in \mathbb{R},\ \mathcal{N}_*(\lambda A) = |\lambda| \mathcal{N}_*(A) ;$$

$$\forall A_1 \in \mathcal{M}_n(\mathbb{R}), \forall A_2 \in \mathcal{M}_n(\mathbb{R}),\ \mathcal{N}_*(A_1 + A_2) \leqslant \mathcal{N}_*(A_1) + \mathcal{N}_*(A_2).$$

Comme $\ll A, M \gg \leqslant \mathcal{N}_*(A)\mathcal{N}(M)$ pour tout $(A, M) \in \mathcal{M}_n(\mathbb{R}) \times \mathcal{M}_n(\mathbb{R})$,

$$(\mathcal{N}_*(A) = 0) \Rightarrow (\ll A, M \gg \leqslant 0 \text{ pour tout } M \in \mathcal{M}_n(\mathbb{R})),\ \text{soit } A = 0.$$

$\mathcal{N}_*$ est bien une *norme* sur $\mathcal{M}_n(\mathbb{R})$.

Il est clair que $\mathcal{N}_*(A)$ est également le maximum de $\ll A, \cdot \gg$ sur $\mathcal{B}$. En fait $\mathcal{N}_*$ est ce qu'on appelle la *fonction d'appui* de $\mathcal{B}$ (ou de $\mathcal{S}$) ; c'est la norme duale de $\mathcal{N}$. Mais attention, $\mathcal{N}$ n'est pas $(\ll \cdot, \cdot \gg)^{1/2}$!

1°) Si $0 \neq M \in \mathcal{M}_n(\mathbb{R})$, $M/\mathcal{N}(M) \in \mathcal{S}$ de sorte que

$$f\left(\frac{M}{\mathcal{N}(M)}\right) = \frac{1}{[\mathcal{N}(M)]^n} f(M) \leqslant \max_{S \in \mathcal{S}} f(S).$$

Ainsi, dans tous les cas, $f(M) \leqslant [\mathcal{N}(M)]^n \max_{S \in \mathcal{S}} f(S)$.

En particulier, $f(M) \leqslant \max_{S \in \mathcal{S}} f(S)$ pour tout $M \in \mathcal{B}$, d'où $\max_{M \in \mathcal{B}} f(M)$ est égal à $\max_{S \in \mathcal{S}} f(S)$.

2°) f est continue, $\mathcal{S}$ est compact, donc f atteint bien son maximum en un point $\overline{X}$ de $\mathcal{S}$.

Il est clair que $f(\overline{X}) \geqslant 0$. Pour s'assurer du caractère inversible de $\overline{X}$, il suffit de montrer que $\det \overline{X} = f(\overline{X}) > 0$.

La matrice $\frac{I_n}{\mathcal{N}(I_n)} \in \mathcal{S}$ et $f\left(\frac{I_n}{\mathcal{N}(I_n)}\right) = \frac{1}{[\mathcal{N}(I_n)]^n} > 0$, donc $f(\overline{X}) > 0$.

On a

$$\mathcal{N}_*\left[\left(\overline{X}^{-1}\right)^\top\right] \geqslant \ll \left(\overline{X}^{-1}\right)^\top, M \gg = \operatorname{tr}\left(\overline{X}^{-1} M\right) \text{ pour tout } M \in \mathcal{S}.$$

En particulier, puisque $\overline{X} \in \mathcal{S}$, $\mathcal{N}_*\left[\left(\overline{X}^{-1}\right)^\top\right] \geqslant \operatorname{tr} I_n = n$.

3°) a) Rappelons que $Df(X) : H \in \mathcal{M}_n(\mathbb{R}) \longmapsto Df(X)\cdot H = \mathrm{tr}((\mathrm{cof}\, X)^\top H)$, où $\mathrm{cof}\, X$ désigne la matrice des cofacteurs de X. D'où $\nabla f(X) = \mathrm{cof}\, X$.

b) Une condition nécessairement vérifiée par $\overline{X}$ maximisant f sur $\mathcal{S}$ (ou sur $\mathcal{B}$) est :

$$\ll \nabla f(\overline{X}), M - \overline{X} \gg \leqslant 0 \text{ pour tout } M \in \mathcal{B}.$$

En X inversible, $X^{-1} = \frac{1}{\det X}(\mathrm{cof}\, X)^\top$, d'où $\nabla f(X) = \det X \left(X^{-1}\right)^\top$.
La condition nécessaire de maximalité évoquée plus haut devient :

$$\ll \left(\overline{X}^{-1}\right)^\top, M - \overline{X} \gg \leqslant 0 \text{ pour tout } M \in \mathcal{B} \text{ (car } \det \overline{X} > 0)$$

soit

$$\ll \left(\overline{X}^{-1}\right)^\top, M \gg \leqslant \ll \left(\overline{X}^{-1}\right)^\top, \overline{X} \gg = \mathrm{tr}\left(\overline{X}^{-1}\overline{X}\right) = n \text{ pour tout } M \in \mathcal{B}.$$

Par suite, $\mathcal{N}_*\left[\left(\overline{X}^{-1}\right)^\top\right] \leqslant n$.

***Exercice III.14.** Soit S un fermé non vide de $\mathbb{R}^n$ et $x \notin S$. On désigne par $P_S(x)$ l'ensemble des $\overline{x} \in S$ tels que $\| x - \overline{x} \| = d_S(x)$. Montrer que

$$x - \overline{x} \in [T(S, \overline{x})]^\circ \text{ pour tout } \overline{x} \in P_S(x).$$

Solution : Les points $\overline{x}$ de $P_S(x)$ sont les solutions du problème de minimisation suivant :

$$\begin{cases} \mathrm{Min}\, f(x) := \frac{1}{2} \| x - s \|^2 \\ s \in S. \end{cases}$$

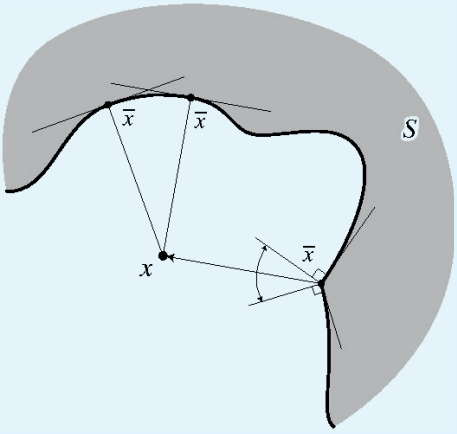


FIGURE 8.

En une solution $\overline{x}$ de ce problème, on écrit la condition nécessaire d'optimalité du 1er ordre, à savoir : $-\nabla f(\overline{x}) \in [T(S,\overline{x})]^\circ$.

Comme $\nabla f(\overline{x}) = -(x-\overline{x})$, le résultat annoncé est démontré.

Remarque : Si g est la fonction $\frac{1}{2}d_S^2$, on peut montrer que g admet en tout $x \notin S$ une dérivée directionnelle (tangentielle) qui est

$$d \longmapsto g'(x,d) = \min\{\langle x-\overline{x}, d\rangle \mid \overline{x} \in P_S(x)\}.$$

On en déduit que g est différentiable en x si et seulement si $P_S(x)$ est réduit à un seul élément.

***Exercice III.15.** *Le lemme de H. Everett*

Soit le problème d'optimisation suivant :

$$(\mathcal{P}) \begin{cases} \text{Minimiser } f(x) \text{ sous les contraintes} \\ h_i(x) = 0 \text{ pour } i = 1,\dots,m \text{ et } g_j(x) \leqslant 0 \text{ pour } j = 1,\dots,p. \end{cases}$$

On définit le lagrangien usuel

$$\mathcal{L} : (x,\lambda,\mu) \in \mathbb{R}^n \times \mathbb{R}^m \times \left(\mathbb{R}^+\right)^p \longmapsto$$
$$\mathcal{L}(x,\lambda,\mu) := f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{j=1}^{p} \mu_j g_j(x).$$

Soit $(\overline{\lambda},\overline{\mu})$ un élément quelconque de $\mathbb{R}^m \times (\mathbb{R}^+)^p$ et soit $x_{\overline{\lambda},\overline{\mu}}$ un point de $\mathbb{R}^n$ minimisant $\mathcal{L}(\cdot,\overline{\lambda},\overline{\mu})$ sur $\mathbb{R}^n$.

Vérifier que $x_{\overline{\lambda},\overline{\mu}}$ est solution du problème d'optimisation $(\mathcal{P}_\varepsilon)$ suivant :

$$(\mathcal{P}_\varepsilon) \quad \begin{cases} \text{Minimiser } \ f(x) \ \text{ sous les contraintes} \\ h_i(x) = \varepsilon_i \text{ pour } i \in I \text{ et } g_j(x) \leqslant \varepsilon_j \text{ pour } j \in J, \end{cases}$$

où

$$I := \{i : \overline{\lambda}_i \neq 0\} \ \text{ et } \varepsilon_i := h_i\left(x_{\overline{\lambda},\overline{\mu}}\right) \ \text{ pour } i \in I,$$
$$J := \{j : \overline{\mu}_j > 0\} \ \text{ et } \varepsilon_j := g_j\left(x_{\overline{\lambda},\overline{\mu}}\right) \ \text{ pour } j \in J.$$

Solution : Par définition de $x_{\overline{\lambda},\overline{\mu}}$, on a :

$$f\left(x_{\overline{\lambda},\overline{\mu}}\right)+\sum_{i\in I}\overline{\lambda}_i h_i\left(x_{\overline{\lambda},\overline{\mu}}\right)+\sum_{j\in J}\overline{\mu}_j g_j\left(x_{\overline{\lambda},\overline{\mu}}\right)$$
$$\leqslant f(x)+\sum_{i\in I}\overline{\lambda}_i h_i(x)+\sum_{j\in J}\overline{\mu}_j g_j(x) \text{ pour tout } x\in\mathbb{R}^n. \qquad (3.6)$$

Soit x vérifiant : $h_i(x)=h_i(x_{\overline{\lambda},\overline{\mu}})$ pour $i\in I$, $g_j(x)\leqslant g_j(x_{\overline{\lambda},\overline{\mu}})$ pour $j\in J$. Il est trivial que $x_{\overline{\lambda},\overline{\mu}}$, entre autres, vérifie ces inégalités.
Au vu de l'inégalité (3.6), et sachant que $\overline{\mu}_j\geqslant 0$, on a bien $f(x_{\overline{\lambda},\overline{\mu}})\leqslant f(x)$. Donc, $x_{\overline{\lambda},\overline{\mu}}$ vérifie les contraintes de $(\mathcal{P}_\varepsilon)$ et est une solution de $(\mathcal{P}_\varepsilon)$.

***Exercice III.16.** Étant donné $Q\in\mathcal{S}_n(\mathbb{R})$ définie positive, $c\in\mathbb{R}^n$, $A\in\mathcal{M}_{m,n}(\mathbb{R})$ de rang m $(m\leqslant n)$ et $b\in\mathbb{R}^m$, on considère le problème de minimisation suivant :

$$(\mathcal{P}_b)\begin{cases}\text{Min } f(x):=\frac{1}{2}\langle Qx,x\rangle+\langle c,x\rangle\\ Ax=b.\end{cases}$$

1°) Indiquer rapidement pourquoi $(\mathcal{P}_b)$ a toujours une et une seule solution.

2°) Déterminer explicitement la solution $\overline{x}$ de $(\mathcal{P}_b)$ ainsi que le vecteur-multiplicateur $\overline{\lambda}\in\mathbb{R}^m$ associé.

3°) Pour $u\in\mathbb{R}^m$, on pose

$$\varphi(u):=\inf\{f(x)\mid Ax=b+u\}.$$

Quelles sont les propriétés essentielles de la fonction φ (convexité, différentiabilité, etc.) ?

Solution : 1°) L'application linéaire $A:\mathbb{R}^n\to\mathbb{R}^m$ étant surjective, l'ensemble-contrainte de $(\mathcal{P}_b)$ est un convexe fermé *non vide* de $\mathbb{R}^n$, et ce quel que soit $b\in\mathbb{R}^m$ (c'est en fait un sous-espace affine de $\mathbb{R}^n$ dont la direction est le sous-espace vectoriel $\mathrm{Ker}A$).

La fonction f est convexe (fortement convexe même) sur $\mathbb{R}^n$ et $\lim\limits_{\|x\|\to+\infty} f(x)=+\infty$.

En conséquence, il existe un et un seul $\overline{x} \in \mathbb{R}^n$ minimisant $f(x)$ sous la contrainte $Ax = b$.

2°) La solution $\overline{x}$ de $(\mathcal{P}_b)$ est *caractérisée* par les relations suivantes :

$$\begin{cases} A\overline{x} = b \ ; \\ \text{Il existe } \overline{\lambda} \in \mathbb{R}^m \text{ tel que } \nabla f\,(\overline{x}) + A^\top \overline{\lambda} = 0 \\ \text{soit } Q\overline{x} + c + A^\top \overline{\lambda} = 0. \end{cases}$$

Comme Q est inversible, ceci est équivalent à :

$$A\overline{x} = b \text{ et } \overline{x} + Q^{-1}c + Q^{-1}A^\top \overline{\lambda} = 0$$

ou encore

$$A\overline{x} = b \text{ et } AQ^{-1}A^\top \overline{\lambda} + AQ^{-1}c + b = 0. \tag{3.7}$$

Or l'application linéaire $AQ^{-1}A^\top : \mathbb{R}^m \to \mathbb{R}^m$ est symétrique et définie positive ; en effet,

$$\left\langle AQ^{-1}A^\top y, y \right\rangle = \left\langle Q^{-1}\left(A^\top y\right), A^\top y \right\rangle \geqslant 0 \ \text{ pour tout } \ y \in \mathbb{R}^m,$$

et

$$\left\langle AQ^{-1}A^\top y, y \right\rangle = \left\langle Q^{-1}\left(A^\top y\right), A^\top y \right\rangle = 0$$

si et seulement si $A^\top y = 0$, soit $y = 0$ ($A^\top$ étant injective puisque A est surjective).

On a ainsi à partir de (3.7) l'expression explicite de $\overline{x}$ et $\overline{\lambda}$:

$$\begin{aligned} \overline{x} &= Q^{-1}A^\top B\left(AQ^{-1}c + b\right) - Q^{-1}c, \\ \overline{\lambda} &= -B\left(AQ^{-1}c + b\right), \end{aligned}$$

où $B := \left(AQ^{-1}A^\top\right)^{-1}$.

3°) On peut expliciter $\varphi(u)$ complètement. Sans aller faire ce calcul, nous savons que $\varphi : \mathbb{R}^m \to \mathbb{R}$ est convexe, et nous sommes dans les conditions assurant que φ est différentiable en 0 avec $\nabla\varphi(0) = -\overline{\lambda}$.

***Exercice III.17.** Considérons le problème de la minimisation de f, supposée de classe C^1, sous les contraintes $h_1(x) = 0, \ldots, h_m(x) = 0$ et $g_1(x) \leqslant 0, \ldots, g_p(x) \leqslant 0$, où les fonctions h_i et g_j sont aussi supposées de classe C^1.

Étant donné $x \in \mathbb{R}^n$ et $c = (\alpha_1, \dots, \alpha_m, \beta_1, \dots \beta_p) \in \mathbb{R}^m \times \mathbb{R}^p$, on pose

$$F(x,c) := \begin{pmatrix} \nabla f(x) + \sum\limits_{i=1}^{m} \alpha_i \nabla h_i(x) + \\ \qquad\qquad \sum\limits_{j=1}^{p} \beta_j^+ \nabla g_j(x) \\ h_1(x) \\ h_2(x) \\ \vdots \\ h_m(x) \\ -g_1(x) + \beta_1^- \\ -g_2(x) + \beta_2^- \\ \vdots \\ -g_p(x) + \beta_p^- \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$$

où $\beta_j^+ := \max(0, \beta_j)$ et $\beta_j^- := \min(0, \beta_j)$.

1°) F est-elle continue ? différentiable ? Donner des conditions suffisantes portant sur les données du problème de minimisation pour que F soit localement lipschitzienne.

2°) Montrer que $\overline{x}$ vérifie les conditions de KKT si, et seulement si, il existe $\overline{c} = (\overline{\alpha}_1, \dots, \overline{\alpha}_m, \overline{\beta}_1, \dots, \overline{\beta}_p)$ tel que $F(\overline{x}, \overline{c}) = 0$.

Solution : 1°) F est continue, mais n'est pas différentiable (à cause des fonctions $\beta_j \longmapsto \beta_j^+$ et $\beta_j \longmapsto \beta_j^-$ qui ne sont pas différentiables en 0).

Les fonctions f, h_i, g_j étant de classe C^1, elles sont localement lipschitziennes (comme fonctions de x, donc de (x,c)) ; par ailleurs les fonctions $\beta_j \longmapsto \beta_j^+$ et $\beta_j \longmapsto \beta_j^-$ sont lipschitziennes. Par conséquent, F sera localement lipschtzienne sur $\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ dès lors que les ∇f, ∇h_i, et ∇g_j sont localement lipschitziennes sur $\mathbb{R}^n$.

2°) Rappelons que $\overline{x}$ vérifie les conditions de KKT si $h_i(\overline{x}) = 0$ pour tout i, $g_j(\overline{x}) \leqslant 0$ pour tout j, et s'il existe $(\overline{\lambda}, \overline{\mu}) \in \mathbb{R}^m \times \mathbb{R}^p$ tel que :

(i) $\nabla f(\overline{x}) + \sum\limits_{i=1}^{m} \overline{\lambda}_i \nabla h_i(\overline{x}) + \sum\limits_{j=1}^{p} \overline{\mu}_j \nabla g_j(\overline{x}) = 0$

(ii) $\overline{\mu}_j \geqslant 0$ pour tout $j = 1, \dots, p$

(iii) $\overline{\mu}_j = 0$ lorsque $g_j(\overline{x}) < 0$, $j = 1, \dots, p$.

Si $\overline{x}$ vérifie les conditions de KKT, on définit $\overline{c} = (\overline{\alpha}_1, \ldots, \overline{\alpha}_m, \overline{\beta}_1, \ldots \overline{\beta}_p)$ comme suit :

$$\overline{\alpha}_i = \overline{\lambda}_i \text{ pour tout } i = 1, \ldots, m$$
$$\overline{\beta}_j = \overline{\mu}_j \text{ si } g_j(\overline{x}) = 0, \quad \overline{\beta}_j = g_j(\overline{x}) \text{ si } g_j(\overline{x}) < 0.$$

Alors $F(\overline{x}, \overline{c}) = 0$.

Réciproquement, soit $(\overline{x}, \overline{c})$, avec $\overline{c} = (\overline{\alpha}, \overline{\beta})$, tel que $F(\overline{x}, \overline{c}) = 0$. On a tout d'abord :

$$h_i(\overline{x}) = 0 \text{ pour tout } i = 1, \ldots, m$$
$$g_j(\overline{x}) = \overline{\beta_j^-} \leqslant 0 \text{ pour tout } j = 1, \ldots, p.$$

Ensuite, posons

$$\overline{\lambda}_i = \overline{\alpha}_i \text{ pour tout } i = 1, \ldots, m$$
$$\overline{\mu}_j = \overline{\beta}_j^+ \text{ pour tout } j = 1, \ldots, p.$$

Alors $(\overline{\lambda}, \overline{\mu})$ vérifie les conditions dans (i), (ii), (iii) énoncées plus haut.

Commentaire : On ramène ainsi la recherche des points vérifiant les conditions de KKT à la résolution d'un système d'équations non linéaires, mais les fonctions non linéaires intervenant dans les équations ne sont pas deux fois différentiables, ce qui oblige à repenser des méthodes usuelles de résolution comme celle de Newton.

***Exercice III.18.** Soit $\mathbb{R}^n$ muni de son produit scalaire usuel $\langle \cdot, \cdot \rangle$ et de la norme euclidienne $\| \cdot \|$ associée. Étant donné $a \in \mathbb{R}^n$, on considère

$$f_a : \Omega := \{x \in \mathbb{R}^n : \| x \| < 1\} \to \mathbb{R}$$
$$x \longmapsto f_a(x) := -\ln\left(1 - \| x \|^2\right) + \langle a, x \rangle .$$

1°) Montrer que f_a est strictement convexe sur Ω.

2°) On considère le problème de minimisation suivant :

$$(\mathcal{P}_a) \quad \begin{cases} \text{Minimiser } f_a(x) \\ x \in C_a := \left\{x \in \mathbb{R}^n : \| x \| \leqslant \frac{1}{2} \text{ et } \langle a, x \rangle \leqslant 0\right\}. \end{cases}$$

a) Résoudre $(\mathcal{P}_a)$ pour $a = 0$.

b) On suppose $a \neq 0$ et on désigne par $\overline{x}$ la solution de $(\mathcal{P}_a)$.

Montrer que $\langle a, \overline{x}\rangle < 0$ nécessairement et donc que le multiplicateur $\overline{\mu}_2$ associé à cette contrainte est nul.

En distinguant selon que la contrainte $\| x \| \leqslant \dfrac{1}{2}$ est active ou pas en $\overline{x}$, déterminer $\overline{x}$ et le multiplicateur $\overline{\mu}_1$ associé à cette contrainte (on sera amené à discuter sur les valeurs prises par $\| a \|$).

Solution : 1°) Il y a au moins deux façons de montrer que f_a est strictement convexe sur l'ouvert convexe Ω.

(i) La fonction $x \longmapsto \varphi(x) := 1 - \| x \|^2$ est strictement concave sur Ω ; la fonction $y \longmapsto \psi(y) := -\ln y$ est strictement décroissante et convexe sur $\mathbb{R}_*^+$. Par suite, la fonction $\psi \circ \varphi : x \longmapsto -\ln(1 - \| x \|^2)$ est strictement convexe sur Ω ; en effet :

$\forall \alpha \in]0,1[, \forall x, x' \in \Omega,\ x \neq x'$,

$$\varphi\left(\alpha x + (1-\alpha)\, x'\right) > \alpha\varphi(x) + (1-\alpha)\,\varphi\left(x'\right) \ ;$$

d'où

$$\begin{aligned}(\psi \circ \varphi)\left(\alpha x + (1-\alpha)\, x'\right) &\leqslant \psi\left[\alpha\varphi(x) + (1-\alpha)\,\varphi\left(x'\right)\right] \\ &\leqslant \alpha\,(\psi \circ \varphi)(x) + (1-\alpha)\,(\psi \circ \varphi)\left(x'\right).\end{aligned}$$

(ii) La fonction f_a est deux fois différentiable sur Ω avec :

$$\forall x \in \Omega, \quad \nabla^2 f_a(x) = \frac{4}{(1 - \| x \|^2)^2} x x^\top + \frac{2}{1 - \| x \|^2} I_n.$$

Ainsi :

$$\forall d \in \mathbb{R}^n,\ \left\langle \nabla^2 f_a(x)\, d, d\right\rangle = \tfrac{4}{(1-\|x\|^2)^2}\langle x, d\rangle^2 + \tfrac{2}{1-\|x\|^2}\| d \|^2 \geqslant 0 \ ;$$

$$\left\langle \nabla^2 f_a(x)\, d, d\right\rangle > 0 \text{ si } d \neq 0.$$

$\nabla^2 f_a(x)$ est définie positive pour tout $x \in \Omega$: f_a est donc strictement convexe sur Ω. S'il y a une solution dans $(\mathcal{P}_a)$ elle est unique.

2°) a) Il s'agit de minimiser $f_0(x) := -\ln(1 - \| x \|^2)$ sur le convexe compact $C_0 := \left\{x \in \mathbb{R}^n \mid \| x \| \leqslant \frac{1}{2}\right\} \subset \Omega$. On peut considérer que C_0 est représenté sous forme d'une inégalité : $g_1(x) := \| x \|^2 - \frac{1}{4} \leqslant 0$. Comme il existe x_0 tel que $g_1(x_0) < 0$ (prendre $x_0 = 0$ par exemple), la solution $\overline{x}$ de $(\mathcal{P}_0)$ est caractérisée par l'existence de $\overline{\mu}_1 \geqslant 0$ tel que :

$$\begin{cases} \nabla f_0(\overline{x}) + \overline{\mu}_1 \nabla g_1(\overline{x}) = 0, \\ \overline{\mu}_1 = 0 \text{ si } g_1(\overline{x}) < 0. \end{cases}$$

La solution $\overline{x}$ peut-elle être sur la frontière de C_0 ? Dans ce cas on aurait :

$$\| \overline{x} \| = \frac{1}{2} \ \nabla f_0(\overline{x}) + \overline{\mu}_1 \nabla g_1(\overline{x}) = \left(\frac{4}{3} + \overline{\mu}_1 \right) \overline{x} = 0 \text{ avec } \overline{\mu}_1 \geqslant 0.$$

Ceci est impossible.

Donc $\overline{x}$ est à l'intérieur de C_0, et vérifie $\nabla f_0(\overline{x}) = 0$. Il en résulte que $\overline{x} = 0$.

Évidemment, dans ce cas particulier, on aurait pu déduire directement que $f_0(x) \geqslant f_0(0)$ pour tout $x \in C_0$.

b) La fonction f_a est continue, strictement convexe sur le convexe compact $C_a \subset \Omega$; le problème $(\mathcal{P}_a)$ a une et une seule solution $\overline{x}$ qu'il s'agit de déterminer. On voit facilement qu'il existe x_0 tel que $g_1(x_0) < 0$ et $g_2(x_0) := \langle a, x_0 \rangle < 0$ (hypothèse de Slater). En conséquence, $\overline{x} \in C_a$ est solution de $(\mathcal{P}_a)$ si et seulement si : $\exists\ \overline{\mu}_1 \geqslant 0,\ \overline{\mu}_2 \geqslant 0$ tels que

$$\begin{cases} \nabla f_a(\overline{x}) + \overline{\mu}_1 \nabla g_1(\overline{x}) + \overline{\mu}_2 \nabla g_2(\overline{x}) = 0, \\ \overline{\mu}_1 g_1(\overline{x}) = \overline{\mu}_2 g_2(\overline{x}) = 0 \ ; \end{cases}$$

ce qui donne :

$$\frac{2\overline{x}}{1 - \| \overline{x} \|^2} + a + 2\overline{\mu}_1 \overline{x} + \overline{\mu}_2 a = 0, \tag{3.8}$$

$$\overline{\mu}_1 \left(\| \overline{x} \|^2 - \frac{1}{4} \right) = \overline{\mu}_2 \langle a, \overline{x} \rangle = 0. \tag{3.9}$$

Supposons $g_2(\overline{x}) = \langle a, \overline{x} \rangle = 0$. En faisant le produit scalaire avec a dans (3.8), on obtient $(1 + \overline{\mu}_2) \| a \|^2 = 0$, ce qui oblige à avoir $a = 0$.

Donc $g_2(\overline{x}) < 0$ et $\overline{\mu}_2 = 0$. Les conditions (3.8) et (3.9) simplifiées deviennent :

$$\frac{2\overline{x}}{1 - \| \overline{x} \|^2} + a + 2\overline{\mu}_1 \overline{x} = 0, \tag{3.10}$$

$$\overline{\mu}_1 = 0 \text{ si } \| \overline{x} \| < \frac{1}{2}. \tag{3.11}$$

Dans tous les cas, $\overline{x}$ et a sont colinéaires : $\overline{x} = ra$ avec $r < 0$.

1^{re} *possibilité pour* $\overline{x}$: $\| \overline{x} \| < \dfrac{1}{2}$, et $\dfrac{2\overline{x}}{1 - \| \overline{x} \|^2} + a = 0$.

Ceci ne peut se produire que si $\| a \| < \dfrac{4}{3}$, auquel cas

$$\overline{x} = -\frac{a}{\| a \|^2} \left(\sqrt{1 + \| a \|^2} - 1 \right).$$

2^e *possibilité pour* $\overline{x}$: $\| \overline{x} \| = \dfrac{1}{2}$, et (3.10) pour un certain $\overline{\mu}_1 \geqslant 0$.

Après quelques calculs, on constate que ceci n'a lieu que pour $\| a \| \geqslant \dfrac{4}{3}$, auquel cas :

$\overline{\mu}_1 = \| a \| - \dfrac{4}{3}$ et $\overline{x} = -\dfrac{a}{2 \| a \|}$.

En résumé :

– Si $0 < \| a \| < \dfrac{4}{3}$, la solution de $(\mathcal{P}_a)$ est $\overline{x} = -\dfrac{a}{\| a \|^2}\left(\sqrt{1+ \| a \|^2} - 1\right)$, à l'intérieur de C_a ;

– Si $\| a \| \geqslant \dfrac{4}{3}$, la solution de $(\mathcal{P}_a)$ est $\overline{x} = -\frac{a}{2\|a\|}$, sur la frontière de C_a.

***Exercice III.19.** On considère dans $\mathbb{R}^2$ le problème de minimisation suivant :

$$(\mathcal{P}) \begin{cases} \text{Min } f(\xi_1, \xi_2) := \xi_1^3 + \xi_2^2 \\ g(\xi_1, \xi_2) := \xi_1^2 + \xi_2^2 - 9 \leqslant 0. \end{cases}$$

1°) Déterminer les points vérifiant les conditions nécessaires de minimalité du 1er ordre.

2°) En déduire les solutions de $(\mathcal{P})$.

Solution : La fonction f n'étant pas convexe, $(\mathcal{P})$ n'est pas un problème de minimisation convexe.

1°) La fonction g définissant la seule contrainte de type inégalité est convexe ; de plus, il existe x_0 tel que $g(x_0) < 0$. En conséquence, une solution $\overline{x} = (\overline{\xi}_1, \overline{\xi}_2)$ de $(\mathcal{P})$ – et il y en a – vérifie les conditions nécessaires de minimalité du 1er ordre, à savoir :

$$\exists\, \overline{\mu} \geqslant 0 \text{ tel que } \nabla f(\overline{x}) + \overline{\mu} \nabla g(\overline{x}) = 0 \text{ et } \overline{\mu} g(\overline{x}) = 0,$$

soit

$$3\overline{\xi}_1^2 + 2\overline{\mu}\overline{\xi}_1 = 0,\ (1 + \overline{\mu})\,\overline{\xi}_2 = 0,\ \overline{\mu} = 0 \text{ si } \overline{\xi}_1^2 + \overline{\xi}_2^2 - 9 < 0.$$

Il en ressort deux possibilités :

$$(\overline{\xi}_1, \overline{\xi}_2) = (0, 0) \text{ avec } \overline{\mu} = 0,$$
$$(\overline{\xi}_1, \overline{\xi}_2) = (-3, 0) \text{ avec } \overline{\mu} = \frac{9}{2}.$$

2°) Comme $f(0,0) = 0$ et $f(-3,0) = -27$, la seule solution de $(\mathcal{P})$ est $\overline{x} = (-3,0)$, et la valeur optimale correspondante est $\overline{f} = -27$.

On notera que $\nabla^2 f(0,0) = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}$, mais $\overline{x}$ n'est pas un minimum local de f ; pas plus qu'il n'est un point-selle de f d'ailleurs.

***Exercice III.20.** Soit C l'ensemble de $\mathbb{R}^2$ défini par

$$C := \{(\xi_1, \xi_2) \mid \xi_1 + \xi_2 \leqslant 1,\ \xi_1 \geqslant 0,\ \xi_2 \geqslant 0\},$$

et $f : \mathbb{R}^2 \to \mathbb{R}$ définie par $f(\xi_1, \xi_2) := -\xi_1 - 2\xi_2 - 2\xi_1\xi_2 + \frac{\xi_1^2}{2} + \frac{\xi_2^2}{2}$.

1°) La fonction f est-elle convexe ? concave ?

2°) On considère le problème de la minimisation de f sur C.

(i) Montrer que tout minimum (même local) se trouve sur la frontière $\operatorname{fr} C$ de C.

(ii) En explicitant $T(C, \overline{x})$ (où $[T(C,\overline{x})]^\circ = N(C,\overline{x})$) en tout point $\overline{x}$ de $\operatorname{fr} C$, montrer qu'il n'existe qu'un point de C vérifiant la condition nécessaire de minimalité du 1$^{\text{er}}$ ordre. En déduire l'unique solution du problème de la minimisation de f sur C.

3°) Résoudre à présent le problème de la maximisation de f sur C.

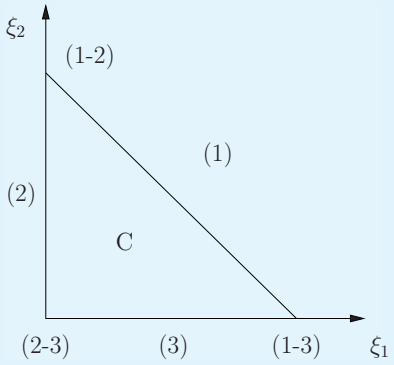
Solution : 1°) On a affaire à une fonction f quadratique qui n'est ni convexe ni concave.

En effet, $\nabla^2 f(\xi_1, \xi_2) = \begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix}$ n'est ni semi-définie positive ni semi-définie négative.

2°) (i) Un minimum local (ou maximum local) $\overline{x}$ de f qui se trouverait à l'intérieur de C devrait vérifier $\nabla f(\overline{x}) = 0$. Or le seul point $\overline{x} = (\overline{\xi}_1, \overline{\xi}_2)$ vérifiant $\nabla f(\overline{x}) = \begin{pmatrix} -1 + \overline{\xi}_1 - 2\overline{\xi}_2 \\ -2 - 2\overline{\xi}_1 + \overline{\xi}_2 \end{pmatrix} = 0$ est $(-\frac{5}{3}, -\frac{4}{3})$ qui est à l'extérieur de C.

Mais f étant continue et C compact, le problème de la minimisation de f sur C (comme celui de la maximisation de f sur C) *admet* des solutions. Ces solutions se trouvent donc nécessairement sur la frontière de C.

(ii) Examinons successivement tous les cas de figure : $\overline{x}$ est sur l'une des arêtes (1), (2), (3), puis $\overline{x}$ est l'un des sommets $(1-2)$, $(2-3)$, $(1-3)$.



$$(1) : T(C,\overline{x}) = \{(d_1,d_2) \mid d_1 + d_2 \leqslant 0\}\,,\ [T(C,\overline{x})]^\circ = N(C,\overline{x}) = \mathbb{R}^+ \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

$$(2) : T(C,\overline{x}) = \{(d_1,d_2) \mid d_1 \geqslant 0\}\,,\ N(C,\overline{x}) = \mathbb{R}^+ \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

$$(3) : T(C,\overline{x}) = \{(d_1,d_2) \mid d_2 \geqslant 0\}\,,\ N(C,\overline{x}) = \mathbb{R}^+ \begin{pmatrix} 0 \\ -1 \end{pmatrix}.$$

$$(1-2) : T(C,\overline{x}) = \{(d_1,d_2) \mid d_1 + d_2 \leqslant 0 \text{ et } d_1 \geqslant 0\}\,,$$
$$N(C,\overline{x}) = \left\{ \mu_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \mu_2 \begin{pmatrix} -1 \\ 0 \end{pmatrix} \mid \mu_1 \geqslant 0,\ \mu_2 \geqslant 0 \right\}.$$

$$(1-3) : T(C,\overline{x}) = \{(d_1,d_2) \mid d_1 + d_2 \leqslant 0 \text{ et } d_2 \geqslant 0\}\,,$$
$$N(C,\overline{x}) = \left\{ \mu_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \mu_2 \begin{pmatrix} 0 \\ -1 \end{pmatrix} \mid \mu_1 \geqslant 0,\ \mu_2 \geqslant 0 \right\}.$$

$$(2-3) : T(C,\overline{x}) = \{(d_1,d_2) \mid d_1 \geqslant 0 \text{ et } d_2 \geqslant 0\}\,,$$
$$N(C,\overline{x}) = \left\{ \mu_1 \begin{pmatrix} -1 \\ 0 \end{pmatrix} + \mu_2 \begin{pmatrix} 0 \\ -1 \end{pmatrix} \mid \mu_1 \geqslant 0,\ \mu_2 \geqslant 0 \right\}.$$

La condition nécessaire du 1$^{\text{er}}$ ordre que doit satisfaire un minimum (même local) de f sur C est : $-\nabla f(\overline{x}) \in N(C,\overline{x})$. Or un seul point vérifie cette condition : il se trouve sur l'arête (1) de C et c'est $\overline{x} = \left(\dfrac{1}{3}, \dfrac{2}{3}\right)$.

Par conséquent, $\overline{x} = \left(\frac{1}{3}, \frac{2}{3}\right)$ est le seul point de C minimisant f sur C. La valeur minimale de f sur C est $f(\overline{x}) = -\frac{11}{6}$.

3°) Un maximum local $\overline{x}$ de f sur C doit nécessairement vérifier : $\nabla f(\overline{x}) \in N(C, \overline{x})$. Deux points sont candidats : $\overline{x}_1 = (1, 0)$ et $\overline{x}_2 = (0, 0)$.

Comme $f(\overline{x}_1) = -\frac{1}{2} < 0 = f(\overline{x}_2)$, on en déduit que $\overline{x}_2 = (0, 0)$ est le maximum global de f sur C.

****** Exercice III.21.*** Soit P un polyèdre convexe fermé de $\mathbb{R}^n$, $n \geqslant 2$, symétrique par rapport à l'origine, décrit comme suit :

$$P := \{x \in \mathbb{R}^n \mid -1 \leqslant \langle a_i, x\rangle \leqslant 1 \text{ pour tout } i = 1, \ldots, m\},$$

où les a_i sont des vecteurs de $\mathbb{R}^n$.

On suppose que P est borné et d'intérieur non vide.

Soit $\mathcal{E}$ un ellipsoïde plein (ou convexe compact elliptique) de $\mathbb{R}^n$, centré en l'origine, décrit comme suit :

$$\mathcal{E} := \left\{x \in \mathbb{R}^n \mid \left\langle A^{-1}x, x\right\rangle \leqslant 1\right\}, \tag{3.12}$$

où $A \in \mathcal{S}_n(\mathbb{R})$ est définie positive.

1°) Montrer que $\mathcal{E}$ décrit en (3.12) est aussi $\{Bu \mid \| u \| \leqslant 1\}$, avec $B := A^{1/2}$. En déduire que le volume de $\mathcal{E}$ est proportionnel à $\sqrt{\det A}$.

2°) On considère le problème qui consiste à chercher l' (ou les) ellipsoïde(s) $\mathcal{E}$ contenu(s) dans P de volume maximal.

a) Montrer que l'inclusion $\mathcal{E} \subset P$ est traduite par les contraintes de type inégalité

$$\langle Aa_i, a_i\rangle \leqslant 1 \text{ pour tout } i = 1, \ldots, m.$$

b) Vérifier que le problème de la recherche d'un ellipsoïde $\mathcal{E}$ contenu dans P de volume maximal revient au problème de minimisation convexe différentiable suivant (posé dans $\mathcal{S}_n(\mathbb{R})$) :

$$(\mathcal{P}) \begin{cases} \text{Minimiser } \ln(\det A^{-1}) \\ A \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R}) \\ \langle Aa_i, a_i\rangle \leqslant 1 \text{ pour tout } i = 1, \ldots, m \end{cases}$$

où $\overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ désigne le cône convexe ouvert de $\mathcal{S}_n(\mathbb{R})$ constitué des matrices définies positives.

c) Montrer que $(\mathcal{P})$ a une et une seule solution.

d) Montrer que la solution $\overline{A}$ de $(\mathcal{P})$ est caractérisée par les conditions suivantes :

$$(\mathcal{C}) \quad \begin{cases} \overline{A} \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R}) \ ; \\ \langle \overline{A} a_i, a_i \rangle \leqslant 1 \text{ pour tout } i = 1, \dots, m \ ; \\ \text{Il existe des réels positifs } \overline{\mu}_1, \dots, \overline{\mu}_m \text{ tels que :} \\ \left(\overline{A}\right)^{-1} = \sum_{i=1}^{m} \overline{\mu}_i a_i a_i^\top \\ \overline{\mu}_i \left(\langle \overline{A} a_i, a_i \rangle - 1 \right) = 0 \quad \text{pour tout } i = 1, \dots, m. \end{cases}$$

e) $\overline{A}$, $\overline{\mu}_1, \dots, \overline{\mu}_m$ vérifiant les conditions de $(\mathcal{C})$. Montrer que

$$\sum_{i=1}^{m} \overline{\mu}_i = n.$$

f) Soit $\overline{\mathcal{E}}$ l'ellipsoïde contenu dans P de volume maximal.
Démontrer l'encadrement suivant :

$$\overline{\mathcal{E}} \subset P \subset \sqrt{n}\, \overline{\mathcal{E}}.$$

Solution : 1°) Si $B := A^{1/2}$, $\langle A^{-1}x, x \rangle = \| A^{-1/2}x \|^2$, de sorte que l'inégalité $\langle A^{-1}x, x \rangle \leqslant 1$ équivaut à $\| A^{-1/2}x \| \leqslant 1$. Grâce au changement de variable $u := A^{-1/2}x$, on voit que

$$\left(\| A^{-1/2}x \| \leqslant 1 \right) \Leftrightarrow \left(x = Bu, \ \| u \| \leqslant 1 \right).$$

$\mathcal{E}$ est ainsi l'image par B de la boule-unité fermée de $\mathbb{R}^n$.
Le volume de $\mathcal{E}$ est donc

$$\det B \pi^{n/2} / \Gamma\left(\frac{n}{2} + 1\right),$$

$\pi^{n/2}/\Gamma\left(\frac{n}{2}+1\right)$ étant le volume de la boule-unité fermée de $\mathbb{R}^n$. Comme $\det B = \sqrt{\det A}$, le volume de $\mathcal{E}$ est bien proportionnel à $\sqrt{\det A}$.

2°) a) $\mathcal{E} \subset P$ signifie

$$| \langle a_i, Bu \rangle | \leqslant 1 \text{ pour tout } u \text{ tel que } \| u \| \leqslant 1,$$

ce qui équivaut à

$$\sup_{\|u\| \leqslant 1} | \langle Ba_i, u \rangle | = \| Ba_i \| \leqslant 1.$$

Mais $\| Ba_i \|^2 \leqslant 1$ revient à $\langle Ba_i, Ba_i \rangle = \langle Aa_i, a_i \rangle \leqslant 1$.

b) Maximiser le volume de $\mathcal{E}$ équivaut à maximiser $\sqrt{\det A}$, ou encore à maximiser $\ln(\det A)$.

Les contraintes du problème sont : $A \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ et

$$g_i(A) := \langle Aa_i, a_i\rangle - 1 \leqslant 0 \text{ pour tout } i = 1, \dots, m.$$

En observant que $\langle Aa_i, a_i\rangle = \ll A, a_i a_i^\top \gg$, où $\ll \cdot, \cdot \gg$ désigne le produit scalaire usuel dans $\mathcal{S}_n(\mathbb{R})$ (rappel : $\ll U, V \gg = \operatorname{tr}(UV)$), il est clair que les g_i sont des fonctions affines (de A), avec

$$\nabla g_i(A) = a_i a_i^\top \text{ en tout } A \in \mathcal{S}_n(\mathbb{R}).$$

La fonction $f : A \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R}) \longmapsto f(A) := -\ln(\det A) = \ln(\det A^{-1})$ est convexe (strictement même) et différentiable sur $\overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ (*cf.* Exercice 1.13 et Exercice 1.4), avec

$$\nabla f(A) = -A^{-1} \text{ en tout } A \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R}).$$

Le problème de la recherche d'un ellipsoïde $\mathcal{E}$ contenu dans P de volume maximal revient donc à

$$(\mathcal{P}) \quad \begin{cases} \text{Minimiser } f(A) \\ A \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R}) \\ g_i(A) \leqslant 0 \text{ pour tout } i = 1, \dots, m. \end{cases}$$

C'est un problème de minimisation convexe différentiable avec contraintes du type inégalités affines.

c) Soit $\tilde{\mathcal{E}}$ un ellipsoïde contenant P (un tel ellipsoïde existe puisque P est borné) ; il est clair que les ellipsoïdes-candidats $\mathcal{E}$ à être solutions de $(\mathcal{P})$ ont un volume majoré par celui de $\tilde{\mathcal{E}}$. De plus,

$$\ln(\det A^{-1}) = -\ln(\det A) \to +\infty \text{ lorsque } A \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R}) \to A_0 \in \operatorname{fr}\left(\overset{\circ}{\mathcal{P}}_n(\mathbb{R})\right).$$

(la fonction $A \longmapsto -\ln(\det A)$ joue le rôle de « fonction-barrière » ou de « pénalisation intérieure » pour la contrainte « $A \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ »).

En conséquence (revoir l'Exercice II.1 si nécessaire), le problème $(\mathcal{P})$ a une solution au moins.

La stricte convexité de la fonction-objectif fait que cette solution est unique.

d) En raison de ce qui a été dit en b), la solution $\overline{A}$ de $(\mathcal{P})$ est caractérisée par les conditions de KKT :

$$(\mathcal{C})\begin{cases}\overline{A} \text{ vérifie les contraintes du problème } (\mathcal{P})\,;\\ \text{Il existe des réels positifs } (\overline{\mu}_1,\dots,\overline{\mu}_m) \text{ (les multiplicateurs}\\ \text{de Lagrange – KKT) tels que :}\\ -\left(\overline{A}\right)^{-1}+\sum_{i=1}^{m}\overline{\mu}_i a_i a_i^{\top}=0,\\ \overline{\mu}_i\left(\left\langle \overline{A}a_i,a_i\right\rangle -1\right)=0 \text{ pour tout } i=1,\dots,m.\end{cases}$$

e) En pré-multipliant et en post-multipliant par $\left(\overline{A}\right)^{1/2}$ la relation $\left(\overline{A}\right)^{-1}=\sum_{i=1}^{m}\overline{\mu}_i a_i a_i^{\top}$, il vient :

$$I_n=\sum_{i=1}^{m}\overline{\mu}_i\left(\overline{A}\right)^{1/2}a_i a_i^{\top}\left(\overline{A}\right)^{1/2},$$

et, en prenant la trace,

$$\begin{aligned} n&=\sum_{i=1}^{m}\overline{\mu}_i\left\langle \overline{A}a_i,a_i\right\rangle\\ &=\sum_{i=1}^{m}\overline{\mu}_i \text{ d'après la dernière relation de } (\mathcal{C}).\end{aligned}$$

f) $\overline{\mathcal{E}}\subset P$ évidemment ; démontrons l'inclusion $P\subset\sqrt{n}\,\overline{\mathcal{E}}$.

Soit $x\in P$, c'est-à-dire vérifiant $|\langle x,a_i\rangle|\leqslant 1$ pour tout $i=1,\dots,m$; il nous faut démontrer que

$$\left\langle \left(\overline{A}\right)^{-1}\frac{x}{\sqrt{n}},\frac{x}{\sqrt{n}}\right\rangle\leqslant 1 \;\; i.e. \;\left\langle \left(\overline{A}\right)^{-1}x,x\right\rangle\leqslant n.$$

Or $\left(\overline{A}\right)^{-1}=\sum_{i=1}^{m}\overline{\mu}_i a_i a_i^{\top}$ de sorte que

$$\left\langle \left(\overline{A}\right)^{-1}x,x\right\rangle=\sum_{i=1}^{m}\overline{\mu}_i\langle a_i,x\rangle^2\leqslant n\quad\left(\begin{array}{l}\text{puisque } \langle a_i,x\rangle^2\leqslant 1\\ \text{pour tout } i=1,...,m \text{ et } \sum_{i=1}^{m}\overline{\mu}_i=n\end{array}\right).$$

Commentaire :

– La relation $\left(\overline{A}\right)^{-1} = \sum_{i=1}^{m} \overline{\mu}_i a_i a_i^\top$, alliée à $\sum_{i=1}^{m} \overline{\mu}_i = n$, indique que $\left(\overline{A}\right)^{-1}/n$ est une combinaison convexe de matrices symétriques semi-définies positives de rang 1 construites à partir des vecteurs a_i normaux aux facettes de P. Elle traduit une « tangence simultanée » de l'ellipsoïde optimal $\overline{\mathcal{E}}$ à certaines facettes de P, ce que l'intuition de la géométrie du problème laissait supposer.

– La constante $\sqrt{n}$ obtenue dans l'encadrement de f est indépendante de P, notamment du nombre m de doubles inégalités le décrivant. Un exemple simple dans $\mathbb{R}^2$, en considérant un carré P et donc une boule fermée pour $\overline{\mathcal{E}}$, montre que cette constante ne peut être améliorée.

– Le centre des ellipsoïdes-candidats $\mathcal{E}$ avait été fixé à l'origine, mais on peut aisément imaginer que s'il avait été libre (c'est-à-dire un paramètre additionnel dans le problème), la symétrie de P aurait de toute façon conduit à un $\overline{\mathcal{E}}$ optimal centré à l'origine.

****** Exercice III.22.*** Soit P un polyèdre convexe fermé de $\mathbb{R}^n$ décrit de la manière suivante

$$P := \{x \in \mathbb{R}^n \mid \langle a_i, x\rangle \leqslant b_i \ \text{ pour tout } i = 1, \dots, m\}$$

où les a_i sont des vecteurs de $\mathbb{R}^n$ et les b_i des réels.
On suppose que P est borné et d'intérieur non vide.
Soit $\mathcal{E}$ un ellipsoïde plein de $\mathbb{R}^n$ décrit de la manière suivante

$$\mathcal{E} := c + \{Bu \mid \| u \| \leqslant 1\}, \tag{3.13}$$

où $c \in \mathbb{R}^n$ et $B \in \mathcal{S}_n(\mathbb{R})$ est définie positive. $\mathcal{E}$ est ainsi centré en c et de volume proportionnel à $\det B$.

On considère le problème qui consiste à chercher l' (ou les) ellipsoïde(s) $\mathcal{E}$ contenu(s) dans P de volume maximal.

1°) Décrire l'inclusion $\mathcal{E} \subset P$ sous la forme d'une conjonction d'inégalités

$$g_i(c, B) \leqslant 0 \ \text{ pour tout } i = 1, \dots, m$$

où les g_i sont des fonctions convexes de $(c, B) \in \mathbb{R}^n \times \mathcal{S}_n(\mathbb{R})$.

2°) Formaliser le problème de la recherche d'un ellipsoïde $\mathcal{E}$ contenu dans P de volume maximal comme un problème de minimisation convexe.

Solution : 1°) L'inclusion $\mathcal{E} \subset \{x \in \mathbb{R}^n \mid \langle a_i, x\rangle \leqslant b_i\}$ se traduit par

$$\langle a_i, c + Bu\rangle \leqslant b_i \text{ pour tout } u \text{ vérifiant } \| u \| \leqslant 1 \ ;$$

ce qui équivaut à

$$\langle a_i, c\rangle + \sup_{\|u\|\leqslant 1} \langle a_i, Bu\rangle \leqslant b_i$$

soit encore, puisque $B^\top = B$ et que $\sup_{\|u\|\leqslant 1} \langle v, u\rangle = \| v \|$

$$\langle a_i, c\rangle + \| Ba_i \| \leqslant b_i.$$

Soit $g_i : \mathbb{R}^n \times \mathcal{S}_n(\mathbb{R}) \to \mathbb{R}$ la fonction qui à $(c, B) \in \mathbb{R}^n \times \mathcal{S}_n(\mathbb{R})$ associe $g_i(c, B) := \langle a_i, c\rangle + \| Ba_i \| - b_i$. Il est immédiat que g_i est convexe (d'accord ?). Donc l'inclusion $\mathcal{E} \subset P$ équivaut au jeu d'inégalités convexes

$$g_i(c, B) \leqslant 0 \text{ pour tout } i = 1, \ldots, m.$$

2°) Soit $\overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ l'ouvert convexe de $\mathcal{S}_n(\mathbb{R})$ constitué des $B \in \mathcal{S}_n(\mathbb{R})$ qui sont définies positives. La fonction

$$f : B \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R}) \longmapsto f(B) := -\ln(\text{dét } B)$$

est convexe (strictement convexe même) sur $\overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ (*cf.* Exercice I.13 par exemple).

Le volume de l'ellipsoïde décrit comme en (3.13) est dét B, à une constante multiplicative près (fixe, indépendante de B ; c'est le volume de la boule-unité fermée de $\mathbb{R}^n$). Maximiser dét B équivaut à maximiser $\ln(\text{dét } B)$, ou encore à minimiser $-\ln(\text{dét } B)$.

Le problème de la recherche d'un ellipsoïde $\mathcal{E}$ (décrit comme en (3.13)) contenu dans P de volume maximal se formalise donc en le problème de minimisation convexe suivant (posé dans $\mathbb{R}^n \times \mathcal{S}_n(\mathbb{R})$) :

$$\begin{cases} \text{Minimiser } f(B) = -\ln(\text{dét } B) \\ B \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R}), \\ g_i(c, B) \leqslant 0 \text{ pour tout } i = 1, \ldots, m. \end{cases}$$

Commentaire : – Prolongement de l'exercice : on peut montrer, dans l'esprit de l'exercice précédent, qu'il y a un et un seul ellipsoïde contenu dans P de volume maximal.

– Autre prolongement : Montrer qu'il existe un et un seul ellipsoïde $\mathcal{E}$ *contenant* P de volume *minimal*.

****** Exercice III.23.*** Soit $B \in \mathcal{S}_n(\mathbb{R})$, s et y dans $\mathbb{R}^n$, $s \neq 0$. On considère le problème de minimisation suivant :

$$(\mathcal{P}) \qquad \text{Minimiser } \frac{1}{2} \parallel X - B \parallel^2 \text{ parmi les } X \in \mathcal{S}_n(\mathbb{R}) \text{ vérifiant } Xs = y,$$

où $\| \cdot \|$ désigne la norme euclidienne associée au produit scalaire usuel $\ll \cdot, \cdot \gg$ sur $\mathcal{S}_n(\mathbb{R})$ (rappel : $\ll U, V \gg = \mathrm{tr}(UV)$).

1°) Soit $l : \mathcal{S}_n(\mathbb{R}) \to \mathbb{R}^n$ définie par $l(X) := Xs$.

a) Vérifier que l est linéaire surjective.

b) Déterminer l'application linéaire adjointe l^* de l (rappel : $l^* : \mathbb{R}^n \to \mathcal{S}_n(\mathbb{R})$ est définie par $\langle l(X), u \rangle = \ll X, l^*(u) \gg$ pour tout $X \in \mathcal{S}_n(\mathbb{R})$ et $u \in \mathbb{R}^n$).

2°) a) Indiquer pourquoi le problème $(\mathcal{P})$ a une et une seule solution.

b) Montrer que $\overline{X} \in \mathcal{S}_n(\mathbb{R})$ est solution de $(\mathcal{P})$ si et seulement si :

$$(\mathcal{L}) \qquad \begin{cases} \overline{X}s = y \ ; \\ \exists\ \overline{u} \in \mathbb{R}^n \text{ tel que } \overline{X} - B = \dfrac{\overline{u}s^\top + s\overline{u}^\top}{2}. \end{cases}$$

c) Vérifier que $\overline{X} := B + \dfrac{(y - Bs)s^\top + s(y - Bs)^\top}{\| s \|^2} - \dfrac{\langle y - Bs, s \rangle}{\| s \|^4} ss^\top$ est la solution de $(\mathcal{P})$.

3°) On suppose ici qu'il existe $W \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ telle que $Wy = W^{-1}s$, et on considère le problème de minimisation suivant :

$(\mathcal{Q})$ Minimiser $\frac{1}{2} \parallel W(X - B)W \parallel^2$ parmi les $X \in \mathcal{S}_n(\mathbb{R})$ vérifiant $Xs = y$.

a) Montrer que la solution X^* de $(\mathcal{Q})$ est donnée par

$$X^* := B + \frac{(y - Bs)y^\top + y(y - Bs)^\top}{\langle y, s \rangle} - \frac{\langle y - Bs, s \rangle}{(\langle y, s \rangle)^2} yy^\top.$$

b) On suppose de plus que B est inversible. Montrer que X^* est inversible et

$$(X^*)^{-1} = B^{-1} + \frac{ss^\top}{\langle y, s \rangle} - \frac{B^{-1}yy^\top B^{-1}}{\langle B^{-1}y, y \rangle}.$$

En déduire que X^* est définie positive lorsque B est définie positive.

Indication. Pour répondre à la question 3°) b), on pourra utiliser sur $(X^*)^{-1}$ le résultat de la 5e question de l'Exercice I.8.

Solution : 1°) a) l est linéaire, c'est clair. Désignons par E_{ij} la matrice carrée de taille n dont tous les éléments sont nuls sauf le terme (i,j) qui vaut 1. La famille $\{E_{ii} \mid i = 1,\dots,n\} \cup \{E_{ij} + E_{ji} \mid i \neq j\}$ constitue une base de $\mathcal{S}_n(\mathbb{R})$; et :

$l(E_{ii})$ = vecteur de $\mathbb{R}^n$ dont toutes les composantes sont nulles sauf la i^{e} qui vaut s_i ;

$l(E_{ij} + E_{ji})$ = vecteur de $\mathbb{R}^n$ dont toutes les composantes sont nulles sauf la i^{e} qui vaut s_j et la j^{e} qui vaut s_i.

Puisque $s = (s_1,\dots,s_n) \neq 0$, l'image par l de la base ci-dessus est une famille génératrice de $\mathbb{R}^n$: l est donc surjective.

On peut être plus explicite en proposant une matrice symétrique X telle que $Xs = y$:

– Si $\langle y, s\rangle \neq 0$, $X := \dfrac{yy^\top}{\langle y, s\rangle}$ répond à la question ;

– Si $\langle y, s\rangle = 0$, $X := \dfrac{sy^\top + ys^\top}{\| s \|^2}$ fait l'affaire.

b) $l^*(u)$ est la seule matrice symétrique vérifiant

$$\langle Xs, u\rangle = \ll X, l^*(u) \gg = \operatorname{tr}(l^*(u)X) \quad \text{pour tout } X \in \mathcal{S}_n(\mathbb{R}) \text{ et } u \in \mathbb{R}^n.$$

Puisque $\langle Xs, u\rangle = (Xs)^\top u$ et que $s^\top X^\top u = \operatorname{tr}(us^\top X) = \operatorname{tr}(su^\top X)$, on a : $l^*(u) = \frac{us^\top + su^\top}{2}$ (d'accord ?).

2°) a) Puisque l est linéaire (continue) surjective, $\{X \in \mathcal{S}_n(\mathbb{R}) \mid l(X) = y\}$ est un sous-espace affine (fermé) de $\mathcal{S}_n(\mathbb{R})$. Le problème $(\mathcal{P})$ est donc celui de trouver la projection orthogonale (dans le contexte de l'espace euclidien $(\mathcal{S}_n(\mathbb{R}), \ll \cdot, \cdot \gg)$ de B sur ce sous-espace affine : $(\mathcal{P})$ a une et une seule solution.

b) En tout point $\overline{X}$ de l'ensemble-contrainte de $(\mathcal{P})$, le sous-espace (vectoriel) normal est le même, c'est $(\operatorname{Ker} l)^\perp = \operatorname{Im} l^*$. Ainsi, $\overline{X}$ est solution de $(\mathcal{P})$ si et seulement si :

$$\overline{X} \in \mathcal{S}_n(\mathbb{R}), \quad \overline{X}s = y \text{ et } \overline{X} - B \in \operatorname{Im} l^*.$$

Connaissant l^*, nous avons donc la caractérisation $(\mathcal{L})$ annoncée (il s'agit, bien sûr, des conditions nécessaires et suffisantes d'optimalité de Lagrange).

c) l^* étant injective (puisque l est surjective), il n'y a qu'un $\overline{u} \in \mathbb{R}^n$ répondant à la question. On vérifie que

$$\overline{u} := 2\frac{y - Bs}{\| s \|^2} - \frac{\langle y - Bs, s\rangle}{\| s \|^4} s$$

fait l'affaire. La matrice $\overline{X}$ qui s'ensuit vérifie bien $\overline{X}s = y$ (c'est facile à vérifier) ; c'est donc la solution de $(\mathcal{P})$.

3°) a) Le problème $(\mathcal{Q})$ consiste à minimiser $\frac{1}{2} \| WXW - WBW \|^2$ parmi les $X \in \mathcal{S}_n(\mathbb{R})$ vérifiant $(WXW)(W^{-1}s) = Wy$.

En posant $B' := WBW,\ s' := W^{-1}s,\ y' := Wy$, et en faisant le changement de variable $X' := WXW$, le problème $(\mathcal{Q})$ est équivalent à :

$(\mathcal{Q}')$ Minimiser $\frac{1}{2} \| X' - B' \|^2$ parmi les $X' \in \mathcal{S}_n(\mathbb{R})$ vérifiant $X's' = y'$.

D'après le résultat de la question précédente, la solution de ($\mathcal{Q}$') est fournie par

$$\overline{X}' = WBW + \frac{W(y - Bs)s^\top W^{-1} + W^{-1}s(y - Bs)^\top W}{\| W^{-1}s \|^2} - \frac{\langle W(y - Bs), W^{-1}s\rangle}{\| W^{-1}s \|^4} W^{-1}ss^\top W^{-1} ;$$

par suite, la solution du problème $(\mathcal{Q})$ est $X^* = W^{-1}(\overline{X}')W^{-1}$, soit

$$X^* = B + \frac{(y - Bs)s^\top W^{-2} + W^{-2}s(y - Bs)^\top}{\| W^{-1}s \|^2} - \frac{\langle W(y - Bs), W^{-1}s\rangle}{\| W^{-1}s \|^4} W^{-2}ss^\top W^{-2}.$$

Mais puisque $Wy = W^{-1}s$ et que W est symétrique, on a :

$$W^{-2}s = y,\ \| W^{-1}s \|^2 = \langle W^{-1}s, W^{-1}s\rangle = \langle Wy, W^{-1}s\rangle = \langle y, s\rangle ,$$
$$\langle W(y - Bs), W^{-1}s\rangle = \langle y - Bs, s\rangle .$$

D'où l'expression annoncée de X^*.

b) Partons de $Y^* := B^{-1} + \frac{ss^\top}{\langle y, s\rangle} - \frac{B^{-1}yy^\top B^{-1}}{\langle B^{-1}y, y\rangle}$ et posons $A := B^{-1}$, $u := s,\ v := B^{-1}y,\ \alpha := \frac{1}{\langle y, s\rangle}$ et $\beta := -\frac{1}{\langle B^{-1}y, y\rangle}$; nous avons donc à inverser

$(A+\alpha uu^\top+\beta vv^\top)$. Dans le but d'appliquer le résultat de la 5$^{\text{e}}$ question de l'Exercice I.8, nous évaluons les expressions suivantes :

$$1+\alpha\left\langle A^{-1}u,u\right\rangle = 1+\frac{\langle Bs,s\rangle}{\langle y,s\rangle}, \qquad \begin{pmatrix}\text{rappelons que}\\ \langle y,s\rangle = \parallel W^{-1}s\parallel^2>0\end{pmatrix}$$

$$1+\beta\left\langle A^{-1}v,v\right\rangle = 0,$$

$$d := \left(1+\alpha\left\langle A^{-1}u,u\right\rangle\right)\left(1+\beta\left\langle A^{-1}v,v\right\rangle\right)-\alpha\beta\left(\left\langle A^{-1}u,v\right\rangle\right)^2 = \frac{\langle s,y\rangle}{\langle B^{-1}y,y\rangle}.$$

Par suite

$$\begin{aligned}(Y^*)^{-1} &= A^{-1}-\frac{1}{d}\left\{\beta(1+\alpha\left\langle A^{-1}u,u\right\rangle)A^{-1}vv^\top A^{-1}\right.\\ &\qquad\left.-\alpha\beta\left\langle A^{-1}u,v\right\rangle\left(A^{-1}vu^\top A^{-1}+A^{-1}uv^\top A^{-1}\right)\right\}\\ &= B-\frac{ys^\top B+Bsy^\top}{\langle y,s\rangle}+\frac{\langle Bs+y,s\rangle}{\left(\langle s,y\rangle\right)^2}yy^\top,\end{aligned}$$

ce qui n'est autre que X^*.

Démontrons la définie positivité de X^* en montrant la définie positivité de $Y^*=(X^*)^{-1}$. Pour cela, soit x non nul dans $\mathbb{R}^n$ et considérons $\langle Y^*x,x\rangle$. On a :

$$\langle Y^*x,x\rangle = \left\langle B^{-1}x,x\right\rangle+\frac{(\langle s,x\rangle)^2}{\langle s,y\rangle}-\frac{\left(\left\langle B^{-1}y,x\right\rangle\right)^2}{\langle B^{-1}y,y\rangle}.$$

En posant $\xi := B^{-1/2}x$ et $\zeta := B^{-1/2}y$, la formulation ci-dessus devient

$$\langle Y^*x,x\rangle = \frac{\parallel\xi\parallel^2\parallel\zeta\parallel^2-(\langle\xi,\zeta\rangle)^2}{\parallel\zeta\parallel^2}+\frac{(\langle s,x\rangle)^2}{\langle s,y\rangle}.$$

Il est clair à présent (grâce à l'inégalité de Schwarz $\langle\xi,\zeta\rangle\leqslant\parallel\xi\parallel\cdot\parallel\zeta\parallel$) que $\langle Y^*x,x\rangle\geqslant 0$. Sachant que Y^* est inversible, cela suffit à démontrer le caractère de définie positivité de Y^*. Faisons-en néanmoins une démonstration directe.

Si le 1$^{\text{er}}$ terme dans l'expression de $\langle Y^*x,x\rangle$ s'annule, c'est qu'il existe $\alpha\neq 0$ tel que $\xi=\alpha\zeta$ (cas d'égalité dans l'inégalité de Schwarz), soit encore $x=\alpha y$. Mais alors $\langle s,x\rangle=\alpha\langle s,y\rangle$ et le 2$^{\text{e}}$ terme dans l'expression de $\langle Y^*x,x\rangle$ est $\alpha^2\langle s,y\rangle>0$. Donc $\langle Y^*x,x\rangle>0$.

Commentaire : – La matrice $\overline{X}$ obtenue comme solution de $(\mathcal{P})$ n'est pas nécessairement définie positive, pour cela il faudrait qu'au moins $\langle y, s\rangle$ $(= \langle \overline{X}s, s\rangle)$ soit > 0. C'est le cas dans la 3e question où $\langle y, s\rangle = \langle W^{-2}s, s\rangle > 0$.

– Cet exercice trouve ses racines dans la nécessité d'expliquer « de manière variationnelle » les formules de mise à jour dans les méthodes de minimisation sans contraintes dites de quasi-Newton. Ainsi, dans la 2e question, parmi les $X \in \mathcal{S}_n(\mathbb{R})$ vérifiant $Xs = y$ (appelée équation de la sécante ou de quasi-Newton), on cherche celle qui est la plus proche de A (au sens de la distance dérivée de $\ll \cdot, \cdot \gg$). La formule de mise à jour qui en résulte est **PSB** (pour Powell-symétrique-Broyden).

– Le résultat de la 3e question a un pendant que voici.

Soit $W \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ telle que $Ws = W^{-1}y$, et considérons le problème de minimisation suivant :

$$(\mathcal{R}) \quad \text{Minimiser } \frac{1}{2} \parallel W(X-B)W \parallel^2 \text{ parmi les } X \in \mathcal{S}_n(\mathbb{R}) \text{ vérifiant } Xy = s.$$

Alors la solution X_* de $(\mathcal{R})$ est donnée par

$$X_* := B + \frac{(s-By)s^\top + s(s-By)^\top}{\langle s, y\rangle} - \frac{\langle s-By, y\rangle}{(\langle s, y\rangle)^2} ss^\top.$$

Si B est définie positive, il en est de même de X_* et

$$(X_*)^{-1} = B^{-1} + \frac{yy^\top}{\langle y, s\rangle} - \frac{B^{-1}ss^\top B^{-1}}{\langle B^{-1}s, s\rangle}.$$

Les expressions de X^* et X_* sont à la base des formules de mise à jour de la méthode **DFP** (pour Davidon-Fletcher-Powell) et **BFGS** (pour Broyden-Fletcher-Goldfarb-Shanno).

****** Exercice III.24.*** Étant donné $u = (u_1, \ldots, u_n) \in \mathbb{R}^n$, on cherche un élément $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ vérifiant $x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n$ le plus proche possible de u (au sens de la norme euclidienne usuelle de $\mathbb{R}^n$).

1°) Formaliser cette question comme un problème de minimisation convexe, ou comme un problème de projection sur un cône convexe fermé.

2°) Décrire les conditions caractérisant l'unique élément $\overline{x} = (\overline{x}_1, \ldots, \overline{x}_n)$ répondant à la question.

3°) On traite ici un exemple dans $\mathbb{R}^4$: étant donné $u = (2, 1, 5, 4)$, trouver l'unique $\overline{x} = (\overline{x}_1, \overline{x}_2, \overline{x}_3, \overline{x}_4)$ vérifiant $\overline{x}_1 \leqslant \ldots \leqslant \overline{x}_4$ à distance minimale de u.

Solution : 1°) Soit $K := \{x = (x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n\}$. Il est clair que K est un cône convexe fermé de $\mathbb{R}^n$, polyédral même, qui peut être décrit par $n-1$ inégalités $g_i(x) \leqslant 0$, avec

$$g_i(x) := x_i - x_{i+1} \quad \text{pour tout } i = 1, \ldots, n-1.$$

Le problème posé peut être formalisé comme suit :

– Minimiser $\| u - x \|$ (ou $\frac{1}{2} \| u - x \|^2$) sous les contraintes $g_i(x) \leqslant 0,\ i = 1, \ldots, n-1$;

ou bien comme ceci :

– Trouver la projection $\overline{x}$ de u sur K.

2°) Les fonctions g_i définissant les contraintes du type inégalité sont linéaires, avec même les $\nabla g_i(\overline{x}) = e_i - e_{i+1},\ i = 1, \ldots, n-1$ ($\{e_i\}$ base canonique de $\mathbb{R}^n$) linéairement indépendants. Donc, $\overline{x} = (\overline{x}_1, \ldots, \overline{x}_n)$ est solution du problème posé si, et seulement si, $\overline{x} \in K$ et il existe $(\overline{\mu}_1, \ldots, \overline{\mu}_{n-1}) \in (\mathbb{R}^+)^{n-1}$ (unique d'ailleurs) tel que

$$\begin{aligned} &\overline{x} - u + \sum_{i=1}^{n-1} \overline{\mu}_i (e_i - e_{i+1}) = 0, \\ &\overline{\mu}_i (\overline{x}_{i+1} - \overline{x}_i) = 0 \text{ pour tout } i = 1, \ldots, n-1. \end{aligned}$$

Détaillons ces conditions ; cela devient : $\overline{x}_1 \leqslant \overline{x}_2 \leqslant \ldots \leqslant \overline{x}_n$, et il existe $\overline{\mu}_1, \ldots, \overline{\mu}_{n-1}$ tels que :

$$\begin{cases} \overline{x}_1 - u_1 + \overline{\mu}_1 = 0 \\ \overline{x}_2 - u_2 + \overline{\mu}_2 - \overline{\mu}_1 = 0 \\ \vdots \\ \overline{x}_i - u_i + \overline{\mu}_i - \overline{\mu}_{i-1} = 0 \\ \vdots \\ \overline{x}_{n-1} - u_{n-1} + \overline{\mu}_{n-1} - \overline{\mu}_{n-2} = 0 \\ \overline{x}_n - u_n - \overline{\mu}_{n-1} = 0 \ ; \\ \overline{\mu}_1 \geqslant 0, \ldots, \overline{\mu}_{n-1} \geqslant 0 \text{ et} \\ \overline{\mu}_i (\overline{x}_i - \overline{x}_{i+1}) = 0 \text{ pour } i = 1, \ldots, n-1. \end{cases} \tag{3.14}$$

Transformons ces conditions en observant :

$$u_1 - \overline{x}_1 \ (= \overline{\mu}_1) \geqslant 0, \ (u_1 + u_2) - (\overline{x}_1 + \overline{x}_2) \ (= \overline{\mu}_2) \geqslant 0, \ \ldots$$
$$(u_1 + \ldots + u_{n-1}) - (\overline{x}_1 + \ldots + \overline{x}_{n-1}) \ (= \overline{\mu}_{n-1}) \geqslant 0,$$
$$(u_1 + \ldots + u_n) - (\overline{x}_1 + \ldots + \overline{x}_n) = 0 \ ;$$
$$\left(\sum_{j=1}^{i} u_j - \sum_{j=1}^{i} \overline{x}_j \right) (\overline{x}_i - \overline{x}_{i+1}) = 0 \text{ pour tout } i = 1, \ldots, n-1$$
$$(u_n - \overline{x}_n)\,(\overline{x}_{n-1} - \overline{x}_n) = 0.$$

Réciproquement, $(\overline{x}_1, \ldots, \overline{x}_n)$ vérifiant les conditions ci-dessus vérifie les conditions (3.14) où on a posé $\overline{\mu}_i = \sum_{j=1}^{i} u_j - \sum_{j=1}^{i} \overline{x}_j$ pour tout $i = 1, \ldots, n-1$.

En définitive nous avons la caractérisation suivante, exempte de toute référence à des multiplicateurs $\overline{\mu}_i$:

$\overline{x} = (\overline{x}_1, \ldots, \overline{x}_n)$ est solution du problème posé si, et seulement si,

$$\begin{cases} \overline{x}_1 \leqslant \overline{x}_2 \leqslant \ldots \leqslant \overline{x}_n \\ \displaystyle\sum_{j=1}^{i} \overline{x}_j \leqslant \sum_{j=1}^{i} u_j \text{ pour tout } i = 1, \ldots, n-1 \\ \displaystyle\sum_{j=1}^{n} \overline{x}_j = \sum_{j=1}^{n} u_j \\ \displaystyle\left(\sum_{j=1}^{i} u_j - \sum_{j=1}^{i} \overline{x}_j \right) (\overline{x}_i - \overline{x}_{i+1}) = 0 \text{ pour tout } i = 1, \ldots, n-1 \\ (u_n - \overline{x}_n)\,(\overline{x}_{n-1} - \overline{x}_n) = 0. \end{cases}$$

3°) Dans l'exemple proposé, on cherche $\overline{x}_1 \leqslant \overline{x}_2 \leqslant \overline{x}_3 \leqslant \overline{x}_4$ tels que

$$\overline{x}_1 \leqslant 2, \ \overline{x}_1 + \overline{x}_2 \leqslant 3, \ \overline{x}_1 + \overline{x}_2 + \overline{x}_3 \leqslant 8, \ \overline{x}_1 + \overline{x}_2 + \overline{x}_3 + \overline{x}_4 = 12,$$
$$(2 - \overline{x}_1)\,(\overline{x}_1 - \overline{x}_2) = 0, \ (3 - \overline{x}_1 - \overline{x}_2)\,(\overline{x}_2 - \overline{x}_3) = 0,$$
$$(8 - \overline{x}_1 - \overline{x}_2 - \overline{x}_3)\,(\overline{x}_3 - \overline{x}_4) = 0, \ (4 - \overline{x}_4)\,(\overline{x}_3 - \overline{x}_4) = 0.$$

Après quelques calculs algébriques (on commence par supposer $\overline{x}_1 = 2$ et montrer que cela conduit à une contradiction ; donc $\overline{x}_1 < 2$ et par conséquent $\overline{x}_1 = \overline{x}_2$, etc.), on arrive à : $\overline{x}_1 = \overline{x}_2 = \frac{3}{2}$, $\overline{x}_3 = \overline{x}_4 = \frac{9}{2}$.

Évidemment le point $\overline{x}$ est plus proche de u que le point $\overline{u}$ de K obtenu par simple réarrangement par ordre croissant des coordonnées de u. Dans l'exemple traité, la distance de u à $\overline{x}$ est 1 tandis que celle de u à $\overline{u}$ est 2.

***Exercice III.25.** Dans $\mathbb{R}^2$ on considère le problème de minimisation suivant :

$$(\mathcal{P}_\alpha) \quad \begin{cases} \text{Min} \quad f(x) := (\xi_1 - 1)^2 + \xi_2^2 \\ h(x) := -\xi_1 + \alpha\xi_2^2 = 0. \end{cases}$$

1°) Observer que $\overline{x} = (0,0)$ vérifie les conditions nécessaires d'optimalité du 1er ordre de Lagrange.

2°) À l'aide des conditions nécessaires de minimalité du 2e ordre, décider en fonction de α quand $\overline{x}$ est un minimum local et quand il ne l'est pas.

Solution : 1°) La fonction h de la contrainte du type égalité est continûment différentiable et $\nabla h(x) = \begin{pmatrix} -1 \\ 2\alpha\xi_2 \end{pmatrix} \neq (0,0)$. Si $\overline{x} = (0,0)$ est un minimum local (ou un maximum local) de f sous la contrainte $h(x) = 0$, il existe $\overline{\lambda} \in \mathbb{R}$ tel que $\nabla f(\overline{x}) + \overline{\lambda}\nabla h(\overline{x}) = 0$. C'est bien le cas ici avec $\overline{\lambda} = -2$.

2°) Le sous-espace tangent à l'ensemble-contrainte en $\overline{x} = (0,0)$ est $H := \{0\} \times \mathbb{R}$.

De plus

$$\left\langle \left(\nabla^2 f(\overline{x}) + \overline{\lambda}\nabla^2 h(\overline{x})\right) d, d \right\rangle = 2(1 - 2\alpha)d_2^2$$

pour tout $d = (0, d_2) \in H$.

Par suite :

– si $\alpha > \frac{1}{2}$, $\overline{x} = (0,0)$ ne peut être un minimum local de f sous la contrainte $h(x) = 0$;

– si $\alpha < \frac{1}{2}$, $\overline{x} = (0,0)$ est un minimum local strict de f sous la contrainte $h(x) = 0$;

– si $\alpha = \frac{1}{2}$, on ne peut décider à l'aide des seules conditions du 1er et 2e ordre.

***Exercice III.26.** On considère le problème de la minimisation de $f(\xi_1,\xi_2,\xi_3) := -\xi_1 - \xi_2 - \xi_2\xi_3 - \xi_1\xi_3$ sous la contrainte $\xi_1+\xi_2+\xi_3-3=0$.

1°) Ce problème est-il convexe ?

2°) Déterminer tous les points vérifiant les conditions nécessaires de minimalité du 1er ordre.

Parmi ces points quels sont ceux qui vérifient les conditions nécessaires de minimalité du 2e ordre ? ceux qui vérifient les conditions suffisantes de minimalité du 2e ordre ?

Solution : 1°) La fonction-objectif f est quadratique, avec

$$x=(\xi_1,\xi_2,\xi_3)\longmapsto \nabla^2 f(x)=\begin{bmatrix} 0 & 0 & -1\\ 0 & 0 & -1\\ -1 & -1 & 0\end{bmatrix}.$$

Mais $\nabla^2 f(x)$ n'étant pas semi-définie positive, f n'est pas convexe. Le problème posé n'est donc pas un problème de minimisation convexe.

2°) Un point $\overline{x}=(\overline{\xi}_1,\overline{\xi}_2,\overline{\xi}_3)$ vérifie les conditions nécessaires de minimalité du 1er ordre lorsque :

$$\begin{cases}\overline{\xi}_1+\overline{\xi}_2+\overline{\xi}_3=3\ ;\\ \exists\ \overline{\lambda}\in\mathbb{R}\ \text{tel que}\ -1-\overline{\xi}_3+\overline{\lambda}=0\ \text{et}\ -\overline{\xi}_2-\overline{\xi}_1+\overline{\lambda}=0.\end{cases}$$

Ce système d'équations conduit à $\overline{\lambda}=2$ et aux points $\overline{x}=(\overline{\xi}_1,2-\overline{\xi}_1,1)$, $\overline{\xi}_1\in\mathbb{R}$.

Le sous-espace H tangent à l'ensemble-contrainte (noté S) en $\overline{x}$ est l'hyperplan d'équation $d_1+d_2+d_3=0$. Ensuite :

$$\forall\,(d_1,d_2,d_3)\in H,\ \ \langle\nabla^2_{xx}\mathcal{L}(\overline{x},\overline{\lambda})d,d\rangle=2(d_1+d_2)^2=2d_3^2.$$

Donc, tous les points $\overline{x}=(\overline{\xi}_1,2-\overline{\xi}_1,1)$ vérifient les conditions nécessaires de minimalité du 2e ordre.

Soit $d=(d_1,-d_1,0)$ avec $d_1\neq 0$; il est clair que $d\in H\setminus\{0\}$ et $\langle\nabla^2_{xx}\mathcal{L}(\overline{x},\overline{\lambda})d,d\rangle=0$. Donc aucun des points $(\overline{\xi}_1,2-\overline{\xi}_1,1)$ ne vérifie les conditions suffisantes de minimalité (stricte) du 2e ordre.

Par ailleurs, $f(\overline{\xi}_1,2-\overline{\xi}_1,1)=-4$ constamment et
$f(\xi_1,\xi_2,\xi_3)-(-4)=(\xi_1+\xi_2-2)^2\geqslant 0$ pour
tout (ξ_1,ξ_2,ξ_3) dans l'ensemble-contrainte S.

Donc, en fait, tous les points $\left(\overline{\xi}_1, 2-\overline{\xi}_1, 1\right)$ sont des minima globaux de f sur S.

Le fait qu'on n'ait pu vérifier la condition suffisante de minimalité (stricte) en un point $\overline{x} = \left(\overline{\xi}_1, 2-\overline{\xi}_1, 1\right)$ s'explique aisément par le fait que f est constante sur toute la droite $\{(\xi_1, 2-\xi_1, 1) \mid \xi_1 \in \mathbb{R}\}$.

***Exercice III.27.** Soit dans $\mathbb{R}^2$ le problème de minimisation suivant :

$$(\mathcal{P}_a) \begin{cases} \text{Min } f_a(\xi_1,\xi_2) := \xi_1^2 + a\xi_2^2 + \xi_1\xi_2 + \xi_1, \\ \xi_1 + \xi_2 - 1 \leqslant 0. \end{cases} \qquad (a \in \mathbb{R})$$

1°) Quand $(\mathcal{P}_a)$ est-il un problème de minimisation convexe ?

2°) Résoudre $(\mathcal{P}_a)$ suivant les valeurs de a.

Solution : 1°) La fonction $g : (\xi_1,\xi_2) \longmapsto g(\xi_1,\xi_2) := \xi_1 + \xi_2 - 1$ définissant la seule contrainte du type inégalité dans $(\mathcal{P}_a)$ est convexe, affine même.

Pour étudier la convexité de f_a, considérons $\nabla^2 f_a(\xi_1,\xi_2)$:

$$\nabla^2 f_a(\xi_1,\xi_2) = \begin{bmatrix} 2 & 1 \\ 1 & 2a \end{bmatrix} \text{ pour tout } (\xi_1,\xi_2) \in \mathbb{R}^2.$$

$\nabla^2 f_a(\xi_1,\xi_2)$ est semi-définie positive si et seulement si $a \geqslant \frac{1}{4}$.

Donc, le problème $(\mathcal{P}_a)$ n'est un problème de minimisation convexe que lorsque $a \geqslant \frac{1}{4}$.

2°) Observations préliminaires :

– Si $a < 0$, sachant que $(0,\xi_2) \in C := \left\{(\xi_1,\xi_2) \in \mathbb{R}^2 \mid \xi_1 + \xi_2 - 1 \leqslant 0\right\}$ avec $\xi_2 < 0$ arbitraire, f_a n'est pas bornée inférieurement sur C.

– Si $a = 0$, sachant qu'il y a dans C des éléments $\left(\frac{1}{2},\xi_2\right)$ avec $\xi_2 \to -\infty$, f_a n'est pas bornée inférieurement sur C.

On n'étudiera donc $(\mathcal{P}_a)$ que pour $a > 0$.

1^{er} cas : $a \geqslant \dfrac{1}{4}$.

Ici $\overline{x} = \left(\overline{\xi}_1, \overline{\xi}_2\right) \in C$ est solution de $(\mathcal{P}_a)$ si et seulement si :

$$\exists\, \overline{\mu} \geqslant 0 \text{ tel que } \begin{cases} \nabla f(\overline{x}) + \overline{\mu}\nabla g(\overline{x}) = 0, \\ \overline{\mu} g(\overline{x}) = 0, \end{cases}$$

soit encore :

$$(1)\qquad \begin{cases} \overline{\xi}_1 + \overline{\xi}_2 - 1 < 0, \\ \begin{pmatrix} 2\overline{\xi}_1 + \overline{\xi}_2 + 1 \\ \overline{\xi}_1 + 2a\overline{\xi}_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{cases}$$

ou bien

$$(2)\qquad \begin{cases} \overline{\xi}_1 + \overline{\xi}_2 - 1 = 0, \\ \begin{pmatrix} 2\overline{\xi}_1 + \overline{\xi}_2 + 1 + \overline{\mu} \\ \overline{\xi}_1 + 2a\overline{\xi}_2 + \overline{\mu} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ \overline{\mu} \geqslant 0. \end{cases}$$

L'éventualité (1) ne se produit que lorsque $a > \frac{1}{3}$, auquel cas $\overline{x} = \left(-\frac{2a}{4a-1}, \ \frac{1}{4a-1}\right)$.

L'éventualité (2) ne se produit que pour $\frac{1}{4} \leqslant a \leqslant \frac{1}{3}$, auquel cas $\overline{x} = \left(1 - \frac{1}{a}, \ \frac{1}{a}\right)$ et $\overline{\mu} = \frac{1-3a}{a}$.

2^e *cas :* $0 < a < \frac{1}{4}$. Il y a deux points de C vérifiant les conditions nécessaires de minimalité du 1er ordre, à savoir :

$$\overline{x}_1 = \left(1 - \frac{1}{a}, \ \frac{1}{a}\right) \text{ sur la frontière de } C \ ;$$

$$\overline{x}_2 = \left(\frac{2a}{1-4a}, \ -\frac{1}{1-4a}\right) \text{ à l'intérieur de } C.$$

Quid des conditions nécessaires de minimalité du 2e ordre ?

– En $\overline{x}_1$, le sous-espace à considérer est $H = \{(d_1, d_2) \in \mathbb{R}^2 \mid : \langle \nabla g(\overline{x}_1), d\rangle = 0\}$, soit $H = \{(d_1, d_2) \in \mathbb{R}^2 \mid d_2 = -d_1\}$. Alors :

$$\forall d = (d_1, -d_1) \in H \quad \left\langle \nabla^2_{xx} \mathcal{L}(\overline{x}_1, \overline{\mu})d, d\right\rangle = \left\langle \nabla^2 f_a(\overline{x}_1)d, d\right\rangle = 2ad_1^2.$$

D'où $\left\langle \nabla^2 f_a(\overline{x}_1)d, d\right\rangle > 0$ pour tout $d \neq 0$ dans H : le point $\overline{x}_1$ est donc un minimum local strict de f_a sur C.

– En $\overline{x}_2$, le sous-espace à considérer est $H = \mathbb{R}^2$, et on a déjà vu que $\nabla^2 f_a(\overline{x}_2)$ n'était pas semi-définie positive. Donc $\overline{x}_2$ ne saurait être un minimum local de f_a sur C.

Reste à savoir si $\overline{x}_1$ est un minimum global de f_a sur C.

En fait, f_a est une fonction quadratique dont la valeur en $x = (\xi_1, \xi_2)$ peut être décomposée comme suit :

$$f_a(\xi_1, \xi_2) = \left(\xi_1 + \frac{\xi_2}{2}\right)^2 + \left(a - \frac{1}{4}\right)\xi_2^2 + \xi_1.$$

Quand $\xi_1 \to +\infty$ $(\xi_1, -2\xi_1)$ reste dans C et $f_a(\xi_1, -2\xi_1) = 4(a - \frac{1}{4})\xi_1^2 + \xi_1$ tend vers $-\infty$. Donc f_a n'est pas bornée inférieurement sur C.

Résumé :

a	solution	valeur optimale
$a < \frac{1}{4}$	$\emptyset$	$-\infty$
$\frac{1}{4} \leqslant a \leqslant \frac{1}{3}$	$\left(1 - \frac{1}{a},\ \frac{1}{a}\right)$	$\frac{2a-1}{a}$
$\frac{1}{3} < a$	$\left(\frac{2a}{1-4a},\ \frac{1}{4a-1}\right)$	$\frac{a}{1-4a}$

***Exercice III.28.** On se propose dans cet exercice d'obtenir les conditions nécessaires de minimalité de Karush-Kuhn-Tucker pour un problème de minimisation avec contraintes du type égalité et inégalité à partir des conditions nécessaires de minimalité du 1er et 2e ordre pour un problème de minimisation avec contraintes du type égalité seulement.

Soit donc

$$(\mathcal{P}) \quad \begin{cases} \text{Min } f(x) \\ x \in S \end{cases}, \quad \text{où} \quad \begin{aligned} S := \{x \in \mathbb{R}^n \mid h_i(x) = 0 \quad \text{pour} \quad i = 1, \ldots, m \\ \text{et } g_j(x) \leqslant 0 \quad \text{pour} \quad j = 1, \ldots, p\}. \end{aligned}$$

On supposera $f,\ h_1, \ldots, h_m,\ g_1, \ldots, g_p$ deux fois différentiables. Au point $\overline{x} \in S$ minimum local de f sur S, il est supposé

$(QC)'_{\overline{x}}$: $\nabla h_1(\overline{x}), \ldots, \nabla h_m(\overline{x}),\ \nabla g_j(\overline{x}),\ j \in J(\overline{x})$, sont linéairement indépendants.

On considère le problème de minimisation $(\hat{\mathcal{P}})$ suivant, plongement du problème $(\mathcal{P})$ dans $\mathbb{R}^n \times \mathbb{R}^p$:

$$\left(\hat{\mathcal{P}}\right) \quad \begin{cases} \text{Min } f(x) \\ h_i(x) = 0 \text{ pour } i = 1, \ldots, m \\ g_j(x) + y_j^2 = 0 \text{ pour } j = 1, \ldots, p. \end{cases}$$

1°) Vérifier que si $\overline{x}$ est un minimum local dans $(\mathcal{P})$ et si $\overline{y} = (\overline{y}_1, \ldots, \overline{y}_p) \in \mathbb{R}^p$ est tel que $g_j(\overline{x}) + \overline{y}_j^2 = 0$ pour $j = 1, \ldots, p$, alors $(\overline{x}, \overline{y})$ est un minimum local dans $(\hat{\mathcal{P}})$.

2°) Écrire les conditions nécessaires de minimalité du 1er et 2e ordre en $(\overline{x}, \overline{y})$ minimum local dans $(\hat{\mathcal{P}})$, et en déduire les conditions nécessaires de minimalité du 1er ordre en $\overline{x}$ minimum local dans $(\mathcal{P})$.

Commentaire : Cette manière de transformer des contraintes du type égalité et inégalité en contraintes du type égalité exclusivement est appelée parfois *transformation de Valentine* (non non, ce n'est pas le prénom de l'étudiante voisine, mais le nom d'un mathématicien du XX$^{\text{e}}$ siècle qui a travaillé sur les problèmes variationnels).

Solution : On définit les fonctions $\hat{f},\ \hat{h}_1, \dots, \hat{h}_{m+p} : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ comme suit :

$$\begin{aligned} \forall\, (x,y) \in \mathbb{R}^n \times \mathbb{R}^p, \quad & \hat{f}(x,y) := f(x), \\ & \hat{h}_i(x,y) := h_i(x) \text{ pour } i = 1, \dots, m, \\ & \hat{h}_{m+j}(x,y) := g_j(x) + y_j^2 \text{ pour } j = 1, \dots, p. \end{aligned}$$

1°) Pour $\overline{x} \in S$ considérons $(\overline{y}_1, \dots, \overline{y}_p) \in \mathbb{R}^p$ tel que $\hat{h}_{m+j}(\overline{x}, \overline{y}) = g_j(\overline{x}) + \overline{y}_j^2 = 0$ pour $j = 1, \dots, p$. Notons que $\overline{y}_j = 0$ *exactement pour les* $j \in J(\overline{x})$. L'implication suivante est alors claire :

$$\begin{pmatrix} \overline{x} \text{ est un minimum} \\ \text{local de } f \text{ sur } S \end{pmatrix} \Rightarrow \begin{pmatrix} (\overline{x}, \overline{y}) \text{ est un minimum local de } \hat{f} \text{ sur} \\ \hat{S} := \Big\{ (x,y) \in \mathbb{R}^n \times \mathbb{R}^p \mid \hat{h}_k(x,y) = 0 \\ \text{pour } k = 1, \dots, m+p \Big\} \end{pmatrix}.$$

2°) Observons la structure particulière de la matrice jacobienne de $H := (\hat{h}_1, \dots, \hat{h}_{m+p})$ en $(\overline{x}, \overline{y})$

$$[JH(\overline{x}, \overline{y})]^T = \begin{array}{|ccc|ccc|} \hline \multicolumn{3}{|c|}{\overbrace{}^{m}} & \multicolumn{3}{c|}{\overbrace{}^{p}} & \\ \nabla h_1(\overline{x}) & \dots & \nabla h_m(\overline{x}) & \nabla g_1(\overline{x}) & \dots & \nabla g_p(\overline{x}) & \Big\} n \\ \vdots & & \vdots & \vdots & & \vdots & \\ \vdots & & \vdots & \vdots & & \vdots & \\ \hline \vdots & & \vdots & 2\overline{y}_1 & & 0 & \\ 0 & & 0 & & \ddots & & \Big\} p \\ \vdots & & \vdots & 0 & & 2\overline{y}_p & \\ \hline \end{array}$$

pour constater que l'indépendance linéaire supposée des $\nabla h_1(\overline{x}), \dots, \nabla h_m(\overline{x}), \nabla g_j(\overline{x}),\ j \in J(\overline{x})$, entraîne celle des $\nabla \hat{h}_1(\overline{x}, \overline{y}), \dots, \nabla \hat{h}_{m+p}(\overline{x}, \overline{y})$.

En effet :

$$0 = \sum_{k=1}^{m+p} \alpha_k \nabla \hat{h}_k(\overline{x},\overline{y}) = \sum_{i=1}^{m} \alpha_i \begin{pmatrix} \nabla h_i(\overline{x}) \\ 0 \end{pmatrix} + \sum_{j=1}^{p} \alpha_{m+j} \begin{pmatrix} \nabla g_j(\overline{x}) \\ \vdots \\ 2\overline{y}_j \\ \vdots \\ 0 \end{pmatrix}$$

implique

$$\alpha_{m+j}\ \overline{y}_j = 0 \text{ pour } j = 1,\dots,p, \text{ d'où } \alpha_{m+j} = 0 \text{ si } j \notin J(\overline{x})$$

puis

$$0 = \sum_{i=1}^{m} \alpha_i \nabla h_i(\overline{x}) + \sum_{j \in J(\overline{x})} \alpha_{m+j} \nabla g_j(\overline{x})\ ;$$

et ceci n'est possible que si tous les coefficients α_k sont nuls.

Par conséquent, il existe $\left(\overline{\lambda}_1,\dots,\overline{\lambda}_m,\ \overline{\mu}_1,\dots,\overline{\mu}_p\right) \in \mathbb{R}^{m+p}$ unique tel que

$$0 = \nabla \hat{f}(\overline{x},\overline{y}) + \sum_{i=1}^{m} \overline{\lambda}_i \nabla \hat{h}_i(\overline{x},\overline{y}) + \sum_{j=1}^{p} \overline{\mu}_j \nabla \hat{h}_{m+j}(\overline{x},\overline{y}).$$

C'est la condition de minimalité du 1$^{\text{er}}$ ordre de Lagrange écrite en $(\overline{x},\overline{y})$ minimum local dans $(\hat{\mathcal{P}})$.

Cette relation se décompose en deux :

$$0 = \nabla f(\overline{x}) + \sum_{i=1}^{m} \overline{\lambda}_i \nabla h_i(\overline{x}) + \sum_{j=1}^{p} \overline{\mu}_j \nabla g_j(\overline{x})$$

et

$$0 = \overline{\mu}_j \overline{y}_j \quad \text{pour tout} \quad j = 1,\dots,p.$$

Cette dernière relation signifie encore : $\overline{\mu}_j = 0$ dès lors que $j \notin J(\overline{x})$.

Il ne reste plus qu'à démontrer que les $\overline{\mu}_j$ sont positifs. Cela résultera des conditions nécessaires de minimalité du 2$^{\text{e}}$ ordre écrites en $(\overline{x},\overline{y})$ minimum local de $\hat{f}$ sur $\hat{S}$.

Étant donné que les $\nabla \hat{h}_1(\overline{x},\overline{y}),\dots,\nabla \hat{h}_{m+p}(\overline{x},\overline{y})$ sont linéairement indépendants, et sachant que $\overline{y}_j = 0$ lorsque $j \in J(\overline{x})$, nous avons :

$$(d,\delta) \in T\left(\hat{S},(\overline{x},\overline{y})\right) \Leftrightarrow \begin{cases} \langle \nabla h_i(\overline{x}), d\rangle = 0 & \text{pour } i = 1,\dots,m \\ \langle \nabla g_j(\overline{x}), d\rangle = 0 & \text{pour } j \in J(\overline{x}) \\ \langle \nabla g_j(\overline{x}), d\rangle + 2\overline{y}_j\delta_j = 0 & \text{pour } j \notin J(\overline{x}). \end{cases}$$

Concernant le lagrangien

$$\begin{aligned} \hat{\mathcal{L}} : (\mathbb{R}^n \times \mathbb{R}^p) \times (\mathbb{R}^m \times \mathbb{R}^p) &\to \mathbb{R} \\ (x, y, \lambda, \mu) &\longmapsto \hat{\mathcal{L}}\,(x, y\,; \lambda, \mu) = \hat{f}(x,y) + \sum_{i=1}^{m} \lambda_i \hat{h}_i(x,y) + \sum_{j=1}^{p} \mu_j \hat{h}_{m+j}(x,y) \end{aligned}$$

associé au problème de minimisation $(\hat{\mathcal{P}})$, nous avons :

$$\nabla^2_{(x,y;x,y)}\hat{\mathcal{L}}\left(\overline{x},\overline{y};\overline{\lambda},\overline{\mu}\right) = \begin{array}{r} n\,\{ \\ p\,\{ \end{array} \underbrace{\begin{array}{|c|c|} \hline \nabla^2_{xx} f(\overline{x}) & 0 \\ \hline 0 & 0 \\ \hline \end{array}}_{n \quad\quad p} + \sum_{i=1}^{m} \overline{\lambda}_i \begin{array}{|c|c|} \hline \nabla^2_{xx} h_i(\overline{x}) & 0 \\ \hline 0 & 0 \\ \hline \end{array}$$

$$+ \sum_{j=1}^{p} \overline{\mu}_j \begin{array}{|c|cccc|} \hline \nabla^2_{xx} g_j(\overline{x}) & & & 0 & \\ \hline & 0 & & & \\ & & \ddots & & \\ 0 & & 2 & \cdots & \cdots \\ & & \vdots & \ddots & \\ & & \vdots & & 0 \\ \hline \end{array} \; j$$

$$\hspace{4em} j$$

de sorte que

$$\left\langle \nabla^2_{(x,y;x,y)}\hat{\mathcal{L}}\left(\overline{x},\overline{y};\overline{\lambda},\overline{\mu}\right)(d,\delta)\,,(d,\delta) \right\rangle =$$

$$\left\langle \left(\nabla^2_{xx} f\left(\overline{x}\right) + \sum_{i=1}^{m} \overline{\lambda}_i \nabla^2_{xx} h_i\left(\overline{x}\right) + \sum_{j \in J(\overline{x})} \overline{\mu}_j \nabla^2_{xx} g_j\left(\overline{x}\right) \right) d, d \right\rangle + 2 \sum_{j \in J(\overline{x})} \overline{\mu}_j \delta_j^2.$$

La positivité de cette forme quadratique sur $T\left(\hat{S}, (\overline{x},\overline{y})\right)$ implique alors la positivité des $\overline{\mu}_j$.

**** *Exercice III.29.*** On considère le problème :
$(\mathcal{P})$ Minimiser $f(x)$ sous les contraintes $g_1(x) \leqslant 0, \ldots, g_m(x) \leqslant 0$, avec les hypothèses suivantes sur les données :

$$f : \mathbb{R}^n \to \mathbb{R} \text{ est continue et 0-coercive } (\textit{i.e.} \lim_{\|x\| \to +\infty} f(x) = +\infty)\ ;$$

$g_1, \ldots, g_m : \mathbb{R}^n \to \mathbb{R}$ sont continues ; l'ensemble-contrainte de $(\mathcal{P})$ n'est pas vide.

Pour tout $k \in \mathbb{N}^*$, soit $P_k : \mathbb{R}^n \to \mathbb{R}$ définie par

$$x \in \mathbb{R}^n \longmapsto P_k(x) := f(x) + k \sum_{i=1}^{m} \left[g_i^+(x)\right]^2,$$

où $g_i^+(x)$ désigne $\max(g_i(x), 0)$. (P_k est une *version pénalisée* de f.)

1°) Montrer qu'il existe $\overline{x}_k$ minimisant globalement P_k sur $\mathbb{R}^n$.

2°) Vérifier que la suite $\{P_k(\overline{x}_k)\}$ est croissante et majorée par la valeur optimale $\overline{f}$ de $(\mathcal{P})$.

3°) Montrer que

$$\lim_{k \to +\infty} \sum_{i=1}^{m} \left[g_i^+(\overline{x}_k)\right]^2 = 0.$$

4°) Établir que la suite $\{\overline{x}_k\}$ est bornée et que toute limite de sous-suite convergente de $\{\overline{x}_k\}$ est solution de $(\mathcal{P})$.

5°) Montrer que

$$\lim_{k \to +\infty} k \sum_{i=1}^{m} \left[g_i^+(\overline{x}_k)\right]^2 = 0.$$

6°) On suppose que les fonctions $f,\ g_1, \ldots, g_m$ sont convexes et différentiables.

a) P_k est-elle convexe différentiable ?

b) Comment caractériser les $\overline{x}_k$ minimisant P_k sur $\mathbb{R}^n$?

Solution : f est continue et 0-coercive sur $\mathbb{R}^n$, l'ensemble-contrainte de $(\mathcal{P})$ est un fermé non vide de $\mathbb{R}^n$; donc $(\mathcal{P})$ a des solutions.

1°) P_k est continue, et comme $P_k \geqslant f$, elle est également 0-coercive sur $\mathbb{R}^n$: il existe donc $\overline{x}_k$ minimisant (globalement) P_k sur $\mathbb{R}^n$.

2°) Par définition même des $\overline{x}_k$,

$$\begin{aligned} P_k(\overline{x}_k) = \min_{x \in \mathbb{R}^n} P_k(x) \leqslant P_k(\overline{x}_{k+1}) &= f(\overline{x}_{k+1}) + k \sum_{i=1}^{m} \left[g_i^+(\overline{x}_{k+1})\right]^2 \\ &\leqslant \underbrace{f(\overline{x}_{k+1}) + (k+1) \sum_{i=1}^{m} \left[g_i^+(\overline{x}_{k+1})\right]^2}_{= P_{k+1}(\overline{x}_{k+1})}. \end{aligned}$$

Soit x vérifiant les contraintes de $(\mathcal{P})$: on a $g_i^+(x) = 0$ pour tout $i = 1, \ldots, m$. D'où

$$P_k(\overline{x}_k) \leqslant f(x) + k \sum_{i=1}^{m} \left[g_i^+(x)\right]^2 \text{ pour tout } x \in \mathbb{R}^n,$$
$$\leqslant f(x) \text{ pour tout } x \text{ vérifiant les contraintes de } (\mathcal{P}).$$

Par suite, $P_k(\overline{x}_k) \leqslant \inf\{f(x) \mid g_i(x) \leqslant 0 \text{ pour } i = 1, \ldots, m\} =: \overline{f}$.

3°) Soit r un minorant de f sur $\mathbb{R}^n$ (rappelons que f est bornée inférieurement sur $\mathbb{R}^n$ car continue et 0-coercive ; elle atteint même sa borne inférieure). Alors

$$r + k \sum_{i=1}^{m} \left[g_i^+(\overline{x}_k)\right]^2 \leqslant f(\overline{x}_k) + k \sum_{i=1}^{m} \left[g_i^+(\overline{x}_k)\right]^2 = P_k(\overline{x}_k) \leqslant \overline{f},$$

d'où

$$0 \leqslant \sum_{i=1}^{m} \left[g_i^+(\overline{x}_k)\right]^2 \leqslant \frac{\overline{f} - r}{k},$$

et donc

$$g_i^+(\overline{x}_k) \to 0 \text{ quand } k \to +\infty, \text{ pour tout } i = 1, \ldots, m.$$

4°) Si $\lim\limits_{l \to +\infty} \| \overline{x}_{k_l} \| = +\infty$ pour une sous-suite $\{\overline{x}_{k_l}\}_l$ de $\{\overline{x}_k\}$, on aurait $\lim\limits_{l \to +\infty} f(\overline{x}_{k_l}) = +\infty$ puisque f est 0-coercive. Mais

$$f(\overline{x}_{k_l}) \leqslant P_{k_l}(\overline{x}_{k_l}) \leqslant \overline{f},$$

d'où une contradiction. La suite $\{\overline{x}_k\}$ est donc bornée.
Considérons une sous-suite $\{\overline{x}_{k_l}\}$ convergente, de limite $\overline{x}$. On a :
– d'une part

$$g_i^+(\overline{x}) = \lim_{l \to +\infty} g_i^+(\overline{x}_{k_l}) = 0, \text{ pour tout } i = 1, \ldots, m$$
$$(\overline{x} \text{ vérifie les contraintes de } (\mathcal{P})) \ ;$$

– d'autre part

$$f(\overline{x}_{k_l}) \leqslant P_{k_l}(\overline{x}_{k_l}) \leqslant \overline{f},$$

d'où $f(\overline{x}) \leqslant \overline{f}$.
Le point-limite $\overline{x}$ est bien une solution de $(\mathcal{P})$.

5°) Raisonnons par l'absurde et supposons que la limite supérieure de $k\sum_{i=1}^{m}\left[g_i^+(\overline{x}_k)\right]^2$ quand $k\to+\infty$ est >0. Il s'ensuit alors une sous-suite $\{\overline{x}_{k_l}\}_l$ de $\{\overline{x}_k\}$ et $\alpha>0$ tels que

$$P_{k_l}(\overline{x}_{k_l})-f(\overline{x}_{k_l})=k_l\sum_{i=1}^{m}\left[g_i^+(\overline{x}_{k_l})\right]^2\geqslant\alpha \text{ pour tout } l.$$

Quitte à prendre une sous-suite de la suite (bornée) $\{\overline{x}_{k_l}\}$, on peut supposer que $\{\overline{x}_{k_l}\}_l$ est convergente. Sa limite $\overline{x}$ est solution de $(\mathcal{P})$, d'après ce qui a été démontré dans la 4e question. On a donc :

$$\overline{f}-f(\overline{x}_{k_l})\geqslant P_{k_l}(\overline{x}_{k_l})-f(\overline{x}_{k_l})\geqslant\alpha>0 \text{ pour tout } l$$

et

$$f(\overline{x}_{k_l})\to f(\overline{x})=\overline{f} \text{ quand } l\to+\infty,$$

d'où une contradiction.

6°) a) La fonction g_i étant convexe, il en est de même de $g_i^+=\max(0,g_i)$ et donc de $(g_i^+)^2$ (d'accord ?). La fonction g_i étant différentiable, il en est de même de $(g_i^+)^2$.

Donc P_k est une fonction convexe différentiable.

b) $\overline{x}_k$ minimise P_k sur $\mathbb{R}^n$ si, et seulement si, $\nabla P_k(\overline{x}_k)=0$. Cela s'exprime par

$$\nabla f(\overline{x}_k)+2k\sum_{i=1}^{m}g_i^+(\overline{x}_k)\nabla g_i(\overline{x}_k)=0.$$

****** Exercice III.30.*** On considère le problème d'optimisation suivant :

$$(\mathcal{P})\begin{cases}\text{Minimiser } f(x) \text{ sous les contraintes}\\ h_i(x)=0,\ \ i=1,\dots,m\\ g_i(x)\leqslant 0,\ \ i=m+1,\dots,p,\end{cases}$$

où $f,\ h_i,\ g_i:\mathbb{R}^n\to\mathbb{R}$ sont des fonctions deux fois différentiables sur $\mathbb{R}^n$.

Pour tout $c=(c_1,\dots,c_p)\in(\mathbb{R}^+)^p$, on définit la fonction p_c par :

$$x\in\mathbb{R}^n\longmapsto p_c(x):=f(x)+\sum_{i=1}^{m}c_i\,|h_i(x)|+\sum_{i=m+1}^{p}c_i\ g_i^+(x).$$

1°) Calculer la dérivée directionnelle $p_c'(x,d)$ de p_c en un point x de l'ensemble-contrainte de $(\mathcal{P})$ dans une direction quelconque d de $\mathbb{R}^n$.

2°) Soit $\overline{x}$ un point vérifiant les conditions nécessaires de minimalité du 1er ordre du type Karush-Kuhn-Tucker pour le problème $(\mathcal{P})$ (avec $\overline{\lambda}_1, \ldots, \overline{\lambda}_m, \overline{\lambda}_{m+1}, \ldots, \overline{\lambda}_p$ comme multiplicateurs). Montrer que si $c_i \geqslant \left|\overline{\lambda}_i\right|$ pour tout $i = 1, \ldots, p$, alors

$$p_c'(\overline{x}, d) \geqslant 0 \text{ pour tout } d.$$

3°) Pour tout x vérifiant les contraintes du problème $(\mathcal{P})$ on pose :

$$\begin{aligned} L(x) := \{d \in \mathbb{R}^n \mid & \langle \nabla h_i(x), d\rangle = 0 \text{ pour tout } i = 1, \ldots, m, \\ & \langle \nabla g_i(x), d\rangle = 0 \text{ pour tout } i \in I(x) \text{ tel que } \overline{\lambda}_i > 0, \\ & \langle \nabla g_i(x), d\rangle \leqslant 0 \text{ pour tout } i \in I(x) \text{ tel que } \overline{\lambda}_i = 0\big\}, \end{aligned}$$

où $I(x) := \{i \mid m+1 \leqslant i \leqslant p,\ \ g_i(x) = 0\}$.

On considère un minimum local $\overline{x}$ pour $(\mathcal{P})$ et on fait les hypothèses suivantes :

(H_1) Les gradients $\{\nabla h_i(\overline{x}),\ i = 1, \ldots, m\}$ et $\{\nabla g_i(\overline{x}),\ i \in I(\overline{x})\}$ sont linéairement indépendants (ce qui assure l'existence et l'unicité des multiplicateurs $\overline{\lambda}_i$ associés à $\overline{x}$) ;

(H_2) $\forall d \in L(\overline{x}), \quad \exists\, i \leqslant m$ pour lequel $\left\langle \nabla^2 h_i(\overline{x})d, d\right\rangle \neq 0$
ou bien
$\exists\, i \in I(\overline{x})$ pour lequel $\overline{\lambda}_i > 0$ et $\left\langle \nabla^2 g_i(\overline{x})d, d\right\rangle \neq 0$.

Montrer que si $c_i > \left|\overline{\lambda}_i\right|$ pour tout i, alors, pour tout $d \in \mathbb{R}^n$, il existe $\varepsilon_d > 0$ tel que : $p_c(\overline{x} + td) \geqslant p_c(\overline{x})$ pour tout $t \in]-\varepsilon_d, \varepsilon_d[$ ($\overline{x}$ est un minimum local de p_c dans toute direction d).

Indication. On distinguera le cas $d \notin L(\overline{x})$ et le cas $d \in L(\overline{x})$.

Solution : 1°) Si φ_1 et φ_2 sont deux fonctions de classe C^1 et si $\varphi_1(x) = \varphi_2(x)$, la fonction $\varphi_3 := \max\{\varphi_1, \varphi_2\}$ admet une dérivée directionnelle en x avec :

$$\forall d \in \mathbb{R}^n,\ \ \varphi_3'(x, d) = \max\{\langle \nabla\varphi_1(x), d\rangle, \langle \nabla\varphi_2(x), d\rangle\}.$$

Par application de ce résultat nous avons :

$$\forall d \in \mathbb{R}^n,\ \ p_c'(x, d) = \langle \nabla f(x), d\rangle + \sum_{i=1}^{m} c_i \left|\langle \nabla h_i(x), d\rangle\right| + \sum_{i \in I(x)} c_i \left[\langle \nabla g_i(x), d\rangle\right]^+.$$

2°) Soit $\overline{x}$ un point vérifiant les conditions nécessaires de minimalité du 1$^{\text{er}}$ ordre du type KKT, *i.e.* : Il existe $\overline{\lambda}_1, \ldots, \overline{\lambda}_m, \overline{\lambda}_{m+1}, \ldots, \overline{\lambda}_p$ tels que

$$\nabla f(\overline{x}) + \sum_{i=1}^{m} \overline{\lambda}_i \nabla h_i(\overline{x}) + \sum_{i=m+1}^{p} \overline{\lambda}_i \nabla g_i(\overline{x}) = 0,$$
$$\overline{\lambda}_i \geqslant 0 \text{ et } \overline{\lambda}_i g_i(\overline{x}) = 0 \quad \text{pour tout } i = m+1, \ldots, p.$$

Supposons $c_i \geqslant |\overline{\lambda}_i|$ pour tout i. Alors :
– pour $1 \leqslant i \leqslant m$, $\overline{\lambda}_i \langle \nabla h_i(\overline{x}), d\rangle \leqslant |\overline{\lambda}_i| \cdot |\langle \nabla h_i(\overline{x}), d\rangle| \leqslant c_i |\langle \nabla h_i(\overline{x}), d\rangle|$;
– pour $i \in I(\overline{x})$, $\overline{\lambda}_i \geqslant 0$ de sorte que

$$\overline{\lambda}_i \langle \nabla g_i(\overline{x}), d\rangle \leqslant \overline{\lambda}_i [\langle \nabla g_i(\overline{x}), d\rangle]^+ \leqslant c_i [\langle \nabla g_i(\overline{x}), d\rangle]^+ .$$

Ainsi

$$p_c'(\overline{x}, d) \geqslant \left\langle \nabla f(\overline{x}) + \sum_{i=1}^{m} \overline{\lambda}_i \nabla h_i(\overline{x}) + \sum_{i \in I(\overline{x})} \overline{\lambda}_i \nabla g_i(\overline{x}), d \right\rangle = 0.$$

3°) L'hypothèse (H_1) assure l'existence (et l'unicité) des multiplicateurs $\overline{\lambda}_i$ associés à $\overline{x}$.

– Considérons $d \notin L(\overline{x})$. L'une des inégalités dans la preuve au-dessus de la positivité de $p_c'(x, d)$ est stricte (d'accord ?) ; donc $p_c'(\overline{x}, d) > 0$.

– Considérons $d \in L(\overline{x})$. Notons que nous avons le développement suivant :

$$p_c'(\overline{x}, d) = 0 \ ;$$
$$p_c(\overline{x} + td) - p_c(\overline{x}) = \frac{t^2}{2} \left\{ \langle \nabla^2 f(\overline{x}) d, d\rangle + \sum_{i=1}^{m} c_i \left|\langle \nabla^2 h_i(\overline{x}) d, d\rangle\right| + \sum_{i \in I(\overline{x})} c_i \left[\langle \nabla^2 g_i(\overline{x}) d, d\rangle\right]^+ \right\} + o(t^2).$$

Alors :
– pour $1 \leqslant i \leqslant m$, $\overline{\lambda}_i \langle \nabla^2 h_i(\overline{x}) d, d\rangle \leqslant c_i \left|\langle \nabla^2 h_i(\overline{x}) d, d\rangle\right|$;
– pour $i \in I(\overline{x})$, $\overline{\lambda}_i \langle \nabla^2 g_i(\overline{x}) d, d\rangle \leqslant c_i \left[\langle \nabla^2 g_i(\overline{x}) d, d\rangle\right]^+ .$

L'hypothèse (H_2) implique que l'une des inégalités ci-dessus est stricte. Par conséquent :

$$\begin{aligned}\lim_{t\to 0}\frac{p_c(\overline{x}+td)-p_c(\overline{x})}{t^2/2} &= \left\langle \nabla^2 f(\overline{x})d,d\right\rangle + \sum_{i=1}^{m} c_i \left|\left\langle \nabla^2 h_i(\overline{x})d,d\right\rangle\right| \\ &\quad + \sum_{i\in I(\overline{x})} c_i \left[\left\langle \nabla^2 g_i(\overline{x})d,d\right\rangle\right]^+ \\ &> \left\langle \nabla^2 f(\overline{x})d,d\right\rangle + \sum_{i=1}^{m} \overline{\lambda}_i \left\langle \nabla^2 h_i(\overline{x})d,d\right\rangle \\ &\quad + \sum_{i\in I(\overline{x})} \overline{\lambda}_i \left\langle \nabla^2 g_i(\overline{x})d,d\right\rangle \\ &\geqslant 0 \left(\begin{array}{c}\text{de par les conditions nécessaires}\\ \text{de minimalité du } 2^{\text{e}} \text{ ordre}\end{array}\right).\end{aligned}$$

Donc $\lim\limits_{t\to 0}\dfrac{p_c(\overline{x}+td)-p_c(\overline{x})}{t^2/2}$ est >0. Dans les deux cas de figure ($d\notin L(\overline{x})$ et $d\in L(\overline{x})$), $\overline{x}$ est un minimum local de p_c dans la direction d.

***** Exercice III.31.*** Soit f et h définies sur $\mathbb{R}^2$ comme suit :

$$\begin{aligned}\forall x=(\xi_1,\xi_2)\in\mathbb{R}^2,\quad f(x) &:= -\xi_2^3-2\xi_2^2-\xi_2\\ h(x) &:= \xi_1^2+\xi_2^2+\xi_2.\end{aligned}$$

On considère le problème d'optimisation $(\mathcal{P})$ qui consiste à minimiser $f(x)$ sous la contrainte $h(x)=0$.

1°) Déterminer les points vérifiant les conditions nécessaires d'optimalité du 1^{er} ordre.
Parmi ces points quels sont ceux qui sont minima globaux dans $(\mathcal{P})$?
(Il y en a deux, que nous noterons $\overline{x}_1$ et $\overline{x}_2$.)

2°) Les conditions suffisantes du 2^{e} ordre pour être un minimum local strict dans $(\mathcal{P})$ sont-elles satisfaites en $\overline{x}_1$ et $\overline{x}_2$?

3°) On pose

$$\overline{f}(\alpha) := \inf_{h(x)=\alpha} f(x),\ \text{où } \alpha \text{ est un paramètre réel.}$$

Montrer, sans calculer son expression explicite, que $\overline{f}$ ne saurait être dérivable en 0.

Solution : 1°) Les fonctions f et h sont continûment différentiables sur $\mathbb{R}^2$, de plus $\nabla h(x) = \begin{pmatrix} 2\xi_1 \\ 2\xi_2 + 1 \end{pmatrix} \neq 0$ pour tout x de l'ensemble-contrainte. Donc les points $\overline{x} = (\overline{\xi}_1, \overline{\xi}_2)$ vérifiant les conditions nécessaires d'optimalité du 1er ordre (conditions de Lagrange) sont ceux pour lesquels

$$\begin{cases} \overline{\xi}_1^2 + \overline{\xi}_2^2 + \overline{\xi}_2 = 0 \ ; \\ \exists\ \overline{\lambda} \in \mathbb{R} \text{ tel que } \begin{pmatrix} 0 \\ -3\overline{\xi}_2^2 - 4\overline{\xi}_2 - 1 \end{pmatrix} + \overline{\lambda} \begin{pmatrix} 2\overline{\xi}_1 \\ 2\overline{\xi}_2 + 1 \end{pmatrix} = 0. \end{cases}$$

Des calculs simples conduisent aux quatre possibilités suivantes :

$$\overline{x}_1 = (0,0) \text{ et } \overline{\lambda}_1 = 1 \ ; \ \overline{x}_2 = (0,-1) \text{ et } \overline{\lambda}_2 = 0 \ ;$$
$$\overline{x}_3 = \left(\frac{\sqrt{2}}{3}, -\frac{1}{3}\right) \text{ et } \overline{\lambda}_3 = 0 \ ; \ \overline{x}_4 = \left(-\frac{\sqrt{2}}{3}, -\frac{1}{3}\right) \text{ et } \overline{\lambda}_4 = 0.$$

Les points $\overline{x}_1$ et $\overline{x}_2$ sont les minima globaux dans $(\mathcal{P})$; la valeur minimale dans $(\mathcal{P})$ est $\overline{f} = 0$.

2°) Le sous-espace tangent à l'ensemble-contrainte en $\overline{x}_1$ comme en $\overline{x}_2$ est $H = \mathbb{R} \times \{0\}$. Ensuite

$$\nabla^2_{xx}\mathcal{L}\left(\overline{x}_1, \overline{\lambda}_1\right) := \nabla^2 f(\overline{x}_1) + \overline{\lambda}_1 \nabla^2 h(\overline{x}_1) = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$$

vérifie

$$\left\langle \nabla^2_{xx}\mathcal{L}\left(\overline{x}_1, \overline{\lambda}_1\right) d, d \right\rangle > 0 \text{ pour tout } d \neq 0 \text{ dans } H.$$

Les conditions suffisantes du second ordre de minimalité locale sont bien satisfaites en $\overline{x}_1$.

En ce qui concerne $\overline{x}_2$, $\nabla^2_{xx}\mathcal{L}\left(\overline{x}_2, \overline{\lambda}_2\right) = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}$ de sorte que $\left\langle \nabla^2_{xx}\mathcal{L}\left(\overline{x}_2, \overline{\lambda}_2\right) d, d \right\rangle = 0$ pour tout $d \in H$. Les conditions suffisantes du second ordre de minimalité locale ne sont pas satisfaites en $\overline{x}_2$.

3°) Les conditions suffisantes du second ordre de minimalité locale étant satisfaites en $\overline{x}_1$, on a le résultat suivant : il existe un intervalle ouvert A contenant 0 et une application $\alpha \in A \longmapsto \overline{x}_{1,\alpha}$ dérivable en 0 tels que :

(i) $\overline{x}_{1,0} = \overline{x}_1$ et, pour tout $\alpha \in A$, $\overline{x}_{1,\alpha}$ est un minimum local strict de f sous la contrainte $h(x) = \alpha$;

(ii) $\alpha \longmapsto \varphi(\alpha) := f(\overline{x}_{1,\alpha})$ est dérivable en 0 et $\varphi'(0) = -\overline{\lambda}_1$.

Mais ce qui est demandé ici concerne la valeur minimale *globale* $f(\alpha)$ et non $\varphi(\alpha)$. Puisque $\overline{x}_1$ et $\overline{x}_2$ sont des minima globaux dans $(\mathcal{P})$,

$$f(\overline{x}_1) = f(\overline{x}_2) = \overline{f}(0) = \overline{f}(h(\overline{x}_1)) = \overline{f}(h(\overline{x}_2)).$$

Soit v quelconque dans $\mathbb{R}^n$ et considérons le problème (perturbé) de la minimisation de f sous la contrainte $h(x) = h(v)$. Il est évident que v vérifie cette contrainte et donc $\overline{f}(h(v)) \leqslant f(v)$. On est donc dans la situation suivante :

$$\begin{aligned} &f(v) - \overline{f}(h(v)) \geqslant 0 \text{ pour tout } v \in \mathbb{R}^n \ ; \\ &f(\overline{x}_1) - \overline{f}(h(\overline{x}_1)) = 0, \ \ f(\overline{x}_2) - \overline{f}(h(\overline{x}_2)) = 0. \end{aligned}$$

Ainsi $\overline{x}_1$ et $\overline{x}_2$ minimisent (sans contraintes) la fonction $x \longmapsto g(x) := f(x) - \overline{f}(h(x))$.

Supposons $\overline{f}$ dérivable en 0. Il vient de ce qui précède :

$$\begin{aligned} 0 = \nabla g(\overline{x}_1) &= \nabla f(\overline{x}_1) - \overline{f}'(0)\nabla h(\overline{x}_1) \\ &= \nabla f(\overline{x}_2) - \overline{f}'(0)\nabla h(\overline{x}_2). \end{aligned}$$

Mais – conditions nécessaires d'optimalité du 1er ordre – on a :

$$\nabla f(\overline{x}_1) + \overline{\lambda}_1 \nabla h(\overline{x}_1) = 0 \text{ et } \nabla f(\overline{x}_2) + \overline{\lambda}_2 \nabla h(\overline{x}_2) = 0.$$

Par suite, $\overline{f}'(0) = -\overline{\lambda}_1$ et $\overline{f}'(0) = -\overline{\lambda}_2$. Ceci est impossible puisque $\overline{\lambda}_1 \neq \overline{\lambda}_2$. Donc $\overline{f}$ ne saurait être dérivable en 0.

Commentaire : Sous des hypothèses raisonnables on arrive à faire en sorte que la fonction valeur minimale $\overline{f}$ ait une dérivée directionnelle $\delta \longmapsto \overline{f}'(0, \delta)$; suivant la direction δ dans laquelle on regarde, $\overline{f}'(0, \delta)$ « choisit » dans son expression les multiplicateurs $\overline{\lambda}_i$ associés aux différents minima $\overline{x}$ dans $(\mathcal{P})$. Dans le cas de l'exemple, suivant que l'on se déplace vers la droite $(\delta = 1)$ ou vers la gauche $(\delta = -1)$, l'expression de $\overline{f}'(0, \delta)$ (c'est-à-dire de la dérivée à droite ou de la dérivée à gauche ici) fait appel au multiplicateur $\overline{\lambda}_1$ associé à $\overline{x}_1$ ou à celui $\overline{\lambda}_2$ associé à $\overline{x}_2$.

***** Exercice III.32.*** Étant données des fonctions $f_1, f_2, \ldots, f_p$ de $\mathbb{R}^n$ dans $\mathbb{R}$, on définit

$$x \in \mathbb{R}^n \longmapsto g(x) := \max\{f_1(x), f_2(x), \ldots, f_p(x)\}.$$

1°) Vérifier que si les f_i sont toutes continues, alors g est continue.

On pose, pour $x \in \mathbb{R}^n$,

$$I(x) := \{1 \leqslant i \leqslant p \mid f_i(x) = g(x)\}\,.$$

2°) On suppose que les f_i sont toutes différentiables en $\overline{x}$. À quelle condition (nécessaire et suffisante) portant sur les $\nabla f_i(\overline{x})$, $i \in I(\overline{x})$, la fonction g est-elle différentiable en $\overline{x}$?

3°) On définit $\varphi_0, \varphi_1, \ldots, \varphi_p : \mathbb{R}^n \times \mathbb{R} \longmapsto \mathbb{R}$ de la manière suivante :

$$\begin{aligned} \forall (x,r) \in \mathbb{R}^n \times \mathbb{R}, \quad & \varphi_0(x,r) := r, \\ & \varphi_i(x,r) := f_i(x) - r \quad \text{pour} \quad i = 1, \ldots, p. \end{aligned}$$

– Vérifier que si $\overline{x}$ est un minimum local de g, alors $(\overline{x}, g(\overline{x}))$ est un minimum local de φ_0 sur l'ensemble-contrainte

$$S := \{(x,r) \in \mathbb{R}^n \times \mathbb{R} \mid \varphi_i(x,r) \leqslant 0 \text{ pour tout } i = 1, \ldots, p\}\,.$$

– Quelle condition nécessaire de minimalité (en $\overline{x}$) peut-on déduire de cette observation ?

Solution : 1°) En posant $g_j := \max\{f_1, f_2, \ldots, f_j\}$ pour $j = 1, \ldots, p$, on a :

$$g_{j+1} = \max\{g_j, f_{j+1}\} = \frac{1}{2}\,(g_j + f_{j+1} + |g_j - f_{j+1}|)\,.$$

Grâce à la continuité de $g_1 = f_1$, de $f_2, \ldots, f_p$, et celle de $|h|$ lorsque h est continue, la formule ci-dessus conduit par récurrence à la continuité de $g_p = g$.

2°) Lorsque $i \notin I(\overline{x})$, on a $f_i(\overline{x}) < g(\overline{x})$. De par la continuité de g et des f_i, il existe un voisinage V de $\overline{x}$ tel que :

$$f_i(x) < g(x) \text{ pour tout } i \notin I(\overline{x}) \text{ et tout } x \in V.$$

Une autre manière de dire la même chose est :

$$I(x) \subset I(\overline{x}) \text{ pour tout } x \in V.$$

En conséquence,

$$g(x) = \max\{f_i(x) \mid i \in I(\overline{x})\} \text{ pour tout } x \in V. \tag{3.15}$$

Qu'est-ce qui fait à présent que g est différentiable en $\overline{x}$? L'examen attentif de quelques exemples (même de fonctions d'une seule variable) suggère le résultat suivant :

(g est différentiable en $\overline{x}$) $\Leftrightarrow$ ($\nabla f_i(\overline{x})$ est constant pour tout $i \in I\,(\overline{x})$)\,.

Auquel cas, $\nabla g(\overline{x}) = \nabla f_i(\overline{x})$ (i étant pris quelconque dans $I(\overline{x})$).

[⇐]. Posons $v = \nabla f_i(\overline{x}),\ i \in I(\overline{x})$. Soit $i \in I(\overline{x})$ et écrivons que f_i est différentiable en $\overline{x}$: $\forall \varepsilon > 0,\ \exists\ \delta_i > 0$ tel que

$$(\|x - \overline{x}\| \leqslant \delta_i) \Rightarrow (|f_i(x) - f_i(\overline{x}) - \langle v, x - \overline{x}\rangle| \leqslant \varepsilon \|x - \overline{x}\|). \tag{3.16}$$

Pour x assez voisin de $\overline{x}$, $g(x) = f_i(x)$ où i est un certain indice de $I(\overline{x})$ (cela résulte de (3.15)). En conséquence, il vient de (3.16) : $\exists\ \delta > 0$ tel que

$$(\|x - \overline{x}\| \leqslant \delta) \Rightarrow (|g(x) - g(\overline{x}) - \langle v, x - \overline{x}\rangle| \leqslant \varepsilon \|x - \overline{x}\|).$$

On a ainsi démontré que g est différentiable en $\overline{x}$ et $\nabla g(\overline{x}) = v$.

[⇒]. Supposons g différentiable en $\overline{x}$. Soit i quelconque dans $I(\overline{x})$. Puisque $g \geqslant f_i$ (par définition même de g), que $(g - f_i)(\overline{x}) = 0$ (car $i \in I(\overline{x})$), et que $g - f_i$ est différentiable en $\overline{x}$, on a $\nabla(g - f_i)(\overline{x}) = 0$. D'où $\nabla g(\overline{x}) = \nabla f_i(\overline{x})$.

3°) – Soit V un voisinage de $\overline{x}$ sur lequel $g(x) \geqslant g(\overline{x})$. Un simple jeu d'inégalités permet de voir que l'on a

$$r \geqslant g(\overline{x}) \ \text{ pour tout } (x, r) \in (V \times \mathbb{R}) \cap S.$$

Donc $(\overline{x}, g(\overline{x}))$ est un minimum local de φ_0 sur S. (Il est d'ailleurs intéressant de visualiser sur une figure ce passage de $\mathbb{R}^n$ à $\mathbb{R}^n \times \mathbb{R}$).

– Concernant φ_0 et φ_i nous avons :

$$\nabla\varphi_0(\overline{x}, g(\overline{x})) = \begin{pmatrix} 0 \\ 1 \end{pmatrix},\ \nabla\varphi_i(\overline{x}, g(\overline{x})) = \begin{pmatrix} \nabla f_i(\overline{x}) \\ -1 \end{pmatrix} \text{ dans } \begin{matrix} \mathbb{R}^n \\ \times \\ \mathbb{R} \end{matrix};$$

par ailleurs, dans le problème de minimisation de φ_0 sur S, les indices des contraintes-inégalités actives en $(\overline{x}, g(\overline{x}))$ sont exactement ceux pour lesquels $f_i(\overline{x}) = g(\overline{x})$, c'est-à-dire ceux de $I(\overline{x})$. Comme l'hypothèse de qualification des contraintes $(QC)_{(\overline{x}, g(\overline{x}))}$, qui requiert l'existence de $(d, \delta) \in \mathbb{R}^n \times \mathbb{R}$ tel que

$$\left\langle \begin{pmatrix} \nabla f_i(\overline{x}) \\ -1 \end{pmatrix}, \begin{pmatrix} d \\ \delta \end{pmatrix} \right\rangle < 0 \text{ pour tout } i \in I(\overline{x}),$$

est trivialement vérifiée (comme d'ailleurs l'hypothèse de qualification des contraintes $(QC)'_{(\overline{x}, g(\overline{x}))}$), les conditions nécessaires de minimalité de KKT induisent qu'il existe des réels positifs $\overline{\mu}_1, \ldots, \overline{\mu}_p$ tels que :

$$\begin{cases} \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \sum_{i=1}^{p} \overline{\mu}_i \begin{pmatrix} \nabla f_i(\overline{x}) \\ -1 \end{pmatrix} = 0, \\ \overline{\mu}_i = 0 \text{ si } i \notin I(\overline{x}). \end{cases}$$

Une condition nécessaire de minimalité locale (du 1er ordre) vérifiée en $\overline{x}$ est donc :

– Une combinaison convexe des $\nabla f_i(\overline{x}),\ i \in I(\overline{x})$, est égale (au vecteur) 0.

– Autre manière de dire la même chose : (le vecteur) 0 est dans le plus petit polyèdre convexe construit à partir des $\nabla f_i(\overline{x}),\ i \in I(\overline{x})$.

Commentaire : – Si on a affaire à une famille *infinie* de fonctions continues f_i, on ne peut seulement assurer que la *semi-continuité inférieure* de $g := \sup\limits_i f_i$. Prendre comme exemple

$$x \in \mathbb{R} \longmapsto f_i(x) := |x|^{1/i},\ i \in \mathbb{N}^* ;$$

il y a alors un « décrochement vers le bas » de $g := \sup_i f_i$ au point 0.

– Lorsque les f_i sont différentiables, l'objet mathématique adéquat qui joue le rôle de gradient de $g := \max\{f_1, \ldots, f_p\}$ en $\overline{x}$ est

$$\hat{\nabla} g(\overline{x}) := \left\{ \sum_{i \in I(\overline{x})} \alpha_i \nabla f_i(\overline{x}) \mid \sum_{i \in I(\overline{x})} \alpha_i = 1 \text{ et } \alpha_i \geqslant 0 \text{ pour tout } i \in I(\overline{x}) \right\}.$$

Résultats démontrés dans l'exercice :

– g est différentiable en $\overline{x}$ si et seulement si $\hat{\nabla} g(\overline{x})$ est un singleton ;

– Si $\overline{x}$ est un minimum local de g, alors $0 \in \hat{\nabla} g(\overline{x})$.

– Prolongement de l'exercice : réfléchir à ce que pourrait être (ou ce que devrait être) l'objet mathématique jouant le rôle de matrice hessienne de $g := \max\{f_1, \ldots, f_p\}$ en $\overline{x}$ lorsque les fonctions f_i sont deux fois différentiables en $\overline{x}$, et par suite à une condition nécessaire de minimalité (du 2e ordre) vérifiée en un minimum local de g.

IV

MINI-MAXIMISATION. DUALISATION DE PROBLÈMES DE MINIMISATION CONVEXE

Rappels

IV.1. Points-selles (ou cols) ; problèmes de mini-maximisation

Soit $l : X \times Y \to \mathbb{R}$. Un couple $(\overline{x}, \overline{y}) \in X \times Y$ est appelé *point-selle* (ou col) de l sur $X \times Y$ lorsque

$$l(\overline{x}, y) \leqslant l(\overline{x}, \overline{y}) \leqslant l(x, \overline{y}) \text{ pour tout } (x, y) \in X \times Y.$$

L'ensemble des points-selles de l sur $X \times Y$ a nécessairement une structure très particulière : c'est un produit cartésien d'ensembles (qui peut être vide, cela va sans dire). La valeur $l(\overline{x}, \overline{y})$ est constante pour tous les points-selles $(\overline{x}, \overline{y})$ de l sur $X \times Y$: elle est appelée *valeur-selle*.

Un exemple de résultat d'existence de points-selles est fourni par le théorème qui suit.

Théorème. *Soient $X \subset \mathbb{R}^n$ et $Y \subset \mathbb{R}^m$ deux convexes fermés non vides et $l : X \times Y \to \mathbb{R}$. On suppose :*

– Pour tout $y \in Y$, la fonction $l(\cdot, y) : X \to \mathbb{R}$ est convexe ; pour tout $x \in X$, la fonction $l(x, \cdot) : Y \to \mathbb{R}$ est concave (on dit alors que l est convexe-concave sur $X \times Y$) ;

– X est borné, ou bien il existe $y_0 \in Y$ tel que $l(x, y_0) \to +\infty$ quand $\| x \| \to +\infty$, $x \in X$;

– *Y est borné, ou bien il existe $x_0 \in X$ tel que $l(x_0, y) \to -\infty$ quand $\| y \| \to +\infty$, $y \in Y$;*

alors l'ensemble des points-selles de l sur $X \times Y$ est un convexe compact non vide de $X \times Y$.

Posons à présent

$$\varphi : x \in X \longmapsto \varphi(x) := \sup_{y \in Y} l(x, y) \quad (\text{à valeurs } +\infty \text{ éventuellement}),$$
$$\psi : y \in Y \longmapsto \psi(y) := \inf_{x \in X} l(x, y) \quad (\text{à valeurs } -\infty \text{ éventuellement}).$$

Les définitions mêmes d'inf et de sup font que l'on a toujours les inégalités : $\psi(y) \leqslant l(x, y) \leqslant \varphi(x)$ pour tout $(x, y) \in X \times Y$. Et $(\overline{x}, \overline{y})$ est un point-selle de l sur $X \times Y$ si et seulement si $\psi(\overline{y}) \geqslant \varphi(\overline{x})$; dans ce cas $\psi(\overline{y}) = \varphi(\overline{x})$ est la valeur-selle de l sur $X \times Y$.

Il est naturel d'associer à l les deux problèmes d'optimisation suivants :

$$(1) \quad \begin{cases} \text{Minimiser } \left\{ \sup_{y \in Y} l(x, y) \right\} \text{ (problème de « minimisation de sup »)} \\ \text{pour } x \in X; \end{cases}$$

$$(2) \quad \begin{cases} \text{Maximiser } \left\{ \inf_{x \in X} l(x, y) \right\} \text{ (problème de « maximisation d'inf »)} \\ \text{pour } y \in Y. \end{cases}$$

Alors, une condition nécessaire et suffisante pour que l ait des points-selles sur $X \times Y$ est

$$\min_{x \in X} \varphi(x) = \max_{y \in Y} \psi(y) \qquad (i.e. \text{ optima égaux et atteints}).$$

Dans ce cas, $(\overline{x}, \overline{y})$ est un point-selle exactement lorsque $\overline{x}$ minimise φ sur X et $\overline{y}$ maximise ψ sur Y.

IV.2. Points-selles de lagrangiens

Considérons le problème (basique) de la minimisation d'une fonction-objectif sous des contraintes exprimées par des égalités et des inégalités :

$$(\mathcal{P}) \quad \begin{cases} \text{Minimiser } f(x) \\ h_i(x) = 0 \text{ pour } i = 1, \ldots, m \text{ et} \\ g_j(x) \leqslant 0 \text{ pour } j = 1, \ldots, p \quad (x \in C \text{ en bref}). \end{cases}$$

Nous sommes plus particulièrement intéressés, ici, par les points-selles du lagrangien (usuel) sur $\mathbb{R}^n \times \left[\mathbb{R}^m \times (\mathbb{R}^+)^p\right]$; les données sont donc les suivantes :

$$X = \mathbb{R}^n,\ Y = \mathbb{R}^m \times \left(\mathbb{R}^+\right)^p,$$

$$\mathcal{L} : (x, (\lambda, \mu)) \in X \times Y \longmapsto \mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{j=1}^{p} \mu_j g_j(x).$$

Théorème. *Les points-selles de $\mathcal{L}$ sur $\mathbb{R}^n \times \left[\mathbb{R}^m \times (\mathbb{R}^+)^p\right]$ sont exactement les $\left(\overline{x}, \left(\overline{\lambda}, \overline{\mu}\right)\right)$ tels que :*

(i) $\overline{x}$ *minimise* $\mathcal{L}\left(\cdot, \left(\overline{\lambda}, \overline{\mu}\right)\right)$ *sur* $\mathbb{R}^n$;

(ii) $\overline{x} \in C$;

(iii) $\overline{\mu}_j g_j(\overline{x}) = 0$ *pour tout* $j = 1, \ldots, p$.

En particulier, *si* $\left(\overline{x}, \left(\overline{\lambda}, \overline{\mu}\right)\right)$ *est un point-selle de* $\mathcal{L}$ *sur* $\mathbb{R}^n \times \left[\mathbb{R}^m \times (\mathbb{R}^+)^p\right]$ *,alors* $\overline{x}$ *est une solution du problème* $(\mathcal{P})$.

Supposons maintenant que $(\mathcal{P})$ soit un problème de *minimisation convexe* : f et les g_j sont convexes, les h_i sont affines.

Théorème. *Sous les hypothèses de convexité décrites ci-dessus, les deux énoncés suivants sont équivalents :*

(i) $\left(\overline{x}, \left(\overline{\lambda}, \overline{\mu}\right)\right)$ *est un point-selle de* $\mathcal{L}$ *sur* $\mathbb{R}^n \times \left[\mathbb{R}^m \times (\mathbb{R}^+)^p\right]$;

(ii) $\overline{x}$ *est solution de* $(\mathcal{P})$ *et* $\left(\overline{\lambda}, \overline{\mu}\right)$ *est un multiplicateur de Lagrange.*

Dans tous les exercices considérés dans ce chapitre, le problème de minimisation convexe $(\mathcal{P})$ aura des solutions (l'ensemble S des solutions de $(\mathcal{P})$ ne sera pas vide), et des hypothèses de qualification des contraintes seront faites pour qu'il y ait des multiplicateurs de Lagrange (l'ensemble M des multiplicateurs de Lagrange ne sera pas vide). Donc l'ensemble des points-selles du lagrangien $\mathcal{L}$ sur $\mathbb{R}^n \times \left[\mathbb{R}^m \times (\mathbb{R}^+)^p\right]$ sera le produit cartésien des deux convexes fermés S et M.

IV.3. Premiers pas dans la théorie de la dualité

Revenons au problème de minimisation $(\mathcal{P})$ et au lagrangien $\mathcal{L}$ qui lui est associé, et voyons ce que sont alors les problèmes de mini-maximisation introduits au premier paragraphe. Le premier problème de mini-maximisation est celui de la minimisation sur $\mathbb{R}^n$ de

$$x \longmapsto \varphi(x) := \sup_{(\lambda,\mu) \in \mathbb{R}^m \times (\mathbb{R}^+)^p} \mathcal{L}(x, \lambda, \mu).$$

Le calcul explicite de $\varphi(x)$ est aisé : $\varphi(x) = f(x)$ si $x \in C$, $+\infty$ sinon.

Donc, notre premier problème de mini-maximisation est celui de la minimisation sur C de $f(x)$, c'est-à-dire ni plus ni moins que le problème $(\mathcal{P})$ originel. Regardons maintenant ce qu'est le second problème de mini-maximisation. Il faut considérer

$$(\lambda, \mu) \in \mathbb{R}^m \times \left(\mathbb{R}^+\right)^p \longmapsto \psi(\lambda, \mu) := \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda, \mu).$$

La fonction ψ est concave et semi-continue supérieurement (finie sur une partie convexe de $\mathbb{R}^m \times (\mathbb{R}^+)^p$). On appelle *problème dual de $(\mathcal{P})$*, et on le notera $(\mathcal{D})$, le problème de la maximisation de ψ sur $\mathbb{R}^m \times (\mathbb{R}^+)^p$. Cette fonction duale est également notée Θ parfois. Si $\overline{f}$ et $\overline{\psi}$ désignent les valeurs optimales dans $(\mathcal{P})$ et $(\mathcal{D})$ respectivement, on a toujours $\overline{f} \geqslant \overline{\psi}$ (la différence $\overline{f} - \overline{\psi}$ s'appelle le *saut de dualité*[(1)]).

Si $(\mathcal{P})$ est un problème de minimisation convexe, avec un ensemble S $(\neq \emptyset)$ de solutions et un ensemble M $(\neq \emptyset)$ de multiplicateurs de Lagrange, on a :

$\overline{f} = \overline{\psi}$ et les solutions du problème dual $(\mathcal{D})$ sont exactement les multiplicateurs de Lagrange du problème $(\mathcal{P})$.

Si on a accès à un $\left(\overline{\lambda}, \overline{\mu}\right) \in M$, comment en déduit-on les solutions de $(\mathcal{P})$?

Réponse : Ce sont les $\overline{x}$ qui

– minimisent $\mathcal{L}\left(\cdot, \overline{\lambda}, \overline{\mu}\right)$ sur $\mathbb{R}^n$;

– sont dans C *et* vérifient $\overline{\mu}_j g_j(\overline{x}) = 0$ pour tout $j = 1, \ldots, p$.

Si $\mathcal{L}\left(\cdot, \overline{\lambda}, \overline{\mu}\right)$ est minimisée en un seul point, comme c'est le cas lorsque f est strictement convexe par exemple, ce point est automatiquement solution de $(\mathcal{P})$.

Références. Section 4 du chapitre VII de [12].

***Exercice IV.1.** Soit $l : (x, y) \in \mathbb{R}^n \times \mathbb{R}^m \longmapsto l(x, y)$ une fonction différentiable sur $\mathbb{R}^n \times \mathbb{R}^m$, $X \subset \mathbb{R}^n$ et $Y \subset \mathbb{R}^m$ deux convexes fermés non vides.

1°) – Montrer l'équivalence des deux assertions suivantes :

(i) l est convexe-concave sur $X \times Y$;

(ii) $\begin{cases} \langle \nabla_x l(x_1, y_1) - \nabla_x l(x_2, y_2), x_1 - x_2 \rangle \geqslant \\ \qquad\qquad \langle \nabla_y l(x_1, y_1) - \nabla_y l(x_2, y_2), y_1 - y_2 \rangle \\ \text{pour tout } (x_1, y_1) \text{ et } (x_2, y_2) \text{ dans } X \times Y. \end{cases}$

(1) D'autres appellations sont parfois utilisées ; ainsi entendre parler de *fossé de dualité* avec l'accent québécois est charmant...

– *Exemple*. Soit $P \in \mathcal{S}_n(\mathbb{R})$, $Q \in \mathcal{S}_m(\mathbb{R})$, et $R \in \mathcal{M}_{m,n}(\mathbb{R})$.
Donner une condition suffisante portant sur P et Q, assurant que la fonction $l : (x,y) \longmapsto l(x,y) := \frac{1}{2}\langle Px, x\rangle + \frac{1}{2}\langle Qy, y\rangle + \langle Rx, y\rangle$ est convexe-concave sur $\mathbb{R}^n \times \mathbb{R}^m$.

2°) On suppose l convexe-concave sur $X \times Y$. Montrer l'équivalence des assertions suivantes concernant $(\overline{x}, \overline{y}) \in X \times Y$:

(j) $(\overline{x}, \overline{y})$ est un point-selle de l sur $X \times Y$;

(jj) $\begin{cases} \langle \nabla_x l(\overline{x}, \overline{y}), x - \overline{x}\rangle \geqslant 0 \text{ pour tout } x \in X \\ \text{et} \\ \langle \nabla_y l(\overline{x}, \overline{y}), y - \overline{y}\rangle \leqslant 0 \text{ pour tout } y \in Y \ ; \end{cases}$

(jjj) $\langle \nabla_x l(\overline{x}, \overline{y}), x - \overline{x}\rangle - \langle \nabla_y l(\overline{x}, \overline{y}), y - \overline{y}\rangle \geqslant 0$ pour tout $(x,y) \in X \times Y$.

Remarque : Les deux questions de l'exercice sont indépendantes.

Solution : 1°) – $[(i) \Rightarrow (ii)]$. Soient (x_1, y_1) et (x_2, y_2) dans $X \times Y$. Puisque $l(\cdot, y_2)$ est convexe sur X et $l(x_1, \cdot)$ concave sur Y, on a :

$$l(x_1, y_2) - l(x_2, y_2) \geqslant \langle \nabla_x l(x_2, y_2), x_1 - x_2\rangle$$
et
$$l(x_1, y_2) - l(x_1, y_1) \leqslant \langle \nabla_y l(x_1, y_1), y_2 - y_1\rangle .$$
Il s'ensuit :

$$(*) \qquad l(x_1, y_1) - l(x_2, y_2) \geqslant \langle \nabla_x l(x_2, y_2), x_1 - x_2\rangle + \langle \nabla_y l(x_1, y_1), y_1 - y_2\rangle .$$

Comme $l(\cdot, y_1)$ est aussi convexe sur X et $l(x_2, \cdot)$ concave sur Y on a, de la même manière,

$$(**) \qquad l(x_2, y_2) - l(x_1, y_1) \geqslant \langle \nabla_x l(x_1, y_1), x_2 - x_1\rangle + \langle \nabla_y l(x_2, y_2), y_2 - y_1\rangle .$$

L'addition membre à membre des inégalités $(*)$ et $(**)$ conduit à :

$$0 \geqslant \langle \nabla_x l(x_2, y_2) - \nabla_x l(x_1, y_1), x_1 - x_2\rangle + \langle \nabla_y l(x_1, y_1) - \nabla_y l(x_2, y_2), y_1 - y_2\rangle ,$$

ce qui est l'inégalité de (ii).

$[(ii) \Rightarrow (i)]$. Soit $y \in Y$. Faisons $y_1 = y_2 = y$ dans (ii) ; il vient :

$$\langle \nabla_x l(x_1, y) - \nabla_x l(x_2, y), x_1 - x_2\rangle \geqslant 0 \text{ pour tout } x_1 \text{ et } x_2 \text{ dans } X.$$

Ceci implique (en fait caractérise) la convexité de la fonction $l(\cdot, y)$ sur X.

Parallèlement, soit $x \in X$. En faisant $x_1 = x_2 = x$ dans (ii), il vient :

$$0 \geqslant \langle \nabla_y l(x, y_1) - \nabla_y l(x, y_2), y_1 - y_2 \rangle \text{ pour tout } y_1 \text{ et } y_2 \text{ dans } Y,$$

ce qui implique la concavité de $l(x, \cdot)$ sur Y.
– Dans l'exemple proposé, en tout $(x_i, y_i) \in \mathbb{R}^n \times \mathbb{R}^m$,

$$\nabla_x l(x_i, y_i) = Px_i + R^\top y_i \text{ et } \nabla_y l(x_i, y_i) = Qy_i + Rx_i.$$

La condition (ii) de convexe-concavité de l sur $\mathbb{R}^n \times \mathbb{R}^m$ se traduit par :

$$\begin{cases} \langle P(x_1 - x_2), x_1 - x_2 \rangle - \langle Q(y_1 - y_2), y_1 - y_2 \rangle \geqslant 0 \\ \text{pour tout } (x_1, y_1) \text{ et } (x_2, y_2) \text{ dans } \mathbb{R}^n \times \mathbb{R}^m. \end{cases}$$

Ceci est certainement vrai lorsque P est semi-définie positive et Q semi-définie négative.

2°) $[(j) \Leftrightarrow (jj)]$. $(\overline{x}, \overline{y}) \in X \times Y$ est un point-selle de l sur $X \times Y$ signifie :
$\overline{x}$ est un minimum sur X de la fonction convexe différentiable $l(\cdot, \overline{y})$
et
$\overline{y}$ est un maximum sur Y de la fonction concave différentiable $l(\overline{x}, \cdot)$.

Ce qu'expriment les inégalités de (jj) ne sont que les conditions nécessaires et suffisantes d'optimalité correspondantes.

$[(jj) \Leftrightarrow (jjj)]$. Immédiat.

***Exercice IV.2.** On note $\langle \cdot, \cdot \rangle$ le produit scalaire usuel dans $\mathbb{R}^n$ et $\| \cdot \|$ la norme euclidienne associée. Soit $a \in \mathbb{R}^n$ et $A \in \mathcal{S}_n(\mathbb{R})$ définie positive. On définit :

$$\begin{aligned} X &:= \{x \in \mathbb{R}^n \mid \langle A(x - a), x - a \rangle \leqslant 1\} \ ; \\ Y &:= \{x \in \mathbb{R}^n \mid \| y \| \leqslant 1\} \ ; \\ l &: (x, y) \in \mathbb{R}^n \times \mathbb{R}^n \longmapsto l(x, y) := \langle x, y \rangle . \end{aligned}$$

1°) a) Indiquer rapidement pourquoi l a des points-selles sur $X \times Y$.

b) Soit $(\overline{x}, \overline{y})$ un point-selle de l sur $X \times Y$.

– Montrer que $\overline{x}$ est la projection de l'origine sur X.

– Quelle est la valeur-selle de l sur $X \times Y$?

– En déduire $\overline{y}$.

2°) Étant donné $y \in Y$, on considère le problème de minimisation suivant

$$(\mathcal{P}_y) \begin{cases} \text{Minimiser } l(x,y) \\ x \in X. \end{cases}$$

Résoudre complètement $(\mathcal{P}_y)$.

3°) On suppose que $0 \notin X$. Déduire de ce qui précède :

- $\min_{x\in X} \| x \| = \max_{\|y\|\leqslant 1} \left[\langle a, y\rangle - \sqrt{\langle A^{-1}y, y\rangle} \right]$;
- le maximum dans l'expression de droite ci-dessus est atteint en $\overline{y} = \dfrac{\overline{x}}{\| \overline{x} \|}$, où $\overline{x}$ désigne la projection de 0 sur X.

Solution : 1°) a) X et Y sont des convexes compacts non vides de $\mathbb{R}^n$, la fonction l est convexe-concave sur $\mathbb{R}^n \times \mathbb{R}^n$; par conséquent il y a bien des points-selles de l sur $X \times Y$.

b) Soit $(\overline{x}, \overline{y})$ un point-selle de l sur $X \times Y$. Alors :

$$(*) \quad l(\overline{x}, \overline{y}) = \min_{x\in X} \left[\max_{y\in Y} l(x,y) \right] = \max_{y\in Y} \left[\min_{x\in X} l(x,y) \right] ;$$

$(**)$ $\overline{x}$ minimise la fonction $\varphi(x) := \max_{y\in Y} l(x,y)$ sur X
et
$\overline{y}$ maximise la fonction $\psi(y) := \min_{x\in X} l(x,y)$ sur Y.

Ici $\varphi(x) = \max_{\|y\|\leqslant 1} \langle x, y\rangle$ n'est autre que $\| x \|$; par suite $\overline{x}$ est le point de X minimisant la fonction $\| x \|$ sur X, c'est-à-dire la projection de l'origine sur X.

La valeur-selle de l sur $X \times Y$ est $\| \overline{x} \|$, c'est-à-dire la distance de l'origine à X.

Par suite $\langle \overline{x}, \overline{y}\rangle = \| \overline{x} \|$, ce qui conduit à :

$$\overline{y} = \frac{\overline{x}}{\| \overline{x} \|} \text{ si } \| \overline{x} \| \neq 0, \quad \text{c'est-à-dire lorsque } 0 \notin X;$$

$\overline{y}$ élément quelconque de Y lorsque $0 \in X$.

2°) L'ensemble-contrainte du problème de minimisation $(\mathcal{P}_y)$ est décrit sous la forme $g(x) \leqslant 0$ où $g : x \longmapsto g(x) := \langle A(x-a), x-a\rangle - 1$. La fonction g est convexe différentiable (quadratique même) et il existe x_0 tel que $g(x_0) < 0$ ($x_0 = a$ par exemple). La fonction-objectif dans $(\mathcal{P}_y)$ est linéaire, ce qui fait que, lorsque $y \neq 0$, les solutions $\overline{x}_y$ de $(\mathcal{P}_y)$ sont forcément sur la frontière de X (d'accord ?).

En somme, si $y \neq 0$, une solution $\overline{x}_y$ de $(\mathcal{P}_y)$ est *caractérisée* par :

$$\begin{cases} \langle A(\overline{x}_y - a), \overline{x}_y - a\rangle = 1, \\ \exists\ \overline{\mu} > 0 \text{ tel que } y + 2\overline{\mu}A(\overline{x}_y - a) = 0. \end{cases}$$

La deuxième relation ci-dessus donne $\overline{x}_y - a = -\frac{A^{-1}y}{2\overline{\mu}}$ qui, reportée dans la première, induit $\overline{\mu} = \frac{\sqrt{\langle A^{-1}y, y\rangle}}{2}$. Puis

$$\overline{x}_y = a - \frac{A^{-1}y}{\sqrt{\langle A^{-1}y, y\rangle}}.$$

Si $y = 0$, tout point de X est solution de $(\mathcal{P}_0)$.

3°) D'après ce qui précède,

$$\psi(y) := \min_{x\in X} l(x, y) = \langle \overline{x}_y, y\rangle = \langle a, y\rangle - \sqrt{\langle A^{-1}y, y\rangle},$$

ce qui avec $(*)$ donne l'égalité annoncée.

Les points $\overline{y} \in Y$ maximisant ψ sur Y sont exactement les $\overline{y}$ tels que $(\overline{x}, \overline{y})$ soit un point-selle de l sur $X \times Y$. Donc, dans le cas présent, $\overline{y} = \dfrac{\overline{x}}{\| \overline{x} \|}$.

***Exercice IV.3.** Soient $f_1, \ldots, f_m$ m fonctions convexes différentiables sur $\mathbb{R}^n$. On suppose que l'une d'entre elles est 0-coercive sur $\mathbb{R}^n$, c'est-à-dire vérifie : $\lim\limits_{\|x\|\to+\infty} f_i(x) = +\infty$.

On définit :

$$l : (x, y) \in \mathbb{R}^n \times \mathbb{R}^m \longmapsto l(x, y) := \sum_{i=1}^{m} y_i f_i(x)\ ;$$

$X := \mathbb{R}^n$;

Y est le simplexe-unité de $\mathbb{R}^m$, c'est-à-dire

$$\left\{ (y_1, \ldots, y_m) \in \mathbb{R}^m \mid y_i \geqslant 0 \ \text{ pour tout } i, \text{ et } \sum_{i=1}^{m} y_i = 1 \right\}.$$

1°) Indiquer rapidement pourquoi l a des points-selles sur $X \times Y$.

2°) Montrer que $(\overline{x},\overline{y})$ est un point-selle de l sur $X\times Y$ si et seulement si :

$$\begin{cases}\overline{x} \ \text{ minimise la fonction } \varphi := \max_{i=1,\dots,m} f_i \text{ sur } \mathbb{R}^n\\ \text{et}\\ \overline{y}\in Y \ \text{ avec } \ \overline{y}_i = 0 \ \text{ dès que } f_i(\overline{x}) < \varphi(\overline{x})\,.\end{cases}$$

Solution : 1°) l est convexe-concave sur $X\times Y$. Supposons par exemple f_1 0-coercive sur $\mathbb{R}^n$; en considérant $y_0 = (1,0,\dots,0)\in Y$, on a $l(x,y_0)\to+\infty$ quand $\|\ x\ \|\to+\infty$; par ailleurs Y est convexe compact. Il en résulte que l'ensemble des points-selles de l sur $X\times Y$ est un convexe compact non vide de $X\times Y$, de la forme $S\times T$. Reste à déterminer S et T précisément.

2°) Considérons $\varphi : x\in X \longmapsto \varphi(x) := \sup_{y\in Y} l(x,y)$, la première des fonctions intervenant dans la mini-maximisation de l sur $X\times Y$. En raison de la structure particulière de l et Y ici, on voit aisément que

$$\varphi(x) = \max_{i=1,\dots,m} f_i(x) \text{ pour tout } x\in X.$$

La valeur-selle $\overline{l}$ de l sur $X\times Y$ est donc

$$\overline{l} = \min_{x\in\mathbb{R}^n}\left[\max_{i=1,\dots,m} f_i(x)\right].$$

Les $\overline{x}$ d'un point-selle $(\overline{x},\overline{y})$ de l sur $X\times Y$ sont ceux pour lesquels $\varphi(\overline{x}) = \overline{l}$, c'est-à-dire ceux qui minimisent $\max_{i=1,\dots,m} f_i$ sur $\mathbb{R}^n$.

De même, les $\overline{y}$ d'un point-selle $(\overline{x},\overline{y})$ sont ceux de Y pour lesquels $l(\overline{x},\overline{y}) = \varphi(\overline{x})$. Sachant que $l(\overline{x},\overline{y}) = \sum_{i=1}^{m}\overline{y}_i f_i(\overline{x})$ et $\varphi(\overline{x}) = \max_{i=1,\dots,m} f_i(\overline{x})$, cela conduit aux $\overline{y}$ d'un « sous-simplexe-unité » de Y, à savoir les $\overline{y}\in Y$ dont les composantes $\overline{y}_i$ sont nulles lorsque $f_i(\overline{x}) < \max_{i=1,\dots,m} f_i(\overline{x})$.

***Exercice IV.4.** Étant donnés $a_1,\dots,a_n$ des réels différents de zéro, on considère l'ellipsoïde plein de $\mathbb{R}^n$ défini comme suit :

$$\mathcal{E} := \left\{u = (u_1,\dots,u_n)\in\mathbb{R}^n \mid \sum_{i=1}^{n}\left(\frac{u_i}{a_i}\right)^2 \leqslant 1\right\}.$$

Soient $x\notin\mathcal{E}$ fixé et $f : \mathbb{R}^n\to\mathbb{R}$ définie par :

$$\forall u\in\mathbb{R}^n,\ f(u) := \|\ u-x\ \|^2\,.$$

On cherche à résoudre le problème de minimisation suivant :

$$(\mathcal{P}) \begin{cases} \text{Minimiser } f(u) \\ u \in \mathcal{E}. \end{cases}$$

1°) Déterminer la fonction duale $\mu \in \mathbb{R}^+ \longmapsto \psi(\mu)$ associée au lagrangien usuel $(u, \mu) \in \mathbb{R}^n \times \mathbb{R} \longmapsto \mathcal{L}(u, \mu)$ dans $(\mathcal{P})$.

2°) Montrer que ψ est maximisée sur $\mathbb{R}^+$ en un point $\overline{\mu} > 0$ unique solution de l'équation (en μ) $\sum_{i=1}^{n} \frac{a_i^2 x_i^2}{(a_i^2+\mu)^2} = 1$.

Commentaire : Exercice à relier, évidemment, à l'Exercice 3.8.

Solution : 1°) Pour tout $\mu \geqslant 0$, la fonction de Lagrange $u \in \mathbb{R}^n \longmapsto \mathcal{L}(u, \mu) = \sum_{i=1}^{n} (u_i - x_i)^2 + \mu \left[\sum_{i=1}^{n} \left(\frac{u_i}{a_i} \right)^2 - 1 \right]$ est minimisée sur $\mathbb{R}^n$ en $\overline{u}(\mu)$ dont les composantes sont

$$\overline{u}(\mu)_i = \frac{a_i^2 x_i}{a_i^2 + \mu}, \ i = 1, \ldots, n.$$

La valeur optimale $\psi(\mu) := \inf_{u \in \mathbb{R}^n} \mathcal{L}(u, \mu)$ est

$$\psi(\mu) = \sum_{i=1}^{n} \frac{x_i^2 \mu}{a_i^2 + \mu} - \mu.$$

ψ est notre fonction duale.

ψ est strictement concave continue sur $\mathbb{R}^+$ et $\psi(\mu) \to -\infty$ quand $\mu \to +\infty$. Il existe donc un et un seul $\overline{\mu}$ maximisant ψ sur $\mathbb{R}^+$.

2°) ψ est dérivable sur (un intervalle ouvert contenant) $\mathbb{R}^+$, avec

$$\psi'(\mu) = \sum_{i=1}^{n} \frac{a_i^2 x_i^2}{(a_i^2 + \mu)^2} - 1 \text{ pour tout } \mu \geqslant 0.$$

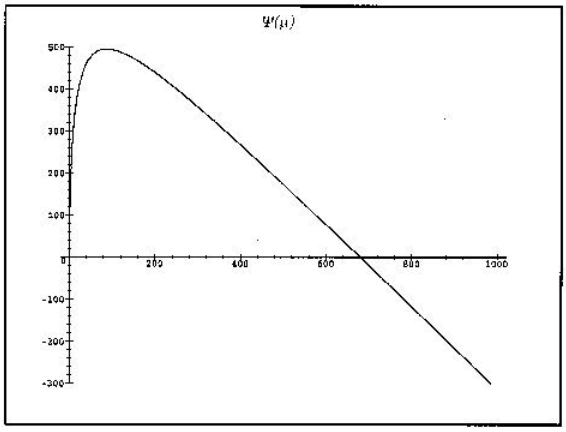
$\psi'(0) > 0$ car $x \notin \mathcal{E}$; ψ est donc maximisée au seul point $\overline{\mu} > 0$ annulant ψ'.

Exemple. $a_i = i$ pour $i = 1, \ldots, 7$; $\quad x_i = 10$ pour $i = 1, \ldots, 7$.
Valeur approchée de $\overline{\mu}$: $\quad 87,67846859$.
Valeur approchée de $\psi(\overline{\mu})$: $\quad 494,9439304$.

Commentaire : Soit $g : u \longmapsto g(u) := \sum_{i=1}^{n} \left(\frac{u_i}{a_i}\right)^2 - 1$, de sorte que $\mathcal{E} = \{u \in \mathbb{R}^n \mid g(u) \leqslant 0\}$. On vérifie sur le cas traité dans l'exercice que la seule manière d'assurer

$$\overline{u}(\mu) \in \mathcal{E} \text{ et } \mu g(\overline{u}(\mu)) = 0$$

(ce qui revient ici à : $\mu > 0$ et $g(\overline{u}(\mu)) = 0$), est d'avoir $\mu = \overline{\mu}$.



***Exercice IV.5.** Étant donnés $s \in \mathbb{R}^n \setminus \{0\}$ et $c \in \mathbb{R}^n$, on considère le problème de minimisation convexe suivant :

$$(\mathcal{P}) \quad \text{Minimiser } \left\{\frac{1}{2} \parallel x \parallel^2 - \langle c, x\rangle\right\} \text{ sous la contrainte } \langle s, x\rangle \leqslant 0.$$

1°) Déterminer la fonction duale $\mu \in \mathbb{R}^+ \longmapsto \psi(\mu)$ associée au lagrangien

$$\mathcal{L} : (x, \mu) \longmapsto \mathcal{L}(x, \mu) = \frac{1}{2} \parallel x \parallel^2 - \langle c, x\rangle + \mu \langle s, x\rangle .$$

2°) Résoudre le problème dual $(\mathcal{D})$ associé à $(\mathcal{P})$, c'est-à-dire celui de la maximisation de ψ sur $\mathbb{R}^+$ (on sera amené à considérer plusieurs cas, suivant le signe de $\langle c, s\rangle$).

3°) Utiliser le résultat précédent pour résoudre effectivement $(\mathcal{P})$.

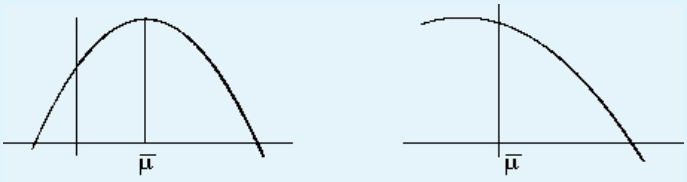
Solution : 1°) $\psi(\mu) = \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \mu)$ par définition.

$\mathcal{L}(\cdot, \mu)$ est une fonction quadratique strictement convexe sur $\mathbb{R}^n$. Son minimum est la solution de l'équation $\nabla_x \mathcal{L}(x, \mu) = 0$, soit

$$x - c + \mu s = 0, \text{ c'est-à-dire } x(\mu) = c - \mu s.$$

Donc $\psi(\mu) = \frac{1}{2} \parallel c - \mu s \parallel^2 - \langle c, c - \mu s\rangle + \mu \langle s, c - \mu s\rangle = -\frac{1}{2} \parallel c - \mu s \parallel^2$.

2°) ψ est une fonction *concave* que l'on doit maximiser sur $\mathbb{R}^+$ (c'est notre problème dual $(\mathcal{D})$). Ici, ψ est une fonction polynomiale du 2e degré en μ, avec $\psi'(\mu) = \langle c - \mu s, s\rangle = \langle c, s\rangle - \mu \parallel s \parallel^2$. Deux cas sont à distinguer :



- $\langle c, s\rangle \geqslant 0$, auquel cas la seule solution de $(\mathcal{D})$ est $\overline{\mu} = \frac{\langle c,s\rangle}{\|s\|^2}$. Alors

$$\psi(\overline{\mu}) = -\frac{1}{2} \parallel c - \overline{\mu} s \parallel^2 = \frac{(\langle c, s\rangle)^2}{2 \parallel s \parallel^2} - \frac{1}{2} \parallel c \parallel^2 .$$

- $\langle c, s\rangle < 0$, auquel cas le supremum de ψ sur $\mathbb{R}^+$ est atteint en $\overline{\mu} = 0$. Alors

$$\psi(\overline{\mu}) = -\frac{1}{2} \parallel c \parallel^2 .$$

3°) Le problème $(\mathcal{P})$ a une seule solution $\overline{x}$. La minimisation de la fonction strictement convexe $\mathcal{L}(\cdot, \overline{\mu})$ sur $\mathbb{R}^n$ donnera automatiquement la solution $\overline{x}$ cherchée.

- $\langle c, s\rangle < 0$. Ici $\mathcal{L}(x, \overline{\mu}) = \frac{1}{2} \parallel x \parallel^2 - \langle c, x\rangle$ et le minimum de $\mathcal{L}(\cdot, \overline{\mu})$ est atteint en $\overline{x} = c$. La valeur optimale est

$$\mathcal{L}(\overline{x}, \overline{\mu}) = f(\overline{x}) = \psi(\overline{\mu}) = -\frac{1}{2} \parallel c \parallel^2 .$$

- $\langle c, s\rangle \geqslant 0$. Alors $\mathcal{L}(x, \overline{\mu}) = \frac{1}{2} \parallel x \parallel^2 - \langle c, x\rangle + \frac{\langle c,s\rangle}{\|s\|^2} \langle s, x\rangle$. La fonction $\mathcal{L}(\cdot, \overline{\mu})$ est minimisée en

$$\overline{x} = c - \frac{\langle c, s\rangle}{\parallel s \parallel^2} s = c - \overline{\mu} s.$$

Finalement

$$f(\overline{x}) = \mathcal{L}(\overline{x}, \overline{\mu}) = \psi(\overline{\mu}) = \frac{(\langle c, s\rangle)^2}{2 \parallel s \parallel^2} - \frac{1}{2} \parallel c \parallel^2 .$$

***Exercice IV.6.** Étant donnés $r_1, \dots, r_n$ des réels tous strictement positifs, $A \in \mathcal{M}_{m,n}(\mathbb{R})$ et $c \in \mathbb{R}^m$, on considère le problème de minimisation suivant :

$$(\mathcal{P}) \begin{cases} \text{Minimiser } \displaystyle\sum_{i=1}^{n} x_i \ln\left(\frac{x_i}{r_i}\right) \\ \text{sous les contraintes } x \in \Lambda_n \text{ et } Ax \leqslant c. \end{cases}$$

Dans cette formulation, Λ_n désigne le simplexe-unité de $\mathbb{R}^n$, et $\alpha \ln(\alpha)$ est prise égale à 0 pour $\alpha = 0$.

Soient $(x, \mu) \in \Lambda_n \times (\mathbb{R}^+)^m \longmapsto \mathcal{L}(x, \mu) = \displaystyle\sum_{i=1}^{n} x_i \ln\left(\frac{x_i}{r_i}\right) + \langle \mu, Ax - c \rangle$ le lagrangien dans le problème $(\mathcal{P})$ et

$$\mu \in \left(\mathbb{R}^+\right)^m \longmapsto \psi(\mu) = \inf_{x \in \Lambda_n} \mathcal{L}(x, \mu)$$

la fonction duale associée.

Déterminer la forme explicite de ψ et formuler (le plus simplement possible) le problème dual $(\mathcal{D})$ de $(\mathcal{P})$.

Solution : Pour déterminer $\psi(\mu)$, $\mu \in (\mathbb{R}^+)^m$, on a à minimiser $\mathcal{L}(\cdot, \mu)$ sur Λ_n. En raison de la compacité de Λ_n, de la continuité de $\mathcal{L}(\cdot, \mu)$ sur Λ_n et de la stricte convexité de $\mathcal{L}(\cdot, \mu)$, il existe un et un seul $\overline{x}(\mu)$ minimisant $\mathcal{L}(\cdot, \mu)$ sur Λ_n. Un tel $\overline{x}(\mu)$, s'il est dans l'intérieur relatif de Λ_n (*i.e.* vérifie $\overline{x}(\mu)_i > 0$ pour tout i), est caractérisé par la propriété :

$$\exists\, \overline{\lambda} \in \mathbb{R}^n \text{ tel que } \nabla_x \mathcal{L}(\overline{x}(\mu), \mu) + \overline{\lambda}(1, \dots, 1)^\top = 0. \tag{4.1}$$

Cette condition se détaille en

$$\ln\left(\frac{\overline{x}(\mu)_i}{r_i}\right) + 1 + \left(A^\top \mu\right)_i + \overline{\lambda} = 0 \text{ pour tout } i = 1, \dots, n.$$

Sachant que $\displaystyle\sum_{i=1}^{n} \overline{x}(\mu)_i = 1$, cela conduit à

$$\overline{x}(\mu)_i = \frac{r_i \mathrm{e}^{-\left(A^\top \mu\right)_i}}{\displaystyle\sum_{i=1}^{n} r_i \mathrm{e}^{-\left(A^\top \mu\right)_i}}.$$

Cet $\overline{x}(\mu)$ vérifie (4.1) et est donc bien l'unique minimum de $\mathcal{L}(\cdot, \mu)$ sur Λ_n.

Par suite, on obtient par de simples calculs

$$\mathcal{L}\left(\overline{x}\left(\mu\right),\mu\right) = -\ln\left(\sum_{i=1}^{n} r_i \mathrm{e}^{-\left(A^{\top}\mu\right)_i + \langle \mu, c\rangle}\right) = \psi\left(\mu\right).$$

Une forme équivalente du problème dual $(\mathcal{D})$ de $(\mathcal{P})$ est donc :

$$(\mathcal{D}) \quad \begin{cases} \text{Minimiser} & \displaystyle\sum_{i=1}^{n} r_i \mathrm{e}^{-\left(A^{\top}\mu\right)_i + \langle \mu, c\rangle} \\ \text{sur le cône} & \left(\mathbb{R}^{+}\right)^{m}. \end{cases}$$

***Exercice IV.7.** *Problème de minimisation de J. Gibbs :*

Soit $f : x = (x_1, \ldots, x_n) \in \mathbb{R}^n \longmapsto f(x) := \displaystyle\sum_{i=1}^{n} f_i(x_i)$, où les n fonctions $f_i : \mathbb{R} \to \mathbb{R}$ sont supposées dérivables sur $\mathbb{R}$. On considère le problème d'optimisation suivant :

$$(\mathcal{P}) \quad \begin{cases} \text{Minimiser } f(x) \\ \text{sous les contraintes : } x_i \geqslant 0 \ \text{ pour tout } i = 1, \ldots, n \ \text{ et } \displaystyle\sum_{i=1}^{n} x_i = 1. \end{cases}$$

1°) a) Le problème $(\mathcal{P})$ a-t-il des solutions ?

b) Est-il licite de dire qu'un point solution de $(\mathcal{P})$ vérifie les conditions nécessaires de minimalité du premier ordre de Karush-Kuhn-Tucker ?

c) Montrer que si $\overline{x}$ est une solution de $(\mathcal{P})$, il existe alors un réel $\overline{\lambda}$ tel que :

$$f_i^{'}\left(\overline{x}_i\right) + \overline{\lambda} = 0 \ \text{ pour les } i \text{ tels que } \overline{x}_i > 0,$$
$$f_i^{'}\left(\overline{x}_i\right) + \overline{\lambda} \geqslant 0 \ \text{ pour les } i \text{ tels que } \overline{x}_i = 0.$$

d) Sous quelle hypothèse additionnelle les conditions précédentes caractérisent-elles les solutions de $(\mathcal{P})$?

2°) $a_1, \ldots, a_n, b_1, \ldots, b_n$ étant des réels (fixés) tous strictement positifs, on pose ici : $f_i(x_i) = a_i(\mathrm{e}^{-b_i x_i} - 1)$ pour tout $i = 1, \ldots, n$.

On considère le lagrangien usuel

$$\mathcal{L} : (x, \lambda) \in \left(\mathbb{R}^{+}\right)^n \times \mathbb{R} \longmapsto \mathcal{L}(x, \lambda) = \sum_{i=1}^{n} f_i(x_i) + \lambda\left(\sum_{i=1}^{n} x_i - 1\right)$$

et la fonction duale ψ associée, *i.e.* :

$$\forall \lambda \in \mathbb{R},\ \psi(\lambda) = \inf_{x \in (\mathbb{R}^+)^n} \mathcal{L}(x, \lambda)$$

(dualisation de la contrainte du type égalité seulement).

a) Vérifier que $\psi(\lambda) = -\infty$ si $\lambda < 0$, et que pour tout $\lambda > 0$ il existe un et un seul $\overline{x}(\lambda)$ minimisant $\mathcal{L}(\cdot, \lambda)$ sur $(\mathbb{R}^+)^n$.

En utilisant la caractérisation de $\overline{x}(\lambda)$ fournie par les conditions nécessaires et suffisantes d'optimalité, montrer :

$$\forall\ \lambda \geqslant 0, \quad \psi(\lambda) = -\sum_{i=1}^{n} a_i \left(1 - \frac{\lambda}{a_i b_i}\right)^+ + \lambda \left(\sum_{i=1}^{n} \frac{1}{b_i} \left[\ln\left(\frac{a_i b_i}{\lambda}\right)\right]^+ - 1\right).$$

b) En étudiant les propriétés de ψ sur $\mathbb{R}^+$ (notamment sa dérivabilité) montrer qu'il existe un et un seul $\overline{\lambda} > 0$ maximisant ψ sur $\mathbb{R}^+$.

Exprimer alors les coordonnées $\overline{x}_i$ de la solution de $(\mathcal{P})$ en fonction de $\overline{\lambda}$ et $a_1, \ldots, a_n, b_1, \ldots, b_n$.

Solution : 1°) a) L'ensemble-contrainte est compact et f est continue : le problème de minimisation $(\mathcal{P})$ a donc des solutions.

b) Les contraintes du type égalité ou inégalité sont exprimées à l'aide de fonctions affines ; il n'y a donc pas d'hypothèse de qualification des contraintes qu'il faudrait vérifier. Ainsi, toute solution de $(\mathcal{P})$ vérifie nécessairement les conditions de minimalité du 1er ordre de Karush-Kuhn-Tucker.

c) Puisque $\nabla f(x) = \left(f_1'(x_1), \ldots, f_n'(x_n)\right)^\top$ et que la fonction

$h : x = (x_1, \ldots, x_n) \longmapsto h(x) := \sum_{i=1}^{n} x_i - 1$ définissant la contrainte de type égalité a un gradient constant $(1, \ldots, 1)^\top$, nous avons le résultat suivant : si $\overline{x}$ est une solution de $(\mathcal{P})$, il existe $(\overline{\mu}_1, \ldots, \overline{\mu}_n) \in (\mathbb{R}^+)^n$ et $\overline{\lambda} \in \mathbb{R}$ tels que

$$(KKT) \quad \begin{cases} f_i'(\overline{x}_i) - \overline{\mu}_i + \overline{\lambda} = 0 \\ \overline{\mu}_i \overline{x}_i = 0 \end{cases} \text{ pour tout } i = 1, \ldots, n.$$

Si i est tel que $\overline{x}_i > 0$ (il y a nécessairement de tels i), $\overline{\mu}_i = 0$ de sorte que $f_i'(\overline{x}_i) + \overline{\lambda} = 0$.

Si i est tel que $\overline{x}_i = 0$,

$$f_i'(\overline{x}_i) + \overline{\lambda} = \overline{\mu}_i \geqslant 0.$$

D'où le résultat escompté.

d) Si les f_i sont convexes, f est convexe et les conditions de KKT précédentes caractérisent les solutions de $(\mathcal{P})$.

2°) La fonction f est strictement convexe. Il existe un et un seul $\overline{x}$ solution de $(\mathcal{P})$, caractérisé par :

$$\begin{cases} \overline{x}_i \geqslant 0 \ \text{ pour tout } \ i, \ \displaystyle\sum_{i=1}^{n} \overline{x}_i = 1 \ ; \\ \exists\, \overline{\lambda} \in \mathbb{R} \text{ tel que : } - \ a_i b_i \mathrm{e}^{-b_i \overline{x}_i} + \overline{\lambda} = 0 \ \text{ pour tout } i \text{ tel que } \overline{x}_i > 0, \\ - \ a_i b_i + \overline{\lambda} \geqslant 0 \text{ pour tout } i \text{ tel que } \overline{x}_i = 0. \end{cases}$$

On considère $(x, \lambda) \in (\mathbb{R}^+)^n \times \mathbb{R} \longmapsto \mathcal{L}(x, \lambda) = f(x) + \lambda\ h(x) = \displaystyle\sum_{i=1}^{n} a_i \left(\mathrm{e}^{-b_i x_i} - 1\right) + \lambda \left(\sum_{i=1}^{n} x_i - 1\right)$.

a) $\mathcal{L}(\cdot, \lambda)$ a une expression « séparée » en les coordonnées x_i, ce qui fait que sa minimisation sous contraintes « séparées » ($x_i \geqslant 0$ pour tout i) s'en trouve facilitée.

Supposons $\lambda < 0$. Puisque x_i peut être pris positif arbitrairement grand, il vient que $\psi(\lambda) = -\infty$.

Si $\lambda = 0$, on a $\mathcal{L}(x, 0) = f(x)$ et $\psi(0) = -\displaystyle\sum_{i=1}^{n} a_i$.

Supposons $\lambda > 0$. La fonction $\mathcal{L}(\cdot, \lambda)$ est continue, 0-coercive, strictement convexe sur $(\mathbb{R}^+)^n$: il existe donc un et un seul point $\overline{x}(\lambda) \in (\mathbb{R}^+)^n$ minimisant $\mathcal{L}(\cdot, \lambda)$ sur $(\mathbb{R}^+)^n$. Ce point $\overline{x}(\lambda)$ est caractérisé par le fait que

$$-\nabla_x \mathcal{L}(\overline{x}(\lambda), \lambda) \in N_{(\mathbb{R}^+)^n}(\overline{x}(\lambda)),$$

c'est-à-dire :

$$\begin{cases} - \ a_i b_i \mathrm{e}^{-b_i \overline{x}(\lambda)_i} + \lambda = 0 \text{ pour tout } i \text{ tel que } \overline{x}(\lambda)_i > 0, \\ \ a_i b_i - \lambda \leqslant 0 \text{ pour tout } i \text{ tel que } \overline{x}(\lambda)_i = 0. \end{cases}$$

On constate que $(\overline{x}(\lambda)_i > 0)$ équivaut à $(\lambda < a_i b_i)$. En clair, $\overline{x}(\lambda)$ est le vecteur de $\mathbb{R}^n$ de composantes $\frac{1}{b_i}\left[\ln\left(\frac{a_i b_i}{\lambda}\right)\right]^+$, $i = 1, \ldots, n$.

Par suite

$$\begin{aligned} \psi(\lambda) &= \mathcal{L}(\overline{x}(\lambda), \lambda) \\ &= \sum_{\{i \mid \overline{x}(\lambda)_i > 0\}} a_i \left(\frac{\lambda}{a_i b_i} - 1\right) + \lambda \left(\sum_{\{i \mid \overline{x}(\lambda)_i > 0\}} \frac{1}{b_i} \ln\left(\frac{a_i b_i}{\lambda}\right) - 1\right) \\ &= -\sum_{i=1}^{n} a_i \left(1 - \frac{\lambda}{a_i b_i}\right)^+ + \lambda \left(\sum_{i=1}^{n} \frac{1}{b_i} \left[\ln\left(\frac{a_i b_i}{\lambda}\right)\right]^+ - 1\right). \end{aligned}$$

La fonction ψ, on le sait, est concave semi-continue supérieurement sur $\mathbb{R}$. Le problème dual $(\mathcal{D})$ de $(\mathcal{P})$ consiste ici à maximiser $\psi(\lambda)$, $\lambda \in \mathbb{R}^+$.

b) Supposons, pour alléger l'écriture et les notations, que nous avons la disposition suivante : $a_1b_1 \leqslant a_2b_2 \leqslant \ldots \leqslant a_nb_n$. Alors :

– Si $\lambda \geqslant a_nb_n$, $\psi(\lambda) = -\lambda$;

– Si $\lambda \leqslant a_1b_1$, $\psi(\lambda) = -\sum_{i=1}^{n} a_i\left(1 - \frac{\lambda}{a_ib_i}\right) + \lambda\left(\sum_{i=1}^{n} \frac{1}{b_i}\ln\left(\frac{a_ib_i}{\lambda}\right) - 1\right)$;

– Si $\lambda \in [a_kb_k, a_{k+1}b_{k+1}[$, $\psi(\lambda)$ a l'expression suivante :

$$\psi(\lambda) = \sum_{i=k+1}^{n}\left(\frac{\lambda}{b_i} - a_i\right) + \lambda\left(\sum_{i=k+1}^{n}\frac{1}{b_i}ln\left(\frac{a_ib_i}{\lambda}\right) - 1\right).$$

Ainsi ψ est dérivable sur $]a_kb_k, a_{k+1}b_{k+1}[$, de dérivée

$$\psi'(\lambda) = \sum_{i=k+1}^{n}\frac{1}{b_i}\ln\left(\frac{a_ib_i}{\lambda}\right) - 1.$$

En observant que $\lim_{\lambda\to(a_kb_k)^-}\psi'(\lambda) = \lim_{\lambda\to(a_kb_k)^+}\psi'(\lambda)$, on constate que ψ est aussi dérivable en a_kb_k. En fait ψ est dérivable sur $\mathbb{R}_*^+$ avec

$$\psi'(\lambda) = \sum_{i=1}^{n}\frac{1}{b_i}\left[\ln\left(\frac{a_ib_i}{\lambda}\right)\right]^+ - 1 \text{ pour tout } \lambda > 0.$$

On note (sans surprise) que $\psi'(\lambda)$ est égal à $h(\overline{x}(\lambda)) = \frac{d\mathcal{L}}{d\lambda}\,|_{x\,=\,\overline{x}(\lambda)}$.

La fonction ψ' est continue sur $\mathbb{R}_*^+$, strictement décroissante sur $[0, a_nb_n]$, et de plus :

$$\lim_{\lambda\to 0^+}\psi'(\lambda) = +\infty ;$$
$$\psi'(\lambda) = -1 \text{ pour } \lambda \geqslant a_nb_n.$$

L'équation $\psi'(\lambda) = 0$ a donc une seule solution $\overline{\lambda}\,(>0)$, c'est évidemment le point maximisant ψ sur $\mathbb{R}^+$.

La solution de $(\mathcal{P})$ est donc $\overline{x} = \overline{x}\left(\overline{\lambda}\right)$, c'est-à-dire :

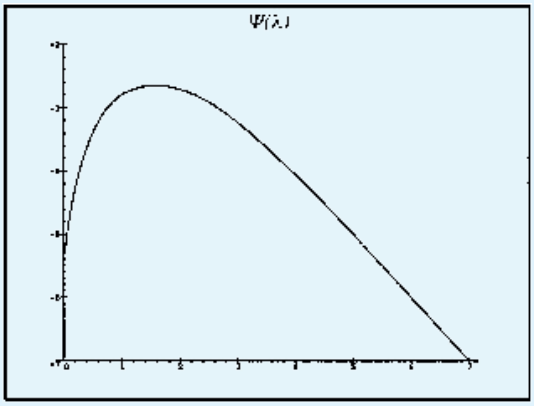
$$\begin{cases} \overline{x}_i = \frac{1}{b_i}\ln\left(\dfrac{a_ib_i}{\overline{\lambda}}\right) \text{ pour les } i \text{ tels que } a_ib_i > \overline{\lambda}, \\ \overline{x}_i = 0 \text{ pour les } i \text{ tels que } a_ib_i \leqslant \overline{\lambda}. \end{cases}$$

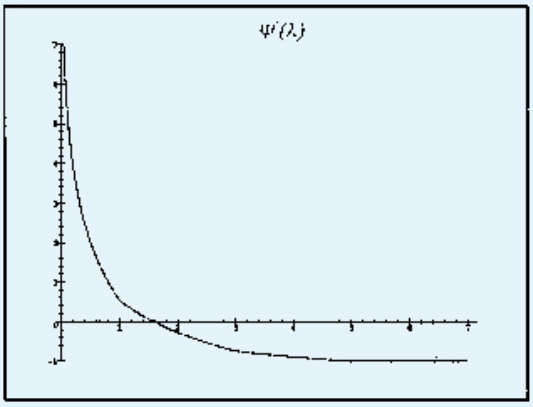
Exemple. $a_1 = 1,\ a_2 = 2,\ a_3 = 5/2$
$b_1 = 1,\ b_2 = 3/2,\ b_3 = 2.$

Voir ci-après le tracé de la fonction duale ψ :

$$\forall \lambda \geqslant 0,\ \psi(\lambda) = -(1-\lambda)^+ - 2\left(1-\frac{\lambda}{3}\right)^+ - \frac{5}{2}\left(1-\frac{\lambda}{5}\right)^+ + \lambda\left(\left[\ln\frac{1}{\lambda}\right]^+ + \frac{2}{3}\left[\ln\frac{3}{\lambda}\right]^+ + \frac{1}{2}\left[\ln\frac{5}{\lambda}\right]^+ - 1\right) ;$$

$$\forall \lambda > 0,\ \psi'(\lambda) = \left[\ln\frac{1}{\lambda}\right]^+ + \frac{2}{3}\left[\ln\frac{3}{\lambda}\right]^+ + \frac{1}{2}\left[\ln\frac{5}{\lambda}\right]^+ - 1.$$





Valeur approchée de $\overline{\lambda}$: $1,58469934$.
Valeur approchée de $\psi(\overline{\lambda})$: $-2,65118410$.

***Exercice IV.8.** $a_1, \ldots, a_n, b_1, \ldots, b_n$ étant des réels (fixés) tous strictement positifs, on pose $f : x = (x_1, \ldots, x_n) \in \mathbb{R}^n \longmapsto f(x) := \sum_{i=1}^{n} \frac{a_i}{1+x_i}$ et

$$C := \left\{ x = (x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_i \geqslant 0 \ \text{ pour tout } i = 1, \ldots, n \text{ et } \sum_{i=1}^{n} b_i x_i = 1 \right\}.$$

On considère le problème d'optimisation suivant :

$$(\mathcal{P}) \quad \begin{cases} \text{Minimiser } f(x) \\ x \in C. \end{cases}$$

1°) a) Quelles propriétés de $(\mathcal{P})$, utiles pour sa résolution, peut-on observer ? Quelles conséquences peut-on en tirer quant à l'existence et l'unicité des solutions de $(\mathcal{P})$?

b) Montrer que $\overline{x} = (\overline{x}_1, \ldots, \overline{x}_n) \in C$ est solution de $(\mathcal{P})$ si et seulement si on a la propriété suivante :

Il existe un réel $\overline{\lambda}$ tel que

$$-\frac{a_i}{(1+\overline{x}_i)^2} + \overline{\lambda} b_i = 0 \ \text{ pour tout } i \ \text{ tel que } \overline{x}_i > 0,$$

$$-a_i + \overline{\lambda} b_i \geqslant 0 \ \text{ pour tout } i \text{ tel que } \overline{x}_i = 0.$$

2°) On considère le lagrangien usuel

$$\mathcal{L} : (x, \lambda) \in \left(\mathbb{R}^+\right)^n \times \mathbb{R} \longmapsto \mathcal{L}(x, \lambda) = f(x) + \lambda \left(\sum_{i=1}^{n} b_i x_i - 1 \right)$$

et la fonction duale ψ associée, *i.e.*

$$\forall \lambda \in \mathbb{R}, \ \psi(\lambda) = \inf_{x \in (\mathbb{R}^+)^n} \mathcal{L}(x, \lambda).$$

a) Vérifier que $\psi(\lambda) = -\infty$ si $\lambda < 0$, et que pour tout $\lambda > 0$ il existe un et un seul $\overline{x}(\lambda)$ minimisant $\mathcal{L}(\cdot, \lambda)$ sur $(\mathbb{R}^+)^n$.

En utilisant la caractérisation de $\overline{x}(\lambda)$ fournie par les conditions nécessaires et suffisantes d'optimalité, montrer que :

$$\forall \lambda \geqslant 0, \ \psi(\lambda) = \sum_{i=1}^{n} a_i \left\{ 1 + \left(\sqrt{\frac{b_i \lambda}{a_i}} - 1 \right) \left(1 - \sqrt{\frac{b_i \lambda}{a_i}} \right)^+ \right\} - \lambda.$$

b) En étudiant les propriétés de ψ sur $\mathbb{R}^+$ (notamment sa dérivabilité) montrer qu'il existe un et un seul $\overline{\lambda} > 0$ maximisant ψ sur $\mathbb{R}^+$.

Exprimer alors les coordonnées $\overline{x}_i$ de la solution $(\mathcal{P})$ en fonction de $\overline{\lambda}$ et $a_1, \ldots, a_n, b_1, \ldots, b_n$.

Solution : 1°) a) L'ensemble-contrainte est un polyèdre convexe compact non vide (le caractère borné de C se déduit par exemple de l'observation suivante : si $x = (x_1, \ldots, x_n) \in C$, alors $0 \leqslant x_i \leqslant 1/(\min_i b_i)$), f est continue sur C : le problème de minimisation $(\mathcal{P})$ a donc des solutions.

La fonction f est strictement convexe sur l'ouvert convexe

$$\Omega := (]-1, +\infty[)^n \text{ de } \mathbb{R}^n \text{ (car } \nabla^2 f(x) = \operatorname{diag}(\ldots, \frac{2a_i}{(1+x_i)^3}, \ldots)$$

est définie positive en tout $x = (x_1, \ldots, x_n)$ de Ω).

Par conséquent, le problème $(\mathcal{P})$ a une et une seule solution.

$(\mathcal{P})$ est un problème de minimisation différentiable et convexe, les contraintes du type égalité ou inégalité sont exprimées à l'aide de fonctions affines ; par suite, la solution de $(\mathcal{P})$ est *caractérisée* par les conditions du 1er ordre de Karush-Kuhn-Tucker.

b) $\overline{x} = (\overline{x}_1, \ldots, \overline{x}_n) \in C$ est solution de $(\mathcal{P})$ si et seulement si les conditions de KKT suivantes sont vérifiées : il existe $(\overline{\mu}_1, \ldots, \overline{\mu}_n) \in (\mathbb{R}^+)^n$ et $\overline{\lambda} \in \mathbb{R}$ tels que

$$(KKT) \begin{cases} -\dfrac{a_i}{(1+\overline{x}_i)^2} - \overline{\mu}_i + \overline{\lambda} b_i = 0 \\ \overline{\mu}_i \overline{x}_i = 0 \end{cases} \qquad \text{pour tout } i = 1, \ldots, n.$$

Si i est tel que $\overline{x}_i > 0$ (il y a nécessairement de tels i), $\overline{\mu}_i = 0$ de sorte que $-\frac{a_i}{(1+\overline{x}_i)^2} + \overline{\lambda} b_i = 0$.

Si i est tel que $\overline{x}_i = 0$, on a $-a_i + \overline{\lambda} b_i = \overline{\mu}_i \geqslant 0$.

D'où la caractérisation annoncée.

2°) On considère

$$(x, \lambda) \in (\mathbb{R}^+)^n \times \mathbb{R} \longmapsto \mathcal{L}(x, \lambda) = \sum_{i=1}^n \tfrac{a_i}{1+x_i} + \lambda \left(\sum_{i=1}^n b_i x_i - 1 \right).$$

a) $\mathcal{L}(\cdot, \lambda)$ a une expression « séparée » en les coordonnées x_i, ce qui fait que sa minimisation sous contraintes « séparées » ($x_i \geqslant 0$ pour tout i) s'en trouve facilitée.

Supposons $\lambda < 0$. Puisque x_i peut être pris positif arbitrairement grand et que $b_i > 0$, il s'ensuit $\psi(\lambda) = -\infty$.

Si $\lambda = 0$, $\mathcal{L}(x, 0) = f(x)$ et $\psi(0) = 0$.

Supposons $\lambda > 0$. La fonction $\mathcal{L}(\cdot, \lambda)$ est continue, 0-coercive, strictement convexe sur $(\mathbb{R}^+)^n$: il existe donc un et un seul point $\overline{x}(\lambda) \in (\mathbb{R}^+)^n$ minimisant $\mathcal{L}(\cdot, \lambda)$ sur $(\mathbb{R}^+)^n$. Ce point $\overline{x}(\lambda)$ est caractérisé par la propriété

$$-\nabla_x \mathcal{L}(\overline{x}(\lambda), \lambda) \in N_{(\mathbb{R}^+)^n}(\overline{x}(\lambda)),$$

c'est-à-dire

$$\begin{cases} -\dfrac{a_i}{(1+\overline{x}(\lambda)_i)^2} + b_i\lambda = 0 \text{ pour tout } i \text{ tel que } \overline{x}(\lambda)_i > 0, \\ a_i - b_i\lambda \leqslant 0 \text{ pour tout } i \text{ tel que } \overline{x}(\lambda)_i = 0. \end{cases}$$

On constate que $(\overline{x}(\lambda)_i > 0)$ équivaut à $\left(\lambda < \frac{a_i}{b_i}\right)$. Ainsi $\overline{x}(\lambda)$ est le vecteur de $\mathbb{R}^n$ de composantes $\left(\sqrt{\frac{a_i}{b_i\lambda}} - 1\right)^+$, $i = 1, \ldots, n$.

Par suite

$$\psi(\lambda) = \mathcal{L}(\overline{x}(\lambda), \lambda) = \sum_{\{i \mid \overline{x}(\lambda)_i = 0\}} a_i + \sum_{\{i \mid \overline{x}(\lambda)_i > 0\}} \sqrt{a_i b_i \lambda} + \lambda \left(\sum_{\{i \mid \overline{x}(\lambda)_i > 0\}} b_i \left(\sqrt{\frac{a_i}{b_i\lambda}} - 1 \right) - 1 \right).$$

Or

$$\sum_{\{i \mid \overline{x}(\lambda)_i = 0\}} a_i + \sum_{\{i \mid \overline{x}(\lambda)_i > 0\}} \sqrt{a_i b_i \lambda} = \sum_{i=1}^n a_i \left[1 - \left(1 - \sqrt{\frac{b_i\lambda}{a_i}} \right)^+ \right]$$

et

$$\sum_{\{i \mid \overline{x}(\lambda)_i > 0\}} b_i \left(\sqrt{\frac{a_i}{b_i\lambda}} - 1 \right) = \sum_{i=1}^n b_i \left(\sqrt{\frac{a_i}{b_i\lambda}} - 1 \right)^+,$$

de sorte que

$$\begin{aligned} \psi(\lambda) &= \sum_{i=1}^n \left\{ a_i \left[1 - \left(1 - \sqrt{\frac{b_i\lambda}{a_i}} \right)^+ \right] + \lambda b_i \left(\sqrt{\frac{a_i}{b_i\lambda}} - 1 \right)^+ \right\} - \lambda \\ &= \sum_{i=1}^n a_i \left\{ 1 + \left(\sqrt{\frac{b_i\lambda}{a_i}} - 1 \right) \left(1 - \sqrt{\frac{b_i\lambda}{a_i}} \right)^+ \right\} - \lambda. \end{aligned}$$

b) Quitte à renuméroter les indices, on peut supposer que nous avons la disposition suivante : $\frac{a_1}{b_1} \leqslant \frac{a_2}{b_2} \leqslant \ldots \leqslant \frac{a_n}{b_n}$. Ainsi :

– Si $\lambda \geqslant \frac{a_n}{b_n}$, $\psi(\lambda) = \sum_{i=1}^n a_i - \lambda$;

– Si $\lambda < \frac{a_1}{b_1}$, $\psi(\lambda) = \sum_{i=1}^n \sqrt{a_i b_i \lambda} \left(2 - \sqrt{\frac{\lambda b_i}{a_i}} \right) - \lambda$;

– Si $\lambda \in \left[\frac{a_k}{b_k},\ \frac{a_{k+1}}{b_{k+1}}\right[$, $\psi(\lambda)$ a pour expression

$$\sum_{i=1}^{k} a_i + \sum_{i=k+1}^{n} \sqrt{a_i b_i \lambda}\left(2 - \sqrt{\frac{\lambda b_i}{a_i}}\right) - \lambda.$$

La fonction ψ est finie sur $\mathbb{R}^+$ et continue sur $\mathbb{R}_*^+$ (puisque concave sur $\mathbb{R}^+$).
Étudions sa dérivabilité sur $\mathbb{R}_*^+$.
Il est clair que ψ est dérivable sur $\left]\frac{a_k}{b_k}\ ,\ \frac{a_{k+1}}{b_{k+1}}\right[$, de dérivée

$$\psi'(\lambda) = \sum_{i=k+1}^{n}\left(\sqrt{\frac{a_i b_i}{\lambda}} - b_i\right) - 1.$$

Notons que $\lim\limits_{\lambda \to \left(\frac{a_k}{b_k}\right)^-} \psi'(\lambda) = \lim\limits_{\lambda \to \left(\frac{a_k}{b_k}\right)^+} \psi'(\lambda)$; donc ψ est dérivable sur $\mathbb{R}_*^+$.
D'ailleurs la formule générale donnant ψ' sur $\mathbb{R}_*^+$ est :

$$\psi'(\lambda) = \sum_{i=1}^{n} b_i \left(\sqrt{\frac{a_i}{b_i \lambda}} - 1\right)^+ - 1 \text{ pour tout } \lambda > 0.$$

Observons – et ce n'est pas surprenant – que $\psi'(\lambda)$ est la dérivée par rapport à λ de $\lambda \longmapsto \mathcal{L}(x,\lambda) = f(x) + \lambda\left(\sum\limits_{i=1}^{n} b_i x_i - 1\right)$ prise au point optimal $\overline{x}(\lambda)$.

La fonction ψ' est continue sur $\mathbb{R}_*^+$, strictement décroissante sur $\left[0, \frac{a_n}{b_n}\right]$, et de plus :

$$\lim_{\lambda \to 0^+} \psi'(\lambda) = +\infty\ ;$$
$$\psi'(\lambda) = -1 \text{ pour } \lambda \geqslant \frac{a_n}{b_n}\cdot$$

L'équation $\psi'(\lambda) = 0$ a donc une et une seule solution $\overline{\lambda}\,(>0)$, et c'est l'unique point maximisant ψ sur $\mathbb{R}^+$.
La solution de $(\mathcal{P})$ est donc $\overline{x} = \overline{x}\left(\overline{\lambda}\right)$, c'est-à-dire :

$$\begin{cases} \overline{x}_i = \sqrt{\dfrac{a_i}{\overline{\lambda} b_i}} - 1 \text{ pour tous les } i \text{ tels que } \dfrac{a_i}{b_i} > \overline{\lambda}, \\ \overline{x}_i = 0 \text{ pour tous les } i \text{ tels que } \dfrac{a_i}{b_i} \leqslant \overline{\lambda}. \end{cases}$$

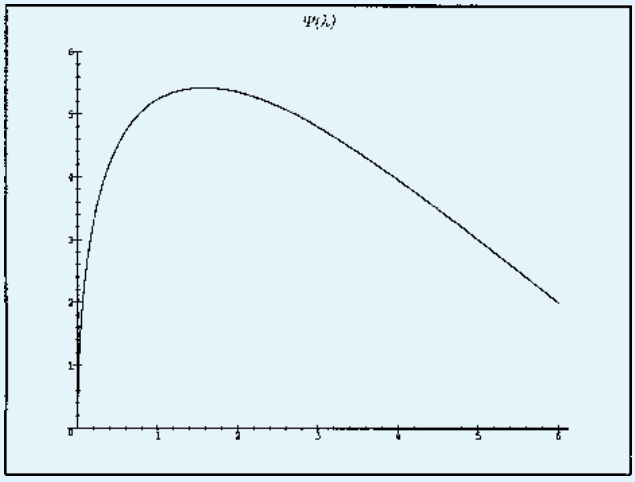
Exemple. $a_1 = 1,\ a_2 = 3,\ a_3 = 4$
$b_1 = 1,\ b_2 = 1,\ b_3 = 4/5.$

Voir ci-dessous le tracé de la fonction duale ψ :

$$\forall \lambda \geqslant 0,\ \psi(\lambda) = 1 + \left(\sqrt{\lambda} - 1\right)\left(1 - \sqrt{\lambda}\right)^{+}$$
$$+3\left[1 + \left(\sqrt{\frac{\lambda}{3}} - 1\right)\left(1 - \sqrt{\frac{\lambda}{3}}\right)^{+}\right] + 4\left[1 + \left(\sqrt{\frac{\lambda}{5}} - 1\right)\left(1 - \sqrt{\frac{\lambda}{5}}\right)^{+}\right] - \lambda\ ;$$

$$\forall \lambda > 0,\ \psi'(\lambda) = \left(\sqrt{\frac{1}{\lambda}} - 1\right)^{+} + \left(\sqrt{\frac{3}{\lambda}} - 1\right)^{+} + \frac{4}{5}\left(\sqrt{\frac{5}{\lambda}} - 1\right)^{+} - 1.$$

Valeur approchée de $\overline{\lambda}$: $1,58122109$.
Valeur approchée de $\psi(\overline{\lambda})$: $5,42941905$.



***Exercice IV.9.** Dans $\mathbb{R}^n$ muni du produit scalaire usuel $\langle \cdot, \cdot \rangle$, on considère le problème de minimisation suivant :

$$(\mathcal{P}) \quad \begin{cases} \text{Minimiser} \quad f(x) := \dfrac{1}{2}\langle Mx, x\rangle + \langle q, x\rangle \\ \text{sous la contrainte} \quad Ax + b \leqslant 0, \end{cases}$$

où $M \in \mathcal{M}_n(\mathbb{R})$ est symétrique définie positive, $q \in \mathbb{R}^n$, $A \in \mathcal{M}_{m,n}(\mathbb{R})$ et $b \in \mathbb{R}^m$.

1°) Déterminer la fonction duale Θ associée au lagrangien

$$\mathcal{L} : (x, \mu) \longmapsto \mathcal{L}(x, \mu) := \frac{1}{2}\langle Mx, x\rangle + \langle q, x\rangle + \langle \mu, Ax + b\rangle .$$

2°) (*i*) Formuler le problème dual $(\mathcal{D})$ de $(\mathcal{P})$.

(*ii*) Comment caractériser une solution $\overline{\mu}$ de $(\mathcal{D})$?

(*iii*) Donner une condition suffisante pour que $(\mathcal{D})$ ait une solution unique.

(*iv*) Comment une solution $\overline{\mu}$ de $(\mathcal{D})$ permettrait-elle de retrouver la solution de $(\mathcal{P})$?

Solution : f est une fonction (quadratique) fortement convexe, l'ensemble $C := \{x \in \mathbb{R}^n \mid Ax + b \leqslant 0\}$ défini par la conjonction des contraintes $\langle a_j, x\rangle + b_j \leqslant 0$ pour $j = 1, \dots, m$ (inégalités définies par des fonctions affines) est un convexe fermé. Donc, si $C \neq \emptyset$, $(\mathcal{P})$ a une et une seule solution $\overline{x}$. Les contraintes étant qualifiées (puisque les fonctions les définissant sont affines), $\overline{x}$ est *caractérisée* par :

$$\exists\, \overline{\mu} = (\overline{\mu}_1, \dots, \overline{\mu}_m) \in \left(\mathbb{R}^+\right)^m \text{ tel que } \begin{cases} M\overline{x} + q + A^\top \overline{\mu} = 0, & (1) \\ \langle \overline{\mu}, A\overline{x} + b\rangle = 0 & (2) \\ \left(\overline{\mu}_j = 0 \text{ si } \langle a_j, \overline{x}\rangle + b_j < 0\right). \end{cases}$$

1°) $\Theta(\mu)$ est par définition $\inf_{x\in\mathbb{R}^n} \mathcal{L}(x, \mu)$.

À μ fixé, le minimum de $\mathcal{L}(\cdot, \mu)$ sur $\mathbb{R}^n$ est atteint en $\overline{x}(\mu) = -M^{-1}\left(q + A^\top \mu\right)$.

D'où :

$$\begin{aligned} \Theta(\mu) &= -\frac{1}{2}\left\langle M^{-1}\left(A^\top\mu + q\right), A^\top\mu + q\right\rangle + \langle b, \mu\rangle \\ &= -\frac{1}{2}\left\langle \left(AM^{-1}A^\top\right)\mu, \mu\right\rangle + \langle b - AM^{-1}q, \mu\rangle - \frac{1}{2}\langle M^{-1}q, q\rangle. \end{aligned}$$

2°) (*i*) Le problème dual $(\mathcal{D})$ de $(\mathcal{P})$ est :

$$(\mathcal{D}) \quad \begin{cases} \text{Maximiser } \Theta(\mu) \\ \text{sous la contrainte } \mu \in (\mathbb{R}^+)^m. \end{cases}$$

On sait que ce problème de maximisation d'une fonction quadratique concave sur $(\mathbb{R}^+)^m$ *a* des solutions (ce qui ne se voit pas directement puisque $AM^{-1}A^\top$ est certes semi-définie positive, mais pas nécessairement définie positive).

(*ii*) Une solution $\overline{\mu}$ de $(\mathcal{D})$ est caractérisée par :

$$\overline{\mu} \in \left(\mathbb{R}^+\right)^m \text{ et } \nabla\Theta(\overline{\mu}) \text{ est normal à } \left(\mathbb{R}^+\right)^m \text{ en } \overline{\mu} \text{ ;}$$

soit encore

$$\overline{\mu} \in \left(\mathbb{R}^+\right)^m, \left(AM^{-1}A^\top\right)\overline{\mu} + AM^{-1}q - b \in \left(\mathbb{R}^+\right)^m$$

et

$$\left\langle \left(AM^{-1}A^\top\right)\overline{\mu} + AM^{-1}q - b, \overline{\mu} \right\rangle = 0.$$

(*iii*) Si $m \leqslant n$ et A est *surjective*, alors $A^\top$ est injective de sorte que $AM^{-1}A^\top$ est définie positive. En effet

$$\left\langle \left(AM^{-1}A^\top\right) y, y \right\rangle = \left\langle M^{-1}\left(A^\top y\right), \left(A^\top y\right) \right\rangle \geqslant 0$$

et

$$\left(\left\langle \left(AM^{-1}A^\top\right) y, y \right\rangle = 0 \right) \Leftrightarrow \left(A^\top y = 0\right) \Leftrightarrow (y = 0).$$

Alors, $(\mathcal{D})$ a une et une seule solution $\overline{\mu}$.

Cela se voit aussi à partir de (1) : les solutions $\overline{\mu}$ de $(\mathcal{D})$ sont les multiplicateurs de $(\mathcal{P})$, ils vérifient (1) et (2) ; de (1) on tire

$$A^\top \overline{\mu} = -M\overline{x} - q.$$

Un tel $\overline{\mu}$ – dont on est assuré de l'existence – est unique.

(*iv*) Connaissant une solution $\overline{\mu}$ de $(\mathcal{D})$, *la solution de* $(\mathcal{P})$ *est le point minimisant* $\mathcal{L}(\cdot, \overline{\mu})$ *sur* $\mathbb{R}^n$, *i.e.* $\overline{x}(\overline{\mu})$. Il n'y a pas lieu de distinguer les minima de $\mathcal{L}(\cdot, \overline{\mu})$ qui sont admissibles et ceux qui ne le sont pas : $\overline{x}(\overline{\mu})$ est forcément admissible.

***Exercice IV.10.** Données :

A_0 symétrique définie positive, $A_1, \ldots, A_m$ symétriques semi-définies positives ; $b_0, b_1, \ldots, b_m$ dans $\mathbb{R}^n$; $c_0, c_1, \ldots, c_m$ dans $\mathbb{R}$.

On considère le problème d'optimisation suivant :

$$(\mathcal{P}) \quad \begin{cases} \text{Minimiser } f(x) := \frac{1}{2}\langle A_0 x, x\rangle + \langle b_0, x\rangle + c_0 \\ \text{sous les contraintes} \\ g_j(x) := \frac{1}{2}\langle A_j x, x\rangle + \langle b_j, x\rangle + c_j \leqslant 0, \ j = 1, \ldots, m. \end{cases}$$

1°) Quelles propriétés de $(\mathcal{P})$, utiles pour sa résolution, peut-on noter ?

2°) Soit

$$\mathcal{L} : (x, \mu) \in \mathbb{R}^n \times \mathbb{R}^m \longmapsto \mathcal{L}(x, \mu) = f(x) + \sum_{j=1}^{m} \mu_j g_j(x)$$

le lagrangien usuel dans $(\mathcal{P})$, et

$$\mu \in \left(\mathbb{R}^+\right)^m \longmapsto \psi(\mu) = \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \mu)$$

la fonction duale associée.
Pour $\mu \in (\mathbb{R}^+)^m$ on pose

$$\begin{aligned} A(\mu) &:= A_0 + \mu_1 A_1 + \ldots + \mu_m A_m \\ b(\mu) &:= b_0 + \mu_1 b_1 + \ldots + \mu_m b_m \\ c(\mu) &:= c_0 + \mu_1 c_1 + \ldots + \mu_m c_m. \end{aligned}$$

Déterminer l'expression de $\psi(\mu)$ en fonction de $A(\mu)$, $b(\mu)$ et $c(\mu)$, et formuler (le plus simplement possible) le problème dual $(\mathcal{D})$ de $(\mathcal{P})$.

Solution : 1°) La fonction-objectif f est (quadratique) strictement convexe, les fonctions g_j définissant les contraintes du type inégalité sont (quadratiques) convexes. Donc, si l'ensemble-contrainte n'est pas vide, le problème de minimisation convexe $(\mathcal{P})$ a une et une seule solution.

2°) On a

$$\psi(\mu) = \inf_{x \in \mathbb{R}^n} \left\{ \frac{1}{2} \langle A(\mu) x, x \rangle + \langle b(\mu), x \rangle + c(\mu) \right\}.$$

$A(\mu)$ est symétrique définie positive pour tout $\mu \in (\mathbb{R}^+)^m$. La solution du problème de minimisation de $\mathcal{L}(\cdot, \mu)$ sur $\mathbb{R}^n$ est

$$\overline{x}(\mu) = -\left[A(\mu)\right]^{-1} b(\mu),$$

d'où

$$\psi(\mu) = -\frac{1}{2} \left\langle \left[A(\mu)\right]^{-1} b(\mu), b(\mu) \right\rangle + c(\mu).$$

Le problème dual $(\mathcal{D})$ de $(\mathcal{P})$ consiste à maximiser la fonction (concave) ψ sur $(\mathbb{R}^+)^m$.

***Exercice IV.11.** Soient $a_1, \dots, a_m$ dans $\mathbb{R}^n$, $b_1, \dots, b_m$ dans $\mathbb{R}$, et soit $\left(\hat{\mathcal{P}}\right)$ le problème qui consiste à minimiser $\hat{f}(x) := \ln\left(\sum_{j=1}^{m} \mathrm{e}^{\langle a_j, x\rangle - b_j}\right)$ pour $x \in \mathbb{R}^n$.

1°) Montrer que $\left(\hat{\mathcal{P}}\right)$ est équivalent (en des termes qu'on précisera) au problème (avec contraintes) $(\mathcal{P})$ posé dans $\mathbb{R}^n \times \mathbb{R}^m$ suivant :

$$(\mathcal{P}) \begin{cases} \text{Minimiser} f(x,y) := \ln\left(\sum_{j=1}^{m} \mathrm{e}^{y_j}\right) \\ \text{sous les contraintes } Ax - b = y, \end{cases}$$

où $A \in \mathcal{M}_{m,n}(\mathbb{R})$ est la matrice dont les lignes sont $a_1, \dots, a_m$ et b le vecteur de $\mathbb{R}^m$ de composantes $b_1, \dots, b_m$.

2°) Expliciter la fonction duale

$$\lambda \in \mathbb{R}^m \longmapsto \psi(\lambda) = \inf_{(x,y)\in\mathbb{R}^n\times\mathbb{R}^m} \{f(x,y) + \langle \lambda, Ax - b - y\rangle\}.$$

En déduire la forme du problème dual $(\mathcal{D})$ de $(\mathcal{P})$.

Solution : 1°) Si $\overline{x}$ est une solution de $\left(\hat{\mathcal{P}}\right)$, on pose $\overline{y} = A\overline{x} - b$ et $(\overline{x}, \overline{y})$ devient solution de $(\mathcal{P})$. Réciproquement, si $(\overline{x}, \overline{y})$ est solution de $(\mathcal{P})$, on a $A\overline{x} - b = \overline{y}$ et $\overline{x}$ est solution de $\left(\hat{\mathcal{P}}\right)$.

2°) Il s'agit de calculer

$$\inf_{(x,y)\in\mathbb{R}^n\times\mathbb{R}^m} \left\{ \ln\left(\sum_{j=1}^{m} \mathrm{e}^{y_j}\right) - \langle \lambda, y\rangle - \langle \lambda, b\rangle + \left\langle A^\top \lambda, x\right\rangle \right\}.$$

Plusieurs cas sont à distinguer :
a) Si $A^\top \lambda \neq 0$, l'infimum ci-dessus est clairement égal à $-\infty$.
b) Si $A^\top \lambda = 0$, on a à déterminer en fait

$$\inf_{y\in\mathbb{R}^m} \left\{ \ln\left(\sum_{j=1}^{m} \mathrm{e}^{y_j}\right) - \langle \lambda, y\rangle \right\}. \tag{4.2}$$

– Situation où $\sum_{j=1}^{m} \lambda_j \neq 1$.

De l'inégalité $\mathrm{e}^{y_1} + \ldots + \mathrm{e}^{y_m} \leqslant m\ \mathrm{e}^{\max_j y_j}$, on déduit

$$\ln\left(\sum_{j=1}^{m} \mathrm{e}^{y_j}\right) - \langle \lambda, y\rangle \leqslant \ln\ m + \max_j y_j - \sum_{j=1}^{m} \lambda_j y_j.$$

En prenant par exemple $y_j = \frac{k}{m}$ pour tout j (ou $-\frac{k}{m}$ pour tout j) et en faisant $k \to +\infty$, on constate aisément qu'alors

$$\inf_{y\in\mathbb{R}^m} \left\{ \ln\left(\sum_{j=1}^{m} \mathrm{e}^{y_j}\right) - \langle \lambda, y\rangle \right\} = -\infty. \tag{4.3}$$

– Situation où l'un des λ_j (disons λ_{j_0}) est < 0.

En prenant $y_j = 0$ pour $j \neq j_0$ et $y_{j_0} = -k$ et en faisant $k \to +\infty$, on arrive également à la relation (4.3) .

– Situation où $\sum\limits_{j=1}^{m} \lambda_j = 1$ et $\lambda_j \geqslant 0$ pour tout $j = 1, \ldots, m$.

Observons tout d'abord que si J_+ dénote $\{j = 1, \ldots, m \mid \lambda_j > 0\}$,

$$\inf_{y\in\mathbb{R}^m} \left\{ \ln\left(\sum_{j=1}^{m} \mathrm{e}^{y_j}\right) - \langle \lambda, y\rangle \right\} = \inf_{y\in\mathbb{R}^m} \left\{ \ln\left(\sum_{j\in J_+} \mathrm{e}^{y_j}\right) - \sum_{j\in J_+} \lambda_j y_j \right\}.$$

Il suffit donc de déterminer la borne inférieure de (4.2) dans le cas où tous les λ_j sont > 0. C'est en fait la situation où la borne inférieure de (4.2) est finie et atteinte. En effet, la borne inférieure de (4.2) est finie et atteinte si et seulement si le système d'optimalité suivant a une solution

$$\frac{\mathrm{e}^{\overline{y}_j}}{\sum\limits_{j=1}^{m} e^{\overline{y}_j}} = \lambda_j \quad \text{pour tout } j = 1, \ldots, m,$$

c'est-à-dire si et seulement si

$$\sum_{j=1}^{m} \lambda_j = 1 \text{ et } \lambda_j > 0 \ \text{ pour tout } j = 1, \ldots, m.$$

Dans ce cas la borne inférieure en question vaut $-\sum\limits_{j=1}^{m} \lambda_j \ln \lambda_j$.

Résumons tous les cas de figure en l'expression suivante de la fonction duale ψ :

$$\psi(\lambda) = -\sum_{j=1}^{m} \lambda_j \ln \lambda_j - \langle \lambda, b \rangle \text{ si } A^\top \lambda = 0,\ \sum_{j=1}^{m} \lambda_j = 1 \text{ et } \lambda_j \geqslant 0 \text{ pour tout } j\ ;$$
$$= -\infty \quad \text{sinon}$$

(avec le prolongement habituel $0 \ln 0 = 0$).

Le problème dual $(\mathcal{D})$ de $(\mathcal{P})$ est donc :

$$(\mathcal{D}) \quad \begin{cases} \text{Maximiser } -\sum_{j=1}^{m} \lambda_j \ln \lambda_j - \langle \lambda, b \rangle \\ \text{sous les contraintes } A^\top \lambda = 0,\ \sum_{j=1}^{m} \lambda_j = 1 \text{ et } \lambda_j \geqslant 0 \text{ pour tout } j. \end{cases}$$

***Exercice IV.12.** Soit $(\mathcal{P})$ le problème consistant à minimiser $f(x)$ sous la contrainte

$$x \in C := \{x \in \mathbb{R}^n \mid g_1(x) \leqslant 0, \ldots, g_p(x) \leqslant 0\}.$$

On fait les hypothèses suivantes sur les données de $(\mathcal{P})$:

– Les fonctions $f,\ g_1, \ldots, g_p$ sont convexes et différentiables sur $\mathbb{R}^n$;

– L'ensemble-contrainte C est borné ;

– Il existe $x_0 \in C$ tel que $g_j(x_0) < 0$ pour tout $j = 1, \ldots, p$.

1°) a) Montrer que l'intérieur de C est

$$\overset{\circ}{C} = \{x \in \mathbb{R}^n \mid g_j(x) < 0 \text{ pour tout } j = 1, \ldots, p\}.$$

b) Énoncer les propriétés du problème d'optimisation $(\mathcal{P})$ que les hypothèses faites induisent.

2°) Soit

$$\psi : \mu \in \left(\mathbb{R}^+\right)^p \longmapsto \psi(\mu) := \inf_{x \in \mathbb{R}^n} \left(f(x) + \sum_{j=1}^{p} \mu_j g_j(x) \right)$$

la fonction duale usuelle associée au problème de minimisation $(\mathcal{P})$, et pour tout $\alpha > 0$ soit

$$\varphi_\alpha : x \in \overset{\circ}{C} \longmapsto \varphi_\alpha(x) := f(x) - \frac{1}{\alpha}\sum_{j=1}^{p} \ln\left(-g_j(x)\right).$$

a) Vérifier que φ_α est convexe et différentiable sur $\overset{\circ}{C}$.

b) Montrer qu'il existe des points de $\overset{\circ}{C}$ minimisant φ_α sur $\overset{\circ}{C}$.

c) Soit $\overline{x}_\alpha$ un minimum de φ_α sur $\overset{\circ}{C}$ et on désigne par $\overline{\mu}^\alpha$ le vecteur de $(\mathbb{R}^+)^p$ dont les composantes sont $\dfrac{-1}{\alpha g_j(\overline{x}_\alpha)}$.

– Montrer que $\overline{x}_\alpha$ minimise $x \longmapsto f(x) + \sum\limits_{j=1}^{p} \overline{\mu}_j^\alpha g_j(x)$ sur $\mathbb{R}^n$.

– Vérifier que la fonction duale ψ prend une valeur finie en $\overline{\mu}^\alpha$.

– Établir l'encadrement suivant

$$f(\overline{x}_\alpha) \geqslant \overline{f} \geqslant f(\overline{x}_\alpha) - \frac{p}{\alpha},$$

où $\overline{f}$ désigne la valeur minimale dans $(\mathcal{P})$.

d) Commenter la pertinence de l'approche proposée dans cet exercice pour résoudre le problème originel $(\mathcal{P})$.

Solution : 1°) a) Il est clair que $C = \{x \in \mathbb{R}^n \mid g(x) \leqslant 0\}$, où on a posé $g := \max_j g_j$. La fonction g est convexe, mais pas différentiable en général.

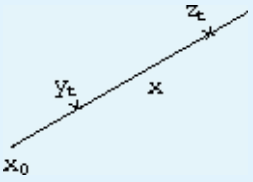
De par la continuité de g, on a :

$$(g(x) < 0) \Rightarrow \left(x \in \overset{\circ}{C}\right) \quad \text{(la convexité de } g \text{ n'est pas essentielle ici).}$$

Soit à présent x à l'intérieur de C et montrons que $g(x) < 0$ nécessairement.

Supposons $g(x) = 0$ et montrons que cela conduit à une contradiction.

Pour $t \in]0,1[$, posons $y_t := x - t(x - x_0)$ et $z_t := x + t(x - x_0)$, où x_0 est un point en lequel $g_j(x_0) < 0$ pour tout $j = 1, \ldots, p$ (et dont l'existence figure en hypothèse).



Comme $x = \dfrac{1}{2}(y_t + z_t)$ et que g est convexe,

$0 = g(x) \leqslant \dfrac{1}{2}g(y_t) + \dfrac{1}{2}g(z_t)$.

Toujours en raison de la convexité de g,

$g(y_t) \leqslant (1-t)\,g(x) + tg(x_0) < 0$.

Par suite $g(z_t) > 0$. On a donc z_t arbitrairement proche de x en lequel la valeur prise par g est strictement positive. Ceci contredit l'hypothèse de départ selon laquelle x est à l'intérieur de $\{x \in \mathbb{R}^n \mid g(x) \leqslant 0\}$.

En définitive

$$\overset{\circ}{C} = \{x \in \mathbb{R}^n \mid g_j(x) < 0 \text{ pour tout } j = 1, \ldots, p\},$$

et donc

$$\operatorname{fr} C = \{x \in C \mid \exists\ j \text{ tel que } g_j(x) = 0\}$$

(qui est aussi $\operatorname{fr} \overset{\circ}{C}$ car C est un convexe d'intérieur non vide).

b) $(\mathcal{P})$ est un problème de minimisation convexe. La fonction-objectif f étant continue et l'ensemble-contrainte C compact, le problème $(\mathcal{P})$ a bien des solutions. Tout ceci conduit à affirmer que l'ensemble S des solutions de $(\mathcal{P})$ est un convexe compact non vide de C.

La condition (de qualification des contraintes) de Slater étant vérifiée pour $(\mathcal{P})$ (cela figure parmi les hypothèses du problème), l'ensemble M des multiplicateurs de Lagrange-KKT est donc un convexe compact non vide de $(\mathbb{R}^+)^p$ (à tout $\overline{x} \in S$ est associé le même ensemble de multiplicateurs M).

La fonction duale

$$\psi : \mu \in \left(\mathbb{R}^+\right)^p \longmapsto \psi(\mu) = \inf_{x \in \mathbb{R}^n} \left(f(x) + \sum_{j=1}^{p} \mu_j g_j(x) \right)$$

est concave semi-continue supérieurement, et on sait que les points maximisant ψ sur $(\mathbb{R}^+)^p$ sont exactement ceux de M.

2°) a) La fonction $g_j : \overset{\circ}{C} \longrightarrow \mathbb{R}_*^-$ est convexe et différentiable ; la fonction $y \in \mathbb{R}_*^- \longmapsto -\ln(-y)$ est convexe, croissante et différentiable. Il s'ensuit que la composée des deux, qui n'est autre que $x \in \overset{\circ}{C} \longmapsto -\ln(-g_j(x))$, est convexe et différentiable sur $\overset{\circ}{C}$.

Par suite, φ_α est convexe et différentiable sur $\overset{\circ}{C}$.

b) f est bornée inférieurement sur C, $-\ln(-g_j(x)) \to +\infty$ quand $g_j(x) \to 0^-$. Il s'ensuit

$$\varphi_\alpha(x) \to +\infty \quad \text{quand} \quad x \in \overset{\circ}{C} \to \tilde{x} \in \operatorname{fr} C.$$

Par conséquent tous les ingrédients sont là pour qu'il existe $\overline{x}_\alpha$ minimisant φ_α sur $\overset{\circ}{C}$. Ces points $\overline{x}_\alpha$ sont caractérisés par :

$$\overline{x}_\alpha \in \overset{\circ}{C} \text{ et } \nabla f(\overline{x}_\alpha) + \sum_{j=1}^{p} \frac{-1}{\alpha g_j(\overline{x}_\alpha)} \nabla g_j(\overline{x}_\alpha) = 0. \tag{4.4}$$

c) La fonction $x \in \mathbb{R}^n \longmapsto f(x) + \sum_{j=1}^{p} \overline{\mu}_j^{\alpha} g_j(x)$ est convexe et différentiable sur $\mathbb{R}^n$. Comme $\nabla f(\overline{x}_\alpha) + \sum_{j=1}^{p} \overline{\mu}_j^{\alpha} \nabla g_j(\overline{x}_\alpha) = 0$, le point $\overline{x}_\alpha$ minimise bien $f + \sum_{j=1}^{p} \overline{\mu}_j^{\alpha} g_j$ sur $\mathbb{R}^n$.

Ainsi $\psi(\overline{\mu}^{\alpha}) := \inf_{x \in \mathbb{R}^n} \left(f(x) + \sum_{j=1}^{p} \overline{\mu}_j^{\alpha} g_j(x) \right) > -\infty$.

On a :

$$f(\overline{x}_\alpha) \geqslant \overline{f} = \inf_{x \in C} f(x) = \sup_{\mu \in (\mathbb{R}^+)^p} \psi(\mu) \geqslant \psi(\overline{\mu}^{\alpha}),$$

et

$$\psi(\overline{\mu}^{\alpha}) = f(\overline{x}_\alpha) + \sum_{j=1}^{p} \overline{\mu}_j^{\alpha} g_j(\overline{x}_\alpha) = f(\overline{x}_\alpha) + \sum_{j=1}^{p} \left(-\frac{1}{\alpha} \right)$$

puisque $\overline{\mu}_j^{\alpha} = \frac{-1}{\alpha g_j(\overline{x}_\alpha)}$ pour tout $j = 1, \ldots, p$.

d) φ_α joue le rôle de fonction-barrière (ou de fonction pénalisée par l'intérieur) réglée par le paramètre $\alpha > 0$.

$\alpha \longmapsto \overline{x}_\alpha$ est un « chemin central » dont on espère qu'il nous conduira à une solution $\overline{x}$ de $(\mathcal{P})$ lorsque $\alpha \to +\infty$.

La recherche de $\overline{x}_\alpha$ est un problème de minimisation sans contraintes, et on peut envisager de trouver $\overline{x}_\alpha$ en résolvant le système d'optimalité (4.4) (par une méthode de Newton par exemple).

L'encadrement $f(\overline{x}_\alpha) \geqslant \overline{f} \geqslant f(\overline{x}_\alpha) - \frac{p}{\alpha}$ est fort utile pour gérer un test d'arrêt sur l'incrémentation en α.

On peut aussi voir $\overline{x}_\alpha$ comme résultat d'une perturbation *ad hoc* des conditions de minimalité. En effet :

$$(\overline{x} \text{ est solution de } (\mathcal{P})) \Leftrightarrow \begin{pmatrix} \overline{x} \in C \text{ et il existe } \overline{\mu} \in (\mathbb{R}^+)^p \text{ tel que :} \\ \nabla f(\overline{x}) + \sum_{j=1}^{p} \overline{\mu}_j \nabla g_j(\overline{x}) = 0 \ ; \\ \overline{\mu}_j g_j(\overline{x}) = 0 \text{ pour tout } j = 1, \ldots, p \end{pmatrix},$$

alors que $\overline{x}_\alpha$ est caractérisé par

$$\begin{pmatrix} \overline{x}_\alpha \in C \text{ et il existe } \overline{\mu}^{\alpha} \in (\mathbb{R}^+)^p \text{ tel que :} \\ \nabla f(\overline{x}_\alpha) + \sum_{j=1}^{p} \overline{\mu}_j^{\alpha} \nabla g_j(\overline{x}_\alpha) = 0 \ ; \\ \overline{\mu}_j^{\alpha} g_j(\overline{x}_\alpha) = -\frac{1}{\alpha} \text{ pour tout } j = 1, \ldots, p. \end{pmatrix}$$

***Exercice IV.13.** *Le problème dual augmenté :*

1°) Considérons le problème de minimisation (de base) suivant :

$$(\mathcal{P}) \quad \begin{cases} \text{Min } f(x) \\ h_1(x) = 0, \ldots, h_m(x) = 0. \end{cases}$$

Pour $r > 0$ on pose $f_r(x) := f + \frac{r}{2} \sum_{i=1}^{m} h_i^2$ et on considère la version modifiée $(\mathcal{P}_r)$ de $(\mathcal{P})$ ci-dessous :

$$(\mathcal{P}_r) \quad \begin{cases} \text{Min } f_r(x) \\ h_1(x) = 0, \ldots, h_m(x) = 0. \end{cases}$$

a) Vérifier que les problèmes $(\mathcal{P})$ et $(\mathcal{P}_r)$ sont équivalents.
b) Soit

$$\begin{aligned} \mathcal{L}_r : (x, \lambda) \in \mathbb{R}^n \times \mathbb{R}^m \longmapsto \mathcal{L}_r(x, \lambda) : &= f_r(x) + \sum_{i=1}^{m} \lambda_i h_i(x) \\ &= f(x) + \frac{r}{2} \sum_{i=1}^{m} [h_i(x)]^2 + \sum_{i=1}^{m} \lambda_i h_i(x) \end{aligned} \tag{4.5}$$

le lagrangien usuel pour le problème $(\mathcal{P}_r)$; on l'appelle *lagrangien augmenté* du problème $(\mathcal{P})$.

– Comparer $\mathcal{L}_r$ et le lagrangien usuel $\mathcal{L}$ du problème $(\mathcal{P})$.

– Écrire le problème dual $(\mathcal{D}_r)$ (dit augmenté) dérivé de $\mathcal{L}_r$.

2°) On désire définir un lagrangien augmenté pour le problème de minimisation avec contraintes du type inégalité ci-dessous :

$$(\mathcal{Q}) \quad \begin{cases} \text{Min } f(x) \\ g_1(x) \leqslant 0, \ldots, g_p(x) \leqslant 0. \end{cases}$$

Pour ce faire, on considère le problème $(\hat{\mathcal{Q}})$ suivant, plongement du problème $(\mathcal{Q})$ dans $\mathbb{R}^n \times \mathbb{R}^p$:

$$\left(\hat{\mathcal{Q}}\right) \quad \begin{cases} \text{Min } f(x) \\ g_j(x) + y_j^2 = 0 \text{ pour } j = 1, \ldots, p. \end{cases}$$

Soit $\hat{\mathcal{L}}_r : ((x,y),\lambda) \in (\mathbb{R}^n \times \mathbb{R}^p) \times \mathbb{R}^p \longmapsto \hat{\mathcal{L}}_r(x,y,\lambda)$ le lagrangien augmenté associé à $(\hat{\mathcal{Q}})$ suivant la définition (4.5) . Comme l'expression de la fonction duale dérivée de $\hat{\mathcal{L}}_r$ conduit à calculer

$$\lambda \in \mathbb{R}^p \longmapsto \inf_{(x,y)\in\mathbb{R}^n\times\mathbb{R}^p} \hat{\mathcal{L}}_r(x,y,\lambda) = \inf_{x\in\mathbb{R}^n} [\inf_{y\in\mathbb{R}^p} \hat{\mathcal{L}}_r(x,y,\lambda)],$$

on propose de définir un lagrangien augmenté directement associé à $(\mathcal{Q})$ en posant :

$$\mathcal{L}_r : (x,\lambda) \in \mathbb{R}^n \times \mathbb{R}^p \longmapsto \mathcal{L}_r(x,\lambda) := \inf_{y\in\mathbb{R}^p} \hat{\mathcal{L}}_r(x,y,\lambda).$$

a) Démontrer que

$$\mathcal{L}_r(x,\lambda) = f(x) + \sum_{j=1}^{p} \varphi_r(\lambda_j, g_j(x)),$$

où $\varphi_r : (\alpha,t) \in \mathbb{R}\times\mathbb{R} \longmapsto \varphi_r(\alpha,t) := \frac{1}{2r}\left\{[\max\{0,\alpha+rt\}]^2 - \alpha^2\right\}$.

b) – Déduire du comportement de $\varphi_r(\alpha,t)$, lorsque $r \to 0^+$, ce que devient $\mathcal{L}_r(x,\lambda)$ quand $r \to 0^+$.

– Donner quelques premiers éléments de comparaison entre la fonction duale Θ_r dérivée de $\mathcal{L}_r$ et la fonction duale usuelle Θ.

c) Quelles propriétés de $\mathcal{L}_r$ peut-on déduire des hypothèses du type convexité ou différentiabilité des données du problème $(\mathcal{Q})$?

Commentaire :

– Le plongement d'un problème d'optimisation avec contraintes du type inégalité dans un problème d'optimisation avec contraintes du type égalité (en ajoutant des variables d'écart) a déjà été considéré dans l'Exercice III.28.

– Concernant l'apport et l'interêt du lagrangien augmenté dans un contexte particulier, on pourra « boucler » sur la partie B du problème I.18.

Solution : 1°) a) Désignons par S l'ensemble-contrainte dans $(\mathcal{P})$, c'est-à-dire

$$S := \{x \in \mathbb{R}^n \mid h_1(x) = 0, \ldots, h_m(x) = 0\}.$$

Il est évident que $f_r(x) = f(x)$ pour tout $x \in S$. Donc le problème de la minimisation (locale, resp. globale) de f_r sur S est équivalent à celui de la minimisation (locale, resp. globale) de f sur S.

b) – Le lagrangien usuel $\mathcal{L}$ pour le problème $(\mathcal{P})$ est

$$(x,\lambda)\in\mathbb{R}^n\times\mathbb{R}^m\longmapsto\mathcal{L}(x,\lambda)=f(x)+\sum_{i=1}^m\lambda_i h_i(x),$$

c'est-à-dire la fonction-limite $\lim\limits_{r\to 0}\mathcal{L}_r(x,\lambda)$. En fait,

$$\mathcal{L}_r=\mathcal{L}+\frac{r}{2}\sum_{i=1}^m h_i^2$$

est la somme du lagrangien usuel et d'un terme positif d'« augmentation » ou de « pénalisation extérieure » mesurant d'une certaine manière la violation des contraintes du problème.

– La fonction duale dérivée de $\mathcal{L}_r$ est

$$\Theta_r:=\lambda\in\mathbb{R}^m\longmapsto\Theta_r(\lambda):=\inf_{x\in\mathbb{R}^n}\mathcal{L}_r(x,\lambda),$$

d'où le problème dual correspondant :

$$(\mathcal{D}_r)\quad\begin{cases}\operatorname{Max}\Theta_r(\lambda)\\ \lambda\in\mathbb{R}^m.\end{cases}$$

2°) a) Par définition,

$$\begin{aligned}\hat{\mathcal{L}}_r(x,y,\lambda)&=f(x)+\frac{r}{2}\sum_{j=1}^p[g_j(x)+y_j^2]^2+\sum_{j=1}^p\lambda_j[g_j(x)+y_j^2]\\&=f(x)+\sum_{j=1}^p\left\{\frac{r}{2}y_j^4+[\lambda_j+rg_j(x)]y_j^2+\lambda_j g_j(x)+\frac{r}{2}g_j(x)^2\right\}.\end{aligned}$$

La structure « séparée » de $\hat{\mathcal{L}}_r(x,y,\lambda)$ en les variables y_j facilite la minimisation de $\hat{\mathcal{L}}_r(x,\cdot,\lambda)$ sur $\mathbb{R}^p$. En posant $u:=y_j^2$, on est en fait réduit à la minimisation de la fonction quadratique convexe $u\longmapsto\frac{r}{2}u^2+[\lambda_j+rg_j(x)]u+\lambda_j g_j(x)+\frac{r}{2}g_j(x)^2$ sur $\mathbb{R}^+$. Deux cas sont alors à envisager :

- $g_j(x)\leqslant-\frac{\lambda_j}{r}$, auquel cas la fonction en question est minimisée en $\overline{u}=-\frac{\lambda_j+rg_j(x)}{r}$;
- $g_j(x)>-\frac{\lambda_j}{r}$, auquel cas la fonction en question est minimisée en $\overline{u}=0$.

En somme, $\overline{u}=\max\left\{0,-\frac{\lambda_j+rg_j(x)}{r}\right\}$.

Par suite, $g_j(x)+\overline{u}=\max\left\{g_j(x),-\frac{\lambda_j}{r}\right\}$ et

$\frac{r}{2}\left(\max\left\{g_j(x),-\frac{\lambda_j}{r}\right\}\right)^2 + \lambda_j \max\left\{g_j(x),-\frac{\lambda_j}{r}\right\} = \frac{1}{2r}\{[\max\{0,\lambda_j + rg_j(x)\}]^2 - \lambda_j^2\}$ (si, si !), d'où l'expression escomptée de $\mathcal{L}_r(x,\lambda)$.

b) – Voici ci-après des exemples de fonctions $t \longmapsto \varphi_r(\alpha,t)$.

Les fonctions $\varphi_r(\alpha,\cdot)$ sont convexes, croissantes et dérivables (mais pas deux fois dérivables). De plus, lorsque $\alpha \geqslant 0$,

$$\varphi_r(\alpha,t) \geqslant \alpha t \quad \text{pour tout} \quad t \in \mathbb{R} \text{ (d'accord ?)},$$

de sorte que

$$\mathcal{L}_r(x,\lambda) \geqslant f(x) + \sum_{j=1}^{p} \lambda_j g_j(x) \text{ si } \lambda = (\lambda_1,\dots,\lambda_p) \in (\mathbb{R}^+)^p.$$

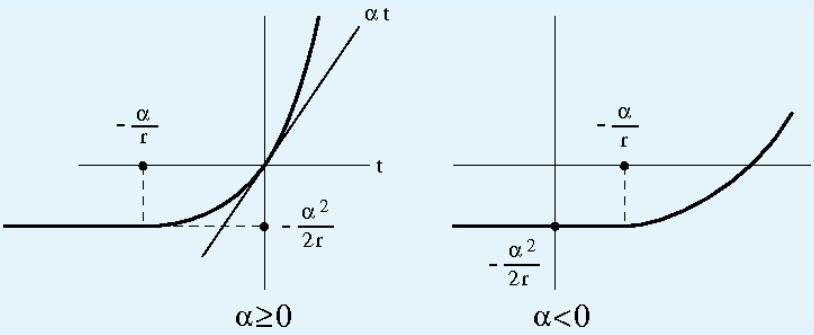


Figure 9.

Par ailleurs, de simples calculs de limites montrent :

– si $\alpha \geqslant 0$, $\varphi_r(\alpha,t) \to \alpha t$ quand $r \to 0^+$, et ce pour tout $t \in \mathbb{R}$;

– si $\alpha < 0$, $\varphi_r(\alpha,t) \to -\infty$ quand $r \to 0^+$, et ce pour tout $t \in \mathbb{R}$.

En conséquence, pour tout $(x,\lambda) \in \mathbb{R}^n \times \mathbb{R}^p$,

$$\lim_{r\to 0^+} \mathcal{L}_r(x,\lambda) = \begin{cases} f(x) + \displaystyle\sum_{j=1}^{p} \lambda_j g_j(x) = \mathcal{L}(x,\lambda) \text{ si } \lambda = (\lambda_1,\dots,\lambda_p) \in (\mathbb{R}^+)^p, \\ -\infty \text{ si l'un des } \lambda_j \text{ est } < 0. \end{cases}$$

– Le lagrangien augmentée $\mathcal{L}_r$, et la fonction duale augmenté Θ_r qui lui est associée, sont naturellement définis pour tout $\lambda \in \mathbb{R}^p$, alors que la fonction duale usuelle Θ n'est à considérer que pour $\lambda \in (\mathbb{R}^+)^p$. Si, néanmoins, on pose $\Theta(\lambda) = -\infty$ pour $\lambda \notin (\mathbb{R}^+)^p$, il est clair que l'on a :

$$\Theta_r(\lambda) \geqslant \Theta(\lambda) \text{ pour tout } \lambda \in \mathbb{R}^p.$$

Problèmes duaux dérivés :

$$(\mathcal{D}_r)\quad \begin{cases} \text{Max } \Theta_r(\lambda) \\ \lambda \in \mathbb{R}^p; \end{cases} \qquad (\mathcal{D})\quad \begin{cases} \text{Max } \Theta(\lambda) \\ \lambda \in \mathbb{R}^p. \end{cases}$$

c) On vérifie aisément que, pour tout $t \in \mathbb{R}$, les fonctions $\varphi_r(\cdot, t)$ sont concaves sur $\mathbb{R}$. Par suite, la fonction $\mathcal{L}_r(x, \cdot)$ est concave pour tout $x \in \mathbb{R}^n$.

Si les fonctions g_j sont convexes, il en est de même des fonctions $x \longmapsto \varphi_r(\lambda_j, g_j(x))$ (d'accord ?) ; par conséquent la convexité de f et des fonctions g_j induisent la convexité de $\mathcal{L}_r(\cdot, \lambda)$, pour tout $\lambda \in \mathbb{R}^p$.

Dans ce cas on a affaire à un lagrangien augmenté $\mathcal{L}_r$ qui est convexe-concave (et partout à valeurs finies) sur $\mathbb{R}^n \times \mathbb{R}^p$.

Si f et les fonctions g_j sont différentiables, il en est de même de la fonction $\mathcal{L}_r(\cdot, \lambda)$ et

$$\nabla_x \mathcal{L}_r(\overline{x}, \overline{\lambda}) = \nabla f(\overline{x}) + \sum_{j=1}^{p} \max\left\{0, \overline{\lambda}_j + r g_j(\overline{x})\right\} \nabla g_j(\overline{x}).$$

Par contre, même si f et les g_j sont deux fois différentiables, la différentiabilité seconde de $\mathcal{L}_r(\cdot, \lambda)$ n'est pas assurée en les points $\overline{x}$ tels que $\overline{\lambda}_j + r g_j(\overline{x}) = 0$ pour un certain j.

Commentaire : Prolongement de l'exercice : Dans le cas d'un problème de minimisation *convexe* $(\mathcal{Q})$, montrer que :

– les points-selles de $\mathcal{L}_r$ sur $\mathbb{R}^n \times \mathbb{R}^p$ sont les mêmes que ceux de $\mathcal{L}$ sur $\mathbb{R}^n \times (\mathbb{R}^+)^p$;

– les solutions du problème dual augmenté $(\mathcal{D}_r)$ sont les mêmes que celles du problème dual ordinaire $(\mathcal{D})$.

This page intentionally left blank

V

POLYÈDRES CONVEXES FERMÉS. OPTIMISATION À DONNÉES AFFINES (PROGRAMMATION LINÉAIRE)

Rappels

$\langle \cdot, \cdot \rangle$ dénote le produit scalaire usuel dans $\mathbb{R}^n$; les formes linéaires sur $\mathbb{R}^n$ sont du type $x \in \mathbb{R}^n \longmapsto \langle c, x \rangle$, où $c \in \mathbb{R}^n$.

Les hyperplans affines de $\mathbb{R}^n$ peuvent être décrits de la manière suivante :

$$\{x \in \mathbb{R}^n \mid \langle a_i, x \rangle = b_i\}, \tag{5.1}$$

où a_i est un vecteur non nul de $\mathbb{R}^n$ et b_i un réel. Les demi-espaces affines fermés de $\mathbb{R}^n$ peuvent être décrits d'une manière similaire à (5.1), une inégalité ($\langle a_i, x \rangle \leqslant b_i$) se substituant à l'égalité $\langle a_i, x \rangle = b_i$.

Naturellement, un hyperplan affine (d'équation $\langle a_i, x \rangle = b_i$) peut être vu comme l'intersection des deux demi-espaces fermés (d'équation $\langle a_i, x \rangle \leqslant b_i$ et $\langle -a_i, x \rangle \leqslant -b_i$ respectivement).

V.1. Polyèdres convexes fermés

Définition. Un polyèdre convexe fermé de $\mathbb{R}^n$ est l'intersection d'un nombre fini de demi-espaces affines fermés de $\mathbb{R}^n$.

Le polyèdre convexe fermé C, intersection des m demi-espaces affines fermés d'équation $\langle a_i, x \rangle \leqslant b_i$, est décrit d'une manière condensée sous la forme

$$Ax \leqslant b, \tag{5.2}$$

où $A \in \mathcal{M}_{m,n}(\mathbb{R})$ est la matrice dont les m lignes successives sont $a_1, \ldots, a_m$, b le vecteur de $\mathbb{R}^m$ de coordonnées $b_1, \ldots, b_m$, et $\leqslant$ l'ordre de $\mathbb{R}^m$ « coordonnée par coordonnée » (*i.e.*, si $x = (x_1, \ldots, x_m)$ et $y = (y_1, \ldots, y_m)$ sont deux vecteurs de $\mathbb{R}^m$, « $x \leqslant y$ » signifie « $x_i \leqslant y_i$ pour tout $i = 1, \ldots, m$ »).

Parmi les classes particulières de polyèdres convexes fermés signalons :

– Les *sous-espaces affines* de $\mathbb{R}^n$, que l'on peut décrire comme intersections d'hyperplans affines fermés :

$$\begin{array}{l} \{x \in \mathbb{R}^n \mid \langle a_i, x\rangle = b_i \text{ pour tout } i = 1, \ldots, m\} \\ (Ax = b \text{ sous forme condensée}). \end{array} \tag{5.3}$$

Lorsqu'un sous-espace affine est représenté comme en (5.3) avec des a_i linéairement indépendants (*i.e.* A est surjective), ce sous-espace est de dimension $n - m$.

– Les *polyèdres convexes compacts* (c'est-à-dire fermés bornés) de $\mathbb{R}^n$, appelés aussi *polytopes* de $\mathbb{R}^n$ quand ils sont d'intérieur non vide. Il y a deux manières équivalentes de représenter un polyèdre convexe compact C de $\mathbb{R}^n$:

$$C = \{x \in \mathbb{R}^n \mid Ax \leqslant b\} \text{ et } C \text{ est borné,}$$

ou bien

$$C = \operatorname{conv}\{v_1, \ldots, v_k\}, \tag{5.4}$$

où $v_1, \ldots, v_k$ sont des vecteurs de $\mathbb{R}^n$.

– Les *cônes convexes fermés polyédriques* de $\mathbb{R}^n$; ce sont les intersections d'un nombre fini de demi-espaces vectoriels fermés de $\mathbb{R}^n$ ($b = 0$ dans la description (5.2)) :

$$\begin{array}{l} \{x \in \mathbb{R}^n \mid \langle a_i, x\rangle \leqslant 0 \text{ pour } i = 1, \ldots, m\} \\ (Ax \leqslant 0 \text{ sous forme condensée}). \end{array} \tag{5.5}$$

Il y a deux manières équivalentes de représenter un cône convexe fermé polyédrique K de $\mathbb{R}^n$: comme en (5.5), ou bien

$$K = \left\{ \sum_{i=1}^{k} t_i v_i \mid t_i \geqslant 0 \text{ pour tout } i = 1, \ldots, k \right\}, \tag{5.6}$$

où $v_1, \ldots, v_k$ sont des vecteurs de $\mathbb{R}^n$.

K, décrit en (5.6), est un cône convexe fermé (le caractère fermé de K fera l'objet d'un exercice), auquel on se référera par la notation *cône* $\{v_1, \ldots, v_k\}$

(plutôt que cône-conv $\{v_1, \ldots, v_k\}$). Le passage d'une description à l'autre se fait par polarité : si K est un cône convexe fermé polyédrique de $\mathbb{R}^n$, son cône polaire

$$K^\circ := \{s \in \mathbb{R}^n \mid \langle s, d\rangle \leqslant 0 \text{ pour tout } d \in K\}$$

est encore un cône convexe fermé polyédrique de $\mathbb{R}^n$; de cette manière :

$$K^\circ = \text{cône } \{a_1, \ldots, a_m\} \text{ lorsque } K \text{ est décrit comme en } (5.5)\,; \tag{5.7}$$

$$\begin{aligned} &K^\circ = \{x \in \mathbb{R}^n \mid \langle v_i, x\rangle \leqslant 0 \ \text{ pour tout } i = 1, \ldots, k\} \\ &\text{lorsque } K \text{ est décrit comme en } (5.6)\,. \end{aligned} \tag{5.8}$$

Ce que dit (5.7) est le *lemme de Minkowski-Farkas* (sa première forme du moins) :

$$\begin{cases} \{x \in \mathbb{R}^n \mid \langle a_i, x\rangle \leqslant 0 \text{ pour } i = 1, \ldots, m\} \text{ est contenu dans} \\ \{x \in \mathbb{R}^n \mid \langle b, x\rangle \leqslant 0\} \end{cases} \tag{5.9}$$

si et seulement si

$$b \in \text{cône } \{a_1, \ldots, a_m\}\,. \tag{5.10}$$

Le plus petit sous-espace affine de $\mathbb{R}^n$ contenant un polyèdre convexe fermé C de $\mathbb{R}^n$ s'appelle le *sous-espace affine engendré par* C ; on appelle dimension de C (et on note dim C) la dimension de ce sous-espace affine engendré. L'intérieur de C dans le sous-espace affine engendré par C est ce qu'on appelle *l'intérieur relatif de* C ; il est noté ir C (ou $\text{int}_r\, C$).

Le caractère borné ou non d'un polyèdre convexe fermé de $\mathbb{R}^n$ est mesuré par le *cône asymptote* (ou asymptotique) de C ; il s'agit de

$$C_\infty := \{d \in \mathbb{R}^n \mid x_0 + td \in C \text{ pour tout } t \geqslant 0\}\,. \tag{5.11}$$

Cet ensemble, qui se trouve être indépendant du x_0 choisi dans C, est un cône convexe fermé polyédrique de $\mathbb{R}^n$; il est réduit à $\{0\}$ si et seulement si C est borné. Dans l'exemple d'une description de C comme en (5.2), C_∞ est $\{x \in \mathbb{R}^n \mid Ax \leqslant 0\}$.

Il y a une manière de décrire un polyèdre convexe fermé C de $\mathbb{R}^n$, complémentaire de celle donnée en (5.2), faisant apparaître la partie bornée et la partie non bornée de C :

$$C = C_0 + C_\infty, \tag{5.12}$$

où C_0 est un polyèdre convexe compact. On peut même poursuivre la décomposition en faisant $C_\infty = K + L$, où K est un cône convexe fermé polyédrique saillant (*i.e.* $K \cap (-K) = \{0\}$) et L un sous-espace vectoriel de $\mathbb{R}^n$.

Autres notions rattachées à un polyèdre convexe fermé C de $\mathbb{R}^n$:

– Si $\overline{x} \in C$, le cône tangent à C en $\overline{x}$ et le cône normal à C en $\overline{x}$, c'est-à-dire

$$T(C,\overline{x}) = \{d \in \mathbb{R}^n \mid d = t(c-\overline{x}) \text{ avec } c \in C \text{ et } t \geqslant 0\}$$

et

$$N(C,\overline{x}) = \{s \in \mathbb{R}^n \mid \langle s, c-\overline{x}\rangle \leqslant 0 \text{ pour tout } c \in C\}$$

respectivement, sont à leur tour des cônes convexes fermés polyédriques.

– Un point $\overline{x}$ de C est appelé *point extrémal* (*ou sommet*) de C s'il n'y a pas deux points distincts y et z de C tels que $\overline{x} = 1/2(y+z)$.

Autre manière équivalente de désigner un tel point : il est impossible d'avoir $\overline{x} = \alpha y + (1-\alpha)\, z$ avec y et z deux points distincts de C et $\alpha \in]0,1[$.

Un polyèdre convexe compact est l'enveloppe convexe de ses points extrémaux.

– Un hyperplan affine de $\mathbb{R}^n$, d'équation $\langle a, x\rangle = b$, est appelé *hyperplan d'appui* à C en $\overline{x} \in C$ lorsque

$$\langle a, \overline{x}\rangle = b \text{ et } \langle a, x\rangle \leqslant b \text{ pour tout } x \in C.$$

$F \subset C$ est une *face* (*exposée*) de C s'il existe un hyperplan d'appui H à C tel que $F = C \cap H$. Si un tel hyperplan a pour équation $\langle a, x\rangle = b$, on dira que F est exposée par a.

Une face F de C est un polyèdre convexe fermé, de dimension comprise entre 0 et $n-1$. Si $\dim F = 0$ on retrouve la notion de point extrémal (ou sommet) de C ; si $\dim F = 1$ on dira que F est une arête de C ; si $\dim F = n-1$ on parlera de F comme d'une facette de C.

Un polyèdre convexe fermé a un nombre fini de faces.

V.2. Optimisation à données affines (Programmation linéaire)

V.2.1. Définitions et notations

– Un problème d'optimisation à données affines (ou programme linéaire) se présente sous la forme suivante :

$$(\mathcal{P}) \begin{cases} \text{Maximiser } \langle c, x\rangle \\ Ax \leqslant b, \\ x \geqslant 0 \end{cases} \qquad \text{où } c \in \mathbb{R}^n,\ b \in \mathbb{R}^m \text{ et } A \in \mathcal{M}_{m,n}(\mathbb{R}). \tag{5.13}$$

Minimiser $\langle c, x\rangle$ sous les contraintes $Ax \leqslant b$ et $x \geqslant 0$ conduit au même type de problème (en changeant c en $-c$). La présentation (5.13) est appelée

forme *canonique* d'un programme linéaire ; une autre présentation est la forme dite *standard* [1] :

$$(\mathcal{P})\begin{cases}\text{Maximiser } \langle c, x\rangle\\ Ax = b\\ x \geqslant 0.\end{cases} \tag{5.14}$$

On passe de la forme canonique (avec A, b, c) à la forme standard, de la manière suivante :

$$\begin{pmatrix}\langle a_i, x\rangle \leqslant b_i\\ \text{dans } \mathbb{R}^n\end{pmatrix} \quad \text{devient} \quad \begin{pmatrix}\text{Il existe } u_i \geqslant 0 \text{ tel que}\\ \langle a_i, x\rangle + u_i = b_i\end{pmatrix};$$

d'où

$$\begin{pmatrix}Ax \leqslant b\\ x \geqslant 0\\ \text{dans } \mathbb{R}^n\end{pmatrix} \quad \text{devient} \quad \begin{pmatrix}Ax + I_m u = b\\ x \geqslant 0, \quad u \geqslant 0\end{pmatrix}.$$

Le programme linéaire de (5.13) est transporté *dans* $\mathbb{R}^n \times \mathbb{R}^m$ à présent :

$$\left(\mathcal{P}'\right)\begin{cases}\text{Maximiser } \langle c', x'\rangle\\ A'x' = b',\\ x' \geqslant 0\end{cases} \qquad \text{où } x' = (x, u) \in \mathbb{R}^n \times \mathbb{R}^m, \tag{5.15}$$

avec $A' = [A \mid I_m] \in \mathcal{M}_{m,n+m}(\mathbb{R})$, $b' = b$ et $c' = (c, 0)$.

Les problèmes d'optimisation (5.13) et (5.15) sont équivalents au sens où résoudre l'un permet de résoudre l'autre. Ainsi tous les résultats sur les programmes linéaires formulés sous la forme standard ont des contreparties sur les programmes linéaires formulés sous une forme canonique, et *vice versa*.

La résolution d'un programme linéaire est interprétée géométriquement comme suit : Si C est le polyèdre convexe fermé définissant les contraintes, maximiser $\langle c, x\rangle$ pour $x \in C$ revient à chercher à s'appuyer sur C avec un hyperplan d'équation $\langle c, x\rangle =$ constante ; l'ensemble des solutions (lorsqu'il y en a) est une face de C exposée par c.

– Soit C un polyèdre convexe fermé de $\mathbb{R}^n$ décrit de la manière suivante :

$$\begin{cases}Ax = b\\ x \geqslant 0\end{cases}, \quad \text{avec } A \in \mathcal{M}_{m,n}(\mathbb{R}) \text{ de rang } m. \tag{5.16}$$

[1] Ces appellations ne sont pas universelles, certains auteurs les permutent même. *Canonique* doit être compris ici au sens de « naturelle, toujours possible dans l'espace des variables x » ; *standard* au sens de « on peut toujours s'y ramener, quitte à ajouter des variables ».

Une *base* est un ensemble $J \subset \{1, \dots, n\}$ d'indices de colonnes de A telle que la sous-matrice A^J constituée des colonnes de A d'indices $j \in J$ soit inversible ; A^J est la *matrice de base* associée à J.

Un *élément (ou point) de base* correspondant à la base J est un vecteur $x = (x_1, \dots, x_n)$ de $\mathbb{R}^n$ tel que

$$x_j = 0 \text{ si } j \notin J,\ x_j \text{ est pris dans } x_J := \left(A^J\right)^{-1} b \text{ sinon.}$$

Cet x a donc des composantes « hors-base » qui sont nulles, et des composantes « de base » qui sont celles de x_J.

Une base J est dite *admissible* lorsque l'élément de base x correspondant est dans l'ensemble-contrainte décrit en (5.16), c'est-à-dire lorsque $(x_J)_j \geqslant 0$ pour tout $j \in J$ (les autres composantes, celles hors-base, étant nulles).

Si (5.16) est la représentation d'un ensemble-contrainte d'un programme linéaire, une base admissible J est appelée *optimale* lorsque l'élément de base correspondant est solution du programme linéaire considéré.

À partir d'une base J, on construit l'élément de base correspondant en complétant par des 0 l'élément x_J ; si l'un des x_j, $j \in J$, est nul, il y a une certaine ambiguïté à reconnaître la base J ; on dira que l'élément de base est *dégénéré* si précisément l'une des composantes de base est nulle.

Le résultat qui suit établit un lien fort entre la géométrie d'un polyèdre convexe fermé et sa représentation sous la forme standard (5.16) .

Théorème. *Soit C décrit comme en* (5.16) *. Alors les points extrémaux (ou sommets) de C sont exactement les éléments de base admissibles.*

V.2.2. Résultats fondamentaux d'existence

Théorème. *Si la forme linéaire $x \longmapsto \langle c, x\rangle$ est majorée sur le polyèdre convexe fermé $C \neq \emptyset$, alors elle y atteint sa borne supérieure.*

Ceci est un résultat d'existence particulier au « monde linéaire » : le seul fait pour $\langle c, \cdot\rangle$ d'être majorée sur C entraîne *l'existence* de $\overline{x}$ maximisant $\langle c, \cdot\rangle$ sur C. Sa démonstration fera l'objet d'un exercice.

Théorème. *– Si l'ensemble-contrainte d'un programme linéaire est un polyèdre convexe fermé non vide décrit comme en* (5.16)*, alors il existe des éléments de base admissibles (i.e. il existe des éléments de base correspondant à des bases admissibles).*

– Si l'ensemble-contrainte d'un programme linéaire est décrit comme en (5.16) *et si ce programme linéaire a des solutions, alors il existe des éléments de base optimaux (i.e. il existe des éléments de base correspondant à des bases optimales).*

Ainsi un polyèdre convexe fermé C décrit comme en (5.16) a toujours des points extrémaux, et si la forme linéaire $\langle c, \cdot \rangle$ est majorée sur C, il y a toujours au moins un point extrémal maximisant $\langle c, \cdot \rangle$ sur C.

V.3. La dualité en programmation linéaire

V.3.1. Formulations de problèmes duaux

Commençons par un programme linéaire écrit sous forme canonique :

$$(\mathcal{P}) \quad \begin{cases} \text{Maximiser } \langle c, x \rangle \\ Ax \leqslant b \\ x \geqslant 0 \end{cases} \rightarrow \quad (\mathcal{D}) \quad \begin{cases} \text{Minimiser } \langle b, y \rangle \\ A^\top y \geqslant c \\ y \geqslant 0 \end{cases} \tag{5.17}$$

$(\mathcal{D})$ est appelé *problème dual* (ou *programme linéaire dual*) de $(\mathcal{P})$.

Dans le cas de programmes linéaires $(\mathcal{P})$ écrits sous forme standard, les problèmes duaux $(\mathcal{D})$ se présentent comme suit :

$$(\mathcal{P}) \quad \begin{cases} \text{Maximiser } \langle c, x \rangle \\ Ax = b \\ x \geqslant 0 \end{cases} \rightarrow \quad (\mathcal{D}) \quad \begin{cases} \text{Minimiser } \langle b, y \rangle \\ \\ A^\top y \geqslant c \end{cases} \tag{5.18}$$

$$(\mathcal{P}) \quad \begin{cases} \text{Minimiser } \langle c, x \rangle \\ Ax = b \\ x \geqslant 0 \end{cases} \rightarrow \quad (\mathcal{D}) \quad \begin{cases} \text{Maximiser } \langle b, y \rangle \\ \\ A^\top y \leqslant c \end{cases} \tag{5.19}$$

À côté de la transformation « symétrique » (5.17), retenons dans les transformations « asymétriques » (5.18) et (5.19) les points suivants : « maximiser » devient « minimiser » et *vice versa* ; les contraintes du type égalité deviennent des contraintes du type inégalité, il n'y a plus de conditions de signe sur les variables duales y.

V.3.2. Relations entre les valeurs optimales et les solutions de programmes linéaires en dualité

Théorème. *Considérons*

$$(\mathcal{P}) \quad \begin{cases} \text{Maximiser } \langle c, x\rangle \\ Ax \leqslant b \\ x \geqslant 0 \end{cases} \quad \text{et} \quad (\mathcal{D}) \quad \begin{cases} \text{Minimiser } \langle b, y\rangle \\ A^\top y \geqslant c \\ y \geqslant 0. \end{cases}$$

(i) Si $\overline{x}$ est admissible pour $(\mathcal{P})$ et si $\overline{y}$ est admissible pour $(\mathcal{D})$, alors $\langle c, \overline{x}\rangle \leqslant \langle b, \overline{y}\rangle$.

(ii) Si $\overline{x}$ est admissible pour $(\mathcal{P})$, si $\overline{y}$ est admissible pour $(\mathcal{D})$ et si $\langle c, \overline{x}\rangle = \langle b, \overline{y}\rangle$, alors $\overline{x}$ est une solution de $(\mathcal{P})$ et $\overline{y}$ est une solution de $(\mathcal{D})$.

Revoyons ce même théorème avec une autre forme de $(\mathcal{P})$, et donc de $(\mathcal{D})$.

Théorème. *Considérons*

$$(\mathcal{P}) \quad \begin{cases} \text{Minimiser } \langle c, x\rangle \\ Ax = b \\ x \geqslant 0 \end{cases} \quad \text{et} \quad (\mathcal{D}) \quad \begin{cases} \text{Maximiser } \langle b, y\rangle \\ \\ A^\top y \leqslant c. \end{cases}$$

(i) Si $\overline{x}$ est admissible pour $(\mathcal{P})$ et si $\overline{y}$ est admissible pour $(\mathcal{D})$, alors $\langle c, \overline{x}\rangle \geqslant \langle b, \overline{y}\rangle$.

(ii) Si $\overline{x}$ est admissible pour $(\mathcal{P})$, si $\overline{y}$ est admissible pour $(\mathcal{D})$ et si $\langle c, \overline{x}\rangle = \langle b, \overline{y}\rangle$, alors $\overline{x}$ et $\overline{y}$ sont solutions de $(\mathcal{P})$ et de $(\mathcal{D})$ respectivement.

Revenons au couple $(\mathcal{P})$ – $(\mathcal{D})$ de problèmes en dualité du premier théorème, c'est-à-dire celui explicité en (5.17) ; on a à leur sujet le résultat fondamental suivant :

Théorème. *(i) Si l'un des problèmes $(\mathcal{P})$ ou $(\mathcal{D})$ a une valeur optimale finie, alors il en est de même de l'autre, et les valeurs optimales sont égales.*

(ii) Si le supremum dans $(\mathcal{P})$ est $+\infty$, alors l'ensemble-contrainte de $(\mathcal{D})$ est vide ; si l'infimum dans $(\mathcal{D})$ est $-\infty$, alors c'est que l'ensemble-contrainte de $(\mathcal{P})$ est vide.

Tous les cas possibles sont rassemblés dans le tableau synoptique ci-dessous ; pour englober tous les cas de valeurs optimales dans $(\mathcal{P})$ ou $(\mathcal{D})$ ($+\infty$ et $-\infty$ éventuellement), on conviendra que $\inf_\phi = +\infty$ et $\sup_\phi = -\infty$.

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="2">L'ensemble-contrainte de $(\mathcal{P})$ n'est pas vide</td><td rowspan="2">L'ensemble contrainte de $(\mathcal{P})$ est vide</td></tr>
<tr><td>$(\mathcal{P})$ a des solutions</td><td>$(\mathcal{P})$ n'a pas de solutions</td></tr>
<tr><td rowspan="2">L'ensemble-contrainte de $(\mathcal{D})$ n'est pas vide</td><td>$(\mathcal{D})$ a des solutions</td><td>O.K. max = min $(\mathcal{P})$ $(\mathcal{D})$</td><td>Impossible</td><td>Impossible</td></tr>
<tr><td>$\mathcal{D}$ n'a pas de solutions</td><td>Impossible</td><td>Impossible</td><td>inf dans $(\mathcal{D})$ = sup dans $(\mathcal{P})$ = $-\infty$</td></tr>
<tr><td colspan="2">L'ensemble-contrainte de $(\mathcal{D})$ est vide</td><td>Impossible</td><td>sup dans $(\mathcal{P})$ = inf dans $(\mathcal{D})$ = $+\infty$</td><td>Cas « pathologique » possible : sup dans $(\mathcal{P}) = -\infty$ inf dans $(\mathcal{D}) = +\infty$</td></tr>
</table>

V.3.3. Caractérisation simultanée des solutions du problème primal et du problème dual

Considérons par exemple le couple de problèmes en dualité de (5.19) :

$$(\mathcal{P}) \quad \begin{cases} \text{Minimiser } \langle c, x\rangle \\ Ax = b \\ x \geqslant 0 \end{cases} \quad \text{et} \quad (\mathcal{D}) \quad \begin{cases} \text{Maximiser } \langle b, y\rangle \\ \\ A^\top y \leqslant c. \end{cases}$$

Théorème. *Soient $\overline{x}$ et $\overline{y}$ des points admissibles pour $(\mathcal{P})$* et *$(\mathcal{D})$ respectivement. Alors :*

$$\begin{pmatrix} \overline{x} \text{ est solution de } (\mathcal{P}) \\ \text{et } \overline{y} \text{ est solution de } (\mathcal{D}) \end{pmatrix} \Leftrightarrow \begin{pmatrix} \left(\langle a_j', \overline{y}\rangle - c_j\right)\overline{x}_j = 0 \\ \text{pour tout } j = 1, \ldots, n \end{pmatrix},$$

où les a_j' désignent les vecteurs-lignes de $A^\top$.

Une manière équivalente de dire ce qui est exprimé dans l'assertion de droite de l'équivalence est :

$$\left(\overline{x}_j > 0 \Rightarrow \left\langle a_j', \overline{y}\right\rangle = c_j\right) \text{ et } \left(\left\langle a_j', \overline{y}\right\rangle < c_j \Rightarrow \overline{x}_j = 0\right).$$

Réécrivons ce résultat pour la formulation symétrique de programmes linéaires en dualité (de (5.17)) :

$$(\mathcal{P}) \quad \begin{cases} \text{Maximiser } \langle c, x\rangle \\ Ax \leqslant b \\ x \geqslant 0 \end{cases} \qquad (\mathcal{D}) \quad \begin{cases} \text{Minimiser } \langle b, y\rangle \\ A^\top y \geqslant c \\ y \geqslant 0. \end{cases}$$

Théorème. *Soient $\overline{x}$ et $\overline{y}$ des points admissibles pour $(\mathcal{P})$ et $(\mathcal{D})$ respectivement. Alors :*

$$\begin{pmatrix} \overline{x} \text{ est solution de } (\mathcal{P}) \\ \text{et } \overline{y} \text{ est solution de } (\mathcal{D}) \end{pmatrix} \Leftrightarrow \begin{pmatrix} \left(\langle a_j', \overline{y}\rangle - c_j\right)\overline{x}_j = 0 \\ \text{pour tout } j = 1, \dots, n \\ \text{et} \\ \left(\langle a_i, \overline{x}\rangle - b_i\right)\overline{y}_i = 0 \\ \text{pour tout } i = 1, \dots, m \end{pmatrix},$$

où les a_i (resp. les a_j') désignent les vecteurs-lignes de A (resp. de $A^\top$).

Toujours pour le couple $(\mathcal{P})$ – $(\mathcal{D})$ de (5.17), considérons le lagrangien

$$\begin{aligned} \mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m &\to \mathbb{R} \\ (x, y) &\longmapsto \mathcal{L}(x, y) := \langle c, x\rangle - \langle y, Ax - b\rangle \\ &= \langle c, x\rangle - \sum_{i=1}^{m} y_i \left(\langle a_i, x\rangle - b_i\right). \end{aligned}$$

(Ici, dans $(\mathcal{P})$, on maximise $\langle c, x\rangle$ – ou on minimise $-\langle c, x\rangle$ – ce qui explique le changement de signe par rapport au formalisme du chapitre IV et induit quelques adaptations dans les définitions et résultats qui y sont rappelés.)

Un point-selle (ou col) de $\mathcal{L}$ sur $(\mathbb{R}^+)^n \times (\mathbb{R}^+)^m$ est un couple $(\overline{x}, \overline{y}) \in (\mathbb{R}^+)^n \times (\mathbb{R}^+)^m$ tel que

$$\mathcal{L}(x, \overline{y}) \leqslant \mathcal{L}(\overline{x}, \overline{y}) \leqslant \mathcal{L}(\overline{x}, y) \text{ pour tout } (x, y) \in \left(\mathbb{R}^+\right)^n \times \left(\mathbb{R}^+\right)^m$$

(d'où $\mathcal{L}(\overline{x}, \overline{y}) = \max_{x \geqslant 0} \mathcal{L}(x, \overline{y}) = \min\limits_{y \geqslant 0} \mathcal{L}(\overline{x}, y)$).

Théorème. *Soient $\overline{x} \in (\mathbb{R}^+)^n$ et $\overline{y} \in (\mathbb{R}^+)^m$. Il y a alors équivalence entre les assertions suivantes :*

(i) $\overline{x}$ est une solution de $(\mathcal{P})$ et $\overline{y}$ est une solution de $(\mathcal{D})$;

(ii) $(\overline{x}, \overline{y})$ est un point-selle de $\mathcal{L}$ sur $(\mathbb{R}^+)^n \times (\mathbb{R}^+)^m$.

Lorsque ceci a lieu

$$\mathcal{L}(\overline{x}, \overline{y}) = \max \quad \textit{dans } (\mathcal{P}) = \min \quad \textit{dans } (\mathcal{D}).$$

Références

[7] Concis et dense. Très bon.

[26] [21], [22] et [24] contiennent de nombreux exemples simples et des illustrations ; elles intègrent la Programmation linéaire dans un domaine plus vaste, répertorié sous le vocable de « Recherche Opérationnelle ».

[27] Description et analyse des principales méthodes de résolution numérique des problèmes de programmation linéaire reposant sur l'algorithme du simplexe, ainsi que les programmes nécessaires à leur mise en œuvre sur micro-ordinateur.

[28] [CS] Très complets sur la question ; de véritables « Bibles ». Longtemps dominée par les algorithmes du type « méthode du simplexe », la résolution numérique des programmes linéaires a subi un véritable révolution avec l'apport de N. Karmarkar (1984). Les techniques du type « points intérieurs » (*cf.* l'Exercice V.25 pour une idée) commencent à prendre place dans les formations du niveau 2e cycle : voir le chapitre 4 de [8] et les chapitres XIII et XIV de [24] par exemple. La 4e partie de [4] et les ouvrages [20] et [25] sont consacrés pour l'essentiel à ces nouvelles approches.

***Exercice V.1.** Soit l'ensemble-contrainte d'un programme linéaire dans $\mathbb{R}^5$ décrit de la manière suivante :

$$\begin{cases} 2x_1 + x_2 + x_3 = 8,\ x_1 + 2x_2 + x_4 = 7,\ x_2 + x_5 = 3 \\ x_i \geqslant 0 \text{ pour tout } i = 1, \ldots, 5. \end{cases}$$

1°) Combien y a-t-il de bases au plus ? Quelles sont ces bases et les éléments de base associés ?

2°) Déterminer toutes les bases admissibles.

Solution : L'ensemble-contrainte est décrit sous la forme

$\begin{cases} Ax = b \\ x \geqslant 0 \end{cases}$, avec $A \in \mathcal{M}_{3,5}(\mathbb{R})$ de rang 3.

Il y a au plus $\binom{5}{3} = 10$ bases.

Bases	Éléments de base associés	Statut (admissible ou non)
$J = \{1, 2, 3\}$	$x = (1, 3, 3, 0, 0)$	admissible
$J = \{1, 2, 4\}$	$x = \left(\frac{5}{2}, 3, 0, -\frac{3}{2}, 0\right)$	non admissible

$J=\{1,2,5\}$	$x=(3,2,0,0,1)$	admissible
$J=\{2,3,4\}$	$x=(0,3,5,1,0)$	admissible
$J=\{2,3,5\}$	$x=\left(0,\frac{7}{2},\frac{9}{2},0,-\frac{1}{2}\right)$	non admissible
$J=\{3,4,5\}$	$x=(0,0,8,7,3)$	admissible
$J=\{1,3,5\}$	$x=(7,0,-6,0,3)$	non admissible
$J=\{1,4,5\}$	$x=(4,0,0,3,3)$	admissible
$J=\{2,4,5\}$	$x=(0,8,0,-9,-5)$	non admissible

$J=\{1,3,4\}$ n'est pas une base.

***Exercice V.2.** Soit Λ_n le simplexe-unité de $\mathbb{R}^n$, c'est-à-dire

$$\Lambda_n := \left\{ x=(x_1,\dots,x_n)\in\mathbb{R}^n \,|\, \sum_{i=1}^n x_i = 1,\ x_i \geqslant 0 \ \text{ pour tout } i=1,\dots,n \right\}.$$

Déterminer tous les points extrémaux de Λ_n de la manière suivante :

– décrire Λ_n sous la forme $\begin{cases} Ax=b \\ x\geqslant 0 \end{cases}$, avec $A\in\mathcal{M}_{m,n}(\mathbb{R})$ de rang m et $b\in\mathbb{R}^m$;

– faire ensuite la liste des éléments de base admissibles.

Solution : Λ_n peut être décrit sous la forme $\begin{cases} Ax=b \\ x\geqslant 0 \end{cases}$, avec $A=[1.....1]\in\mathcal{M}_{1,n}(\mathbb{R})$ (de rang 1) et $b=1$. Il s'ensuit la liste des bases et des éléments de base :

Bases	Éléments de base associés	Statut
$\{i\}$, $1\leqslant i\leqslant n$	$\mathrm{e}_i=(0,\dots,0,1,0,\dots,0)$, $1\leqslant i\leqslant n$ éléments de la base canonique de $\mathbb{R}^n$	Tous admissibles.

$\mathrm{e}_1,\dots,\mathrm{e}_n$ sont les (seuls) points extrémaux de Λ_n.

Commentaire : On peut, bien entendu, démontrer directement les résultats suivants :

– $\Lambda_n = \mathrm{conv}\{\mathrm{e}_1,\dots,\mathrm{e}_n\}$ (il suffit d'écrire $x=(x_1,\dots,x_n)\in\Lambda_n$ comme $\sum_{i=1}^n x_i\mathrm{e}_i$).

– Tout e_i est un point extrémal de Λ_n.

***Exercice V.3.** Soit C un polyèdre convexe compact de $\mathbb{R}^n$ décrit comme $\{x \in \mathbb{R}^n \mid Ax = b,\ x \geqslant 0\}$, avec $A \in \mathcal{M}_{m,n}(\mathbb{R})$ de rang m. Montrer l'équivalence suivante :

$$\begin{pmatrix}\text{Chaque élément de } C \text{ a au}\\ \text{moins } m \text{ composantes} > 0\end{pmatrix} \Leftrightarrow \begin{pmatrix}\text{Chaque sommet de } C \text{ a}\\ \text{exactement } m \text{ composantes} > 0\end{pmatrix}.$$

Commentaire : On illustrera ce résultat avec le simplexe-unité de $\mathbb{R}^n$, c'est-à-dire avec

$$C := \left\{ x = (x_1, \ldots, x_n) \in \mathbb{R}^n \mid \sum_{i=1}^{n} x_i = 1,\ x_i \geqslant 0 \text{ pour tout } i = 1, \ldots, n \right\}.$$

Solution : *[⇒]*. Un sommet $\overline{x}$ de C est un élément de base admissible ; les $n-m$ composantes hors-base sont donc nulles. Comme $\overline{x}$, élément de C, a au moins m composantes > 0, il en résulte que $\overline{x}$ a exactement m composantes > 0.

[⇐]. Un élément x de C est une combinaison convexe de points extrémaux $\overline{x}^i$ de C :

$$x = \sum_{i=1}^{k} \alpha_i \overline{x}^i, \text{ avec } \alpha_i > 0 \text{ pour tout } i \text{ et } \sum_{i=1}^{k} \alpha_i = 1. \tag{5.20}$$

Supposons que x ait $l > n - m$ composantes nulles et montrons que cela conduit à une contradiction. Soit $J \subset \{1, \ldots, k\}$ de cardinal l tel que $x_j = 0$ pour tout $j \in J$. Il vient de (5.20)

$$\sum_{i=1}^{k} \alpha_i \left(\overline{x}^i\right)_j = 0 \text{ pour tout } j \in J,$$

d'où $\left(\overline{x}^i\right)_j = 0$ pour tout $j \in J$, ce qui voudrait dire que $\overline{x}^i$ a $l > n - m$ composantes nulles au moins. Ceci entre en contradiction avec le fait que $\overline{x}^i$ a, par hypothèse, m composantes > 0.

Dans le cas du simplexe-unité C de $\mathbb{R}^n$, les n points extrémaux e_i de C ont exactement 1 composante > 0, mais les éléments de C peuvent avoir 1, 2, ... ou n composantes > 0.

***Exercice V.4.** Soit $C := \{x \in \mathbb{R}^n \mid Ax \leqslant b,\ x \geqslant 0\}$, où $A \in \mathcal{M}_{m,n}(\mathbb{R})$, et

$$\tilde{C} := \{(x,u) \in \mathbb{R}^n \times \mathbb{R}^m \mid Ax + u = b,\ x \geqslant 0,\ u \geqslant 0\}\,.$$

Il y a ainsi une correspondance entre les points x de C et les points $\tilde{x} = (x, u = b - Ax)$ de $\tilde{C}$.

Montrer l'équivalence suivante :

$$(\overline{x} \text{ est extrémal dans } C) \Leftrightarrow \Big((\overline{x}, \overline{u} = b - A\overline{x}) \text{ est extrémal dans } \tilde{C}\Big)\,.$$

Solution : Considérons $\overline{x}$ extrémal dans C. Posons $\overline{u} = b - A\overline{x}$, de sorte que $(\overline{x}, \overline{u})$ soit le point de $\tilde{C}$ correspondant à $\overline{x}$. Montrons que $(\overline{x}, \overline{u})$ est extrémal dans $\tilde{C}$.

On va pour cela raisonner par l'absurde : en supposant que $(\overline{x}, \overline{u})$ n'est pas extrémal dans $\tilde{C}$, on va être conduit à une contradiction.

Si $(\overline{x}, \overline{u})$ n'est pas extrémal dans $\tilde{C}$, il existe $\alpha \in]0,1[$, (x_1, u_1) et (x_2, u_2) dans $\tilde{C}$, $(x_1, u_1) \neq (x_2, u_2)$, tels que

$$(\overline{x}, \overline{u}) = \alpha\,(x_1, u_1) + (1-\alpha)\,(x_2, u_2)\,.$$

Cette dernière relation induit entre autres

$$\overline{x} = \alpha x_1 + (1-\alpha)\,x_2,$$

ce qui, en raison du caractère extrémal de $\overline{x}$ dans C, n'est possible que si $x_1 = x_2$.

Mais alors $u_1 = b - Ax_1 = b - Ax_2 = u_2$, d'où $(x_1, u_1) = (x_2, u_2)$, ce qui est contraire à l'une des propriétés de départ de (x_1, u_1) et (x_2, u_2).

Donc $(\overline{x}, \overline{u})$ est bien extrémal dans $\tilde{C}$.

Réciproquement, soit $(\overline{x}, \overline{u})$ extrémal dans $\tilde{C}$ et montrons que $\overline{x}$ est nécessairement extrémal dans C.

Raisonnons à nouveau par l'absurde. Supposons que $\overline{x}$ ne soit pas extrémal dans C : il existe $\alpha \in]0,1[$, x_1 et x_2 dans C, $x_1 \neq x_2$, tels que $\overline{x} = \alpha x_1 + (1-\alpha)\,x_2$.

Mais alors, puisque $\overline{u} = b - A\overline{x}$,

$$\begin{aligned}(\overline{x}, \overline{u}) &= \alpha\,(x_1, b - Ax_1) + (1-\alpha)\,(x_2, b - Ax_2)\\ &= \alpha z_1 + (1-\alpha)\,z_2,\end{aligned}$$

où z_1 et $z_2 \in \tilde{C}$, $z_1 \neq z_2$. Ceci contredit le caractère extrémal de $(\overline{x}, \overline{u})$ dans $\tilde{C}$.

Donc $\overline{x}$ est bien extrémal dans C.

Commentaire : Cette correspondance entre points extrémaux de $\tilde{C} \subset \mathbb{R}^n \times \mathbb{R}^m$ et de sa projection C sur $\mathbb{R}^n$ est particulière ; elle est sans espoir pour des convexes fermés en général (*cf.* Exercice VI.7).

***Exercice V.5.** Soit C le polyèdre convexe fermé de $\mathbb{R}^2$ décrit à l'aide des inégalités suivantes :

$$x_1 + \frac{8}{3}x_2 \leqslant 4,\ x_1 + x_2 \leqslant 2,\ 2x_1 \leqslant 3,\ x_1 \geqslant 0,\ x_2 \geqslant 0.$$

1°) Écrire C sous la forme standard (c'est-à-dire comme la projection sur $\mathbb{R}^2$ d'un polyèdre convexe fermé $\tilde{C}$ de $\mathbb{R}^5$ d'équation : $A\tilde{x} = \tilde{b},\ \tilde{x} \geqslant 0$).

2°) Quels sont les points extrémaux de $\tilde{C}$? En déduire les points extrémaux de C.

Commentaire : Grâce à une représentation graphique de C, on voit facilement que les points extrémaux de C sont $\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}3/2\\0\end{pmatrix}, \begin{pmatrix}0\\3/2\end{pmatrix}, \begin{pmatrix}4/5\\6/5\end{pmatrix}, \begin{pmatrix}3/2\\1/2\end{pmatrix}$. Le but de l'exercice est de retrouver ces points en utilisant la caractérisation des points extrémaux d'un polyèdre convexe fermé d'équation : $A\tilde{x} = \tilde{b},\ \tilde{x} \geqslant 0$ (revoir l'Exercice V.4 à ce sujet).

Solution : 1°) En introduisant les variables d'écart u_1, u_2, u_3, on peut dire : $(x = (x_1, x_2) \in C) \Leftrightarrow$ (Il existe u_1, u_2, u_3 tels que $(x_1, x_2, u_1, u_2, u_3) \in \tilde{C}$), où $\tilde{C}$ est décrit comme suit :

$$\begin{cases} x_1 + \frac{8}{3}x_2 + u_1 = 4,\ x_1 + x_2 + u_2 = 2,\ 2x_1 + u_3 = 3, \\ x_1 \geqslant 0,\ x_2 \geqslant 0,\ u_1 \geqslant 0,\ u_2 \geqslant 0,\ u_3 \geqslant 0. \end{cases}$$

Les points extrémaux de C s'obtiennent par projection sur $\mathbb{R}^2$ des points extrémaux de $\tilde{C}$:

$(x = (x_1, x_2)$ est extrémal dans $C) \Leftrightarrow (\tilde{x} = (x_1, x_2, u_1, u_2, u_3)$ est extrémal dans $\tilde{C})$ (*cf.* Exercice V.4).

Comme $\tilde{C}$ est décrit sous la forme $A\tilde{x} = \tilde{b},\ \tilde{x} \geqslant 0$, avec $A \in \mathcal{M}_{3,5}(\mathbb{R})$ de rang 3, il suffit, pour avoir les points extrémaux de $\tilde{C}$, de repérer ses éléments de base admissibles. Ce sont :

$x = \left(\frac{3}{2}, \frac{1}{2}, \frac{7}{6}, 0, 0\right)$ correspondant à la base $\{1, 2, 3\}$,

$x = \left(\frac{4}{5}, \frac{6}{5}, 0, 0, \frac{7}{5}\right)$ correspondant à la base $\{1, 2, 5\}$,

$x = \left(\frac{3}{2}, 0, \frac{5}{2}, \frac{1}{2}, 0\right)$ correspondant à la base $\{1, 3, 4\}$,

$x = (0, 0, 4, 2, 3)$ correspondant à la base $\{3, 4, 5\}$,

$x = \left(0, \frac{3}{2}, 0, \frac{1}{2}, 3\right)$ correspondant à la base $\{2, 4, 5\}$.

***Exercice V.6.** Soit $A \in \mathcal{M}_{m,n}(\mathbb{R})$, $b \in \mathbb{R}^m$, et C le polyèdre convexe fermé de $\mathbb{R}^n$ décrit comme suit

$$C := \{x \in \mathbb{R}^n \mid Ax \leqslant b\}\,.$$

On suppose : $m \geqslant n$, aucun des vecteurs-lignes a_i de A n'est nul.

Pour toute partie non vide I de $\{1,\ldots,m\}$ (de cardinal k par exemple), on note :

– A_I la matrice extraite de A en ne conservant que les lignes de numéro $i \in I$ (ainsi $A_I \in \mathcal{M}_{k,n}(\mathbb{R})$) ;

– b_I le vecteur extrait de b en ne conservant que les coordonnées de numéro $i \in I$ (ainsi $b_I \in \mathbb{R}^k$).

Soit $\overline{x}$ un point-frontière de C ; on désigne par $I(\overline{x})$ l'ensemble des indices $i \in \{1,\ldots,m\}$ correspondant aux contraintes-inégalités actives en $\overline{x}$, c'est-à-dire

$$I(\overline{x}) = \{i \mid \langle a_i, \overline{x}\rangle = b_i\}\,.$$

Montrer que $\overline{x}$ est un point extrémal de C si et seulement si le rang de $A_{I(\overline{x})}$ est égal à n.

Solution : De par la structure de C on a :

$$\overset{\circ}{C} = \{x \in \mathbb{R}^n \mid \langle a_i, x\rangle < b_i \text{ pour tout } i = 1,\ldots,m\}\ ;$$

$$\mathrm{fr}C = \left\{x \in C \mid \max_{i=1,\ldots,m}\{\langle a_i, x\rangle - b_i\} = 0\right\} = \{x \in C \mid I(x) \neq \emptyset\}\,.$$

Soient $\overline{x} \in \mathrm{fr}\ C$ et $A_{I(\overline{x})}$ la matrice extraite de A correspondante. Le rang de $A_{I(\overline{x})}$ est un entier compris entre 1 et $\min(n, \mathrm{card}\ I(\overline{x}))$. On se propose de montrer que $\overline{x}$ est un point extrémal de C si et seulement si le rang de $A_{I(\overline{x})}$ est exactement n.

– $\left[\left(\textit{rang de } A_{I(\overline{x})} = n\right) \Rightarrow (\overline{x} \textit{ est un point extrémal de } C)\right]$.

Supposons $A_{I(\overline{x})}$ de rang n et supposons que $\overline{x}$ ne soit pas extrémal dans C.

Il existe alors y et z dans C, $y \neq z$, et $\alpha \in]0,1[$ tels que $\overline{x} = \alpha y + (1-\alpha)\, z$. Prenons $i \in I(\overline{x})$; on a :

$$\langle a_i, \overline{x}\rangle = b_i,\ \langle a_i, y\rangle \leqslant b_i,\ \langle a_i, z\rangle \leqslant b_i.$$

Mais alors $\langle a_i, y\rangle - b_i = \langle a_i, z\rangle - b_i = 0$ nécessairement. Nous avons démontré que $I(\overline{x}) \subset I(y) \cap I(z)$.

Maintenant, puisque $A_{I(\overline{x})}$ est de rang n, il existe $I \subset I(\overline{x})$ de cardinal n telle que la matrice $A_I\,(\in \mathcal{M}_n(\mathbb{R}))$ soit régulière. Le système $A_I u = b_I$ n'a qu'une seule solution, et comme il a été observé que $A_I\overline{x} = b_I$, $A_I y = b_I$, $A_I z = b_I$, il s'ensuit $\overline{x} = y = z$ nécessairement. Ceci entre en contradiction avec une assertion du début du raisonnement. L'hypothèse faite au départ est donc absurde : $\overline{x}$ est un point extrémal de C.

– $\left[\left(\textit{rang de } A_{I(\overline{x})} < n\right) \Rightarrow (\overline{x}\ \textit{n'est pas un point extrémal de } C)\right]$.

Si rang de $A_{I(\overline{x})} < n$, le système $A_{I(\overline{x})}u = 0_{I(\overline{x})}$ a une solution non nulle, que nous notons $\overline{u}$.

Si $i \notin I(\overline{x})$, $\langle a_i, \overline{x}\rangle < b_i$ de sorte que

$$\langle a_i, \overline{x} + \alpha\overline{u}\rangle < b_i \text{ et } \langle a_i, \overline{x} - \alpha\overline{u}\rangle < b_i$$

pour $\alpha \neq 0$ suffisamment petit. On prend $\overline{\alpha} \neq 0$ faisant l'affaire pour tous les $i \notin I(\overline{x})$.

Mais comme $A_{I(\overline{x})}(\overline{x} \pm \overline{\alpha}\,\overline{u}) = A_{I(\overline{x})}(\overline{x}) \pm A_{I(\overline{x})}(\overline{u}) = b_{I(\overline{x})}$, on a $A(\overline{x} \pm \overline{\alpha}\,\overline{u}) \leqslant b$ en tout état de cause. Il en résulte que $\overline{x} + \overline{\alpha}\,\overline{u}$ et $\overline{x} - \overline{\alpha}\,\overline{u}$ sont dans C, et du coup $\overline{x}$ n'est pas extrémal dans C puisque $\overline{x} = 1/2\,(\overline{x} + \overline{\alpha}\,\overline{u}) + 1/2\,(\overline{x} - \overline{\alpha}\,\overline{u})$.

Commentaire : – En un point extrémal $\overline{x}$ de C, on a card $I(\overline{x}) \geqslant n$; lorsque card $I(\overline{x}) = n$, le point $\overline{x}$ est appelé point extrémal *non-dégénéré*.

– Le résultat de l'exercice conduit à un majorant du nombre de points extrémaux de C, c'est $\binom{m}{n}$. Toutefois ce majorant est très grossier en général, et une autre majoration a été donnée par McMullen en 1970 : le nombre de points extrémaux de C est majoré par

$$\mathrm{e}(m,n) := \binom{m - \left[\frac{n+1}{2}\right]}{m-n} + \binom{m - \left[\frac{n+2}{2}\right]}{m-n},$$

où $[k]$ désigne la partie entière de k. Cet entier $e(m,n)$ est en général considérablement plus petit que $\binom{m}{n}$.

– On peut compléter l'exercice en cernant un peu plus les points extrémaux non-dégénérés de C :

– un point extrémal non-dégénéré $\overline{x}$ a exactement n points extrémaux adjacents $\overline{x}_i$ (c'est-à-dire que les segments de droites joignant $\overline{x}$ aux $\overline{x}_i$ sont des arêtes de C) ;

– si $\overline{x}$ est un point extrémal non-dégénéré de C et si $\overline{x}_1, \ldots, \overline{x}_n$ sont ses points extrémaux adjacents, alors C est contenu dans le cône convexe polyédral de sommet $\overline{x}$ et de génératrices $\overline{x}_1 - \overline{x}, \ldots, \overline{x}_n - \overline{x}$ (c'est-à-dire $\overline{x} + \text{cône}\{\overline{x}_1 - \overline{x}, \ldots, \overline{x}_n - \overline{x}\}$).

***Exercice V.7.** Soit C un polyèdre convexe fermé de $\mathbb{R}^n$.

1°) On suppose ici que C est borné ; plus précisément on a

$$C = \text{conv}\,\{v_1, \ldots, v_k\}\,.$$

Montrer qu'un point extrémal de C est nécessairement l'un des v_i.

2°) On suppose C décrit de la manière suivante :

$$C = C_0 + K, \tag{5.21}$$

où C_0 est un polyèdre convexe compact et K un cône convexe fermé polyédrique.

a) Montrer que tout point extrémal de C est nécessairement dans C_0 et qu'il est aussi extrémal dans C_0.

b) Donner un exemple de C décrit comme en (5.21) mais sans point extrémal.

Quelle condition portant sur C (ou sur K) assurerait que C a effectivement des points extrémaux ?

Solution : 1°) Soit $\overline{x}$ un point extrémal de C. Si $\overline{x}$ n'est pas l'un des v_i, il y a l vecteurs v_i, $2 \leqslant l \leqslant k$, tels que :

$$\overline{x} = \sum_{i=1}^{l} \alpha_i v_i,\ 0 < \alpha_i < 1 \ \text{ pour tout } i \text{ et } \sum_{i=1}^{l} \alpha_i = 1.$$

Il s'ensuit

$$\overline{x} = \alpha_{i_0} v_{i_0} + (1 - \alpha_{i_0}) \sum_{i \neq i_0} \frac{\alpha_i}{1 - \alpha_{i_0}} v_i,$$

et, par conséquent, $\overline{x}$ n'est pas extrémal (d'accord ?).

Les points extrémaux de C figurent donc nécessairement parmi les v_i (mais tous les v_i ne sont pas extrémaux dans C).

2°) a) Soit $\overline{x} \in C$, $\overline{x} = c_0 + d$ avec $c_0 \in C_0$ et $d \in K$. Si $d \neq 0$, $\overline{x}$ ne saurait être extrémal dans C ; en effet

$$c_0 + d = \frac{1}{2} c_0 + \frac{1}{2} (c_0 + 2d)\,,$$

où $c_0 \in C$, $c_0 + 2d \in C_0 + 2K = C_0 + K = C$ et $c_0 \neq c_0 + 2d$.

Donc, si $\overline{x}$ est extrémal dans C, il est donc dans C_0 nécessairement ; et comme $C_0 \subset C$, $\overline{x}$ est aussi extrémal dans C_0.

b) Considérons l'exemple suivant dans $\mathbb{R}^2$:

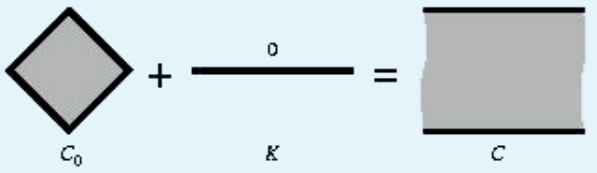


Figure 10.

C n'a pas de point extrémal ici. La raison en est que C contient une droite entière.

On peut en effet montrer que si C ne contient pas de droite, c'est-à-dire en fait si $K \cap (-K) = \{0\}$, alors C a effectivement au moins un point extrémal.

***Exercice V.8.** On considère dans $\mathbb{R}^n$, $n \geqslant 2$, le cône convexe fermé polyédrique suivant :

$$K_1 := \{x = (x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_1 \leqslant x_2 \ldots \leqslant x_n\}\,.$$

1°) Représenter K_1 sous la forme

$$\{x \in \mathbb{R}^n \mid \langle x, d_i\rangle \leqslant 0 \ \text{ pour tout } i = 1, \ldots, p\}\,,$$

où $d_1, \ldots, d_p$ sont des vecteurs de $\mathbb{R}^n$ à déterminer.

2°) Utiliser la représentation obtenue dans la question précédente pour démontrer que le cône polaire de K_1 est

$$K_1^\circ = \left\{ y = (y_1, \ldots, y_n) \in \mathbb{R}^n \mid \sum_{i=1}^{k} y_i \geqslant 0 \ \text{ pour tout } k = 1, \ldots, n-1 \right.$$
$$\left. \text{et } \sum_{i=1}^{n} y_i = 0 \right\}$$

3°) Déduire de ce qui précède des vecteurs $a_1, \ldots, a_q$ tels que $K_1 =$ cône $\{a_1, \ldots, a_q\}$.

Solution : 1°) On observe que $x = (x_1, \ldots, x_n) \in K$ si et seulement si les $n-1$ inégalités linéaires suivantes sont vérifiées :

$$x_1 - x_2 \leqslant 0,\ x_2 - x_3 \leqslant 0, \ldots, x_{n-1} - x_n \leqslant 0,$$

c'est-à-dire encore :

$$\langle x, d_i \rangle \leqslant 0 \ \text{ pour tout } \ i = 1, \ldots, n-1,$$

où $d_i := (0, \ldots, 0, 1, -1, 0, \ldots, 0)$.

$$\begin{array}{ll} & \uparrow \quad \uparrow \\ i^{\text{e}} & (i+1)^{\text{e}} \\ \text{position} & \text{position} \end{array}$$

2°) Il vient de la représentation ci-dessus de K_1 :

$$K_1^\circ = \left\{ \sum_{i=1}^{n-1} \alpha_i d_i \mid \alpha_i \geqslant 0 \text{ pour tout } i = 1, \ldots, n-1 \right\}.$$

Explicitons K_1°. Un élément $y = (y_1, \ldots, y_n)$ de K_1° a pour composantes

$$y_1 = \alpha_1,\ y_2 = \alpha_2 - \alpha_1, \ldots, y_{n-1} = \alpha_{n-1} - \alpha_{n-2},\ y_n = -\alpha_{n-1}.$$

Il s'ensuit : $y_1 \geqslant 0,\ y_1 + y_2 \geqslant 0, \ldots, y_1 + \ldots + y_{n-1} \geqslant 0$ et $y_1 + \ldots + y_n = 0$.

Réciproquement, partant de $y_1, \ldots, y_n$ vérifiant les conditions ci-dessus, posons :

$$\alpha_1 = y_1,\ \alpha_2 = y_1 + y_2, \ldots, \alpha_{n-2} = y_1 + \ldots + y_{n-2} \text{ et } \alpha_{n-1} = -y_n.$$

Il est clair que $\alpha_1 \geqslant 0,\ \alpha_2 \geqslant 0, \ldots, \alpha_{n-2} \geqslant 0,\ \alpha_{n-1} = y_1 + \ldots + y_{n-1} \geqslant 0$, et $(y_1, \ldots, y_n) = \sum_{i=1}^{n-1} \alpha_i d_i$.

On a donc démontré l'expression annoncée de K_1°.

3°) K_1° peut aussi s'exprimer de la manière suivante :

$$K_1^\circ = \{ y \in \mathbb{R}^n \mid \langle y, \delta_i \rangle \leqslant 0 \ \text{ pour } i = 1, \ldots, n+1 \},$$

où les δ_i sont les vecteurs de $\mathbb{R}^n$ suivants :

$$\delta_1 = (-1, 0, \ldots, 0)$$
$$\delta_2 = (-1, -1, 0, \ldots, 0), \ldots, \delta_{n-1} = (-1, \ldots, -1, 0)$$
$$\delta_n = (-1, -1, \ldots, -1),\ \delta_{n+1} = -\delta_n.$$

Par suite

$$K_1 = (K_1^\circ)^\circ = \text{cône}\,\{\delta_1, \ldots, \delta_n, -\delta_n\}.$$

****** Exercice V.9.*** On considère dans $\mathbb{R}^n$, $n \geqslant 2$, les cônes convexes fermés polyédriques suivants :

$$K_2 := \{x = (x_1, \ldots, x_n) \in \mathbb{R}^n \mid 0 \leqslant x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n\}\ ;$$
$$K_3 := \Big\{x = (x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_1 \leqslant \tfrac{x_1+x_2}{2} \leqslant \ldots \leqslant \tfrac{x_1+\ldots+x_k}{k} \leqslant \ldots \leqslant \tfrac{x_1+\ldots+x_n}{n}\Big\}\ ;$$
$$K_4 := \Big\{x = (x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_1 \geqslant \tfrac{x_1+x_2}{2} \geqslant \ldots \geqslant \tfrac{x_1+\ldots+x_k}{k} \geqslant \ldots \geqslant \tfrac{x_1+\ldots+x_n}{n} \geqslant 0\Big\}.$$

Déterminer K_2°, K_3° et K_4°.

Indication. Utiliser le résultat de l'exercice précédent.

Solution : Soit K_1 le cône convexe fermé polyédrique de l'exercice précédent.

Le passage de K_1 à K_2 se fait par l'ajout d'une nouvelle inégalité linéaire « $-x_1 \leqslant 0$ », *i.e.* $\langle x, d_n\rangle \leqslant 0$ avec $d_n = (-1, 0, \ldots, 0)$. Un élément $y = (y_1, \ldots, y_n) = \sum_{i=1}^{n} \alpha_i d_i \in K_2^\circ$ a donc pour composantes

$$y_1 = \alpha_1 - \alpha_n,\ y_2 = \alpha_2 - \alpha_1, \ldots,\ \ y_{n-1} = \alpha_{n-1} - \alpha_{n-2},\ y_n = -\alpha_{n-1}.$$

Sachant que les α_i sont tous positifs ou nuls, il s'ensuit :
$y_n \leqslant 0,\ y_{n-1} + y_n \leqslant 0, \ldots,\ y_k + \ldots + y_n \leqslant 0,\ y_1 + \ldots + y_n \leqslant 0.$

Réciproquement, partant de $y_1, \ldots, y_n$ vérifiant les conditions ci-dessus, posons :

$$\alpha_n = -(y_1 + \ldots + y_n),$$
$$\alpha_{n-1} = -y_n,\ \alpha_{n-2} = -(y_{n-1} + y_n), \ldots,\ \alpha_k = -(y_{k+1} + \ldots + y_n), \ldots,$$
$$\alpha_1 = -(y_2 + \ldots + y_n).$$

On a :

$$\alpha_i \geqslant 0 \ \text{ pour tout } \ i = 1, \ldots, n$$

et

$$(y_1, \ldots, y_n) = \sum_{i=1}^{n} \alpha_i d_i.$$

Nous avons ainsi démontré que

$$K_2^\circ = \left\{ y = (y_1, \ldots, y_n) \in \mathbb{R}^n \mid \sum_{i=k}^{n} y_i \leqslant 0 \ \text{ pour tout } k = 1, \ldots, n \right\}.$$

K_3 a un air de ressemblance avec K_1 ; en effet

$$(x \in K_3) \Leftrightarrow (Ax \in K_1),$$

où

$$A := \begin{bmatrix} 1 & 0 & 0 & \ldots & \ldots \\ \frac{1}{2} & \frac{1}{2} & 0 & \ldots & \ldots \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & \ldots \\ & & & \ddots & \\ \frac{1}{n} & \frac{1}{n} & \ldots & \ldots & \frac{1}{n} \end{bmatrix}.$$

A est visiblement inversible. La relation

$$\langle y, x \rangle \leqslant 0 \text{ pour tout } x \in K_3,$$

caractérisant un élément y de K_3°, est équivalente à

$$\left\langle y, A^{-1} x' \right\rangle \leqslant 0 \ \text{ pour tout } x' \in K_1,$$

soit encore

$$\left\langle \left(A^{-1}\right)^\top y, x' \right\rangle \leqslant 0 \ \text{ pour tout } x' \in K_1,$$

c'est-à-dire

$$\left(A^{-1}\right)^\top y \in K_1^\circ. \tag{5.22}$$

A^{-1} est aisée à déterminer ici :

$$A^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & 0 \\ -1 & 2 & 0 & 0 & \ldots & 0 \\ 0 & -2 & 3 & 0 & \ldots & 0 \\ \vdots & & & & \ddots & \\ 0 & 0 & \ldots & \ldots & -(n-1) & n \end{bmatrix}.$$

Connaissant K_1°, la relation (5.22) se traduit par :

$$\begin{array}{c} y_1 \geqslant y_2,\ y_1+y_2 \geqslant 2y_3,\ y_1+y_2+y_3 \geqslant 3y_4, \dots \\ \dots, y_1+\dots+y_{n-1} \geqslant (n-1)\, y_n \text{ et } y_1+y_2+\dots+y_n = 0. \end{array} \tag{5.23}$$

Un dernier effort pour montrer que (5.23) équivaut à :

$$\begin{array}{c} y_1 \geqslant \frac{y_1+y_2}{2} \geqslant \frac{y_1+y_2+y_3}{3} \geqslant \dots \geqslant \frac{y_1+\dots+y_n}{n} \\ \text{et } y_1+\dots+y_n = 0. \end{array} \tag{5.24}$$

Finalement

$$K_3^\circ = \left\{ y=(y_1,\dots,y_n)\in\mathbb{R}^n \,|\, y_1 \geqslant \frac{y_1+y_2}{2} \geqslant \dots \geqslant \frac{y_1+\dots+y_k}{k} \geqslant \dots \right.$$
$$\left. \dots \geqslant \frac{y_1+\dots+y_n}{n} \text{ et } \sum_{i=1}^n y_i = 0 \right\}.$$

En suivant une démarche similaire, on arrive à l'expression suivante de K_4° :

$$K_4^\circ = \left\{ y=(y_1,\dots,y_n)\in\mathbb{R}^n \,|\, y_1 \leqslant \frac{y_1+y_2}{2} \leqslant \dots \leqslant \frac{y_1+\dots+y_k}{k} \leqslant \dots \right.$$
$$\left. \dots \leqslant \frac{y_1+\dots+y_n}{n} \leqslant 0 \right\}.$$

Commentaire : Les cônes K_i interviennent en Statistique (régression monotone) où à partir d'un échantillon $x_1,\dots,x_n$, on cherche $\overline{x}_1,\dots,\overline{x}_n$ ordonné dans un certain sens (*i.e.* $(\overline{x}_1,\dots,\overline{x}_n)\in K_i$) le plus proche possible de $x_1,\dots,x_n$ au sens d'une distance euclidienne. Voir l'Exercice III.24 par exemple.

***** Exercice V.10.*** Soit C le polyèdre convexe fermé de $\mathbb{R}^n$ représenté de la manière suivante :

$$\begin{array}{rl} C := \{x\in\mathbb{R}^n \mid & \langle a_i, x\rangle \leqslant b_i \text{ pour } i=1,\dots,p, \\ & \text{et } \langle a_i, x\rangle = b_i \text{ pour } i=p+1,\dots,q\}. \end{array}$$

Déterminer en tout $x \in C$ le cône normal à C ainsi que son intérieur relatif (en fonction des données de la représentation de C et de $I(x) := \{i \in \{1,\dots,p\} \mid \langle a_i, x\rangle = b_i\}$).

Solution :

- $N(C,x) = \left\{ \sum_{i\in I(x)} t_i a_i + \sum_{i=p+1}^{q} t_i a_i \mid t_i \geqslant 0 \text{ pour } i \in I(x) \right\}$.

C'est la somme (vectorielle) de deux ensembles : le cône convexe engendré par les a_i, $i \in I(x)$, et le sous-espace vectoriel engendré par les a_i, $i \in \{p+1,\dots,q\}$.

- ir $N(C,x) = \left\{ \sum_{i\in I(x)} t_i a_i + \sum_{i=p+1}^{q} t_i a_i \mid t_i > 0 \text{ pour } i \in I(x) \right\}$.

***Exercice V.11.** Soit $\mathcal{B}_2$ le sous-ensemble des matrices $[a_{ij}]_{\substack{i=1,2\\ j=1,2}} \in \mathcal{M}_2(\mathbb{R})$ défini par :

$$([a_{ij}] \in \mathcal{B}_2) \Leftrightarrow \left(\begin{array}{c} a_{ij} \geqslant 0 \text{ pour tout } (i,j)\,; \\ \sum_{i=1}^{2} a_{ij} = \sum_{j=1}^{2} a_{ij} = 1 \text{ pour tout } (i,j) \end{array} \right).$$

Montrer que $\mathcal{B}_2$ est un segment de droite de $\mathcal{M}_2(\mathbb{R})$ dont on déterminera les extrémités.

Solution : De par les conditions imposées, une matrice M de $\mathcal{B}_2$ est nécessairement de la forme

$$M = \begin{bmatrix} \alpha & 1-\alpha \\ 1-\alpha & \alpha \end{bmatrix}, \text{ avec } \alpha \in [0,1].$$

Ainsi $M = \alpha \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + (1-\alpha) \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, où $\alpha \in [0,1]$.

Les deux matrices distinctes $M_0 := \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ et $M_1 := \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ sont donc les extrémités du segment de droite $\mathcal{B}_2$ (ou encore les points extrémaux de $\mathcal{B}_2$).

Commentaire : Les matrices $[a_{ij}] \in \mathcal{M}_n(\mathbb{R})$ vérifiant

$$a_{ij} \geqslant 0,\ \sum_i a_{ij} = 1,\ \sum_j a_{ij} = 1 \text{ pour tout } (i,j)$$

sont appelées *bistochastiques*. L'ensemble $\mathcal{B}_n$ de toutes les matrices bistochastiques de taille n est un convexe compact de $\mathcal{M}_n(\mathbb{R})$ dont les points extrémaux sont les matrices de permutation (une matrice de permutation est une matrice dont chaque ligne et chaque colonne ne contiennent qu'un seul terme non nul, lequel vaut 1). Ainsi, si Π_n désigne l'ensemble des $n!$ matrices de permutation,

$$\Pi_n = \operatorname{extr} \mathcal{B}_n \text{ et } \mathcal{B}_n = \operatorname{conv} \Pi_n.$$

Ce résultat, dû à G. Birkhoff (1946), est illustré pour $n = 2$ dans l'exercice.

Le plus petit sous-espace affine contenant $\mathcal{B}_n$ (*i.e.* le sous-espace affine engendré par $\mathcal{B}_n$) est de dimension $(n-1)^2$, une droite affine dans le cas de l'exercice.

***Exercice V.12.** On considère le cône convexe K de $\mathbb{R}^n$ engendré par les vecteurs $a_1, \ldots, a_m$, c'est-à-dire $K := \left\{ \sum_{i=1}^m t_i a_i \mid t_i \geqslant 0 \text{ pour tout } i = 1, \ldots, m \right\}$. Montrer que K est nécessairement fermé.

Solution : *1er cas : $a_1, \ldots, a_m$ sont linéairement indépendants.*

On pose $H := \operatorname{vect}\{a_1, \ldots, a_m\}$; H est un sous-espace vectoriel fermé dont $\{a_1, \ldots, a_m\}$ constitue une base. Soit $\{x_k\}$ une suite d'éléments de K ayant pour limite x.

Puisque $x_k = \sum_{i=1}^m t_i^k a_i$ avec $t_i^k \geqslant 0$ pour tout i, k, et que $x_k \to x$ *dans* H, nous avons :

$$t_i^k \to t_i \text{ pour tout } i;\ x = \sum_{i=1}^m t_i a_i.$$

Donc $t_i \geqslant 0$ pour tout i, et $x \in K$ en fait.

2e cas : $a_1, \ldots, a_m$ sont linéairement dépendants.

Il existe donc $\alpha_1, \alpha_2, \ldots, \alpha_m$ non tous nuls tels que $\sum_{i=1}^{m} \alpha_i a_i = 0$. On s'arrange pour que $I := \{i \mid \alpha_i < 0\}$ ne soit pas vide. Tout vecteur x de K peut alors s'écrire

$$x = \sum_{i=1}^{m} \left(t_i + \overline{t}\alpha_i\right) a_i, \text{ où } \overline{t} := \min_{i \in I} \left\{ -\frac{t_i}{\alpha_i} \right\} \ (\geqslant 0)\,.$$

Par choix de $\overline{t}$, $t_i + \overline{t}\alpha_i \geqslant 0$ pour tout i, mais l'un d'entre eux au moins est nul, disons celui correspondant à i_0. Ainsi

$$K = \bigcup_{i_0=1}^{m} \left\{ v \mid v = \sum_{i \neq i_0} \tau_i a_i \mid \tau_i \geqslant 0 \text{ pour tout } i \neq i_0 \right\}.$$

On réitère le procédé jusqu'à exprimer K comme *réunion finie de cônes convexes engendrés par des familles libres de vecteurs*, donc fermés d'après le résultat du 1er point.

***Exercice V.13.** $\mathbb{R}^n$ est muni de la base canonique $\{\mathrm{e}_1, \mathrm{e}_2, \ldots, \mathrm{e}_n\}$ et du produit scalaire canonique noté $\langle ., . \rangle$. Étant donnés $A \in \mathcal{M}_{m,n}(\mathbb{R})$, $b \in \mathbb{R}^m$ et $c \in \mathbb{R}^n$ (de coordonnées c_j), on considère le programme linéaire suivant :

$$(PL) \quad \begin{cases} \text{Minimiser } \langle c, x \rangle \\ x \in C := \{x \geqslant 0 : Ax = b\}\,. \end{cases}$$

On suppose $C \neq \phi$ et $\overline{\alpha} := \inf_{x \in C} \langle c, x \rangle > -\infty$. L'objet de l'exercice est de démontrer directement que (PL) a alors une solution (au moins). On pourra utiliser librement le résultat suivant : si $v_1, \ldots, v_p$ sont des vecteurs de $\mathbb{R}^q$, alors le cône convexe de $\mathbb{R}^q$ engendré par $v_1, \ldots, v_p$, *i.e.*

$$\text{cône}\,\{v_1, \ldots, v_p\} := \left\{ \sum_{j=1}^{p} \lambda_j v_j : \lambda_j \geqslant 0 \ \text{ pour tout } \ j \right\}$$

est fermé.

Soit $\mathcal{A} := \left[\dfrac{c_1, \ldots, c_n}{A}\right] \in \mathcal{M}_{m+1,n}(\mathbb{R})$ et désignons par $\mathcal{K}$ le cône convexe de $\mathbb{R}^{m+1}$ engendré par les vecteurs $\mathcal{A}\mathrm{e}_1, \ldots, \mathcal{A}\mathrm{e}_n$.

On considère $\{x_k\}$ une suite minimisante pour (PL), *i.e.* vérifiant : $x_k \in C$ pour tout k et $\langle c, x_k\rangle \to \overline{\alpha}$ quand $k \to +\infty$.

1°) Vérifier que $\mathcal{A}x_k \in \mathcal{K}$ pour tout k.

2°) En déduire qu'il existe $\lambda \geqslant 0$ dans $\mathbb{R}^n$ tel que $\overline{\alpha} = \langle c, \lambda\rangle$ et $b = A\lambda$. Conclure.

Solution : 1°) Puisque $x_k = \sum_{j=1}^{n} (x_k)_j \, \mathrm{e}_j$ et $\mathcal{A}$ est linéaire, $\mathcal{A}x_k = \sum_{j=1}^{n} (x_k)_j \, \mathcal{A}\mathrm{e}_j$.

Comme tous les $(x_k)_j$ sont dans $\mathbb{R}^+$, on a bien

$$\mathcal{A}x_k \in \mathcal{K} \text{ pour tout } k.$$

2°) Notons que $\mathcal{A}x_k = \begin{pmatrix} \langle c, x_k\rangle \\ Ax_k \end{pmatrix} = \begin{pmatrix} \langle c, x_k\rangle \\ b \end{pmatrix}$ et que la suite $\{\mathcal{A}x_k\}$ converge, quand $k \to +\infty$, vers $\begin{pmatrix} \overline{\alpha} \\ b \end{pmatrix}$.

Le cône $\mathcal{K}$ étant fermé, la limite de $\{\mathcal{A}x_k\}$ se trouve encore dans $\mathcal{K}$.
Ainsi, il existe $\lambda = (\lambda_1, \ldots, \lambda_n) \in (\mathbb{R}^+)^n$ tel que

$$\begin{pmatrix} \overline{\alpha} \\ b \end{pmatrix} = \sum_{j=1}^{n} \lambda_j \mathcal{A}\mathrm{e}_j,$$

c'est-à-dire, puisque $\mathcal{A}\mathrm{e}_j = \begin{pmatrix} c_j \\ A\mathrm{e}_j \end{pmatrix}$,

$$\overline{\alpha} = \sum_{j=1}^{n} \lambda_j c_j \text{ et } b = A\lambda.$$

Donc $\lambda \in C$ et $\langle c, \lambda\rangle = \overline{\alpha} = \inf_{x\in C} \langle c, x\rangle$; en clair λ est une solution de (PL).

***Exercice V.14.** Soient $c = (c_1, \ldots, c_n) \in \mathbb{R}^n$, $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$ et $b_0 \in \mathbb{R}$ tels que :

$$b_0 > 0 \quad \text{et } c_j > 0, \ a_j > 0 \text{ pour tout } j = 1, \ldots, n.$$

On considère le programme linéaire suivant :

$$(\mathcal{P}) \quad \begin{cases} \text{Max } \langle c, x\rangle \\ x \in C := \{x \in \mathbb{R}^n \mid \langle a, x\rangle \leqslant b_0, \ x_j \geqslant 0 \ \text{ pour tout } j = 1, \ldots, n\}. \end{cases}$$

1°) Vérifier que C n'est pas vide et est borné.

2°) Quels sont les points extrémaux de C ?

3°) Décrire la face de C constituée de l'ensemble des solutions de $(\mathcal{P})$.

4°) Illustrer les résultats précédents en faisant $n = 3$, $a_1 = a_2 = a_3 = b_0 = 1$, et en choisissant successivement $c = (0, 0, 1)$, $c = (0, 1, 1)$ et $c = (1, 1, 1)$.

Solution : 1°) C n'est pas vide puisque $(0, \dots, 0)$ y appartient. De plus, C est borné puisque

$$(x = (x_1, \dots, x_n) \in C) \Rightarrow \left(0 \leqslant x_j \leqslant \frac{b_0}{\min_j a_j} \text{ pour tout } j = 1, \dots, n\right).$$

2°) *1re méthode.* Décrivons C sous la forme standard. Pour cela posons

$$\tilde{C} := \{(x, u) \in \mathbb{R}^n \times \mathbb{R} \mid \langle a, x\rangle + u = b_0,\ x \geqslant 0,\ u \geqslant 0\}.$$

Les points extrémaux de $\tilde{C}$ sont faciles à déterminer ; les points extrémaux de C s'ensuivront (*cf.* Exercice V.4).

Dans l'équation

$$[a_1 \dots a_n 1] \begin{bmatrix} x_1 \\ \vdots \\ x_n \\ u \end{bmatrix} = b_0,$$

les éléments de base admissibles sont :

$\left(\overbrace{0, \dots, 0}^{n}, b_0\right)$ correspondant à la base $\{n+1\}$;

$\left(0, \dots, 0, \frac{b_0}{a_j}, 0, \dots, 0\right)$ correspondant à la base $\{j\}$, $1 \leqslant j \leqslant n$.

Ce sont les points extrémaux de $\tilde{C}$.

Par suite, les points extrémaux de C sont

$$(0, \dots, 0), \left(0, \dots, 0, \frac{b_0}{a_j}, 0, \dots, 0\right), \ 1 \leqslant j \leqslant n. \text{ (Il y en a } n+1 \text{ au total.)}$$

2e méthode. Écrivons C sous la forme $Ax \leqslant b$ où

$$A := \begin{bmatrix} a_1 \dots a_n \\ -I_n \end{bmatrix} \in \mathcal{M}_{n+1,n}(\mathbb{R}) \text{ et } b := \begin{pmatrix} b_0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

et utilisons le résultat de l'Exercice V.6.

Pour extraire de A une matrice $A_{I(\overline{x})}$ de rang n, il n'y a pas beaucoup de choix :

- $A_{I(\overline{x})} = -I_n$, ce qui correspond à $\overline{x} = 0$;
- $A_{I(\overline{x})}$ a $[a_1, \ldots, a_n]$ pour première ligne et $n-1$ lignes prises dans $-I_n$;

ceci revient, pour $\overline{x}$, à avoir $n-1$ composantes nulles (mettons, toutes sauf la j^{e}) et $a_j x_j = b_0$ (relation qui vient de $\langle a, x\rangle = b_0$).

3°) Les points extrémaux de C, solutions de $(\mathcal{P})$, sont ceux de la forme $(0, \ldots, 0, \frac{b_0}{a_{j_0}}, 0, \ldots, 0)$, où $\frac{c_{j_0} b_0}{a_{j_0}} = \max_j \frac{c_j b_0}{a_j}$; supposons qu'il y en ait k. Par conséquent, la face-solution de $(\mathcal{P})$ est l'enveloppe convexe de ces k points extrémaux.

4°) Si $a_1 = a_2 = \ldots = a_n = b_0 = 1$, les points extrémaux de

$$C := \left\{ x = (x_1, \ldots, x_n) \in \mathbb{R}^n \mid \sum_{j=1}^{n} x_j \leqslant 1,\ x_j \geqslant 0 \text{ pour tout } j = 1, \ldots, n \right\}$$

sont 0 et les n vecteurs de base canonique $\mathrm{e}_1, \ldots, \mathrm{e}_n$.

Pour $n = 3$, les choix successifs de c conduisent à une face-solution de $(\mathcal{P})$ qui est :

– le sommet $(0, 0, 1)$ lorsque $c_1 = (0, 0, 1)$;
– l'arête de C joignant $(0, 0, 1)$ à $(0, 1, 0)$ lorsque $c_2 = (0, 1, 1)$;
– la facette de C enveloppe convexe de $(0, 0, 1)$, $(0, 1, 0)$ et $(0, 0, 1)$ lorsque $c_3 = (1, 1, 1)$.

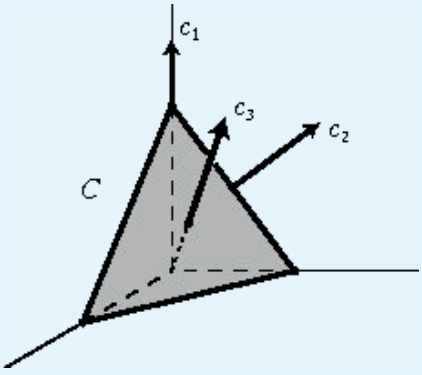


Figure 11.

***Exercice V.15.** Soient n réels fixés $x_1, \ldots, x_n$ et considérons n coefficients réels $\alpha_1, \ldots, \alpha_n$ vérifiant :

$$(I) \qquad 0 \leqslant \alpha_i \leqslant 1 \text{ pour tout } i = 1, \ldots, n \text{ et } \sum_{i=1}^{n} \alpha_i = m,$$

où m est un entier compris entre 1 et n.

Désignons par $x_{i_1}, \ldots, x_{i_m}$ les m plus grands nombres parmi les $x_1, \ldots, x_n$.

1°) Montrer par un calcul direct que

$$\sum_{i=1}^{n} \alpha_i x_i \leqslant x_{i_1} + \ldots + x_{i_m}. \tag{5.25}$$

2°) Soit Π_m le polyèdre convexe compact de $\mathbb{R}^n$ constitué de tous les $\alpha = (\alpha_1, \ldots, \alpha_n)$ vérifiant (I).

a) Déterminer les points extrémaux (ou sommets) de Π_m.

b) En déduire (5.25).

Solution : 1°) Quitte à changer l'indexation, on suppose $x_1 \geqslant x_2 \geqslant \ldots \geqslant x_m \geqslant x_{m+1} \geqslant \ldots \geqslant x_n$.

Si $m = n$, il n'y a rien à démontrer. Supposons donc $m < n$.

Comme $\sum_{i=1}^{m} (1 - \alpha_i) = \sum_{i=m+1}^{n} \alpha_i$, nous avons :

$$\sum_{i=m+1}^{n} \alpha_i x_i \leqslant \left(\sum_{i=m+1}^{n} \alpha_i \right) x_m = \sum_{i=1}^{m} (1 - \alpha_i)\, x_m \leqslant \sum_{i=1}^{m} (1 - \alpha_i)\, x_i.$$

D'où

$$\sum_{i=1}^{n} \alpha_i x_i = \sum_{i=1}^{m} \alpha_i x_i + \sum_{i=m+1}^{n} \alpha_i x_i \leqslant \sum_{i=1}^{m} \alpha_i x_i + \sum_{i=1}^{m} (1 - \alpha_i)\, x_i = \sum_{i=1}^{m} x_i.$$

2°) a) Les points extrémaux de Π_m sont de la forme $(\overline{\alpha}_1, \ldots, \overline{\alpha}_n)$ avec : toutes les coordonnées $\overline{\alpha}_i$ sont nulles sauf m d'entre elles qui sont égales à 1 (d'accord ?).

Illustration pour $n = 3$.

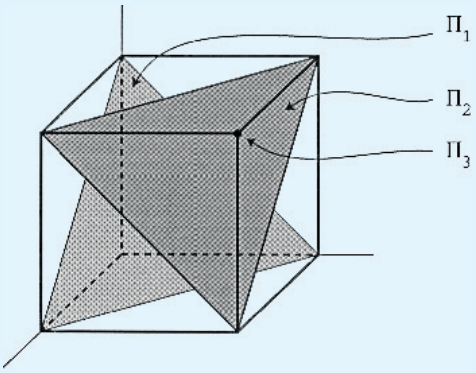


Figure 12.

Π_1 de sommets (0,0,1), (0,1,0), (0,0,1)
Π_2 de sommets (0,1,1), (1,0,1), (1,1,0)
$\Pi_3 = \{(1,1,1)\}$.

b) Appelons $\mathcal{E}_m$ l'ensemble des points extrémaux de Π_m et $\langle x, \cdot \rangle$ la forme linéaire sur $\mathbb{R}^n$ définie par $\langle x, (\alpha_1, \ldots, \alpha_n) \rangle = \sum_{i=1}^{n} \alpha_i x_i$. Le maximum de $\langle x, \cdot \rangle$ est le même sur Π_m et sur $\mathcal{E}_m$:

$$\max_{\alpha \in \Pi_m} \langle x, \alpha \rangle = \max_{\overline{\alpha} \in \mathcal{E}_m} \langle x, \overline{\alpha} \rangle .$$

Au vu de l'expression des $\overline{\alpha} \in \mathcal{E}_m$, ce dernier maximum est la somme des m plus grands nombres parmi les $x_1, \ldots, x_n$.

***Exercice V.16.** Soit C un polyèdre convexe fermé de $\mathbb{R}^n$. Pour tout $d \in \mathbb{R}^n$, on note $F(d)$ la face de C exposée par d, c'est-à-dire

$$F(d) = \left\{ \overline{x} \in C \mid \langle \overline{x}, d \rangle = \max_{x \in C} \langle x, d \rangle \right\}.$$

1°) Soit $\overline{d} \in \mathbb{R}^n$. Montrer qu'il existe un voisinage V de $\overline{d}$ tel que

$$F(d) \subset F(\overline{d}) \text{ pour tout } d \in V.$$

2°) On considère le programme linéaire suivant :

$$(\mathcal{P}) \quad \begin{cases} \text{Max } \langle x, d \rangle \\ x \in C. \end{cases}$$

Quelle conséquence pour la résolution de $(\mathcal{P})$ peut-on tirer du résultat de la 1re question ?

Indication. Pour répondre à la 1re question on pourra
– raisonner par l'absurde,
– utiliser dans le raisonnement le fait que C a un nombre fini de faces.

Solution : 1°) Raisonnons par l'absurde. Supposer le contraire de l'assertion que l'on veut prouver revient à supposer qu'il existe une suite (d_k) convergeant vers $\overline{d}$ telle que :

$$\forall k,\ F(d_k) \text{ n'est pas inclus dans } F(\overline{d}). \tag{5.26}$$

Puisque C a un nombre fini de faces, il existe une sous-suite de (d_k), notée $(d_{k_l})_l$, et une face F de C telles que

$$F(d_{k_l}) = F \text{ pour tout } l.$$

Soit y quelconque dans F. Alors, pour tout l :

$$\langle y, d_{k_l}\rangle \geqslant \langle x, d_{k_l}\rangle \text{ pour tout } x \in C.$$

Un passage à la limite sur l $(l \to +\infty)$ conduit à :

$$\langle y, \overline{d}\rangle \geqslant \langle x, \overline{d}\rangle \text{ pour tout } x \in C,$$

soit encore : $y \in F(\overline{d})$.

Nous avons donc prouvé que $F \subset F(\overline{d})$, ce qui entre en contradiction avec (5.26).

2°) Soit (d_k) convergeant vers d et soit $(\mathcal{P}_k)$ le programme linéaire suivant :

$$(\mathcal{P}_k) \quad \begin{cases} \text{Max } \langle x, d_k\rangle \\ x \in C. \end{cases}$$

Du résultat de la 1re question, il vient ceci : pour k suffisamment grand, l'ensemble des solutions de $(\mathcal{P}_k)$ est *contenu* dans l'ensemble des solutions $(\mathcal{P})$.

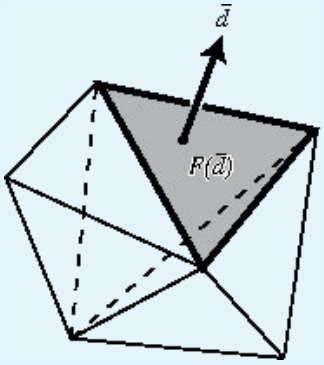


Figure 13.

***Exercice V.17.** Soit C le polyèdre convexe compact de $\mathbb{R}^4$ décrit comme suit :

$$\begin{cases} x_1 + \frac{4}{3}x_2 + 2x_3 = \frac{3}{2} \\ x_2 + 3x_3 = \frac{3}{2} \\ x_1 + x_2 + x_3 + x_4 = 1 \\ x_1 \geqslant 0, \ldots, x_4 \geqslant 0. \end{cases}$$

1°) Lister les points extrémaux de C.

2°) Résoudre

$$(\mathcal{P}) \quad \begin{cases} \text{Min } 3x_1 + x_2 + x_3 + x_4 \\ (x_1, x_2, x_3, x_4) \in C. \end{cases}$$

3°) Pour $\varepsilon > 0$ suffisamment petit (disons $0 < \varepsilon < \frac{1}{10}$), on remplace $\frac{4}{3}$ par $\frac{4}{3} - \varepsilon$ dans la 1re équation définissant C, ce qui donne un nouveau polyèdre C_ε. Résoudre à présent

$$(\mathcal{P}_\varepsilon) \quad \begin{cases} \text{Min } 3x_1 + x_2 + x_3 + x_4 \\ (x_1, x_2, x_3, x_4) \in C_\varepsilon. \end{cases}$$

Conclusions ?

Solution : 1°) C est décrit sous la forme : $Ax = b,\ x \geqslant 0$, avec $A \in \mathcal{M}_{3,4}(\mathbb{R})$ de rang 3 ; les éléments de base admissibles (et donc les points extrémaux de C) sont :

$$\left(0, \frac{3}{4}, \frac{1}{4}, 0\right) \quad \text{correspondant à la base } \{2,3,4\}\,,$$

$$\left(\frac{1}{2}, 0, \frac{1}{2}, 0\right) \quad \text{correspondant à la base } \{1,3,4\}\,.$$

2°) $\overline{x} = \left(0, \frac{3}{4}, \frac{1}{4}, 0\right)$ est la seule solution de $(\mathcal{P})$, et $\text{val}(\mathcal{P}) = 1$.

3°) On a introduit une légère perturbation dans un des coefficients de A, par exemple $\frac{4}{3}$ est remplacé par 1,333333. Il s'ensuit par des calculs similaires à ceux de la question précédente que la (seule) solution de $(\mathcal{P}_\varepsilon)$ est à présent $\overline{x}_\varepsilon = \left(\frac{1}{2}, 0, \frac{1}{2}, 0\right)$, et la valeur optimale correspondante $\text{val}(\mathcal{P}_\varepsilon) = 2$.

Il n'y a donc pas, en général, de « continuité » de la valeur optimale ou de l'ensemble-solution d'un programme linéaire $\begin{cases} \text{Min } \langle c, x\rangle \\ Ax = b,\ x \geqslant 0 \end{cases}$ considérés comme fonctions des coefficients de A.

***Exercice V.18.** Étant donnés $a_1, \ldots, a_m$ dans $\mathbb{R}^n$, $b_1, \ldots, b_m$ dans $\mathbb{R}$, et le système d'équations

$$(\mathcal{S}) \quad \begin{cases} \langle a_1, x\rangle - b_1 = 0 \\ \langle a_2, x\rangle - b_2 = 0 \\ \ldots \\ \langle a_m, x\rangle - b_m = 0, \end{cases}$$

on cherche le (ou les) point(s) minimisant
$x \longmapsto f(x) := \max_{i=1,\ldots,m} |\langle a_i, x\rangle - b_i|$ sur $\mathbb{R}^n$ (problème appelé $(\mathcal{P}_1)$ dans la suite).

1°) Que représente géométriquement $|\langle a_i, x\rangle - b_i|$ lorsque $\| a_i \| = 1$?
Quelle est la signification géométrique du problème posé ci-dessus lorsque tous les a_i sont de norme égale à 1 ?

2°) Formaliser $(\mathcal{P}_1)$ comme un problème de programmation linéaire, et en tirer toutes les conséquences.

Solution : 1°) Lorsque $\| a_i \| = 1$, $|\langle a_i, x\rangle - b_i|$ représente la distance (euclidienne) du point x à l'hyperplan H_i d'équation $\langle a_i, x\rangle - b_i = 0$.

Lorsque tous les a_i sont des vecteurs unitaires, le problème posé revient donc à chercher les points x de $\mathbb{R}^n$ minimisant la plus grande des distances de x aux hyperplans H_i.

2°) Étant donné que $|\langle a_i, x\rangle - b_i| = \max\left(\langle a_i, x\rangle - b_i, b_i - \langle a_i, x\rangle\right)$, chercher $x \in \mathbb{R}^n$ rendant $\max_i |\langle a_i, x\rangle - b_i|$ le plus petit possible revient à chercher $(x, y) \in \mathbb{R}^n \times \mathbb{R}$ vérifiant

$$\begin{aligned} \langle a_i, x\rangle - b_i &\leqslant y \\ -\langle a_i, x\rangle + b_i &\leqslant y \end{aligned} \quad \text{pour tout } i = 1, \ldots, m,$$

avec y le plus petit possible.

Considérons donc le programme linéaire suivant, posé dans $\mathbb{R}^n \times \mathbb{R}$:

$$(\mathcal{P}) \quad \left\{ \begin{array}{l} \text{Min } y \\ \left[\begin{array}{c|c} A & \begin{array}{c} -1 \\ \vdots \\ -1 \end{array} \\ \hline A & \begin{array}{c} -1 \\ \vdots \\ -1 \end{array} \end{array}\right] \begin{bmatrix} x_1 \\ \vdots \\ \vdots \\ \vdots \\ x_n \\ y \end{bmatrix} \leqslant \begin{bmatrix} b_1 \\ \vdots \\ b_m \\ -b_1 \\ \vdots \\ -b_m \end{bmatrix}, \end{array} \right.$$

où $A \in \mathcal{M}_{m,n}(\mathbb{R})$ est la matrice dont les vecteurs-lignes sont les a_i.

Une solution $(\overline{x}, \overline{y})$ de $(\mathcal{P})$ fournit une solution $\overline{x}$ de $(\mathcal{P}_1)$ et sa valeur optimale $\min\limits_{x \in \mathbb{R}^n} \max\limits_i |\langle a_i, x\rangle - b_i| = \max\limits_i |\langle a_i, \overline{x}\rangle - b_i|$.

L'ensemble-contrainte de $(\mathcal{P})$ n'est pas vide (il suffit de prendre $x \in \mathbb{R}^n$ quelconque et $y = \max_i |\langle a_i, x\rangle - b_i|$), et tout y de cet ensemble-contrainte est $\geqslant 0$.

La fonction-objectif (linéaire) est bornée inférieurement sur le polyèdre-contrainte, donc elle y atteint sa borne inférieure : $(\mathcal{P})$ – et donc $(\mathcal{P}_1)$ – a une solution.

Pour résoudre numériquement $(\mathcal{P}_1)$, une possibilité intéressante est donc de résoudre le programme linéaire $(\mathcal{P})$ associé.

***Problème V.19.** Dans tout le problème, C désigne un polyèdre convexe fermé non vide de $\mathbb{R}^n$.

1re partie. On suppose ici que C est représenté de la manière suivante :

$$C := \{x \in \mathbb{R}^n \mid Ax \leqslant b\}\,, \text{ où } A \in \mathcal{M}_{m,n}(\mathbb{R}) \text{ et } b \in \mathbb{R}^m. \tag{5.27}$$

1°) a) Montrer que le cône asymptote C_∞ de C est

$$C_\infty = \{d \in \mathbb{R}^n \mid Ad \leqslant 0\}\,.$$

b) En déduire que C est borné si, et seulement si, le système d'inéquations $Ad \leqslant 0$ n'a que la solution $d = 0$.

2°) Montrer l'équivalence des deux assertions suivantes :
(i) $\{d \in \mathbb{R}^n \mid Ad \leqslant 0\} = \{0\}$;
(ii) Pour tout $c \in \mathbb{R}^n$, le système d'équations $A^\top y = c$ a une solution $y \geqslant 0$.

2e partie. On suppose que C a la représentation suivante :

$$C := C_0 + K + L, \tag{5.28}$$

où

$$C_0 = \operatorname{conv}\{u_1, \ldots, u_s\} \text{ est un convexe compact (de } \mathbb{R}^n),$$
$$K = \text{cône}\,\{v_1, \ldots, v_r\} \text{ un cône convexe fermé,}$$
$$L = \operatorname{vect}\{w_1, \ldots, w_p\} \text{ un sous-espace vectoriel.}$$

On considère le programme linéaire suivant :

$$(\mathcal{P}) \quad \begin{cases} \text{Minimiser } \langle c, x\rangle\,, \\ x \in C \end{cases}$$

où $c \in \mathbb{R}^n$ est donné.

1°) Montrer que la fonction $f : x \longmapsto f(x) := \langle c, x\rangle$ est bornée inférieurement sur C si, et seulement si,

$$(\mathcal{C}) \quad \begin{cases} \langle c, v_i\rangle \geqslant 0 \text{ pour tout } i = 1, \ldots, r \text{ et} \\ \langle c, w_j\rangle = 0 \text{ pour tout } j = 1, \ldots, p. \end{cases}$$

2°) Montrer que sous la condition $(\mathcal{C})$, l'ensemble Π des solutions de $(\mathcal{P})$ est décrit comme suit :

$$\Pi = \Big\{ \overline{x} \in C \mid \overline{x} = \sum_{i\in I_1} \overline{\alpha}_i u_i + \sum_{j\in J_2} \overline{\beta}_j v_j + \sum_{k=1}^{p} \overline{\gamma}_k w_k \ ; \quad \overline{\alpha}_i \geqslant 0, \sum_{i\in I_1} \overline{\alpha}_i = 1,\ \overline{\beta}_j \geqslant 0,\ \overline{\gamma}_k \in \mathbb{R} \Big\}$$

où $I_1 := \{ i \mid \langle c, u_i\rangle = \overline{f} \}$, $J_2 := \{ j \mid \langle c, v_j\rangle = 0 \}$, $\overline{f} := \inf_{x\in C} f(x)$.

3°) On suppose $K = L = \{0\}$, c'est-à-dire $C = C_0$ borné, et $\Pi \neq C$.

On pose alors : $\alpha := \min\limits_{i\notin I_1} \dfrac{\langle c, u_i\rangle - \overline{f}}{d_\Pi(u_i)}$, où d_Π désigne la fonction-distance à Π.

Montrer que : $\quad \forall x \in C_0,\ d_\Pi(x) \leqslant \frac{\langle c,x\rangle - \overline{f}}{\alpha}$ (5.29)

(on pourra utiliser la convexité de la fonction d_Π).

On admettra pour la suite que le résultat de cette question est encore valable pour un C décrit en (5.28) (non nécessairement borné), et que tout C décrit comme en (5.27) peut être représenté sous la forme (5.28).

3e partie. On revient à la représentation (5.27) de C, et on considère le programme linéaire suivant (dans $\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$) :

$$(PL) \quad \begin{cases} \text{Minimiser } \sum\limits_{i=1}^{m} t_i \\ \langle a_i, x\rangle - b_i = t_i - z_i \text{ pour } i = 1, \ldots, m \\ (\text{les } a_i \text{ désignent les vecteurs-lignes de } A) \\ t_i \geqslant 0,\ z_i \geqslant 0 \quad \text{pour} \quad i = 1, \ldots, m. \end{cases}$$

1°) Vérifier que pour tout $x \in \mathbb{R}^n$, le vecteur $(x, (Ax-b)^+, (b-Ax)^+)$ est admissible pour (PL).

2°) Établir que toute solution de (PL) est de la forme $(x, 0, b-Ax)$, où $x \in C$.

En déduire, à l'aide du résultat de la 3e question de la 2e partie, qu'il existe $\alpha > 0$ tel que :

$$\forall x \in \mathbb{R}^n,\ d_C(x) \leqslant \frac{1}{\alpha} \sum_{i=1}^{m} (\langle a_i, x\rangle - b_i)^+ . \tag{5.30}$$

Solution : *1re partie*

1°) *a*) Soit x_0 quelconque dans C. On rappelle :

$$(d \in C_\infty) \Leftrightarrow (x_0 + td \in C \text{ pour tout } t > 0)\,.$$

Dans le cas présent, $d \in C_\infty$ si et seulement si $Ax_0 + tAd \leqslant b$ pour tout $t > 0$.

De façon évidente, $(Ad \leqslant 0) \Rightarrow (d \in C_\infty)$. Réciproquement, soit $d \in C_\infty$ et supposons $(Ad)_{i_0} > 0$ pour un certain i_0. Alors

$$(Ax_0 + td)_{i_0} = (Ax_0)_{i_0} + t\,(Ad)_{i_0} \leqslant b \text{ pour tout } t > 0$$

est contredit.

b) (C borné) $\Leftrightarrow$ ($C_\infty = \{0\}$). Pour obtenir le résultat annoncé, il suffit de traduire ceci avec l'expression de C_∞ obtenue précédemment.

2°) Soient $a_1, \ldots, a_m$ les vecteurs-lignes de A. Dire que « $A^\top y = c$ a, pour tout $c \in \mathbb{R}^n$, une solution $y \geqslant 0$ », revient à dire que « cône$\{a_1, \ldots, a_m\} = \mathbb{R}^n$ ». Grâce au lemme de Minkowski-Farkas, cela équivaut à

$$\{d \in \mathbb{R}^n \mid \langle a_i, d\rangle \leqslant 0 \text{ pour tout } i = 1, \ldots, m\} = \{0\}\,.$$

2e partie. Soit $x \in C$,

$$x = \sum_{i=1}^{s} \alpha_i u_i + \sum_{j=1}^{r} \beta_j v_j + \sum_{k=1}^{p} \gamma_k w_k,$$

avec :

$$\begin{aligned} &\alpha_i \geqslant 0 \quad \text{pour tout } i,\ \alpha_1 + \ldots + \alpha_s = 1\ ; \\ &\beta_j \geqslant 0 \quad \text{pour tout } j\ ;\ \ \gamma_k \in \mathbb{R} \text{ pour tout } k. \end{aligned} \tag{5.31}$$

1°) Puisque $\langle c, x\rangle = \sum\limits_i \alpha_i \langle c, u_i\rangle + \sum\limits_j \beta_j \langle c, v_j\rangle + \sum\limits_k \gamma_k \langle c, w_k\rangle$ avec les contraintes sur α_i et β_j énoncées ci-dessus, il vient facilement :

$$\big(\langle c, x\rangle \geqslant \overline{f} \text{ pour tout } x \in C\big) \Leftrightarrow \begin{pmatrix} \langle c, v_j\rangle \geqslant 0 \text{ pour tout } j = 1, \ldots, r \\ \text{et} \\ \langle c, w_k\rangle = 0 \text{ pour tout } k = 1, \ldots, p \end{pmatrix}.$$

Ceci est une manière de dire que $-c$ doit se trouver dans le cône normal à K et dans le sous-espace orthogonal à L.

2°) On suppose que la condition ($\mathcal{C}$) est vérifiée. Le problème ($\mathcal{P}$) a alors des solutions ; il est, en effet, clair que tout $\overline{x} \in \Pi$ est solution de ($\mathcal{P}$) car $\langle c, \overline{x}\rangle = \sum\limits_{i \in I_1} \overline{\alpha}_i \langle c, u_i\rangle = \overline{f}$.

Réciproquement, soit $\overline{x} = \sum_i \overline{\alpha}_i u_i + \sum_j \overline{\beta}_j v_j + \sum_k \overline{\gamma}_k w_k$ une solution de $(\mathcal{P})$; on a :

$$\sum_i \alpha_i \langle c, u_i\rangle + \sum_j \beta_j \langle c, v_j\rangle \geqslant \sum_i \overline{\alpha}_i \langle c, u_i\rangle + \sum_j \overline{\beta}_j \langle c, v_j\rangle = \overline{f}$$

pour tout $(\alpha_1, \ldots, \alpha_r)$ et $(\beta_1, \ldots, \beta_s)$ vérifiant les conditions décrites en (5.31).

Avoir $\overline{\beta}_j > 0$ et $\langle c, v_j\rangle > 0$ est impossible car cela contredirait le caractère optimal de $\overline{x}$ (on construirait facilement un $\tilde{x} \in C$ qui ferait « mieux » que $\overline{x}$, $\langle c, \tilde{x}\rangle < \overline{f}$).

Ensuite, sachant que $\langle c, u_i\rangle \geqslant \overline{f}$ pour tout i, l'égalité $\sum_i \overline{\alpha}_i \left[\langle c, u_i\rangle - \overline{f}\right] = 0$ ne souffre pas qu'on puisse avoir simultanément $\overline{\alpha}_i > 0$ et $\langle c, u_i\rangle - \overline{f} > 0$.

Donc $\overline{x}$ est bien dans Π.

3°) Ici $C = C_0 = \operatorname{conv}\{u_1, \ldots, u_s\}$. Puisque $\Pi \neq C$, il existe un i pour lequel $\langle c, u_i\rangle > \overline{f}$. De par la convexité de Π, lorsque $x = \sum_{i=1}^{s} \alpha_i u_i \in C_0$,

$$d_\Pi\left(\sum_{i=1}^{s} \alpha_i u_i\right) \leqslant \sum_{i=1}^{s} \alpha_i d_\Pi(u_i).$$

Nous avons deux cas possibles :

$i \in I_1$, ce qui revient à $u_i \in \Pi$, auquel cas $d_\Pi(u_i) = 0$;

$i \notin I_1$, auquel cas $d_\Pi(u_i) \leqslant \frac{\langle c, u_i\rangle - \overline{f}}{\alpha}$.

En conséquence,

$$d_\Pi(x) \leqslant \sum_{i \notin I_1} \alpha_i \frac{\langle c, u_i\rangle - \overline{f}}{\alpha} \leqslant \frac{\langle c, x\rangle - \overline{f}}{\alpha}.$$

3e partie

Le programme linéaire (PL) est posé dans $\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$ en les variables $x_1, \ldots, x_n$, $t_1, \ldots, t_m$, $z_1, \ldots, z_m$.

1°) Pour x quelconque dans $\mathbb{R}^n$, posons pour $i = 1, \ldots, m$

$$t_i := (\langle a_i, x\rangle - b_i)^+, \quad z_i := (b_i - \langle a_i, x\rangle)^+.$$

Le vecteur

$$\big(x_1, \ldots, x_n, (\langle a_1, x\rangle - b_1)^+, \ldots, (\langle a_m, x\rangle - b_m)^+, (b_1 - \langle a_1, x\rangle)^+, \ldots, (b_m - \langle a_m, x\rangle)^+\big)$$

est admissible pour (PL).

2°) On a par définition même de C : $(x \in C) \Leftrightarrow ((Ax - b)^+ = 0)$.

Si $x \in C$, *i.e.* si $t := (Ax-b)^+ = 0$ et $z := b - Ax$, le point $(x, 0, b-Ax)$ est solution de (PL) puisque la valeur de la fonction-objectif (toujours positive) en ce point est nulle. Réciproquement, en une solution $(\overline{x}, \overline{t}, \overline{z})$ de (PL), on doit avoir $\langle a_i, \overline{x}\rangle - b_i = \overline{t}_i - \overline{z}_i \ \left(= (\langle a_i, \overline{x}\rangle - b_i)^+ - (b_i - \langle a_i, \overline{x}\rangle)^+\right)$ pour tout $i = 1, \dots, m$; si $\langle a_i, \overline{x}\rangle - b_i > 0$ pour un certain i, on pourrait faire « mieux » que $(\overline{x}, \overline{t}, \overline{z})$ en prenant $\tilde{x} \in C$ et le point $(\tilde{x}, \tilde{t} := 0, \tilde{z} := b - A\tilde{x})$ correspondant. Donc $\overline{x} \in C$ en fait et $\overline{t} = 0,\ \overline{z} = b - A\overline{x}$.

La valeur du programme (PL) est 0 et la valeur de sa fonction-objectif en un point $\left(x, (Ax-b)^+, (b-Ax)^+\right)$, $x \in \mathbb{R}^n$, est $\sum_{i=1}^{m} (\langle a_i, x\rangle - b_i)^+$. Il résulte alors de la 3e question de la 2e partie, l'existence de $\alpha > 0$ tel que :

$$\forall x \in \mathbb{R}^n, \qquad d_C(x) \leqslant d_\Pi\left(x, (Ax-b)^+, (b-Ax)^+\right) \leqslant \frac{1}{\alpha} \sum_{i=1}^{m} (\langle a_i, x\rangle - b_i)^+ .$$

***Exercice V.20.** Soit $C := \{x \in \mathbb{R}^n \mid Ax \leqslant b,\ x \geqslant 0\}$ un polyèdre convexe que l'on suppose borné. On considère le problème *d'optimisation fractionnaire* suivant :

$$(\mathcal{P}) \quad \begin{cases} \text{Min } \dfrac{\langle c, x\rangle + \gamma}{\langle d, x\rangle + \delta} \\ x \in C \end{cases},$$

où c et d sont des vecteurs de $\mathbb{R}^n$, γ et δ des réels, et où on suppose que $\langle d, x\rangle + \delta > 0$ pour tout $x \in C$.

On se propose de résoudre $(\mathcal{P})$ par l'intermédiaire du programme linéaire (en la variable $(c, z) \in \mathbb{R}^n \times \mathbb{R}$) suivant :

$$\left(\hat{\mathcal{P}}\right) \quad \begin{cases} \text{Min } \langle c, y\rangle + \gamma z \\ Ay - zb \leqslant 0 \\ \langle d, y\rangle + \delta z = 1 \\ y \geqslant 0,\ z \geqslant 0. \end{cases}$$

1°) Montrer que si (y, z) est admissible pour $(\hat{\mathcal{P}})$, alors $z > 0$ nécessairement.

2°) Démontrer que si $(\overline{y}, \overline{z})$ est une solution de $(\hat{\mathcal{P}})$, alors $\overline{x} := \overline{y}/\overline{z}$ est une solution de $(\mathcal{P})$.

Solution : 1°) Soit (y,z) admissible pour $(\hat{\mathcal{P}})$; $(y,z) \neq (0,0)$ en raison de la contrainte $\langle d,y\rangle + \delta z = 1$; donc si on suppose $z=0$, alors $y \neq 0$. Mais alors on aurait un $y \neq 0$ vérifiant $y \geqslant 0$ et $Ay \leqslant 0$, c'est-à-dire une direction non nulle dans le cône asymptote de C. Ce qui contredit le caractère borné de C.

2°) Si $(\overline{y},\overline{z})$ est une solution de $(\hat{\mathcal{P}})$, $\overline{x} := \overline{y}/\overline{z} \geqslant 0$ et $A\overline{x} \leqslant b$. Donc $\overline{x} \in C$. Montrons que $\overline{x}$ est effectivement une solution de $(\mathcal{P})$.

Considérons $x \in C$; alors le couple $\left(\frac{x}{\langle d,x\rangle+\delta}, \frac{1}{\langle d,x\rangle+\delta}\right)$ est admissible pour $(\hat{\mathcal{P}})$, et comme $(\overline{y},\overline{z})$ est une solution de $(\hat{\mathcal{P}})$,

$$\langle c,\overline{y}\rangle + \gamma\overline{z} \leqslant \left\langle c, \frac{x}{\langle d,x\rangle+\delta}\right\rangle + \frac{\gamma}{\langle d,x\rangle+\delta} = \frac{\langle c,x\rangle+\gamma}{\langle d,x\rangle+\delta}.$$

Divisons le membre de gauche par $1 = \langle d,\overline{y}\rangle + \delta\overline{z}$ pour obtenir

$$\langle c,\overline{y}\rangle + \gamma\overline{z} = \frac{\langle c,\overline{y}\rangle+\gamma\overline{z}}{\langle d,\overline{y}\rangle+\delta\overline{z}} = \frac{\langle c,\overline{x}\rangle+\gamma}{\langle d,\overline{x}\rangle+\delta}.$$

D'où le résultat annoncé.

***Exercice V.21.** Considérons le programme linéaire $(\mathcal{P})$ suivant dans $\mathbb{R}^5$:

$$(\mathcal{P}) \quad \begin{cases} \text{Max } \langle c,x\rangle \\ Ax \leqslant b \\ x \geqslant 0 \end{cases}, \text{ où } A = \begin{bmatrix} 2&1&1&0&0\\ 1&2&0&1&0\\ 0&1&0&0&1 \end{bmatrix}, \; c = \begin{pmatrix} 4\\5\\0\\0\\0 \end{pmatrix}, \; b = \begin{pmatrix} 8\\7\\3 \end{pmatrix}.$$

Quel est le problème dual $(\mathcal{D})$ de $(\mathcal{P})$?

Vérifier que $\overline{x} = (3,2,0,0,1)$ et $\overline{y} = (1,2,0)$ sont solutions de $(\mathcal{P})$ et $(\mathcal{D})$ respectivement.

Solution : Le problème dual $(\mathcal{D})$ de $(\mathcal{P})$ s'écrit :

$$(\mathcal{D}) \quad \begin{cases} \text{Min } \langle b,y\rangle \\ A^\top y \geqslant c \\ y \geqslant 0. \end{cases}$$

On constate que les $\overline{x}$ et $\overline{y}$ proposés sont admissibles pour $(\mathcal{P})$ et $(\mathcal{D})$ respectivement, avec, de plus, l'égalité $\langle c,\overline{x}\rangle = \langle b,\overline{y}\rangle = 22$. En conséquence, $\overline{x}$ et $\overline{y}$ sont solutions de $(\mathcal{P})$ et $(\mathcal{D})$ respectivement.

***Exercice V.22.** Montrer que si le programme linéaire

$$(\mathcal{P}) \quad \begin{cases} \text{Max } \langle c, x \rangle \\ Ax = 0, \ x \geqslant 0 \end{cases}$$

a une valeur optimale finie, alors $\overline{x} = 0$ est certainement solution de $(\mathcal{P})$.

Solution : L'ensemble-contrainte de $(\mathcal{P})$ n'est pas vide (il contient 0) et, par hypothèse, $\text{val}(\mathcal{P}) < +\infty$.

Le problème dual de $(\mathcal{P})$ s'écrit :

$$(\mathcal{D}) \quad \begin{cases} \text{Min } \langle 0, y \rangle \\ A^\top y \geqslant c \end{cases},$$

et donc $0 = \text{val}(\mathcal{D}) = \text{val}(\mathcal{P})$.

Par conséquent, $\overline{x} = 0$ est une solution de $(\mathcal{P})$.

****Exercice V.23.** Soit le programme linéaire suivant dans $\mathbb{R}^4$:

$$(\mathcal{P}_\alpha) \quad \begin{cases} \text{Min } 3x_1 + 4x_2 + x_3 + x_4 \\ x_1 + x_2 + x_3 - x_4 \geqslant 2 \\ -2x_1 + 2x_2 + x_3 + x_4 \geqslant \alpha \\ x_1 \geqslant 0, \ldots, x_4 \geqslant 0 \end{cases}, \text{ où } \alpha \text{ est un paramètre réel.}$$

1°) Écrire le problème dual $(\mathcal{D}_\alpha)$ de $(\mathcal{P}_\alpha)$.

2°) Résoudre $(\mathcal{D}_\alpha)$ suivant les valeurs de α ; en déduire les solutions de $(\mathcal{P}_\alpha)$.

Solution : 1°) $(\mathcal{D}_\alpha)$ se formule comme suit :

$$(\mathcal{D}_\alpha) \quad \begin{cases} \text{Max } 2y_1 + \alpha y_2 \\ y_1 - 2y_2 \leqslant 3, \ y_1 + 2y_2 \leqslant 4, \\ y_1 + y_2 \leqslant 1, \ -y_1 + y_2 \leqslant 1, \\ y_1 \geqslant 0, \ y_2 \geqslant 0. \end{cases}$$

2°) $(\mathcal{D}_\alpha)$ est un programme linéaire dans $\mathbb{R}^2$; il se résout graphiquement :

Valeur du paramètre	Solutions	Valeur optimale
$\alpha < 2$	$\overline{y} = (1,0)$	2
$\alpha = 2$	$\{\overline{y} = (y_1, 1-y_1) \mid 0 \leqslant y_1 \leqslant 1\}$	2
$\alpha > 2$	$\overline{y} = (0,1)$	α

On est dans une situation où val$(\mathcal{P}_\alpha)$ = val $(\mathcal{D}_\alpha)$.

Les deux premières contraintes de $(\mathcal{D}_\alpha)$ sont inactives en $\overline{y}$ solution de $(\mathcal{D}_\alpha)$; donc $\overline{x}_1 = \overline{x}_2 = 0$ nécessairement pour toute solution $\overline{x} = (\overline{x}_1, \overline{x}_2, \overline{x}_3, \overline{x}_4)$ de $(\mathcal{P}_\alpha)$.

Sachant cela et connaissant val$(\mathcal{P}_\alpha)$, on détermine $\overline{x}_3$ et $\overline{x}_4$.

• $\alpha < 2$. On cherche $\overline{x}_3$ et $\overline{x}_4$ tels que

$\overline{x}_3 + \overline{x}_4 = 2,\ \overline{x}_3 - \overline{x}_4 \geqslant 2,$

$\overline{x}_3 + \overline{x}_4 \geqslant \alpha$

$\overline{x}_3 \geqslant 0,\ \overline{x}_4 \geqslant 0.$

Il s'ensuit $\overline{x}_3 = 2$ et $\overline{x}_4 = 0$. La seule solution de $(\mathcal{P}_\alpha)$ est donc $\overline{x} = (0,0,2,0)$.

• $\alpha \geqslant 2$. Les solutions $\overline{x}$ de $(\mathcal{P}_\alpha)$ sont de la forme

$$\overline{x} = (0,0,\beta,\alpha-\beta)\text{ , avec } 1 + \frac{\alpha}{2} \leqslant \beta \leqslant \alpha.$$

C'est donc une arête du polyèdre des contraintes, qui se réduit à un sommet pour $\alpha = 2$.

***Exercice V.24.** 0n considère le programme linéaire suivant :

$$(\mathcal{P}) \quad \begin{cases} \text{Minimiser } x_1 + 2x_2 + 3x_3 + \ldots + nx_n \\ x_1 \geqslant 1,\ x_1 + x_2 \geqslant 2,\ x_1 + x_2 + x_3 \geqslant 3, \ldots, x_1 + x_2 + \ldots + x_n \geqslant n, \\ x_i \geqslant 0 \quad \text{pour tout} \quad i = 1, 2, \ldots, n. \end{cases}$$

1°) Décrire d'une manière détaillée le problème dual $(\mathcal{D})$ de $(\mathcal{P})$.

2°) Montrer que tout élément admissible $y = (y_1, \ldots, y_n)$ de $(\mathcal{D})$ (et donc toute solution de $(\mathcal{D})$) vérifie

$$y_k + y_{k+1} + \ldots + y_n < k \quad \text{pour tout } k = 2, \ldots, n.$$

Déduire de ce qui précède et des relations liant les solutions de $(\mathcal{P})$ et de $(\mathcal{D})$ les valeurs $\overline{x}_2, \overline{x}_3, \ldots, \overline{x}_n$ de toute solution $\overline{x} = (\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n)$ de $(\mathcal{P})$.

3°) Résoudre complètement $(\mathcal{P})$.

Solution : Commençons par visualiser le cas où $n = 2$.

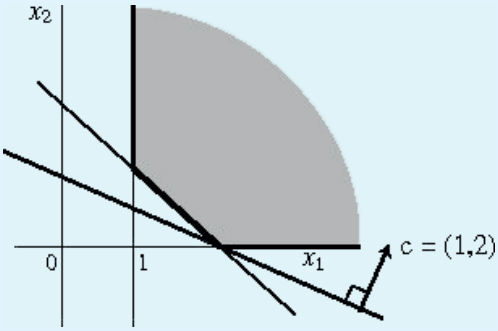


FIGURE 14.

$(\mathcal{P})$ consiste à minimiser $\langle c, x\rangle$ sous les contraintes $Ax \geqslant b,\ x \geqslant 0$, où

$$b = \begin{pmatrix} 1 \\ 2 \\ \vdots \\ n \end{pmatrix},\ c = \begin{pmatrix} 1 \\ 2 \\ \vdots \\ n \end{pmatrix} \quad \text{et} \quad A = \begin{bmatrix} 1 & 0 & 0 & \ldots & \ldots \\ 1 & 1 & 0 & \ldots & \ldots \\ 1 & 1 & 1 & 0 & \ldots \\ \vdots & & & \ddots & \\ 1 & 1 & 1 & \ldots & 1 \end{bmatrix}.$$

1°) Le problème dual de $(\mathcal{P})$ consiste à maximiser $\langle b, y\rangle$ sous les contraintes $A^\top y \leqslant c$ et $y \geqslant 0$, soit :

$$(\mathcal{D}) \begin{cases} \text{Maximiser } y_1 + 2y_2 + \ldots + ny_n \\ y_1 + y_2 + \ldots + y_n \leqslant 1 \\ y_2 + \ldots + y_n \leqslant 2 \\ y_3 + \ldots + y_n \leqslant 3 \\ \ldots \\ y_n \leqslant n \\ y_i \geqslant 0 \text{ pour tout } i = 1, 2, \ldots, n. \end{cases}$$

$(1,1,\dots,1)$ est admissible pour $(\mathcal{P})$, $(0,\dots,0)$ est admissible pour $(\mathcal{D})$: donc

$$\mathrm{val}(\mathcal{P}) = \mathrm{val}(\mathcal{D}) \in \mathbb{R}.$$

2°) Soit $y = (y_1,\dots,y_n)$ admissible pour $(\mathcal{D})$. Pour $k \geqslant 2$,

$$\begin{aligned} y_k + y_{k+1} + \dots + y_n &\leqslant y_1 + \dots + y_n \quad \text{(puisque tous les } y_i \text{ sont } \geqslant 0) \\ &\leqslant 1 \qquad\qquad \text{(d'après la } 1^{\text{re}} \text{ contrainte de type inégalité dans } (\mathcal{D}).) \\ &< k. \end{aligned}$$

D'après les relations de complémentarité liant les solutions du problème primal $(\mathcal{P})$ et celles du dual $(\mathcal{D})$, on a :

$$\begin{pmatrix} \overline{x} = (\overline{x}_1,\dots,\overline{x}_n) \text{ solution de } (\mathcal{P}) \\ \text{et} \\ \overline{y} = (\overline{y}_1,\dots,\overline{y}_n) \text{ solution de } (\mathcal{D}) \end{pmatrix} \Rightarrow \begin{pmatrix} [\langle a'_k, \overline{y}\rangle - c_k] \cdot \overline{x}_k = 0 \\ \text{pour tout } k = 1,\dots,n \end{pmatrix}.$$

Or, ici, $\langle a'_k, \overline{y}\rangle = y_k + y_{k+1} + \dots + y_n < k = c_k$ pour tout $k = 2,\dots,n$. En conséquence, $\overline{x}_k = 0$ pour tout $k = 2,\dots,n$.

3°) Si $\overline{x}$ est solution de $(\mathcal{P})$, $\langle c, \overline{x}\rangle = \overline{x}_1$. Résoudre $(\mathcal{P})$ revient donc à minimiser x_1 sous les contraintes $x_1 \geqslant 1$, $x_1 \geqslant 2,\dots,x_1 \geqslant n$ et $x_1 \geqslant 0$; d'où $\overline{x}_1 = n$.

La solution de $(\mathcal{P})$ est donc $\overline{x} = (n,0,\dots,0)$, et $\mathrm{val}(\mathcal{P}) = \mathrm{val}(\mathcal{D}) = n$.

****** Exercice V.25.*** On considère le programme linéaire suivant

$$(\mathcal{P}) \begin{cases} \text{Minimiser } \langle c, x\rangle \\ Ax = b \\ x \geqslant 0 \end{cases}, \quad \text{où } c \in \mathbb{R}^n,\ b \in \mathbb{R}^m \text{ et } A \in \mathcal{M}_{m,n}(\mathbb{R}),$$

et son problème dual

$$(\mathcal{D}) \quad \begin{cases} \text{Maximiser } \langle b, y\rangle \\ A^\top y \leqslant c \end{cases}$$

présenté sous la forme standard dans $\mathbb{R}^m \times \mathbb{R}^n$:

$$(\tilde{\mathcal{D}}) \quad \begin{cases} \text{Maximiser } \langle b, y\rangle \\ A^\top y + u = c \qquad (y \in \mathbb{R}^m,\ u \in \mathbb{R}^n). \\ u \geqslant 0 \end{cases}$$

On suppose que les ensembles-contraintes de $(\mathcal{P})$ et de $(\mathcal{D})$ ne sont pas vides (hypothèse qui sera renforcée par la suite) et que A est de rang m.

1°) Vérifier que si $\overline{x}$ est admissible pour $(\mathcal{P})$ et si $(\overline{y},\overline{u})$ est admissible pour $(\tilde{\mathcal{D}})$, alors :

$$\langle \overline{x},\overline{u}\rangle \geqslant 0,$$

et

$$(\langle \overline{x},\overline{u}\rangle = 0) \Leftrightarrow (\overline{x} \text{ est solution de } (\mathcal{P}) \text{ et } (\overline{y},\overline{u}) \text{ est solution de } (\tilde{\mathcal{D}})).$$

En déduire la caractérisation suivante des solutions de $(\mathcal{P})$ et de $(\tilde{\mathcal{D}})$:

$$\begin{pmatrix} \overline{x} \text{ est solution de } (\mathcal{P}) \\ \text{et} \\ (\overline{y},\overline{u}) \text{ est solution de } (\tilde{\mathcal{D}}) \end{pmatrix} \Leftrightarrow \begin{pmatrix} A\overline{x} = b,\ \overline{x} \geqslant 0 \\ A^\top \overline{y} + \overline{u} = c,\ \overline{u} \geqslant 0 \\ \langle \overline{x},\overline{u}\rangle = 0 \end{pmatrix}. \qquad (\mathcal{SO})$$

Désignons par C l'association des ensembles-contraintes de $(\mathcal{P})$ et de $(\tilde{\mathcal{D}})$, *i.e.*

$$C := \left\{(x,y,u) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \mid Ax = b,\ x \geqslant 0,\ A^\top y + u = c,\ u \geqslant 0\right\},$$

et par $(\mathcal{P}\tilde{\mathcal{D}})$ le couplage des problèmes $(\mathcal{P})$ et $(\tilde{\mathcal{D}})$ suivant :

$$(\mathcal{P}\tilde{\mathcal{D}}) \quad \begin{cases} \text{Minimiser } \langle x,u\rangle \\ (x,y,u) \in C \end{cases} \quad \text{(appelé problème } \textit{primal-dual}\text{)}.$$

Pour toute la suite on fait les hypothèses que voici : A est de rang m et il existe $(x,y,u) \in C$ tel que $x > 0$ et $u > 0$ ($v > 0$ dans $\mathbb{R}^n$ signifie que toutes les composantes v_i de v sont > 0).

Étant donné $\sigma > 0$, on propose une approximation de $(\mathcal{P}\tilde{\mathcal{D}})$ par pénalisation intérieure (ou addition d'une fonction-barrière) définie comme suit :

$$(\mathcal{P}\tilde{\mathcal{D}})_\sigma \begin{cases} \text{Minimiser } f_\sigma(x,y,u) \\ (x,y,u) \in C_0, \end{cases}$$

où $C_0 := \{(x,y,u) \in C \mid x > 0 \text{ et } u > 0\}$ et $f_\sigma : (x,y,u) \in C_0 \longmapsto f_\sigma(x,y,u) := \langle x,u\rangle - \sigma \sum_{i=1}^{n} \ln(x_i u_i)$.

2°) Montrer que $(\mathcal{P}\tilde{\mathcal{D}})_\sigma$ a au plus une solution et que cette solution est caractérisée comme suit :

$$\begin{pmatrix} (\overline{x}(\sigma),\overline{y}(\sigma),\overline{u}(\sigma)) \text{ est} \\ \text{solution de } (\mathcal{P}\tilde{\mathcal{D}})_\sigma \end{pmatrix} \Leftrightarrow \begin{pmatrix} A\overline{x}(\sigma) = b \\ A^\top \overline{y}(\sigma) + \overline{u}(\sigma) = c \\ \overline{x}(\sigma)_i \overline{u}(\sigma)_i = \sigma \text{ pour tout } i \end{pmatrix}. \qquad (\mathcal{SO})_\sigma$$

On admettra que, sous les hypothèses qui ont été faites, $(\mathcal{P}\tilde{\mathcal{D}})_\sigma$ a effectivement une solution $(\overline{x}(\sigma), \overline{y}(\sigma), \overline{u}(\sigma))$ et que $(\overline{x}(\sigma), \overline{y}(\sigma), \overline{u}(\sigma))$ a une limite quand $\sigma \to 0$, limite qui sera notée (x^*, y^*, u^*).

3°) Montrer que (x^*, y^*, u^*) est solution de $(\mathcal{P}\tilde{\mathcal{D}})$ et que, de plus :
« Pour tout $i = 1, \ldots, n$, l'une des deux composantes du couple (x_i^*, u_i^*) est nulle tandis que l'autre est strictement positive » .

4°) *Illustration.* Considérons le programme linéaire suivant dans $(\mathbb{R}^3)$:

$$(\mathcal{P}) \quad \begin{cases} \text{Minimiser } x_1 + x_2 + x_3 \\ -x_1 + x_2 = 0 \\ x_3 = 1 \\ x_1 \geqslant 0,\ \ x_2 \geqslant 0,\ \ x_3 \geqslant 0. \end{cases}$$

Illustrer sur cet exemple les résultats obtenus dans les questions précédentes en déterminant $(\overline{x}(\sigma), \overline{y}(\sigma), \overline{u}(\sigma))$ ainsi que (x^*, y^*, u^*).

Solution : On rappelle au préalable le lien qui existe entre les solutions de $(\mathcal{D})$ et celles de $(\tilde{\mathcal{D}})$:

$$(\overline{y} \text{ est solution de } (\mathcal{D})) \Rightarrow ((\overline{y}, c - A^\top \overline{y}) \text{ est solution de } (\tilde{\mathcal{D}}))\,;$$
$$((\overline{y}, \overline{u}) \text{ est solution de } (\tilde{\mathcal{D}})) \Rightarrow (\overline{u} = c - A^\top \overline{y} \text{ et } \overline{y} \text{ est solution de } (\mathcal{D})).$$

Notons aussi qu'en raison du caractère injectif de $A^\top$ (dû au fait que A est surjective),

$$\begin{pmatrix} (\overline{y}_1, \overline{u}) \text{ solution de } (\tilde{\mathcal{D}}) \\ \text{et} \\ (\overline{y_2}, \overline{u}) \text{ solution de } (\tilde{\mathcal{D}}) \end{pmatrix} \Rightarrow (\overline{y}_1 = \overline{y}_2)\,.$$

1°) Soit $\overline{x}$ admissible pour $(\mathcal{P})$ et $(\overline{y}, \overline{u})$ admissible pour $(\tilde{\mathcal{D}})$. Alors $\overline{u} = c - A^\top \overline{y}$ de sorte que

$$\begin{aligned} \langle \overline{x}, \overline{u} \rangle &= \left\langle \overline{x}, c - A^\top \overline{y} \right\rangle = \langle \overline{x}, c \rangle - \langle A\overline{x}, \overline{y} \rangle \\ &= \langle \overline{x}, c \rangle - \langle b, \overline{y} \rangle\,. \end{aligned}$$

Mais puisque $\overline{y}$ est admissible pour $(\mathcal{D})$, $\langle c, \overline{x} \rangle \geqslant \langle b, \overline{y} \rangle$, d'où $\langle \overline{x}, \overline{u} \rangle \geqslant 0$.

Avoir $\langle \overline{x}, \overline{u} \rangle = 0$ équivaut à avoir $\langle c, \overline{x} \rangle = \langle b, \overline{y} \rangle$. Ceci traduit le fait que $\overline{x}$ est solution de $(\mathcal{P})$ et $\overline{y}$ est solution de $(\mathcal{D})$ (soit encore $(\overline{y}, \overline{u} = c - A^\top \overline{y})$ est solution de $(\tilde{\mathcal{D}})$).

La caractérisation $(\mathcal{SO})$ des solutions de $(\mathcal{P})$ et de $(\tilde{\mathcal{D}})$ n'est autre que la formulation synthétique de ce qui vient d'être démontré.

2°) Les problèmes $(\mathcal{P}\tilde{\mathcal{D}})$ et $(\mathcal{P}\tilde{\mathcal{D}})_\sigma$ ne sont plus linéaires car les fonctions-objectifs ne le sont pas. Les « vraies » variables y sont x et u car, comme cela a déjà été dit, la connaissance de u induit celle de y.

On a $\langle x, u\rangle = \langle c, x\rangle - \langle b, y\rangle$ lorsque $(x, y, u) \in C_0$ (*cf.* 1re question si nécessaire) de sorte que

$$f_\sigma(x, y, u) = \langle c, x\rangle - \langle b, y\rangle - \sigma\left(\sum_{i=1}^{n} \ln\, x_i + \sum_{i=1}^{n} \ln\, u_i\right).$$

La stricte concavité de la fonction ln fait donc que f_σ est strictement convexe sur C_0. Ainsi le problème de minimisation convexe $(\mathcal{P}\tilde{\mathcal{D}})_\sigma$ a au plus une solution.

Au point $(\overline{x}(\sigma), \overline{y}(\sigma), \overline{u}(\sigma))$ solution de $(\mathcal{P}\tilde{\mathcal{D}})_\sigma$, on a $\overline{x}(\sigma) > 0$ et $\overline{u}(\sigma) > 0$, de sorte que seules les deux contraintes du type égalité (affine) $Ax = b$ et $A^\top y + u = c$ sont à prendre en compte dans l'expression de la condition nécessaire et suffisante de minimalité. À cet effet :

$$\nabla f_\sigma(x, y, u) = \underbrace{\left(u_1 - \frac{\sigma}{x_1}, \ldots, u_n - \frac{\sigma}{x_n}\right)}_{\in \mathbb{R}^n},\ \underbrace{0, \ldots, 0}_{\in \mathbb{R}^m},\ \underbrace{x_1 - \frac{\sigma}{u_1}, \ldots, x_n - \frac{\sigma}{u_n}}_{\in \mathbb{R}^n}.$$

L'ensemble-contrainte (utile) de $(\mathcal{P}\tilde{\mathcal{D}})_\sigma$ s'écrit sous la forme $\mathcal{A}(x, y, u) = (b, c)$, où

$$\mathcal{A} := \begin{array}{|c|c|c|} \multicolumn{1}{c}{n} & \multicolumn{1}{c}{m} & \multicolumn{1}{c}{n} \\ \hline A & 0 & 0 \\ \hline 0 & A^\top & I \\ \hline \end{array} \begin{array}{l} \\ m \\ n \end{array}$$

La condition nécessaire et suffisante de minimalité

$$\nabla f_\sigma(x, y, u) \in \mathrm{Im}\mathcal{A}^\top$$

se décompose en :

$$\begin{cases} \left(u_1 - \frac{\sigma}{x_1}, \dots, u_n - \frac{\sigma}{x_n}\right) \in \operatorname{Im} A^\top \\ \text{et} \\ \left(x_1 - \frac{\sigma}{u_1}, \dots, x_n - \frac{\sigma}{u_n}\right) \in \operatorname{Ker} A. \end{cases} \tag{5.32}$$

Pour alléger l'écriture, notons
$\frac{u}{\sigma} - \frac{1}{x}$ le vecteur de $\mathbb{R}^n$ de composantes $\frac{u_i}{\sigma} - \frac{1}{x_i}$,
$\frac{x}{\sigma} - \frac{1}{u}$ le vecteur de $\mathbb{R}^n$ de composantes $\frac{x_i}{\sigma} - \frac{1}{u_i}$,
w le vecteur de $\mathbb{R}^n$ de composantes $\sqrt{\frac{x_i u_i}{\sigma}}$.

Soit $\triangle := \operatorname{diag}\left(\sqrt{\frac{x_1}{u_1}}, \dots, \sqrt{\frac{x_n}{u_n}}\right)$. De simples calculs font observer que

$$\sqrt{\sigma}\triangle\left(\frac{u}{\sigma} - \frac{1}{x}\right) = w - \frac{1}{w} = \sqrt{\sigma}\triangle^{-1}\left(\frac{x}{\sigma} - \frac{1}{u}\right).$$

Au vu de (5.32), il vient alors

$$w - \frac{1}{w} \in \operatorname{Im}(\triangle A^\top) \text{ et } w - \frac{1}{w} \in \operatorname{Ker}(A\triangle).$$

Mais comme $(A\triangle)^\top = \triangle A^\top$, les deux sous-espaces $\operatorname{Im}(\triangle A^\top)$ et $\operatorname{Ker}(A\triangle)$ sont orthogonaux, d'où $w - \frac{1}{w} = 0$, c'est-à-dire $x_i u_i = \sigma$ pour tout $i = 1, \dots, n$.

3°) Après un passage à la limite ($\sigma \to 0$), on obtient que

$$\begin{gathered} Ax^* = b,\ x^* \geqslant 0 \\ A^\top y^* + u^* = c,\ y^* \geqslant 0 \\ \langle x^*, u^* \rangle = 0, \end{gathered} \tag{5.33}$$

c'est-à-dire (x^*, y^*, u^*) est solution de $(\mathcal{P}\tilde{\mathcal{D}})$ (ou encore, x^* est solution de $(\mathcal{P})$ et (y^*, u^*) est solution de $(\tilde{\mathcal{D}})$).

En combinant $(\mathcal{SO})_\sigma$ et (5.33), on obtient

$$A(\overline{x}(\sigma) - x^*) = 0 \text{ et } A^\top(\overline{y}(\sigma) - y^*) = u^* - \overline{u}(\sigma),$$

d'où

$$\begin{aligned} 0 &= \langle A(\overline{x}(\sigma) - x^*), \overline{y}(\sigma) - y^* \rangle = \left\langle \overline{x}(\sigma) - x^*, A^\top(\overline{y}(\sigma) - y^*) \right\rangle \\ &= \langle \overline{x}(\sigma) - x^*, u^* - \overline{u}(\sigma) \rangle . \end{aligned} \tag{5.34}$$

Par ailleurs

$$\langle \overline{x}(\sigma) - x^*, \overline{u}(\sigma) - u^* \rangle = \langle \overline{x}(\sigma), \overline{u}(\sigma) \rangle - \left(\langle x^*, \overline{u}(\sigma) \rangle + \langle \overline{x}(\sigma), u^* \rangle \right),$$

ce qui conduit, puisque $\langle \overline{x}(\sigma), \overline{u}(\sigma)\rangle = n\sigma$ et $\langle \overline{x}(\sigma) - x^*, \overline{u}(\sigma) - u^*\rangle = 0$ (*cf.* (5.34)), à :

$$\langle x^*, \overline{u}(\sigma)\rangle + \langle \overline{x}(\sigma), u^*\rangle = \sum_{i=1}^{n} [x_i^* \overline{u}(\sigma)_i + \overline{x}(\sigma)_i u_i^*] = n\sigma.$$

Divisons les deux membres de la dernière égalité ci-dessus par σ $(= \overline{x}(\sigma)_i \overline{u}(\sigma)_i$ pour tout $i = 1, \ldots, n)$; il s'ensuit

$$\sum_{i=1}^{n} \left[\frac{x_i^*}{\overline{x}(\sigma)_i} + \frac{u_i^*}{\overline{u}(\sigma)_i} \right] = n. \tag{5.35}$$

Comme $x^* \geqslant 0,\ u^* \geqslant 0$ et $\langle x^*, u^*\rangle = 0$, on a $x_i^* u_i^* = 0$ pour tout i, de sorte que (5.35) se réécrit en

$$\sum_{i\in I} \frac{x_i^*}{\overline{x}(\sigma)_i} + \sum_{i\in J} \frac{u_i^*}{\overline{u}(\sigma)_i} = n \tag{5.36}$$

où $I := \{i \mid x_i^* > 0\}$ et $J := \{i \mid u_i^* > 0\}$ (I et J sont disjoints). Mais

$$\lim_{\sigma\to 0} \frac{x_i^*}{\overline{x}(\sigma)_i} = 1 \text{ si } i \in I \quad \text{et} \quad \lim_{\sigma\to 0} \frac{u_i^*}{\overline{u}(\sigma)_i} = 1 \text{ si } i \notin I,$$

ce qui, avec (5.36), donne bien le résultat escompté.

4°) Le programme dual $(\mathcal{D})$ de $(\mathcal{P})$ est un programme linéaire dans $\mathbb{R}^2$, à savoir :

$$(\mathcal{D}) \begin{cases} \text{Maximiser } y_2 \\ -y_1 \leqslant 1 \\ \ \ y_1 \leqslant 1 \\ \ \ y_2 \leqslant 1. \end{cases}$$

$(\mathcal{P})$ et $(\mathcal{D})$ peuvent être résolus graphiquement : $\overline{x} = (0, 0, 1)$ est la seule solution de $(\mathcal{P})$ tandis que $\{(y_1, 1) \mid -1 \leqslant y_1 \leqslant 1\}$ est l'ensemble-solution de $(\mathcal{D})$.

$(\mathcal{SO})_\sigma$ devient ici :

$$\begin{cases} -x_1 + x_2 = 0,\ x_3 = 1, \quad (x_1 > 0,\ x_2 > 0) \\ -y_1 + u_1 = 1, \\ \ \ y_1 + u_2 = 1, \quad (u_1 > 0,\ u_2 > 0,\ u_3 > 0) \\ \ \ y_2 + u_3 = 1, \\ x_1 u_1 = x_2 u_2 = x_3 u_3 = \sigma. \end{cases}$$

La résolution de ce système d'équations (dont une est non linéaire) fournit

$$\overline{x}(\sigma) = (\sigma, \sigma, 1),\ \overline{y}(\sigma) = (0, 1-\sigma),\ \overline{u}(\sigma) = (1, 1, \sigma).$$

Par suite

$$x^* = (0, 0, 1),\ y^* = (0, 1),\ u^* = (1, 1, 0).$$

Notons que la solution y^* de $(\mathcal{D})$ obtenue par ce procédé est le milieu du segment-solution de $(\mathcal{D})$.

Commentaire : – Voir l'Exercice 4.12 pour une approche analogue dans la résolution d'un problème de minimisation convexe différentiable (moyennant le changement de paramètre $\alpha = 1/\sigma$). Comme ici, $\sigma \longmapsto \overline{x}(\sigma)$ (resp. $\sigma \longmapsto (\overline{y}(\sigma), \overline{u}(\sigma))$ est appelé « chemin central » conduisant à une solution de $(\mathcal{P})$ (resp. à une solution de $(\tilde{\mathcal{D}})$).

– Prolongement de l'exercice : Démontrer que, sous les hypothèses qui ont été faites, l'ensemble $\{(x, y, u) \in C_0 \mid f_\sigma(x, y, u) \leqslant r\}$ est borné pour tout $r \in \mathbb{R}$. Cette propriété, combinée à la continuité de f_σ, déclenche le résultat d'existence admis à la 2e question.

– La propriété particulière de la solution $x^* = (x_1^*, \ldots, x_n^*)$ de $(\mathcal{P})$ et des variables d'écart $u_1^*, \ldots, u_n^*$ de $(\tilde{\mathcal{D}})$, mise en évidence dans la 3e question, est appelée « de Goldman et Tucker ».

This page intentionally left blank

VI

ENSEMBLES ET FONCTIONS CONVEXES. PROJECTION SUR UN CONVEXE FERMÉ

Rappels

VI.1. Ensembles convexes

Bien des notions et propriétés rappelées ci-après ont déjà été vues à propos des *polyèdres convexes fermés* (rappels du Chapitre 5) ; seuls sont soulignés ici les aspects qui en diffèrent.

– $C \subset \mathbb{R}^n$ est dit *convexe* (ou est un convexe) lorsque $\alpha x + (1-\alpha)x' \in C$ dès que x et x' sont dans C et $\alpha \in]0,1[$.

– Stabilité de la convexité par certaines opérations sur les ensembles :

  – Si C_1 et C_2 sont des convexes de $\mathbb{R}^n$, il en est de même de $C_1 + C_2$;

  – Si C_1 est un convexe de $\mathbb{R}^p$ et C_2 un convexe de $\mathbb{R}^q$, alors $C_1 \times C_2$ est un convexe de $\mathbb{R}^p \times \mathbb{R}^q$;

  – Si $(C_i)_{i\in I}$ est une collection de convexes, $\underset{i\in I}{\cap} C_i$ est encore convexe.

VI.1.1. Ensembles convexes associés à un convexe donné

– Le plus petit sous-espace affine de $\mathbb{R}^n$ contenant un convexe C de $\mathbb{R}^n$ s'appelle le *sous-espace affine engendré par* C, et est noté aff C ; on appelle dimension de C la dimension de ce sous-espace affine engendré (notation : dim C). L'intérieur

de C dans le sous-espace affine engendré par C est appelé *l'intérieur relatif de* C ; il est noté ir C (ou $\text{int}_r\, C$).

Chose extraordinaire : ir $C \neq \phi$ dès que (le convexe) $C \neq \phi$.

– Une partie convexe F de C est appelée *face* (ou partie extrémale) de C lorsque la propriété suivante est vérifiée :

$$\left.\begin{array}{l}(x_1, x_2) \in C \times C \text{ et} \\ \exists\ \alpha \in\ ]0,1[\ \text{ tel que } \alpha x_1 + (1-\alpha)x_2 \in F\end{array}\right\} \Rightarrow [x_1, x_2] \subset F.$$

Les *faces exposées* de C (*cf.* Chapitre 5) sont les exemples les plus importants de faces de C ; toutefois une face de C n'est pas nécessairement une face exposée de C.

Le cas « minimal » de face F est lorsque $\dim F = 0$; alors $F = \{\overline{x}\}$ et on parle plutôt de point extrémal $\overline{x}$. En d'autres termes, $\overline{x} \in C$ est appelé *point extrémal*[(1)] *de* C s'il n'est pas possible d'avoir $\overline{x} = \alpha y + (1-\alpha)z$ avec y et z deux points distincts de C et $\alpha \in\ ]0,1[$. On note $\text{ext}\, C$ l'ensemble des points extrémaux de C.

Propriété : Si C est un convexe compact de $\mathbb{R}^n$, alors $\text{ext}\, C \neq \phi$.

VI.1.2. Enveloppe convexe, enveloppe convexe fermée

Soit $S \subset \mathbb{R}^n$. On appelle *enveloppe convexe de* S, et on note $\text{conv}\, S$ (traditionnellement $\text{co}\, S$), le plus petit ensemble convexe (au sens de l'inclusion) contenant S. Il y a deux manières de construire $\text{conv}\, S$:

$\text{conv}\, S =$ intersection de tous les convexes C contenant S ;
$\text{conv}\, S =$ ensemble de toutes les combinaisons convexes d'éléments de S.

Deux théorèmes importants :

***Théorème* (C. Carathéodory).** *Tout élément x de $\text{conv}\, S$, où $S \subset \mathbb{R}^n$, peut être représenté sous forme de combinaison convexe d'au plus $n+1$ éléments de S.*

***Théorème* (H. Minkowski).** *Si C est un convexe compact de $\mathbb{R}^n$, alors $C = \text{conv}(\text{ext}\, C)$.*

On appelle *enveloppe convexe fermée de* S, et on note $\overline{\text{conv}}\, S$, le plus petit ensemble convexe fermé contenant S. Il y a déjà deux manières de voir $\overline{\text{conv}}\, S$:

$\overline{\text{conv}}\, S =$ intersection de tous les convexes fermés C contenant S ;
$\overline{\text{conv}}\, S =$ adhérence de $\text{conv}\, S$ (autrement dit, pour avoir $\overline{\text{conv}}\, S$ on commence par convexifier S, puis on ferme l'ensemble $\text{conv}\, S$ ainsi obtenu).

(1) Dans le cas non polyédral, le vocable *sommet de* C est plutôt réservé à un point extrémal $\overline{x}$ de C tel que $N(C, \overline{x})$ soit de dimension n ($N(C, \overline{x})$ désigne le cône normal à C en $\overline{x}$; *cf.* Chapitre 3).

VI.1.3. Hyperplan d'appui, fonction d'appui

Un hyperplan affine de $\mathbb{R}^n$, d'équation $\langle s, x\rangle = r$, est appelé *hyperplan d'appui* au convexe C en $\overline{x} \in C$ lorsque

$$\langle s, \overline{x}\rangle = r \text{ et } \langle s, x\rangle \leqslant r \text{ pour tout } x \in C.$$

Résultat important : Si C est convexe et si $\overline{x} \in \operatorname{fr} C$, alors *il y a* (au moins) un hyperplan d'appui à C en $\overline{x}$.

Soit $\phi \neq S \subset \mathbb{R}^n$. On appelle *fonction d'appui de* S, et on note σ_S, la fonction $\sigma_S : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ définie de la manière suivante :

$$d \in \mathbb{R}^n \longmapsto \sigma_S(d) := \sup_{s\in S} \langle s, d\rangle .$$

σ_S est partout finie sur $\mathbb{R}^n$ si, et seulement si, S est borné.

Une fonction d'appui ne sait pas distinguer S de son enveloppe convexe fermée : $\sigma_S = \sigma_{\overline{\operatorname{conv}} S}$.

On peut voir $\overline{\operatorname{conv}} S$ à partir de S – ou plutôt à partir de σ_S – de la manière suivante :

$$(s \in \overline{\operatorname{conv}} S) \Leftrightarrow (\langle s, d\rangle \leqslant \sigma_S(d) \text{ pour tout } d \in \mathbb{R}^n).$$

Ceci est la version « fonctionnelle » du résultat « ensembliste » de description de $\overline{\operatorname{conv}} S$ suivant : $\overline{\operatorname{conv}} S$ est l'intersection de tous les demi-espaces affines fermés contenant S.

Une règle de calcul (parmi d'autres) concernant les fonctions d'appui : Si S_1 et S_2 sont deux parties non vides de $\mathbb{R}^n$, alors $\sigma_{S_1+S_2} = \sigma_{S_1} + \sigma_{S_2}$.

VI.1.4. Théorèmes de séparation par un hyperplan affine

***Théorème* de séparation (au sens large).** *Si C_1 et C_2 sont deux convexes disjoints de $\mathbb{R}^n$, il existe un hyperplan affine qui les sépare, i.e. un vecteur $s \neq 0$ de $\mathbb{R}^n$ tel que :*

$$\langle s, x_1\rangle \leqslant \langle s, x_2\rangle \text{ pour tout } x_1 \in C_1 \text{ et tout } x_2 \in C_2.$$

***Théorème* de séparation (au sens strict).** *Si C_1 est un convexe fermé et C_2 un convexe compact de $\mathbb{R}^n$, et s'ils sont disjoints, il existe un hyperplan affine qui les sépare strictement, i.e. un vecteur $s \neq 0$ de $\mathbb{R}^n$ et $r \in \mathbb{R}$ tels que :*

$$\begin{array}{l}\langle s, x_1\rangle < r \text{ pour tout } x_1 \in C_1\\ \text{et}\\ \langle s, x_2\rangle > r \text{ pour tout } x_2 \in C_2.\end{array}$$

VI.2. Projection sur un convexe fermé

Soit C un convexe fermé non vide de $\mathbb{R}^n$. Pour tout $x \in \mathbb{R}^n$, il existe un et un seul élément dans C à distance (euclidienne) minimale de x ; cet élément s'appelle *la projection de x sur C* et est noté $p_C(x)$:

$$\begin{cases} p_C(x) \in C, \\ \|p_C(x) - x\| \leqslant \|c - x\| \quad \text{pour tout} \quad c \in C. \end{cases}$$

Caractérisation de $p_C(x)$ (*cf.* Chapitre 3) : $\overline{x} \in C$ est $p_C(x)$ si, et seulement si,

$$\langle x - \overline{x}, c - \overline{x} \rangle \leqslant 0 \text{ pour tout } c \in C.$$

Dans le cas où C est un cône convexe fermé, la caractérisation ci-dessus se simplifie quelque peu : Soit K un cône convexe fermé de $\mathbb{R}^n$ et K° son cône polaire ; alors $\overline{x} \in K$ est $p_K(x)$ si, et seulement si,

$$x - \overline{x} \in K^\circ \text{ et } \langle x - \overline{x}, \overline{x} \rangle = 0.$$

Une propriété fondamentale de l'application $p_C : \mathbb{R}^n \to C \subset \mathbb{R}^n$ est comme suit :

$$\|p_C(x_1) - p_C(x_2)\|^2 \leqslant \langle p_C(x_1) - p_C(x_2), x_1 - x_2 \rangle$$

pour tout x_1 et x_2 dans $\mathbb{R}^n$.

Il en découle notamment :

$$\|p_C(x_1) - p_C(x_2)\| \leqslant \|x_1 - x_2\| \text{ pour tout } x_1 \text{ et } x_2 \text{ dans } \mathbb{R}^n.$$

VI.3. Fonctions convexes

Soit C un convexe de $\mathbb{R}^n$ et $f : C \to \mathbb{R}$; les définitions de « f est convexe sur C » et de « f est strictement convexe sur C » ont été rappelées au Chapitre 1. Quelques rappels supplémentaires :

– *Inégalité de Jensen*

Soient $C \subset \mathbb{R}^n$ convexe et $f : C \to \mathbb{R}$ convexe. Alors, pour toute collection $\{x_1, \ldots, x_k\}$ de points de C et tout $\alpha = (\alpha_1, \ldots, \alpha_k)$ dans le simplexe-unité de $\mathbb{R}^k$, on a l'inégalité suivante :

$$f\left(\sum_{i=1}^{k} \alpha_i x_i\right) \leqslant \sum_{i=1}^{k} \alpha_i f(x_i).$$

– Stabilité de la convexité par certaines opérations sur les fonctions :

– Si f_1 et f_2 sont convexes sur C, il en est de même de $f_1 + f_2$;

– Si $(f_i)_{i\in I}$ est une collection de fonctions convexes sur C, alors $\sup_{i\in I} f_i$ est encore convexe sur C.

– Exemples de fonctions convexes :

– Si $\|\cdot\|$ est une norme sur $\mathbb{R}^n$ et C un convexe de $\mathbb{R}^n$, alors la fonction-distance

$$d_C : x \in \mathbb{R}^n \longmapsto d_C(x) := \inf\{\| x - c \|, c \in C\}$$

est convexe sur $\mathbb{R}^n$.

– Si $f : x \in \mathbb{R}^+_* \longmapsto f(x)$ est convexe sur $\mathbb{R}^*_+$, il en est de même de la fonction

$$g : x \in \mathbb{R}^+_* \longmapsto g(x) := xf\left(\frac{1}{x}\right).$$

Références. Chapitres III et IV de [12].

***Exercice VI.1.** Soit $S \subset \mathbb{R}^n$ vérifiant la propriété de « demi-somme » suivante :

$$(x \in S,\ y \in S) \Rightarrow \left(\frac{x+y}{2} \in S\right).$$

1°) S est-il convexe ?

2°) Même question si l'on suppose S fermé.

Solution : 1°) Non. Il suffit pour le voir de prendre pour S l'ensemble des rationnels compris entre 0 et 1.

2°) Oui si S est fermé.

Prenons x_1 et x_2 dans S et considérons $x \in [x_1, x_2] := \{(1-t)x_1 + tx_2 \mid 0 \leqslant t \leqslant 1\}$.

On prend des deux segments $\left[x_1, \frac{x_1+x_2}{2}\right], \left[\frac{x_1+x_2}{2}, x_2\right]$ celui qui contient x, et on le note $\left[x_1^{(1)}, x_2^{(2)}\right]$. On réitère le processus de manière à obtenir :

$$x \in \ldots \left[x_1^{(k)}, x_2^{(k)}\right] \subset \ldots \subset [x_1, x_2],$$
$$\lim_{k\to+\infty} x_1^{(k)} = \lim_{k\to+\infty} x_2^{(k)} = x.$$

Mais grâce à la propriété de « demi-somme » de S, toutes les extrémités $x_1^{(k)}, x_2^{(k)}$ sont dans S. Le caractère fermé de S fait qu'ensuite la limite x est dans S.

***Exercice VI.2.** Soient $A_0, A_1, \ldots, A_m$ dans $\mathcal{S}_n(\mathbb{R})$ et

$$A(x) := A_0 + \sum_{i=1}^{m} x_i A_i \quad \text{pour tout} \quad x = (x_1, \ldots, x_m) \in \mathbb{R}^m.$$

On pose $C := \{x \in \mathbb{R}^m \mid A(x) \text{ est semi-définie positive}\}$. Montrer que C est convexe fermé.

Solution : – Vérification directe : toute combinaison convexe d'éléments de C est dans C ; toute limite de suite d'éléments de C est encore dans C.

– Autre manière de faire. L'application

$$x = (x_1, \ldots, x_m) \in \mathbb{R}^m \longmapsto A(x) = A_0 + \sum_{i=1}^{m} x_i A_i \in \mathcal{S}_n(\mathbb{R})$$

est affine, et C n'est autre que l'image inverse par cette application du cône convexe fermé $\mathcal{P}_n(\mathbb{R})$.

Commentaire :

– Contrairement à ce qu'on peut imaginer au premier abord, C n'est pas polyédrique. On peut aussi décrire C par la conjonction d'inégalités polynomiales en x (tous les mineurs principaux d'ordre k de $A(x)$, $k = 1, \ldots, n$, doivent être positifs).

– Le problème d'optimisation

$$(\mathcal{P}) \qquad \begin{cases} \text{Min } \langle c, x \rangle \\ A(x) \text{ semi-définie positive } (x \in C) \end{cases}$$

est un problème-modèle très général ; s'y ramènent des problèmes classiques d'optimisation (optimisation quadratique, de valeurs propres, etc.) ou d'ingénierie (théorie et commande des systèmes).

****Exercice VI.3.** Soient C un convexe compact de $\mathbb{R}^n$ et H un hyperplan d'appui à C. Montrer que H contient nécessairement des points extrémaux de C.

Solution : $C \cap H$ est un convexe compact ; il a donc des points extrémaux. Mais si $\overline{x}$ est un point extrémal de $C \cap H$, il est aussi point extrémal de C. En effet, supposons que $\overline{x} = \alpha y + (1-\alpha)z$ avec y, z dans C et $\alpha \in]0,1[$. Puisque H est un hyperplan d'appui à C en $\overline{x}$, mettons d'équation $\langle s, x\rangle = r$, on a :

$$\langle s, \overline{x}\rangle = r,\ \langle s, y\rangle \leqslant r,\ \langle s, z\rangle \leqslant r.$$

Par suite,

$$0 = r - \langle s, \overline{x}\rangle = \alpha\,[r - \langle s, y\rangle] + (1-\alpha)\,[r - \langle s, z\rangle]\,,$$

d'où $\langle s, y\rangle = \langle s, z\rangle = r$. Ainsi y et z se trouvent aussi dans H. Mais alors le caractère extrémal de $\overline{x}$ dans $C \cap H$ fait que $\overline{x} = y = z$. Le point $\overline{x}$ est donc extrémal dans C.

En conclusion, $C \cap H$ contient bien des points extrémaux de C.

***Exercice VI.4.** Soit C le convexe défini comme étant l'enveloppe convexe de S : $C = \operatorname{conv} S$.

Montrer que tout point extrémal de C est nécessairement dans S.

Solution : Soit x un point extrémal de C et supposons qu'il ne soit pas dans S.

Comme $C = \operatorname{conv} S$, il existe une représentation de x sous la forme $x = \sum_{i=1}^{k} \alpha_i x_i$, avec $x_1, \ldots, x_k$ dans S et $(\alpha_1, \ldots, \alpha_k)$ dans le simplexe-unité de $\mathbb{R}^k$.

Puisque $x \notin S$, on a nécessairement $k \geqslant 2$. On peut supposer, sans perte de généralité, que $0 < \alpha_k < 1$. Ainsi

$$x = \alpha_k x_k + (1-\alpha_k) \sum_{i=1}^{k-1} \frac{\alpha_i}{1-\alpha_k} x_i,$$

où $y := x_k$ et $z := \sum_{i=1}^{k-1} \frac{\alpha_i}{1-\alpha_k} x_i$ sont des éléments de C différents (différents car, sinon, $x = x_k$, ce qui est impossible vu que $x \notin S$). La représentation de x sous la forme

$$x = \alpha_k y + (1-\alpha_k)z, \text{ avec } y \text{ et } z \text{ dans } C,\ y \neq z \text{ et } 0 < \alpha_k < 1,$$

entre alors en contradiction avec le caractère extrémal de x.

À titre d'exemple, soit $S = \{x_1, \ldots, x_k\}$. Alors tout point extrémal du polyèdre convexe compact $C := \operatorname{conv} S$ est nécessairement l'un des points x_i. Ce cas particulier a déjà été vu à l'Exercice V.7.

Le résultat de l'exercice est important et sera utilisé à plusieurs reprises par la suite.

***Exercice VI.5.** Soit C un convexe. Montrer que $x \in C$ est un point extrémal de C si, et seulement si, $C \setminus \{x\}$ est convexe.

A-t-on une caractérisation similaire pour une face de C ?

Solution : Soit x un point extrémal de C. Considérons deux points distincts y et z de $C \setminus \{x\}$ et $\alpha \in]0,1[$; nous devons montrer que $u := \alpha y + (1-\alpha)z \in C \setminus \{x\}$.

Par la convexité de C, le point u est dans C ; mais il est différent de x car

$$u \in C,\ u = x = \alpha y + (1-\alpha)z \text{ avec } y \text{ et } z \text{ dans } C,\ y \neq z,\ \text{et } \alpha \in]0,1[$$

viendrait à contredire le caractère extrémal de x dans C. Donc u est bien dans $C \setminus \{x\}$.

Réciproquement, supposons $C\setminus\{x\}$ convexe et montrons que x est un point extrémal de C. Supposons qu'on puisse écrire x sous la forme

$$x = \frac{1}{2}(y+z) \text{ avec } y \text{ et } z \text{ dans } C,\ y \neq z. \tag{6.1}$$

Alors y et z sont différents de x nécessairement ; mais par la convexité de $C \setminus \{x\}$, cela entraîne que $\frac{1}{2}y + \frac{1}{2}z \in C \setminus \{x\}$. D'où contradiction avec (6.1).

En conséquence, une décomposition de x comme dans (6.1) est impossible : x est un point extrémal de C.

En ce qui concerne les faces de C, l'implication suivante est une généralisation naturelle de ce qui a été établi pour les points extrémaux de C :

$$(F \text{ face de } C) \Rightarrow (C \setminus F \text{ est convexe}).$$

Mais la réciproque est fausse. Pour voir cela, prendre par exemple un triangle C de $\mathbb{R}^2$ et considérer une moitié F de C.

****Exercice VI.6.** On considère dans $\mathbb{R}^n$ les deux boules suivantes :

$$B_1 := \left\{ x = (x_1, \dots, x_n) \in \mathbb{R}^n \mid \sum_{i=1}^{n} |x_i| \leqslant 1 \right\},$$

$$B_\infty := \{ x = (x_1, \dots, x_n) \in \mathbb{R}^n \mid \max_{i=1,\dots,n} |x_i| \leqslant 1 \}.$$

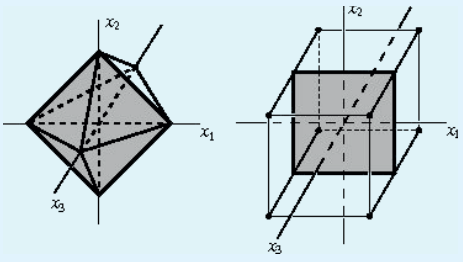
Quels sont les points extrémaux de B_1 ? de B_∞ ?

Solution : Désignons par $\mathrm{e}_1, \dots, \mathrm{e}_n$ les n éléments de la base canonique de $\mathbb{R}^n$.

L'ensemble B_1 peut être vu comme l'enveloppe convexe des $2n$ éléments $\mathrm{e}_1, -\mathrm{e}_1, \dots, \mathrm{e}_i, -\mathrm{e}_i, \dots, \mathrm{e}_n, -\mathrm{e}_n$; de plus chacun de ces e_i ou $-\mathrm{e}_i$ est un point extrémal de B_1 (d'accord ?). En somme, les points extrémaux (ou sommets) de B_1 sont exactement : $\mathrm{e}_1, -\mathrm{e}_1, \dots, \mathrm{e}_n, -\mathrm{e}_n$.

La structure de produit cartésien de $B_\infty = [-1, +1]^n$ facilite la détermination de ses points extrémaux. En effet, il résulte de la définition même de point extrémal que $\mathrm{ext}(C_1 \times C_2) = (\mathrm{ext}C_1) \times (\mathrm{ext}C_2)$. En conséquence, les points extrémaux de B_∞ sont les 2^n points $\overline{x} = (\overline{x}_1, \dots, \overline{x}_n)$, où $\overline{x}_i \in \{-1, +1\}$ pour tout $i = 1, \dots, n$.

Il y a donc une différence notable entre B_1 et B_∞, différence qui ne se voit pas pour $n = 1$ ou 2. Lorsque n croît, le nombre de points extrémaux de B_1 croît de manière *linéaire* (en passant de $\mathbb{R}^n$ à $\mathbb{R}^{n+1}$ on ajoute en fait 2 nouveaux points extrémaux), tandis que le nombre de points extrémaux de B_∞ croît *exponentiellement* (en passant de $\mathbb{R}^n$ à $\mathbb{R}^{n+1}$, on reproduit 2 copies des points extrémaux précédents).



· dans $\mathbb{R}^2$, reproduits dans $\mathbb{R}^3$ · dans $\mathbb{R}^2$,
• ajout dans $\mathbb{R}^3$ • dans $\mathbb{R}^3$, 2 copies de ·

Figure 15.

****Exercice VI.7.** Soit C un convexe de $\mathbb{R}^n$ et $A : \mathbb{R}^n \to \mathbb{R}^m$ une application affine. Quelle relation y a-t-il entre les points extrémaux de $A(C)$ et l'image par A de l'ensemble des points extrémaux de C ?

Solution : Aucune en général.

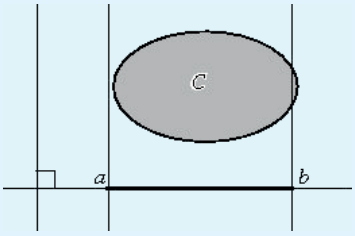


FIGURE 16.

Dans l'exemple ci-contre, A est l'opérateur de projection sur l'axe des x. On note que
$A(\text{ext}\, C)$ est tout le segment $[a,b]$ tandis que
$\text{ext}\,(A(C)) = \{a,b\}$.

Si, à présent, C est la bande $[a,b]\times\mathbb{R}$ de $\mathbb{R}^2$, on a toujours $A(C)=[a,b]$ et donc $\text{ext}(A(C))=\{a,b\}$, mais $\text{ext}\, C=\phi$.

Toutefois, si A est injective et si x est un point extrémal de C, alors $y=A(x)$ est un point extrémal de $A(C)$. En effet, avoir

$$A(x)=\alpha A(x_1)+(1-\alpha)A(x_2) \text{ avec}$$
$$x_1,x_2\in C,\ A(x_1)\neq A(x_2) \text{ et } \alpha\in\,]0,1[$$

est impossible puisque $\alpha A(x_1)+(1-\alpha)A(x_2)=A(\alpha x_1+(1-\alpha)x_2)=A(x)$ implique $x=\alpha x_1+(1-\alpha)x_2$, et que cette dernière décomposition contredit le caractère extrémal de x dans C.

***Exercice VI.8.** Soient k éléments $a_1,\dots,a_k$ de $\mathbb{R}^n$ et K un cône convexe fermé de $\mathbb{R}^n$.

1°) Montrer l'équivalence des deux affirmations ci-dessous :

(i) Il existe $d_0\in K$ tel que $\langle a_i,d_0\rangle<0$ pour tout $i=1,\dots,k$;

(ii) $\text{conv}\{a_1,\dots,a_k\}\cap(-K^\circ)=\emptyset$.

2°) *Applications :*
Traduire l'équivalence démontrée dans les cas particuliers suivants :

(α) $K=\mathbb{R}^n$;

(β) $K=\{d\in\mathbb{R}^n \mid \langle v_i,d\rangle=0 \text{ pour tout } i=1,\dots,m\}$, où $v_1,\dots,v_m$ sont m vecteurs donnés de $\mathbb{R}^n$.

Indication. Dans la 1re question on montrera successivement $[(i)\Rightarrow(ii)]$ et $[non(i)\Rightarrow non(ii)]$; pour cette deuxième implication, on pourra séparer (au sens large) les convexes $\{(\langle d,a_1\rangle,\dots,\langle d,a_k\rangle) \mid d\in K\}$ et $(\mathbb{R}_*^-)^k$.

Solution : 1°) [$(i) \Rightarrow (ii)$]. Considérons un élément quelconque de $\mathrm{conv}\{a_1, \ldots, a_k\}$, c'est-à-dire un élément de la forme $\sum_{i=1}^{k} \alpha_i a_i$, où $(\alpha_1, \ldots, \alpha_k) \in \Delta_k$ (simplexe-unité de $\mathbb{R}^k$). Il découle de (i) : $d_0 \in K$ et

$$\left\langle -\sum_{i=1}^{k} \alpha_i a_i, d_0 \right\rangle = \sum_{i=1}^{k} \alpha_i (-\langle a_i, d_0 \rangle) > 0 \ ;$$

donc $-\sum_{i=1}^{k} \alpha_i a_i$ n'est pas dans K°.

[$non(i) \Rightarrow non(ii)$]. Puisque K est convexe et que l'application $d \in \mathbb{R}^n \longmapsto (\langle d, a_1 \rangle, \ldots, \langle d, a_k \rangle) \in \mathbb{R}^k$ est linéaire, l'ensemble $C_2 := \{(\langle d, a_1 \rangle, \ldots, \langle d, a_k \rangle) \mid d \in K\}$ est un convexe de $\mathbb{R}^k$.

Ce qu'exprime $non(i)$ est exactement $(\mathbb{R}_*^-)^k \cap C_2 = \emptyset$. Si tel est le cas, il existe un hyperplan affine qui sépare $(\mathbb{R}_*^-)^k$ et C_2, *i.e.* un vecteur $s = (s_1, \ldots s_k) \neq 0$ de $\mathbb{R}^k$ tel que :

$$\langle s, x_1 \rangle \leqslant \langle s, x_2 \rangle \text{ pour tout } x_1 \in (\mathbb{R}_*^-)^k \text{ et tout } x_2 \in C_2. \qquad (\mathcal{S})$$

Puisque $(\mathbb{R}^-)^k$ est l'adhérence de $(\mathbb{R}_*^-)^k$, l'inégalité $(\mathcal{S})$ reste vraie pour tout $x_1 \in (\mathbb{R}^-)^k$ et tout $x_2 \in C_2$. Prenons $x_2 = 0$ (qui est dans C_2) dans cette inégalité, il s'ensuit :

$$\langle s, x_1 \rangle \leqslant 0 \text{ pour tout } x_1 \in (\mathbb{R}^-)^k,$$

d'où $s \in (\mathbb{R}^+)^k$ (d'accord ?). Ainsi $s_i \geqslant 0$ pour tout $i = 1, \ldots, k$.

Faisons $x_1 = 0$ dans l'inégalité $(\mathcal{S})$ (prolongée) ; on a alors :

$$\langle s, x_2 \rangle = \sum_{i=1}^{k} s_i \langle d, a_i \rangle = \left\langle \sum_{i=1}^{k} s_i a_i, d \right\rangle \geqslant 0 \text{ pour tout } d \in K,$$

autrement dit : $-\sum_{i=1}^{k} s_i a_i \in K^\circ$.

En résumé : si (i) n'a pas lieu, il existe des réels $s_1 \geqslant 0, \ldots, s_k \geqslant 0$ non tous nuls tels que $-\sum_{i=1}^{k} s_i a_i \in K^\circ$. En posant $\alpha_i := s_i / (\sum_{i=1}^{k} s_i)$ pour tout $i = 1, \ldots, k$, de manière à avoir $(\alpha_1, \ldots, \alpha_k) \in \Delta_k$, on obtient que

$$\sum_{i=1}^{k} \alpha_i a_i \in \mathrm{conv}\,\{a_1, \ldots, a_k\} \cap (-K^\circ).$$

2°) (α) Si $K = \mathbb{R}^n$, alors $K^\circ = \{0\}$; on a donc l'équivalence des énoncés suivants :

(i) $\exists\ d_0 \in \mathbb{R}^n$ tel que $\langle a_i, d_0\rangle < 0$ pour tout $i = 1, \ldots, k$;

(ii) $0 \notin \text{conv}\{a_1, \ldots, a_k\}$.

(β) Si $K = \{d \in \mathbb{R}^n \mid \langle v_i, d\rangle = 0 \text{ pour tout } i = 1, \ldots, m\}$, alors $K^\circ = K^\perp = \text{vect}\,\{v_1, \ldots, v_m\}$. Il y a donc équivalence des énoncés suivants :

(i) $\exists\ d_0 \in \mathbb{R}^n$ tel que $\langle v_i, d_0\rangle = 0$ pour tout $i = 1, \ldots, m$ et $\langle a_i, d_0\rangle < 0$ pour tout $i = 1, \ldots, k$;

(ii) $\text{conv}\{a_1, \ldots, a_k\} \cap \text{vect}\{v_1, \ldots, v_m\} = \emptyset$.

Commentaire : – L'équivalence provenant de l'application (α) de la 2e question constitue le *lemme de Gordan* (1873) ; ce lemme appartient au « royaume des polyèdres convexes » et peut être démontré avec les résultats et techniques du Chapitre 5 ; il sera « revisité » dans l'Exercice VII.1.

– L'équivalence provenant de l'application (β) de la 2e question montre bien la différence entre les conditions de qualification de contraintes $(QC)_{\overline{x}}$ et $(QC)'_{\overline{x}}$ utilisées dans les problèmes de minimisation avec contraintes décrites avec des égalités et inégalités (*cf.* Chapitre 3). En particulier elle indique clairement que $(QC)'_{\overline{x}}$ implique $(QC)_{\overline{x}}$.

****** Exercice VI.9.** Soient A et B deux éléments de $\mathcal{S}_n(\mathbb{R}), n \geq 3$, On suppose que B n'est pas définie négative. Montrer l'équivalence des deux assertions suivantes :

(i) $(\langle Bx, x\rangle \geqslant 0 \text{ et } x \neq 0) \Rightarrow (\langle Ax, x\rangle > 0)$;

(ii) Il existe $\mu \geqslant 0$ tel que $A - \mu B$ soit définie positive.

Commentaire : Résultat à rapprocher du résultat de l'Exercice I.16 qu'il étend.

Indication. On admettra que l'ensemble $C := \{(\langle Ax, x\rangle, \langle Bx, x\rangle) \mid \| x \| = 1\}$ est un convexe de $\mathbb{R}^2$. Pour $[(i) \Rightarrow (ii)]$, penser à séparer strictement C de $\mathbb{R}^- \times \mathbb{R}^+$.

Solution : $[(ii) \Rightarrow (i)]$. Immédiat.

$[(i) \Rightarrow (ii)]$. L'ensemble C défini ci-dessus est un convexe compact de $\mathbb{R}^2$. Observons que C ne rencontre pas $\mathbb{R}^- \times \mathbb{R}^+$. En effet, avoir

$$\langle Ax, x\rangle \leqslant 0 \text{ et } \langle Bx, x\rangle \geqslant 0 \text{ pour un certain } x \text{ de norme } 1$$

est impossible d'après (i).

On peut donc séparer strictement le convexe compact C du convexe fermé $\mathbb{R}^- \times \mathbb{R}^+$: il existe $s = (s_1, s_2) \neq (0,0)$, $r \in \mathbb{R}$, tels que

$$\langle s, x\rangle = s_1x_1 + s_2x_2 > r \text{ pour tout } x = (x_1, x_2) \in C \tag{6.2}$$

et

$$s_1x_1 + s_2x_2 < r \text{ pour tout } x = (x_1, x_2) \in \mathbb{R}^- \times \mathbb{R}^+. \tag{6.3}$$

En faisant $(x_1, x_2) = (0,0)$ dans (6.3), il vient que $r > 0$. De plus, comme $(tx_1, tx_2) \in \mathbb{R}^- \times \mathbb{R}^+$ dès que $(x_1, x_2) \in \mathbb{R}^- \times \mathbb{R}^+$ et avec $t > 0$ arbitrairement grand, il résulte de (6.3) : $s_1 \geqslant 0$, $s_2 \leqslant 0$.

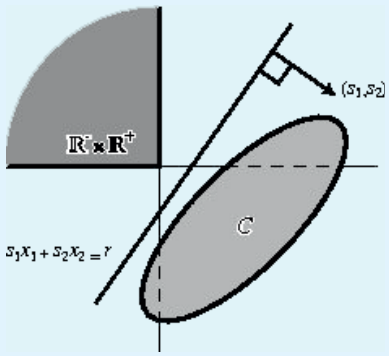


Figure 17.

On ne peut avoir $s_1 = 0$. En effet, si tel était le cas, s_2 serait < 0 et (6.2) impliquerait

$$\langle Bx, x\rangle < \frac{r}{s_2} \text{ pour tout } x \text{ de norme } 1,$$

ce qui est contraire à l'hypothèse faite sur B.

Par suite,

$$x_1 + \frac{s_2}{s_1}x_2 > \frac{r}{s_1} \text{ pour tout } (x_1, x_2) \in C,$$

soit encore

$$\langle Ax, x\rangle + \frac{s_2}{s_1}\langle Bx, x\rangle > \frac{r}{s_1} \text{ pour tout } x \text{ de norme } 1.$$

L'assertion (ii) est ainsi démontrée.

Remarque :

– Curieusement, l'équivalence ne s'étend pas au cas de plus de deux matrices symétriques. Soient $A, B_1, \ldots, B_m$ des éléments de $\mathcal{S}_n(\mathbb{R})$: alors

(i) $(\langle B_1 x, x\rangle \geqslant 0, \ldots, \langle B_m x, x\rangle \geqslant 0 \text{ et } x \neq 0) \Rightarrow (\langle Ax, x\rangle > 0)$

et

(ii) Il existe $\mu_1 \geqslant 0, \ldots, \mu_m \geqslant 0$ tels que $A - \mu_1 B_1 \ldots - \mu_m B_m$ soit définie positive

ne sont pas des assertions équivalentes ; seule $[(ii) \Rightarrow (i)]$ a lieu.

– Par une technique similaire à celle utilisée dans l'Exercice (on sépare (au sens large) C et $(\mathbb{R}_*^-)^2$), on démontre le résultat parent suivant : Soient A et B deux éléments de $\mathcal{S}_n(\mathbb{R})$; alors

(i) $\max\{\langle Ax, x\rangle, \langle Bx, x\rangle\} \geqslant 0$ pour tout $x \in \mathbb{R}^n$

équivaut à

(ii) Il existe $\mu_1 \geqslant 0$ et $\mu_2 \geqslant 0$, non tous deux nuls, tels que $\mu_1 A + \mu_2 B$ soit semi-définie positive.

***Exercice VI.10.** Montrer que $\text{conv}(A \times B) = (\text{conv}A) \times (\text{conv}B)$.

Solution : $A \subset \text{conv}A$, $B \subset \text{conv}B$, et $(\text{conv}A) \times (\text{conv}B)$ est convexe (produit cartésien de deux convexes). Par conséquent, $\text{conv}(A \times B) \subset (\text{conv}A) \times (\text{conv}B)$.

Démontrons l'inclusion réciproque. Soient $u \in \text{conv}A$ et $v \in \text{conv}B$; alors il existe

$$\left.\begin{array}{l} \text{– des entiers non nuls } p \text{ et } q, \\ \text{– des éléments } a_1, \ldots, a_p \text{ de } A \text{ et des éléments } b_1, \ldots, b_q \text{ de } B, \\ \text{– } (\lambda_1, \ldots, \lambda_p) \text{ dans le simplexe-unité de } \mathbb{R}^p, \\ \text{– } (\mu_1, \ldots, \mu_q) \text{ dans le simplexe unité de } \mathbb{R}^q, \end{array}\right\} \tag{6.4}$$

tels que :

$$u = \sum_{i=1}^{p} \lambda_i a_i, \quad v = \sum_{j=1}^{q} \mu_j b_j. \tag{6.5}$$

D'où

$$(u, v) = \sum_{i=1}^{p} \sum_{j=1}^{q} \lambda_i \mu_j (a_i, b_j). \tag{6.6}$$

Mais $(a_i, b_j) \in A \times B$ pour tout (i,j) et $(\lambda_1\mu_1, \ldots, \lambda_1\mu_q, \lambda_2\mu_1, \ldots, \lambda_2\mu_q, \ldots, \lambda_p\mu_1, \ldots, \lambda_p\mu_q)$ est dans le simplexe-unité de $\mathbb{R}^{pq}$ (en effet tous les coefficients $\lambda_i\mu_j$ sont positifs et $\sum\limits_{\substack{1 \leqslant i \leqslant p \\ 1 \leqslant j \leqslant q}} \lambda_i\mu_j = \left(\sum\limits_{i=1}^{p} \lambda_i\right)\left(\sum\limits_{j=1}^{q} \mu_j\right) = 1$).

Il résulte donc de (6.6) que (u,v) est bien dans $\mathrm{conv}(A \times B)$.

**** *Exercice VI.11.*** Soient A et B deux parties (non vides) de $\mathbb{R}^n$. Montrer que :
(i) $\mathrm{conv}(A+B) = \mathrm{conv}A + \mathrm{conv}B$;
(ii) $\overline{\mathrm{conv}}\,(A+B) = \overline{(\overline{\mathrm{conv}}\,A + \overline{\mathrm{conv}}\,B)}$.

Solution : (i) $A \subset \mathrm{conv}A$, $B \subset \mathrm{conv}B$, et $\mathrm{conv}A + \mathrm{conv}B$ est convexe (somme de deux convexes). Par suite, $\mathrm{conv}(A+B) \subset \mathrm{conv}A + \mathrm{conv}B$.

Démontrons l'inclusion réciproque. Soient $u \in \mathrm{conv}A$ et $v \in \mathrm{conv}B$; alors il existe une expression de u et v comme en (6.5) . D'où

$$u + v = \sum_{i=1}^{p} \lambda_i a_i + \sum_{j=1}^{q} \mu_j b_j. \tag{6.7}$$

Comme $\sum\limits_{i=1}^{p} \lambda_i = \sum\limits_{j=1}^{q} \mu_j = 1$, il vient de la relation ci-dessus

$$u + v = \sum_{i=1}^{p} \sum_{j=1}^{q} \lambda_i\mu_j (a_i + b_j). \tag{6.8}$$

Mais $a_i + b_j \in A + B$ pour tout (i,j) et $(\lambda_i\mu_j)_{\substack{1 \leqslant i \leqslant p \\ 1 \leqslant j \leqslant q}}$ est dans le simplexe-unité de $\mathbb{R}^{pq}$. Il en résulte avec (6.8) que $u+v$ est bien dans $\mathrm{conv}(A+B)$.

(ii) Dès qu'il y a une enveloppe convexe fermée en jeu, il faut penser à l'outil « fonction d'appui ».

$\overline{\mathrm{conv}}(A+B)$ est un convexe fermé dont la fonction d'appui est celle de $A+B$, soit

$$d \in \mathbb{R}^n \longmapsto \sigma_{A+B}(d) = \sup_{\substack{a \in A \\ b \in B}} \langle a+b, d\rangle\,. \tag{6.9}$$

$\overline{(\overline{\text{conv}}A + \overline{\text{conv}}B)}$ est un convexe fermé dont la fonction d'appui est la somme des fonctions d'appui de A et de B, soit

$$d \in \mathbb{R}^n \longmapsto (\sigma_A + \sigma_B)(d) = \sup_{a\in A} \langle a, d\rangle + \sup_{b\in B} \langle b, d\rangle \,. \tag{6.10}$$

De l'égalité des fonctions de (6.9) et (6.10) découle l'égalité des convexes fermés dont elles sont les fonctions d'appui.

Commentaire : Prolongement du (i) de l'exercice : Si $S \subset \mathbb{R}^n$ et $U : \mathbb{R}^n \to \mathbb{R}^m$ est une application affine, alors $\text{conv}\, U(S) = U(\text{conv}\, S)$. Ceci, combiné au résultat de l'Exercice 6.10 permet de retrouver (i).

***Exercice VI.12.** Dans $\mathcal{S}_n(\mathbb{R})$ structuré en espace euclidien grâce au produit scalaire $\langle\langle \cdot, \cdot \rangle\rangle$ (rappel : $\langle\langle A, B\rangle\rangle := \text{tr}(AB)$), on considère l'ensemble suivant

$$\Omega_1 := \{M \in \mathcal{P}_n(\mathbb{R}) \mid \text{tr}M = 1\} \,.$$

1°) Montrer que Ω_1 est un convexe compact de $\mathcal{S}_n(\mathbb{R})$.

2°) Montrer que les points extrémaux de Ω_1 sont exactement les matrices de la forme $xx^\top$, où x est un vecteur unitaire de $\mathbb{R}^n$.

Solution : 1°) Ω_1 est par définition l'intersection du cône convexe fermé $\mathcal{P}_n(\mathbb{R})$ avec l'hyperplan affine d'équation $\langle\langle M, I_n\rangle\rangle = 1$: c'est donc un convexe fermé. Par ailleurs, si $M \in \mathcal{S}_n(\mathbb{R})$ et si $\lambda_1, \ldots, \lambda_n$ désignent ses valeurs propres

$$\| M \|^2 := \text{tr}(M^2) = \sum_{i=1}^{n} \lambda_i^2 \quad \text{(d'accord ?)}.$$

Si à présent $M \in \Omega_1$, ses valeurs propres λ_i sont toutes positives (puisque $M \in \mathcal{P}_n(\mathbb{R})$) et inférieures à 1 (puisque $\sum_{i=1}^{n} \lambda_i = \text{tr}M = 1$). Par conséquent

$$\| M \|^2 \leqslant n \text{ pour tout } M \in \Omega_1.$$

Ω_1 est donc borné : c'est un convexe compact de $\mathcal{S}_n(\mathbb{R})$.

2°) Montrons que Ω_1 est l'enveloppe convexe des $xx^\top$, où x est de norme 1 :

$$\Omega_1 = \text{conv}\left\{xx^\top \mid \| x \| = 1\right\}. \tag{6.11}$$

Toute matrice de la forme $xx^\top$, avec $\|x\| = 1$, est symétrique semi-définie positive et de trace 1 (car $\mathrm{tr}(xx^\top) = \| x \|^2$) ; par conséquent, toute combinaison convexe de telles matrices est dans Ω_1.

Réciproquement, soit $M \in \Omega_1$. Par décomposition spectrale, on peut exprimer M sous la forme

$$M = \sum_{i=1}^{n} \lambda_i x_i x_i^\top , \tag{6.12}$$

où les x_i sont des vecteurs unitaires de $\mathbb{R}^n$ et les λ_i les valeurs propres de M.

Mais ces valeurs propres sont positives et de somme égale à 1 ; ce qu'exprime (6.12) est donc que M est une combinaison convexe d'éléments de $\left\{ xx^\top \mid \| x \| = 1 \right\}$.

Visualisation géométrique de (6.11)

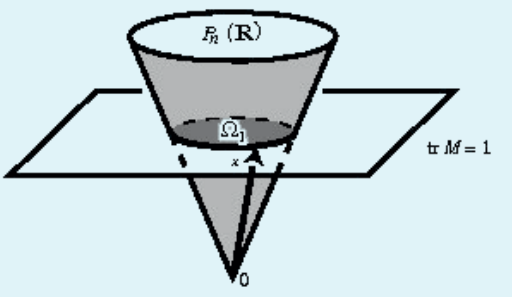


Figure 18.

D'après l'expression (6.11) de Ω_1, les points extrémaux de Ω_1 figurent nécessairement parmi les matrices de la forme $xx^\top$, où $\| x \| = 1$.

Réciproquement, soit $xx^\top$ avec $\| x \| = 1$ et montrons qu'il s'agit d'un point extrémal de Ω_1. Supposons que l'on ait la décomposition

$$\left.\begin{array}{l} xx^\top = \alpha M + (1-\alpha) N, \\ \alpha \in \,]0,1[\,, \\ M \text{ et } N \text{ dans } \Omega_1, \end{array}\right\} \tag{6.13}$$

et montrons que cela conduit nécessairement à $xx^\top = M = N$.

En faisant le produit scalaire $\ll \cdot, \cdot \gg$ avec $xx^\top$ dans chaque membre de la 1re ligne de (6.13), on obtient

$$1 = \alpha \langle Mx, x \rangle + (1-\alpha) \langle Nx, x \rangle . \tag{6.14}$$

Comme $\langle Mx, x\rangle \leqslant \lambda_1(M) \leqslant 1$ ($\lambda_1(M)$ désignant la plus grande valeur propre de M) et $\langle Nx, x\rangle \leqslant \lambda_1(N) \leqslant 1$, il vient de (6.14) que $\lambda_1(M) = \lambda_1(N) = 1$.

Puisque les valeurs propres de M (et de N) sont positives et de somme égale à 1, la seule possibilité pour M (et N) est d'être de rang 1, $M = uu^\top$ avec $\| u \| = 1$ (et $N = vv^\top$, avec $\| v \| = 1$). Avec la 1re ligne de (6.13), on conclut ensuite à $xx^\top = uu^\top = vv^\top$.

Commentaire :

– On aurait pu aussi démontrer que Ω_1 est l'ensemble des matrices de la forme $U\mathrm{diag}(\lambda_1, \ldots, \lambda_n)U^\top$, où U est orthogonale et $(\lambda_1, \ldots, \lambda_n)$ dans le simplexe-unité Λ_n de $\mathbb{R}^n$. La recherche des points extrémaux de Ω_1 revient alors à celle des points extrémaux de Λ_n.

– Conséquence sur la formulation variationnelle usuelle de la plus grande valeur propre d'une matrice symétrique réelle : Soient $A \in \mathcal{S}_n(\mathbb{R})$ et $\lambda_1(A)$ sa plus grande valeur propre ; alors

$$\begin{aligned}\lambda_1(A) &= \max\{\langle Ax, x\rangle \mid \| x \| = 1\} \ \text{(classique, } \textit{cf.} \text{ Exercice III.4)}\\ &= \max\left\{\left\langle\left\langle A, xx^\top\right\rangle\right\rangle \mid \| x \| = 1\right\}\\ &= \max\{\langle\langle A, M\rangle\rangle \mid M \in \Omega_1\}.\end{aligned}$$

Prendre le *max* de $\langle\langle A, \cdot\rangle\rangle$ sur $\left\{xx^\top \mid \| x \| = 1\right\}$ ou sur Ω_1 revient au même ; λ_1 est en fait la fonction d'appui de Ω_1. En somme, la formulation classique de Rayleigh (1re expression au-dessus) nous cache des choses (le plus grand ensemble sur lequel le *max* de $\langle\langle A, \cdot\rangle\rangle$ est $\lambda_1(A)$ est Ω_1), mais dit l'essentiel (on prend le *max* sur les points extrémaux de Ω_1).

– Prolongement de l'exercice avec la généralisation que voici.

Soit m un entier compris entre 1 et n, et soit

$$\Omega_m := \{M \in \mathcal{P}_n(\mathbb{R}) \mid \ \mathrm{tr}M = m \text{ et } \lambda_1(M) \leqslant 1\}.$$

Dans la définition de Ω_1, il n'était pas nécessaire de faire apparaître la contrainte $\lambda_1(M) \leqslant 1$ car elle était automatiquement vérifiée. Évidemment $\Omega_n = \{I_n\}$. Résultat général : Ω_m est un convexe compact de $\mathcal{S}_n(\mathbb{R})$; c'est l'enveloppe convexe des $XX^\top$, X parcourant l'ensemble des matrices de format $n \times m$ telles que $X^\top X = I_m$, ces dernières étant les points extrémaux de Ω_m. En d'autres termes : l'enveloppe convexe des matrices de projections orthogonales de rang m est exactement l'ensemble des matrices symétriques dont les valeurs propres sont entre 0 et 1 et dont la trace vaut m.

Pour démontrer ce résultat, on commence par montrer que Ω_m est l'ensemble des matrices de la forme $U\mathrm{diag}(\lambda_1,\dots,\lambda_n)U^\top$, où U est orthogonale et $(\lambda_1,\dots,\lambda_n)$ dans le polyèdre convexe compact Π_m suivant :

$$\Pi_m := \left\{ (\lambda_1,\dots,\lambda_n) \in \mathbb{R}^n \,|0 \leqslant \lambda_i \leqslant 1 \text{ pour tout } i = 1,\dots,n \right.$$
$$\left. \text{et} \quad \sum_{i=1}^n \lambda_i = m \right\}.$$

La détermination des points extrémaux de Π_m (*cf.* Exercice V.15) conduit alors au résultat annoncé.

La fonction d'appui σ_m de Ω_m

$$A \in \mathcal{S}_n(\mathbb{R}) \longmapsto \sigma_m(A) = \max_{M\in\Omega_m} \langle\langle M, A\rangle\rangle$$

se trouve être

$$A \longmapsto \sigma_m(A) = \text{somme des } m \text{ plus grandes valeurs propres de } A.$$

Cette formulation variationnelle est due à Ky Fan (1949).

****** Exercice VI.13.*** Soit $\mathcal{K}(\mathbb{R}^n)$ l'ensemble des compacts non vides de $\mathbb{R}^n$.

1°) Soient A, B, C dans $\mathcal{K}(\mathbb{R}^n)$. Montrer que :

$$(A + B = A + C) \Rightarrow (\mathrm{conv}\, B = \mathrm{conv}\, C).$$

2°) Soit B arbitraire dans $\mathcal{K}(\mathbb{R}^n)$. Montrer que :

$$n(\mathrm{conv}\, B) + B = n(\mathrm{conv}\, B) + \mathrm{conv}\, B.$$

Montrer sur un exemple que ce résultat ne saurait être vrai avec, au lieu de n, un coefficient strictement inférieur à n.

3°) Déduire de ce qui précède : Si B et C sont dans $\mathcal{K}(\mathbb{R}^n)$,

$$\begin{pmatrix} \forall A \in \mathcal{K}(\mathbb{R}^n), \\ (A + B = A + C) \Rightarrow (B = C) \end{pmatrix} \Leftrightarrow \begin{pmatrix} B \text{ et } C \text{ sont} \\ \text{convexes} \end{pmatrix}.$$

Solution : 1°) Les fonctions d'appui σ_A, σ_B et σ_C de A, B et C sont des fonctions convexes partout finies sur $\mathbb{R}^n$; l'égalité $A + B = A + C$ implique en termes de fonctions d'appui :

$$\sigma_A + \sigma_B = \sigma_A + \sigma_C,$$

d'où $\sigma_B = \sigma_C$ et, par conséquent, $\operatorname{conv} B = \operatorname{conv} C$ (rappelons que lorsque $S \in \mathcal{K}(\mathbb{R}^n)$, $\operatorname{conv} S$ est compact donc fermé).

2°) Ce qu'il faut démontrer est l'inclusion :

$$n(\operatorname{conv} B) + \operatorname{conv} B \subset n(\operatorname{conv} B) + B.$$

Soit $C := \operatorname{conv} B$ et désignons par $\operatorname{ext} C$ l'ensemble des points extrémaux de C. Comme $C := \operatorname{conv}(\operatorname{ext} C)$ et $C \subset \mathbb{R}^n$, tout $c \in C$ peut s'écrire comme combinaison convexe de $n+1$ éléments de $\operatorname{ext} C$:

$c = \sum_{i=1}^{n+1} \alpha_i x_i$, avec $(\alpha_1, \ldots, \alpha_{n+1}) \in \Delta_{n+1}$ (simplexe-unité de $\mathbb{R}^{n+1}$), $x_i \in \operatorname{ext} C$ et $\alpha_i \geqslant 0$ pour tout i.

L'un des α_i est nécessairement supérieur ou égal à $\frac{1}{n+1}$; sans perte de généralité, on peut supposer qu'il s'agit de α_1. Alors :

$$\begin{aligned} c &= \frac{1}{n+1} x_1 + \left(\alpha_1 - \frac{1}{n+1}\right) x_1 + \sum_{i=2}^{n+1} \alpha_i x_i \\ &= \frac{1}{n+1} x_1 + \frac{n}{n+1} \sum_{i=1}^{n+1} \beta_i x_i, \quad \text{où } (\beta_1, \ldots, \beta_{n+1}) \in \Delta_{n+1}. \end{aligned}$$

Par conséquent,

$$(n+1)c = x_1 + n \left(\sum_{i=1}^{n+1} \beta_i x_i \right) \in \operatorname{ext} C + nC.$$

Mais tout point extrémal de $C = \operatorname{conv} B$ est nécessairement dans B (*cf.* Exercice VI.4) ; donc, finalement, $(n+1)c \in B + n(\operatorname{conv} B)$.

Comme $(n+1)\operatorname{conv} B = n(\operatorname{conv} B) + \operatorname{conv} B$ (cela est dû à la convexité de $\operatorname{conv} B$), on a bien démontré l'inclusion $n(\operatorname{conv} B) + \operatorname{conv} B \subset n(\operatorname{conv} B) + B$.

Considérons $B := \{0, \mathrm{e}_1, \ldots, \mathrm{e}_n\}$, où e_i est le i^{me} vecteur de la base canonique de $\mathbb{R}^n$, de sorte que

$$\operatorname{conv} B = \left\{ (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n \mid \sum_{i=1}^{n} \alpha_i \leqslant 1 \text{ et } \alpha_i \geqslant 0 \text{ pour tout } i \right\}.$$

Alors, pour $k < n$, $k(\operatorname{conv} B) + B$ est strictement contenu dans $(k+1)\operatorname{conv} B$.

3°) D'après la 1re question, si B et C sont convexes,

$$(A + B = A + C) \Rightarrow (B = C),$$

et ce quel que soit $A \in \mathcal{K}(\mathbb{R}^n)$.

Réciproquement, si B ou C n'est pas convexe – disons B – on peut avoir $A + B = A + C$ sans avoir $B = C$; il suffit pour cela, d'après la 2e question, de considérer $A = n(\text{conv}B)$.

***Problème VI.14.** Soit C un convexe (non vide) de $\mathbb{R}^n$ et $x_0 \in C$. On note $K_C(x_0)$ le cône engendré par $C - x_0$ (c'est-à-dire $K_C(x_0) = \mathbb{R}^+(C - x_0)$) et $L_C(x_0)$ le plus grand sous-espace vectoriel contenu dans $K_C(x_0)$. On appelle $L_C(x_0)$ le *sous-espace facial de x_0 relatif à C*. L'ensemble

$$F_C(x_0) := [x_0 + L_C(x_0)] \cap C$$

est appelé *face de x_0 dans C*.

1°) Vérifier que

$$L_C(x_0) = \{d \in \mathbb{R}^n \mid \exists\, \varepsilon > 0 \text{ tel que } x_0 + \lambda d \in C \text{ pour } |\, \lambda \,| < \varepsilon\}\,.$$

En déduire que $F_C(x_0)$ est la réunion de tous les segments $[a, b]$ contenus dans C et ayant x_0 dans leur intérieur relatif.

2°) Montrer que x_0 est un point extrémal de C si et seulement si $L_C(x_0)$ se réduit à $\{0\}$ (ou, ce qui revient au même, $F_C(x_0)$ se réduit à $\{x_0\}$).

3°) Soient D un autre convexe de $\mathbb{R}^n$ et $x_0 \in C \cap D$. Établir

$$F_{C\cap D}(x_0) = F_C(x_0) \cap F_D(x_0),$$
$$L_{C\cap D}(x_0) = L_C(x_0) \cap L_D(x_0).$$

4°) Vérifier que

$$\text{aff } (F_C(x_0)) = x_0 + L_C(x_0).$$

(aff S désigne le sous-espace affine engendré par S) et par conséquent

$$\dim F_C(x_0) = \dim L_C(x_0).$$

$\dim L_C(x_0)$ est ce qu'il est convenu d'appeler la *dimension faciale de x_0 relative à C*.

5°) Soient $A : \mathbb{R}^p \to \mathbb{R}^q$ une application affine et C un convexe compact non vide de $\mathbb{R}^p$.

a) On considère $y_0 \in A(C)$. Indiquer pourquoi $A^{-1}(y_0)$ est convexe et fermé.

b) On prend un point extrémal x_0 de $A^{-1}(y_0) \cap C$ (Pourquoi en existe-t-il ?). Soit z un élément de $L_C(x_0)$.

- Montrer que $A(x_0 + z) - y_0 \in L_{A(C)}(y_0)$.
- Établir que $A(x_0 + z) = y_0$ implique $z = 0$.
- En déduire que $A : \text{aff}(F_C(x_0)) \to \text{aff}(F_{A(C)}(y_0))$ ainsi que $A : F_C(x_0) \to F_{A(C)}(y_0)$ sont injectives, et que

$$\dim L_C(x_0) \leqslant \dim L_{A(C)}(y_0).$$

On a ainsi démontré que *pour tout* $y_0 \in A(C)$, il existe $x_0 \in A^{-1}(y_0) \cap C$ *tel que la dimension faciale de* x_0 *relative à* C *soit inférieure à la dimension faciale de* y_0 *relative à* $A(C)$.

6°) Soient $A_1, \ldots, A_k$, k compacts non vides de $\mathbb{R}^n$ et

$$C := \text{conv}\, A_1 \times \text{conv}\, A_2 \times \ldots \times \text{conv}\, A_k.$$

On considère $A : \mathbb{R}^{nk} \to \mathbb{R}^n$ définie par

$$x = (x_1, \ldots, x_k) \in \mathbb{R}^n \times \ldots \times \mathbb{R}^n \longmapsto A(x) := \sum_{i=1}^{k} x_i.$$

a) Rappeler pourquoi $A(C) = \text{conv}(A_1 + \ldots + A_k)$.

b) Soit $y = \sum_{i=1}^{k} x_i$ avec $x_i \in \text{conv}A_i$ pour tout $i = 1, \ldots, k$. Montrer que :

$$\dim L_C(x) = \sum_{i=1}^{k} \dim L_{\text{conv}\, A_i}(x_i).$$

Montrer qu'un point extrémal de $\text{conv}A_i$ appartient nécessairement à A_i. Déduire du résultat de la 5$^{\text{e}}$ question qu'il existe une représentation $y = \sum_{i=1}^{k} \overline{x}_i$ avec :

$\overline{x}_i \in \text{conv}A_i$ pour tout i,

$\text{Card}\,\{i \mid \overline{x}_i \notin A_i\} \leqslant n$ (c'est le théorème de Shapley et Folkman).

On suppose que $\overline{x}_{i_0} \in \text{int}(\text{conv}A_{i_0})$ pour un certain i_0. Que peut-on dire des autres $\overline{x}_i$?

7°) Soit Λ_k le simplexe-unité (ou simplexe élémentaire) de $\mathbb{R}^k$, c'est-à-dire

$$\Lambda_k := \left\{ \lambda = (\lambda_1, \ldots, \lambda_k) \mid \lambda_i \geqslant 0 \ \text{ pour tout } i, \ \sum_{i=1}^{k} \lambda_i = 1 \right\}.$$

Pour tout $\lambda \in \Lambda_k$, on note $n(\lambda) := \mathrm{Card}\{i \mid \lambda_i > 0\}$.

a) Montrer que la dimension faciale de λ relative à Λ_k est $n(\lambda) - 1$.

b) Soit B un sous-ensemble de $\mathbb{R}^n$ et considérons $y \in \mathrm{conv}B$. Il existe donc $\lambda = (\lambda_1, \ldots, \lambda_k) \in \Lambda_k$, $b_1, \ldots, b_k$ dans B tels que $y = \sum_{i=1}^{k} \lambda_i b_i$.

Considérons $A : \mathbb{R}^k \to \mathbb{R}^n$ définie par

$$\alpha = (\alpha_1, \ldots, \alpha_k) \in \mathbb{R}^k \longmapsto A(\alpha) := \sum_{i=1}^{k} \alpha_i b_i.$$

Déduire de la 5e question qu'il existe une représentation $y = \sum_{i=1}^{k} \overline{\lambda}_i b_i$, $\overline{\lambda} = (\overline{\lambda}_1, \ldots, \overline{\lambda}_k) \in \Lambda_k$, telle que $\dim L_{\Lambda_k}(\overline{\lambda}) \leqslant n$. Quel résultat classique d'Analyse convexe retrouve-t-on ainsi ?

8°) On suppose que l'état $x(k)$ d'un système à l'instant $k \in \{0, 1, \ldots, T+1\}$ est décrit de la manière suivante :

$$(E) \begin{cases} x(0) = x_0 \text{ [état initial] ;} \\ \quad \forall k \in \{0, 1, \ldots, T\}, \ x(k+1) = A(k)x(k) + B(k)u(k) \\ \qquad\qquad\qquad\qquad \text{(loi d'évolution du système),} \end{cases}$$

où $A(k)$ est une matrice réelle (n, n), $B(k)$ une matrice (n, m) et $u(0), u(1), \ldots, u(T)$ une suite de vecteurs de $\mathbb{R}^m$ appelés *contrôles* (ou *commandes*). Les contrôles $u(k)$ sont assujettis aux contraintes

$$(\mathcal{A}_d) \qquad u(k) \in U(k) \quad \text{pour tout } k = 0, 1, \ldots, T,$$

où $U(0), U(1), \ldots, U(T)$ sont des convexes compacts non vides de $\mathbb{R}^m$.

À chaque suite $u := (u(0), \ldots, u(T))$ de contrôles admissibles (*i.e.*, vérifiant $(\mathcal{A}_d)$) est associée par (E) une suite d'états $(x_u(0), x_u(1), \ldots, x_u(T), x_u(T+1))$. On dit que l'état

$x_u(T+1)$ est atteint en utilisant la suite u de contrôles.

Soit $y \in \mathbb{R}^n$ un élément qui peut être atteint par une suite u de contrôles.

Montrer que y peut aussi être atteint en utilisant une suite $\overline{u} = (\overline{u}(0), \ldots, \overline{u}(T))$ de contrôles admissibles pour laquelle

$$\text{Card}\,\{k \mid \overline{u}(k) \text{ n'est pas extrémal dans } U(k)\} \leqslant n.$$

On utilisera pour cela le résultat de la 5$^{\text{e}}$ question avec

$$U(0) \times \ldots \times U(T) =: C$$

et $A : \mathbb{R}^{m(T+1)} \to \mathbb{R}^n$ définie par

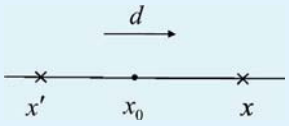
$$u = (u(0), \ldots, u(T)) \longmapsto A(u) := x_u(T+1).$$

Solution : 1°) Le plus grand sous-espace vectoriel contenu dans (le cône convexe) $K_C(x_0)$ est $K_C(x_0) \cap -K_C(x_0) = L_C(x_0)$.

Pour les différentes notions proposées, il est recommandé de faire des dessins pour visualiser les choses ; considérer notamment des polyèdres de $\mathbb{R}^2$ et $\mathbb{R}^3$.

Comme $K_C(x_0) = \bigcup_{\alpha>0} \alpha(C - x_0)$ et $L_C(x_0) = K_C(x_0) \cap -K_C(x_0)$, on a :

$$(d \in L_C(x_0)) \Leftrightarrow \begin{pmatrix} \exists\ \alpha > 0 \text{ tel que } x = x_0 + \frac{d}{\alpha} \in C, \text{ et} \\ \exists\ \alpha' > 0 \text{ tel que } x' = x_0 - \frac{d}{\alpha'} \in C \end{pmatrix}.$$



C étant convexe, le segment $[x, x']$ joignant x à x' est contenu dans C. Par conséquent, en prenant $0 < \varepsilon \leqslant \min\left(\frac{1}{\alpha}, \frac{1}{\alpha'}\right)$, on aura

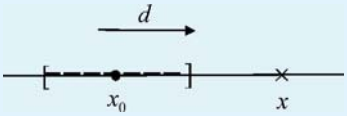
$$x_0 + \lambda d \in C \text{ dès que } \mid \lambda \mid < \varepsilon. \tag{6.15}$$

Réciproquement, si l'on considère d vérifiant (6.15) pour un certain $\varepsilon > 0$, il est immédiat que d et $-d$ appartiennent à $K_C(x_0)$.

Par construction, $F_C(x_0)$ est un convexe contenu dans C et contenant x_0. Tout élément x de $F_C(x_0)$ est un élément *de* C qui peut s'écrire sous la forme

$$x = x_0 + d \text{ avec } d \in L_C(x_0).$$

Puisque $d \in L_C(x_0)$, il existe $1 \geqslant \overline{\varepsilon} > 0$ tel que
$[x_0 - \overline{\varepsilon}d, x_0 + \overline{\varepsilon}d] \subset \underbrace{[x_0 - \overline{\varepsilon}d, x_0 + d]}_{[a,b]} \subset C.$



Ainsi, $x = x_0 + d$ appartient à un segment $[a, b] \subset C$ tel que

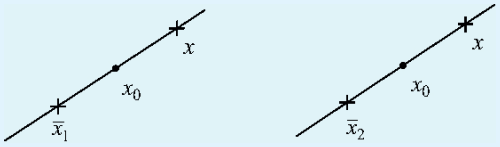
$$x_0 \in]a, b[= \text{ir}\,([a, b]) \text{ lorsque } d \neq 0,$$
$$x_0 \in \{x_0\} = \text{ir}\,(\{x_0\}) \text{ lorsque } d = 0.$$

Réciproquement, si x appartient à un segment contenu dans C et ayant x_0 dans son intérieur relatif, il est évident que $x - x_0 \in L_C(x_0)$.

2°) Par définition, x_0 est extrémal s'il n'est pas intérieur à un segment contenu dans C. Dire ceci est exactement dire que $F_C(x_0) = \{x_0\}$ (ou, d'une manière équivalente, $L_C(x_0) = \{0\}$).

3°) Soit $x \in F_{C\cap D}(x_0),\ x \neq x_0$: il existe alors $\overline{x} \in C \cap D$ et $\rho \in]0, 1[$ tels que $x_0 = \rho\overline{x} + (1 - \rho)x$; on en déduit immédiatement que x appartient à $F_C(x_0)$ et à $F_D(x_0)$.

Inversement, si $x \in F_C(x_0) \cap F_D(x_0),\ x \neq x_0$, on a la situation suivante :



$$x_0 = \rho_1\overline{x}_1 + (1 - \rho_1)x \text{ pour un certain } \overline{x}_1 \in C \text{ et un } \rho_1 \in]0, 1[\ ;$$
$$x_0 = \rho_2\overline{x}_2 + (1 - \rho_2)x \text{ pour un certain } \overline{x}_2 \in D \text{ et un } \rho_2 \in]0, 1[\,.$$

Comme $x_0 \in C \cap D$, on a nécessairement $[\overline{x}_1, x] \subset C \cap D$ ou bien $[\overline{x}_2, x] \subset C \cap D$, donc $x \in F_{C\cap D}(x_0)$.

La deuxième assertion, $L_{C\cap D}(x_0) = L_C(x_0) \cap L_D(x_0)$, se démontre de la même manière.

4°) L'enveloppe affine $\text{aff}(F_C(x_0))$ de $F_C(x_0)$ est, par définition, le plus petit sous-espace affine de $\mathbb{R}^n$ contenant $F_C(x_0)$. C'est l'ensemble de tous les vecteurs x de la forme $x = \sum_{i=1}^{m} \rho_i x_i$, avec $m \geqslant 1,\ x_i \in F_C(x_0)$ pour tout i et $\sum_{i=1}^{m} \rho_i = 1$.

$x_0 + L_C(x_0)$ est un sous-espace affine, d'où $\text{aff}(F_C(x_0)) \subset x_0 + L_C(x_0)$.

Pour l'inclusion inverse, considérons $d \neq 0$ dans $L_C(x_0)$. Il existe alors $\overline{\varepsilon} > 0$ tel que $\overline{x} := x_0 + \overline{\varepsilon} d$ appartienne à C et par conséquent à $F_C(x_0)$. Le point $x := x_0 + d$ se trouvant sur la droite déterminée par x_0 et $\overline{x}$ est donc dans $\mathrm{aff}(F_C(x_0))$. La dimension de $F_C(x_0)$ est celle de $\mathrm{aff}(F_C(x_0))$; c'est par conséquent la dimension de $L_C(x_0)$.

5°) a) $A^{-1}(y_0)$ est convexe fermé comme image réciproque de $\{y_0\}$ par l'application affine (et donc) continue A.

b) C étant compact, il en est de même de $A^{-1}(y_0) \cap C$. Puisque y_0 a été choisi dans $A(C)$, $A^{-1}(y_0) \cap C$ n'est pas vide. $A^{-1}(y_0) \cap C$ étant un convexe compact non vide de $\mathbb{R}^p$, il possède un point extrémal.

Soit $z \in L_C(x_0)$ (si, de plus, $x_0 + z \in C$, on a $x := x_0 + z \in F_C(x_0)$). Rappelons que

$$A(x_0 + \lambda z) = y_0 + \lambda A_0(z) \text{ pour tout } \lambda \in \mathbb{R}, \tag{6.16}$$

où A_0 désigne l'application linéaire de $\mathbb{R}^p$ dans $\mathbb{R}^q$ associée à A.

– Considérons $\varepsilon > 0$ vérifiant

$$x_0 + \lambda z \in C \text{ dès que } |\lambda| < \varepsilon.$$

La relation (6.16) indique immédiatement que $y_0 + \lambda A_0(z) \in A(C)$ lorsque $|\lambda| < \varepsilon$; donc $A(x_0 + z) - y_0 = A_0(z) \in L_{A(C)}(y_0)$. En somme :

$$(z \in L_C(x_0)) \Rightarrow (A_0(z) \in L_{A(C)}(y_0)) \ ;$$
$$(x = x_0 + z \in F_C(x_0)) \Rightarrow (A(x) = y_0 + A_0(z) \in F_{A(C)}(y_0)).$$

– Si z est tel que $A(x_0 + z) = y_0$, il résulte facilement de (6.16) que $z \in L_{A^{-1}(y_0)}(x_0)$. Ainsi,

$$z \in L_C(x_0) \cap L_{A^{-1}(y_0)}(x_0) = L_{A^{-1}(y_0) \cap C}(x_0) = \{0\}$$
$$\text{(puisque } x_0 \text{ est extrémal dans } A^{-1}(y_0) \cap C).$$

– Soient z_1 et z_2 deux éléments de $L_C(x_0)$ pour lesquels $A(x_0 + z_1) = A(x_0 + z_2)$.

Pour $|\lambda|$ suffisamment petit, disons $|\lambda| < \varepsilon$, on a :

$$x_0 + \lambda(z_1 - z_2) \in C, \ A(x_0 + \lambda(z_1 - z_2)) = y_0.$$

Par conséquent, $x_0 + \lambda(z_1 - z_2) \in F_C(x_0)$ et, d'après ce qui a été démontré précédemment, $z_1 - z_2 = 0$ nécessairement.

L'application

$$\begin{array}{ccc} A : \mathrm{aff}\,(F_C(x_0)) & \longrightarrow & \mathrm{aff}(F_{A(C)}(y_0)) \\ (= x_0 + L_C(x_0)) & & (= y_0 + L_{A(C)}(y_0)) \end{array}$$

ou encore

$$A : F_C(x_0) \longrightarrow F_{A(C)}(y_0)$$

est injective et, par conséquent, la dimension de $L_C(x_0)$ est inférieure à la dimension de $L_{A(C)}(y_0)$.

Illustration : A est la projection orthogonale de $\mathbb{R}^3$ sur $\mathbb{R}^2$.

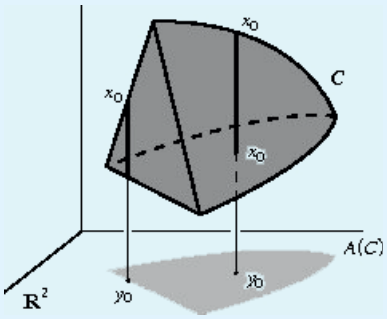


Figure 19.

6°) a) $C := \operatorname{conv} A_1 \times \ldots \times \operatorname{conv} A_k = \operatorname{conv}(A_1 \times \ldots \times A_k)$ (à redémontrer si nécessaire ; *cf.* Exercice VI.10).

L'image par l'application linéaire A de l'enveloppe convexe de $A_1 \times \ldots \times A_k$ est $\sum_{i=1}^{k} \operatorname{conv} A_i$; c'est aussi l'enveloppe convexe de l'image par A de $A_1 \times \ldots \times A_k$, c'est-à-dire $\operatorname{conv}\left(\sum_{i=1}^{k} A_i\right)$. On peut aussi démontrer directement que

$$\operatorname{conv}\left(\sum_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} \operatorname{conv} A_i.$$

(*cf.* Exercice VI.11).

b) Si $x = (x_1, \ldots, x_k) \in C = \operatorname{conv} A_1 \times \ldots \times \operatorname{conv} A_k$, les « perturbations sur x peuvent avoir lieu indépendamment sur chaque coordonnée x_i », de sorte que

$$L_C(x) = (L_{\operatorname{conv} A_1}(x_1)) \times (L_{\operatorname{conv} A_2}(x_2)) \times \ldots \times (L_{\operatorname{conv} A_k}(x_k))$$

et

$$\dim L_C(x) = \sum_{i=1}^{k} \dim L_{\operatorname{conv} A_i}(x_i).$$

Soit $y \in A(C)$; d'après le résultat de la 5e question, il existe $\overline{x} = (\overline{x}_1, \ldots, \overline{x}_k)$ dans C tel que :

$$y = \sum_{i=1}^{k} \overline{x}_i,$$

$$\sum_{i=1}^{k} \dim\, L_{\mathrm{conv}A_i}(\overline{x}_i) \leqslant \dim\, A(C) \leqslant n. \tag{6.17}$$

Par suite,

$$\mathrm{Card}\,\{i \mid \dim\, L_{\mathrm{conv}A_i}(\overline{x}_i) \geqslant 1\} \leqslant n. \tag{6.18}$$

Si $\overline{x}_i$ est un point extrémal de $\mathrm{conv}A_i$, il appartient nécessairement à A_i (résultat classique sur les points extrémaux de l'enveloppe convexe d'un ensemble ; *cf.* Exercice VI.4).

D'où

$$\{i \mid \overline{x}_i \notin A_i\} \subset \{i \mid \dim\, L_{\mathrm{conv}A_i}(\overline{x}_i) \geqslant 1\}$$

et le résultat annoncé découle alors de (6.18).

Si l'on sait que $\overline{x}_{i_0} \in \mathrm{int}(\mathrm{conv}A_{i_0})$ pour un certain indice i_0, alors $\dim\, L_{\mathrm{conv}A_{i_0}}(\overline{x}_{i_0}) = n$ et l'inégalité (6.17) implique

$$\dim\, L_{\mathrm{conv}A_i}(\overline{x}_i) = 0 \text{ pour tout } i \neq i_0,$$

ce qui implique

$$\overline{x}_i \in A_i \quad \text{pour tout} \quad i \neq i_0.$$

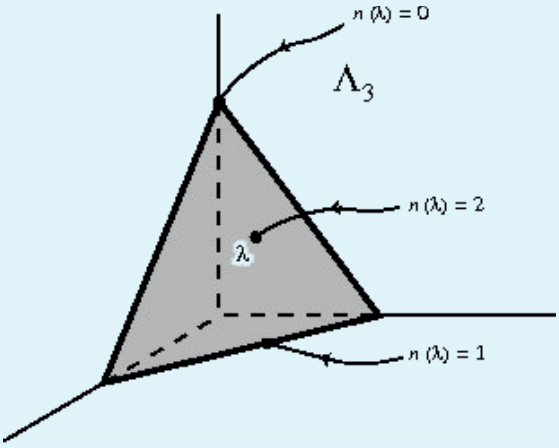


FIGURE 20.

7°) a) Soient $\lambda \in \Lambda_k$ et $d = (d_1, \ldots, d_k)$ un vecteur de $L_{\Lambda_k}(\lambda)$. Il existe $\varepsilon > 0$ tel que

$$\sum_{i=1}^{k} (\lambda_i + \alpha d_i) = 1 \text{ pour } \mid \lambda \mid < \varepsilon, \tag{6.19}$$

$$\lambda_i + \alpha d_i \geqslant 0 \text{ pour tout } i = 1, \ldots, k. \tag{6.20}$$

S'il existe i tel que $\lambda_i = 1$, λ est alors un point extrémal de Λ_k et le résultat escompté est vrai avec $n(\lambda) = 1$.

Supposons donc $\lambda_i < 1$ pour tout i. Il résulte de (6.19) et (6.20) que d_i n'est non nul que lorsque $\lambda_i > 0$. De plus, l'égalité (6.19) induit que $\sum\limits_{i=1}^{k} d_i = 0$.

En bref,

$$(d = (d_1, \ldots, d_k) \in L_{\Lambda_k}(\lambda)) \Rightarrow \begin{pmatrix} d_i = 0 \text{ pour tout } i \text{ tel que } \lambda_i = 0, \\ \sum\limits_{i=1}^{k} d_i = 0 \end{pmatrix}.$$

L'implication inverse est facile à vérifier. En définitive, $L_{\Lambda_k}(\lambda)$ est le sous-espace vectoriel d'équation

$$\begin{cases} d_i = 0 \text{ pour tout } i \text{ tel que } \lambda_i = 0, \\ d_1 + \ldots + d_k = 0, \end{cases}$$

ce qui implique

$$\dim L_{\Lambda_k}(\lambda) = n(\lambda) - 1.$$

b) On prend $C = \Lambda_k$. Avec $A : \mathbb{R}^k \to \mathbb{R}^n$ définie par $A(\lambda) := \sum\limits_{i=1}^{k} \lambda_i b_i$, on a $A(C) = \operatorname{conv}\{b_1, \ldots, b_k\}$. D'après le résultat de la 5e question, il existe $\overline{\lambda} = (\overline{\lambda}_1, \ldots, \overline{\lambda}_k) \in \Lambda_k$ tel que

$$y = \sum_{i=1}^{k} \overline{\lambda}_i b_i \quad \text{et} \quad \dim L_{\Lambda_k}(\overline{\lambda}) \leqslant \dim L_{A(C)}(y).$$

Comme $\dim L_{\Lambda_k}(\overline{\lambda}) = n(\overline{\lambda}) - 1$ et que $\dim L_{A(C)}(y) \leqslant n$, on en déduit

$$\operatorname{Card}\left\{i = 1, \ldots, k \mid \overline{\lambda}_i > 0\right\} \leqslant n + 1.$$

On a démontré que y pouvait s'écrire comme combinaison convexe d'au plus $n + 1$ points de B ; c'est le fameux théorème de Carathéodory.

8°) On pose $C = U(0) \times \ldots \times U(T)$ et on définit $A : \mathbb{R}^{m(T+1)} \to \mathbb{R}^n$ par

$$u = (u(0), \ldots, u(T)) \longmapsto A(u) := x_u(T + 1).$$

A est une application affine (c'est facile à vérifier).

Soit y un état qui peut être atteint en utilisant une séquence de contrôles admissibles. D'après le résultat de la 5$^{\text{e}}$ question, il existe $\overline{u} = (\overline{u}(0), \ldots, \overline{u}(T))$ dans $U(0) \times \ldots \times U(T)$ tel que

$$\dim L_{U(0)\times\ldots\times U(T)}(\overline{u}) \leqslant \dim L_{A(C)}(y),$$

soit

$$\sum_{k=0}^{T} \dim L_{U(k)}(\overline{u}(k)) \leqslant n.$$

Par conséquent,

$$\text{Card}\,\{k \mid \overline{u}(k) \text{ n'est pas extrémal dans } U(k)\} \leqslant n.$$

***Exercice VI.15.** Étant donné un convexe fermé non vide D de $(\mathbb{R}^n, \langle\cdot,\cdot\rangle)$, on note p_D l'opérateur de projection sur D.

Soient C un convexe fermé non vide de $\mathbb{R}^n$ et $f : \mathbb{R}^n \to \mathbb{R}$ une fonction convexe différentiable. Soit $\overline{x} \in C$; montrer l'équivalence des assertions suivantes :

(i) $\overline{x}$ minimise f sur C ;
(ii) $\langle \nabla f(\overline{x}), x - \overline{x}\rangle \geqslant 0$ pour tout $x \in C$;
(iii) $\overline{x} = p_C(\overline{x} - t\nabla f(\overline{x}))$, t étant quelconque dans $\mathbb{R}_*^+$;
(iv) $p_{T(C,\overline{x})}(-\nabla f(\overline{x})) = 0$;
(v) $\big\langle \nabla f(\overline{x}), p_{T(C,\overline{x})}(-\nabla f(\overline{x}))\big\rangle \geqslant 0$.

Solution : On rappelle la caractérisation suivante de l'élément $p_D(x)$:

$$(\hat{x} = p_D(x)) \Leftrightarrow (\hat{x} \in D \text{ et } \langle x - \hat{x}, y - \hat{x}\rangle \leqslant 0 \text{ pour tout } y \in D)\,.$$

Dans le cas particulier où l'on projette sur un cône convexe fermé K,

$$(\hat{x} = p_K(x)) \Leftrightarrow (x - \hat{x} \in K^\circ \text{ et } \langle x - \hat{x}, \hat{x}\rangle = 0)\,.$$

$[(i) \Leftrightarrow (ii)]$. De la définition même de $\nabla f(\overline{x})$ et de l'inégalité $f(x) \geqslant f(\overline{x}) + \langle \nabla f(\overline{x}), x - \overline{x}\rangle$ valable pour tout $x \in \mathbb{R}^n$, il vient :

$$\overline{x} \in C \text{ minimise } f \text{ sur } C \Leftrightarrow \langle \nabla f(\overline{x}), x - \overline{x}\rangle \geqslant 0 \text{ pour tout } x \in C. \qquad (6.21)$$

Cette dernière inégalité a lieu si et seulement si

$$\langle \nabla f(\overline{x}), d\rangle \geqslant 0 \text{ pour tout } d \in \overline{\mathbb{R}^+(C - \overline{x})} = T(C, \overline{x}), \qquad (6.22)$$

soit encore

$$-\nabla f(\overline{x}) \in [T(C,\overline{x})]^\circ = N(C,\overline{x}). \tag{6.23}$$

$[(ii) \Leftrightarrow (iii)]$. Étant donné $t > 0$, dire que $\overline{x} = p_C(\overline{x} - t\nabla f(\overline{x}))$ équivaut (d'après la caractérisation de $\hat{u} = p_C(u)$ rappelée au-dessus) à :

$$\langle x - \overline{x}, (\overline{x} - t\nabla f(\overline{x})) - \overline{x}\rangle \leqslant 0 \text{ pour tout } x \in C,$$

c'est-à-dire

$$-t\,\langle x - \overline{x}, \nabla f(\overline{x})\rangle \leqslant 0 \text{ pour tout } x \in C,$$

ce qui n'est autre que (ii).

$[(iv) \Leftrightarrow (ii)]$. $T(C,\overline{x})$ étant un cône convexe fermé de $\mathbb{R}^n$, utilisons la caractérisation de $\hat{u} = p_K(u)$ rappelée au-dessus ; ainsi :

$$0 = p_{T(C,\overline{x})}(-\nabla f(\overline{x})) \Leftrightarrow -\nabla f(\overline{x}) \in [T(C,\overline{x})]^\circ.$$

D'où l'équivalence de (iv) et (ii) *via* (6.23)−(6.22)−(6.21).

$[(iv) \Leftrightarrow (v)]$. Toujours d'après la caractérisation de $p_{T(C,\overline{x})}(u)$, on a :

$$\left\langle u - p_{T(C,\overline{x})}(u), p_{T(C,\overline{x})}(u)\right\rangle = 0,$$

soit

$$\left\langle u, p_{T(C,\overline{x})}(u)\right\rangle = \| p_{T(C,\overline{x})}(u) \|^2. \tag{6.24}$$

Ce que dit (v) est exactement

$$\left\langle -\nabla f(\overline{x}), p_{T(C,\overline{x})}(-\nabla f(\overline{x}))\right\rangle \leqslant 0,$$

c'est-à-dire, d'après (6.24),

$$\| p_{T(C,\overline{x})}(-\nabla f(\overline{x})) \|^2 = 0,$$

ce qui équivaut à (iv).

***Exercice VI.16.** Soit C un convexe fermé non vide de $(\mathbb{R}^n, \langle\cdot,\cdot\rangle)$. La manière usuelle de caractériser l'élément c_x projection de x sur C est la suivante :

(i) $\quad (c_x = p_C(x)) \Leftrightarrow (c_x \in C \text{ et } \langle x - c_x, c - c_x\rangle \leqslant 0 \text{ pour tout } c \in C)$.

Montrer qu'une autre caractérisation de $p_C(x)$ est comme suit :

(ii) $\quad (c_x = p_C(x)) \Leftrightarrow (c_x \in C \text{ et } \langle c - c_x, x - c\rangle \leqslant 0 \text{ pour tout } c \in C)$.

Solution : Il s'agit de montrer l'équivalence des deux inéquations variationnelles figurant dans les caractérisations (i) et (ii).

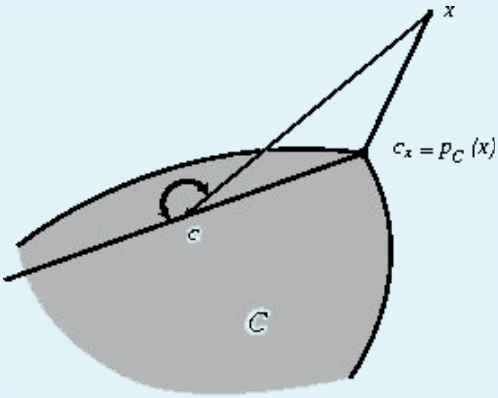


FIGURE 21.

$[(i) \Rightarrow (ii)]$. Observons que pour tout $c \in C$:

$$\langle c - c_x, x - c\rangle = - \parallel c - c_x \parallel^2 + \langle x - c_x, c - c_x\rangle .$$

L'implication voulue s'ensuit.

$[(ii) \Rightarrow (i)]$. Pour $t \in]0,1[$ soit

$$c(t) := c_x + t(c - c_x) ;$$

observant que $c(t) \in C$, on a :

$$t \langle c - c_x, x - c_x\rangle - t^2 \parallel c - c_x \parallel^2 \leqslant 0.$$

En divisant par t, puis en faisant tendre t vers 0, on trouve l'inégalité de (i).

***Exercice VI.17.** Soit C un convexe fermé non vide de $(\mathbb{R}^n, \langle\cdot,\cdot\rangle)$.

1°) Montrer que la fonction $f := d_C^2$ (carré de la fonction-distance euclidienne à C) est différentiable sur $\mathbb{R}^n$ avec

$$\nabla f(x) = 2(x - p_C(x)) \text{ pour tout } x \in \mathbb{R}^n.$$

2°) Déduire de ce qui précède les primitives sur $\mathbb{R}^n$ de l'application p_C, c'est-à-dire les fonctions $g : \mathbb{R}^n \to \mathbb{R}$ différentiables telles que

$$\nabla g(x) = p_C(x) \text{ pour tout } x \in \mathbb{R}^n.$$

Solution : 1°) Considérons $\triangle(h) := d_C^2(x+h) - d_C^2(x)$.

Puisque $d_C^2(x) \leqslant \| x - p_C(x+h) \|^2$, nous avons

$$\triangle(h) \geqslant \| x + h - p_C(x+h) \|^2 - \| x - p_C(x+h) \|^2 = \| h \|^2 + 2\langle x - p_C(x+h), h\rangle .$$

En intervertissant les rôles de x et $x+h$, on obtient de la même manière

$$\triangle(h) \leqslant \| x + h - p_C(x) \|^2 - \| x - p_C(x) \|^2 = \| h \|^2 + 2\langle x - p_C(x), h\rangle .$$

Comme $\| p_C(u) - p_C(v) \| \leqslant \| u - v \|$ pour tout $(u, v) \in \mathbb{R}^n \times \mathbb{R}^n$, on en déduit

$$\triangle(h) = 2\langle x - p_C(x), h\rangle + o(\| h \|) .$$

Par conséquent, $f := d_C^2$ est différentiable en x avec $\nabla f(x) = 2(x - p_C(x))$.

2°) Soit $g_0 : x \in \mathbb{R}^n \longmapsto g_0(x) := \frac{1}{2}[\| x \|^2 - d_C^2(x)]$. D'après ce qui précède, cette fonction est différentiable sur $\mathbb{R}^n$ et

$$\nabla g_0(x) = p_C(x) \text{ pour tout } x \in \mathbb{R}^n.$$

Ainsi, les fonctions g primitives de l'application p_C sont les fonctions de la forme suivante :

$$g = g_0 + r, \ r \in \mathbb{R}.$$

***Exercice VI.18.** Soit $f : \mathbb{R}^n \to \mathbb{R}$ convexe et différentiable, soit C un convexe fermé de $\mathbb{R}^n$. On rappelle que les points $\overline{x} \in C$ minimisant f sur C sont exactement ceux pour lesquels :

$$\langle \nabla f(\overline{x}), c - \overline{x}\rangle \geqslant 0 \text{ pour tout } c \in C. \tag{6.25}$$

On définit alors $F : \mathbb{R}^n \to \mathbb{R}^n$ de la manière suivante :

$$\forall x \in \mathbb{R}^n, \ F(x) := -(\nabla f(p_C(x)) - p_C(x)) .$$

1°) F est-elle continue ? différentiable ?

2°) Montrer que si x est un point fixe de F, alors sa projection $\overline{x}$ sur C minimise f sur C.

Solution : 1°) f étant convexe, $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ est localement lipschitzienne sur $\mathbb{R}^n$.

Comme $p_C : \mathbb{R}^n \to \mathbb{R}^n$ est lipschitzienne sur $\mathbb{R}^n$, il s'ensuit que F est non seulement continue, mais localement lipschitzienne sur $\mathbb{R}^n$.

Même si f est deux fois différentiable sur $\mathbb{R}^n$ (c'est-à-dire si ∇f est différentiable), la non-différentiabilité de p_C (notamment en les points $\overline{x} \in \mathrm{fr}(\mathrm{C})$) fait qu'on ne peut conclure à la différentiabilité de F.

2°) Soient x un point fixe de F et $\overline{x} := p_C(x)$. Puisque $\nabla f(p_C(x)) - p_C(x) = x$, $\nabla f(\overline{x}) - x = p_C(x)$ et la caractérisation de l'élément $p_C(x)$ conduit à :

$$\langle x - (\nabla f(\overline{x}) - x), c - \overline{x}\rangle \leqslant 0 \text{ pour tout } c \in C.$$

Ceci n'est autre que (6.25) .

Commentaire :

– Plus généralement, si $A : \mathbb{R}^n \to \mathbb{R}^n$ est un opérateur (pas nécessairement un gradient), la recherche de $\overline{x} \in C$ vérifiant

$$\langle A(\overline{x}), c - \overline{x}\rangle \geqslant 0 \text{ pour tout } c \in C$$

(*inéquation variationnelle*) peut être faite *via* la recherche de points fixes de $F := -(A \circ p_C - p_C)$.

– Le résultat de l'exercice est à rapprocher de celui de l'Exercice III.17.

*****Exercice VI.19.*** Soient $S \subset \mathbb{R}^n$ et $x \notin S$. On désigne par $P_S(x)$ l'ensemble des $\overline{x} \in S$ tels que $\| x - \overline{x} \| = d_S(x)$ ($P_S(x)$ est l'ensemble des points de S à distance euclidienne minimale de x).

1°) Montrer l'équivalence des assertions suivantes :

(i) $\overline{x} \in P_S(x)$;

(ii) $\overline{x} \in S$ et $\langle x - \overline{x}, c - \overline{x}\rangle \leqslant \frac{1}{2} \| c - \overline{x} \|^2$ pour tout $c \in S$;

(iii) $\overline{x} \in P_S(\overline{x} + t(x - \overline{x}))$ pour tout $t \in [0, 1]$.

2°) Vérifier que si $\overline{x} \in P_S(x)$, alors pour tout $t \in [0, 1[$, $P_S(\overline{x} + t(x - \overline{x}))$ est le singleton $\{\overline{x}\}$, c'est-à-dire $\overline{x}$ est le seul point de S à distance minimale de x.

3°) Commenter les différences avec la caractérisation de $\overline{x} = p_S(x)$ lorsque S est convexe.

Solution : 1°) On a :

$$\overline{x} \in P_S(x) \Leftrightarrow \begin{pmatrix} \overline{x} \in S \\ \text{et} \\ \| x - \overline{x} \|^2 \leqslant \| x - c \|^2 \quad \text{pour tout} \quad c \in S \end{pmatrix}.$$

En développant $\| x - c \|^2 = \| x - \overline{x} + \overline{x} - c \|^2$, l'inégalité ci-dessus est équivalente à

$$2 \langle x - \overline{x}, c - \overline{x} \rangle \leqslant \| c - \overline{x} \|^2 .$$

D'où $(i) \Leftrightarrow (ii)$.

(ii) peut encore s'écrire : $\overline{x} \in S$ et

$$2 \langle x - \overline{x}, c - \overline{x} \rangle \leqslant \frac{1}{t} \| c - \overline{x} \|^2 \quad \text{pour tout } c \in S \text{ et } t \in]0, 1]. \tag{6.26}$$

Reformulons (6.26) comme suit :

$$2 \langle [\overline{x} + t(x - \overline{x})] - \overline{x}, c - \overline{x} \rangle \leqslant \| c - \overline{x} \|^2 \quad \text{pour tout } c \in S \text{ et } t \in [0, 1].$$

D'après l'équivalence de (i) et (ii) déjà vue, nous venons de démontrer que (ii) équivaut à (iii).

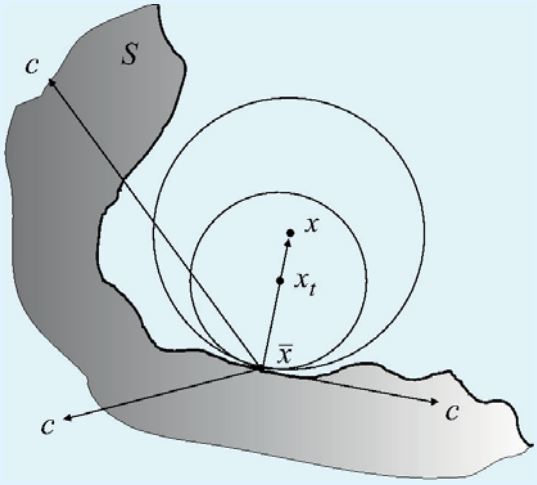


FIGURE 22.

2°) Soit $\tilde{x} \in P_S(\overline{x} + t(x - \overline{x}))$, $0 \leqslant t < 1$, et montrons que $\tilde{x} = \overline{x}$ nécessairement.

D'après la caractérisation (ii) de la 1re question,

$$\langle \overline{x} + t(x - \overline{x}) - \tilde{x}, c - \tilde{x} \rangle \leqslant \frac{1}{2} \| c - \tilde{x} \|^2 \quad \text{pour tout } c \in S.$$

En particulier, pour $c = \overline{x}$, cela donne :

$$\| \overline{x} - \tilde{x} \|^2 + t \langle x - \overline{x}, \overline{x} - \tilde{x} \rangle \leqslant \frac{1}{2} \| \overline{x} - \tilde{x} \|^2 .$$

Or $\overline{x} \in P_S(x)$, d'où

$$\langle x - \overline{x}, c - \overline{x} \rangle \leqslant \frac{1}{2} \| c - \overline{x} \|^2 \text{ pour tout } c \in S,$$

et par conséquent $\langle x - \overline{x}, \tilde{x} - \overline{x} \rangle \leqslant \frac{1}{2} \| \tilde{x} - \overline{x} \|^2$.

En définitive, $\frac{(1-t)}{2} \| \tilde{x} - \overline{x} \|^2 \leqslant 0$, d'où $\tilde{x} = \overline{x}$.

La signification géométrique de ce résultat est claire : si $x_t := \overline{x} + t(x - \overline{x})$ est un point du segment $[\overline{x}, x[$, l'intersection de S et de la boule fermée centrée en x_t et de rayon $\| x_t - \overline{x} \|$ est réduite à $\{\overline{x}\}$.

3°) Contrairement au cas convexe, l'angle de $x - \overline{x}$ et $c - \overline{x}$, $c \in S$, n'est pas nécessairement obtus : on a simplement une majoration de $\langle x - \overline{x}, c - \overline{x} \rangle$ par $\frac{1}{2} \| c - \overline{x} \|^2$. De plus, lorsque $t > 1$, $x_t := \overline{x} + t(x - \overline{x})$ ne se projette pas nécessairement en $\overline{x}$. En clair, on ne peut laisser $t \to +\infty$ dans l'inégalité (6.26) .

Commentaire :

– Les résultats ci-dessus s'étendent sans difficulté au cadre hilbertien. La grande différence est que lorsque S est un fermé de l'espace de Hilbert de dimension infinie H, on n'est pas assuré d'avoir $P_S(x) \neq \phi$. Notons toutefois les résultats suivants :

$$\{x \in H \mid P_S(x) \neq \phi\} \text{ est dense dans } H,$$
$$\{x \in H \mid P_S(x) \text{ est un singleton}\} \text{ est dense dans } H.$$

– Depuis Bunt (1934), on sait que si S est un fermé de $\mathbb{R}^n$ tel que $P_S(x)$ soit un singleton pour tout $x \in \mathbb{R}^n$, alors S est convexe. La même question, posée dans un cadre hilbertien général, reste sans réponse complète à ce jour.

**** Exercice VI.20.** Soient $\{C_k\}$ une suite décroissante de convexes fermés de $\mathbb{R}^n$ et $C := \bigcap_k C_k$ supposé non vide. Montrer que pour tout $x \in \mathbb{R}^n$:

$$p_{C_k}(x) \to p_C(x)$$
$$\text{et} \qquad \text{quand } k \to +\infty.$$
$$d_{C_k}(x) \uparrow d_C(x)$$

($\uparrow$ est une notation pour signifier la convergence en croissant.)

Solution : Puisque $C \subset C_k$, $p_{C_k}(x)$ est élément du compact $B := \{u \mid \| u - x \| \leqslant d_C(x)\}$.

Considérons une sous-suite de $\{p_{C_k}(x)\}$ qui converge vers $\overline{u} \in B$, donc telle que $\| \overline{u} - x \| \leqslant d_C(x)$.

Nous disons que $\overline{u} \in C_k$ pour tout k. En effet, si ce n'était pas le cas, si $\overline{u} \notin C_{k_0}$ pour un certain k_0, alors, en raison de la décroissance de $\{C_k\}$, $\overline{u}$ ne pourrait pas être limite d'une sous-suite de $\{p_{C_k}(x)\}$. Donc $\overline{u} \in \bigcap_k C_k = C$, de sorte que $\| \overline{u} - x \| \geqslant d_C(x)$.

En résumé, $\overline{u} \in C$ et $\| \overline{u} - x \| = d_C(x) : \overline{u} = p_C(x)$.

Le raisonnement ci-dessus étant valable pour toute sous-suite convergente de la suite (bornée) $\{p_{C_k}(x)\}$, on en déduit que (toute) la suite $\{p_{C_k}(x)\}$ converge vers $p_C(x)$.

Comme conséquence, $d_{C_k}(x) = \| x - p_{C_k}(x) \| \uparrow \| x - p_C(x) \| = d_C(x)$.

***Exercice VI.21.** Soit $\mathbb{R}^n$ muni d'un produit scalaire noté $\langle \cdot, \cdot \rangle$ et de la norme euclidienne associée notée $\| \cdot \|$. On considère un convexe fermé non vide C de $\mathbb{R}^n$ et $\overline{x}$ un élément de C.

1re partie

On se propose de démontrer que l'opérateur p_C de projection sur C a en $\overline{x}$ une dérivée directionnelle qui est égale à l'opérateur de projection sur $T(C, \overline{x})$ (cône tangent à C en $\overline{x}$). Soit donc $d \in \mathbb{R}^n$.

1°) Montrer que

$$\frac{p_C(\overline{x} + td) - \overline{x}}{t} = p_{\frac{C-\overline{x}}{t}}(d) \quad \text{pour tout} \quad t > 0.$$

2°) a) Vérifier que si $0 < t_1 < t_2$, on a $\frac{C-\overline{x}}{t_2} \subset \frac{C-\overline{x}}{t_1}$.

b) Qu'est-ce que l'adhérence de $\bigcup_{\tau > 0} \tau(C - \overline{x})$?

c) Soit $\{t_k\}$ une suite de réels strictement positifs de limite 0 ; on pose

$$v_k := p_{\frac{C-\overline{x}}{t_k}}(d).$$

(*i*) Vérifier que la suite $\{v_k\}$ est bornée.

(*ii*) Soit w la limite d'une sous-suite convergente de $\{v_k\}$. Montrer que w est nécessairement la projection de d sur $T(C, \overline{x})$.

d) Déduire de ce qui précède :

$$\frac{p_C(\overline{x} + td) - \overline{x}}{t} \longrightarrow p_{T(C,\overline{x})}(d) \text{ quand } t \longrightarrow 0^+.$$

2^e partie

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

1°) Rappeler, sous ses différentes formes équivalentes, la condition de minimalité du 1^er ordre nécessairement vérifiée en $\overline{x}$ lorsque $\overline{x}$ est un minimum local de f sur C.

2°) Considérons $g_{\overline{x}} : \mathbb{R}^+ \to \mathbb{R}$ définie par $g_{\overline{x}}(t) := f[p_C(\overline{x} - t\nabla f(\overline{x}))]$.

(i) Montrer que la dérivée à droite de $g_{\overline{x}}$ en 0 existe et vaut $- \parallel p_{T(C,\overline{x})}(-\nabla f(\overline{x})) \parallel^2$.

(ii) Vérifier que la dérivée à droite de $g_{\overline{x}}$ en 0 est strictement négative lorsque $\overline{x}$ ne vérifie pas la condition de minimalité du 1^er ordre rappelée à la 1^re question.

3°) On propose l'algorithme suivant (dit du *gradient projeté*) :

$$x_0 \in C,\ x_{k+1} := p_C(x_k - t_k \nabla f(x_k)),$$

où t_k est choisi minimisant $g_{x_k} : t \in \mathbb{R}^+ \longmapsto g_{x_k}(t) := f[p_C(x_k - t\nabla f(x_k))]$ sur $\mathbb{R}^+$ (on suppose qu'un tel t_k existe).

Démontrer que toute limite d'une sous-suite convergente de $\{x_k\}$ est un point $\overline{x}$ de C vérifiant la condition nécessaire de minimalité du 1^er ordre.

Solution : *1^re partie*

1°) D'après la caractérisation de $\overline{u} = p_D(u)$ pour différents D et u (qui est, rappelons-le : $\overline{u} \in D$ et $\langle \delta - \overline{u}, u - \overline{u} \rangle \leqslant 0$ pour tout $\delta \in D$), on a :

$$(\overline{x} + tv = p_C(\overline{x} + td)) \Leftrightarrow \begin{cases} \overline{x} + tv \in C \left(i.e.\ v \in \dfrac{C - \overline{x}}{t} \right) \text{ et} \\ \langle (\overline{x} + td) - (\overline{x} + tv), c - (\overline{x} + tv) \rangle \leqslant 0 \\ \qquad\qquad \text{pour tout } c \in C \\ (i.e., t^2 \left\langle d - v, \frac{c - \overline{x}}{t} - v \right\rangle \leqslant 0 \ \text{ pour tout } c \in C) \ ; \end{cases}$$

$$\left(v = p_{\frac{C-\overline{x}}{t}}(d) \right) \Leftrightarrow \left(v \in \frac{C - \overline{x}}{t} \text{ et } \left\langle d - v, \frac{c - \overline{x}}{t} - v \right\rangle \leqslant 0 \text{ pour tout } c \in C \right).$$

Donc :

$$\frac{p_C(\overline{x} + td) - \overline{x}}{t} = p_{\frac{C-\overline{x}}{t}}(d).$$

2°) a) Si $0 < t_1 < t_2$ et $y \in C$, $\frac{y - \overline{x}}{t_2} = \frac{y' - \overline{x}}{t_1}$ avec $y' := \left(1 - \frac{t_1}{t_2}\right)\overline{x} + \frac{t_1}{t_2} y \in C$. Par conséquent, $(C - \overline{x})/t_2 \subset (C - \overline{x})/t_1$.

b) $\overline{\bigcup_{\tau>0} \tau(C-\overline{x})} = T(C,\overline{x})$.

c) (i) $v_k = p_{\frac{C-\overline{x}}{t_k}}(d) = \frac{p_C(\overline{x}+t_k d)-\overline{x}}{t_k}$; sachant que $\overline{x} \in C$, et donc $p_C(\overline{x}) = \overline{x}$, la propriété de Lipschitz de p_C (sur $\mathbb{R}^n$, avec la constante 1) fait que $\| v_k \| \leqslant \| d \|$.

(ii) Soit w la limite d'une sous-suite convergente de $\{v_k\}$ (sous-suite notée encore $\{v_k\}$) ; il nous faut démontrer que $w = p_{T(C,\overline{x})}(d)$.

Puisque $v_k \in \frac{C-\overline{x}}{t_k}$ et que $v_k \to w$, on a : $w \in \overline{\bigcup_{\tau>0} \tau(C-\overline{x})} = T(C,\overline{x})$.

Soit maintenant $u \in \bigcup_{\tau>0} \tau(C-\overline{x})$, *i.e.*, $u \in \tau(C-\overline{x})$ pour un certain $\tau > 0$. Pour k suffisamment grand, $\frac{1}{t_k} \leqslant \tau$, de sorte que $\tau(C-\overline{x}) \subset \frac{C-\overline{x}}{t_k}$; d'où :

$$\langle d - v_k, u - v_k \rangle \leqslant 0 \text{ (puisque } v_k = p_{\frac{C-\overline{x}}{t_k}}(d)).$$

En passant à la limite (sur k) ci-dessus, on obtient : $\langle d-w, u-w\rangle \leqslant 0$.

Par suite, $\langle d-w, u-w\rangle \leqslant 0$ pour tout $u \in \overline{\bigcup_{\tau>0} \tau(C-\overline{x})} = T(C,\overline{x})$. Ceci caractérise le fait que $w = p_{T(C,\overline{x})}(d)$.

d) Puisque toute sous-suite convergente de la suite (bornée) $\{v_k\}$ a pour limite le même $w = p_{T(C,\overline{x})}(d)$, c'est que toute la suite $\{v_k\}$ est convergente et a pour limite $p_{T(C,\overline{x})}(d)$.

Le résultat précédent étant acquis pour toute suite $\{t_k\} \subset \mathbb{R}_*^+$ de limite 0, on a bien : $\frac{p_C(\overline{x}+td)-\overline{x}}{t} \longrightarrow p_{T(C,\overline{x})}(d)$ quand $t \longrightarrow 0^+$.

Notons qu'en raison du caractère lipschitzien de p_C, on a aussi :

$$\frac{p_C(\overline{x}+tv)-\overline{x}}{t} \longrightarrow p_{T(C,\overline{x})}(d) \text{ quand } t \longrightarrow 0^+ \text{ et } v \longrightarrow d.$$

2e partie

1°) Si $\overline{x}$ est un minimum local de f sur C, on a nécessairement :

$$\langle \nabla f(\overline{x}), x - \overline{x}\rangle \geqslant 0 \text{ pour tout } x \in C,$$

ou, d'une manière équivalente,

$$p_{T(C,\overline{x})}(-\nabla f(\overline{x})) = 0, \text{ ou bien } \overline{x} = p_C(\overline{x} - \nabla f(\overline{x})),$$

ou encore $\langle \nabla f(\overline{x}), p_{T(C,\overline{x})}(-\nabla f(\overline{x}))\rangle \geqslant 0$ (*cf.* Exercice VI.15).

2°) (i) f étant de classe C^1, p_C ayant une dérivée directionnelle en $\overline{x}$, la fonction composée $f \circ p_C$ a une dérivée directionnelle en $\overline{x}$ et :

$$\forall d \in \mathbb{R}^n, \ \frac{(f\circ p_C)(\overline{x}+td) - (f\circ p_C)(\overline{x})}{t} \to_{t\to 0^+} \langle \nabla f(\overline{x}), p_{T(C,\overline{x})}(d)\rangle .$$

En particulier pour $d = -\nabla f(\overline{x})$,

$$\frac{f[p_C(\overline{x} - t\nabla f(\overline{x}))] - f(\overline{x})}{t} \to_{t\to 0^+} \left\langle \nabla f(\overline{x}), p_{T(C,\overline{x})}(-\nabla f(\overline{x})) \right\rangle .$$

Lorsque l'on projette sur un cône convexe fermé K (comme c'est le cas ici avec $T(C,\overline{x})$), on a la propriété suivante :

$$(\overline{u} = p_K(u)) \Rightarrow (\overline{u} \in K \text{ et } \langle u - \overline{u}, \overline{u}\rangle = 0) .$$
$$(\langle u, \overline{u}\rangle = \| \overline{u} \|^2)$$

Dans le cas qui nous concerne

$$(g_{\overline{x}})'_+(0) = \left\langle \nabla f(\overline{x}), p_{T(C,\overline{x})}(-\nabla f(\overline{x})) \right\rangle = - \| p_{T(C,\overline{x})}(-\nabla f(\overline{x})) \|^2 .$$

(ii) Lorsque $\overline{x}$ ne vérifie pas l'une des conditions équivalentes du 1°), on a $(g_{\overline{x}})'_+(0) < 0$ comme cela est clair d'après l'expression de $(g_{\overline{x}})'_+(0)$ ci-dessus.

3°) Par construction : $x_k \in C$ et $f(x_{k+1}) \leqslant f(x_k)$ pour tout k.

Soit $\{x_{k_n}\}$, extraite de $\{x_k\}$, convergeant vers $\overline{x}$ ($\overline{x} \in C$ nécessairement). Raisonnons par l'absurde et supposons que $- \| p_{T(C,\overline{x})}(-\nabla f(\overline{x})) \|^2 < 0$.

Si $\overline{t}$ minimise $t \longmapsto g_{\overline{x}}(t) := f[p_C(\overline{x} - t\nabla f(\overline{x}))]$ sur $\mathbb{R}^+$, $\overline{t} > 0$ nécessairement et $g_{\overline{x}}(\overline{t}) < g_{\overline{x}}(0) = f(\overline{x})$.

Ainsi, pour n assez grand,

$$f[p_C(x_{k_n} - \overline{t}\nabla f(x_{k_n}))] < f(\overline{x}). \tag{6.27}$$

Comme $\{f(x_k)\}$ est décroissante (toute la suite !) et f continue, $f(x_k) \to f(\overline{x})$ quand $k \to +\infty$ (convergence de *toute* la suite $\{f(x_k)\}$).

Mais

$$\begin{aligned} f(\overline{x}) \leqslant f(x_{k_n+1}) &= f[p_C(x_{k_n} - t_{k_n}\nabla f(x_{k_n}))] \\ &\leqslant f[p_C(x_{k_n} - \overline{t}\nabla f(x_{k_n}))] \text{ (d'après la définition même de } t_{k_n}) \\ &< f(\overline{x}) \text{ (d'après (6.27)).} \end{aligned}$$

D'où la contradiction.

Donc $p_{T(C,\overline{x})}(-\nabla f(\overline{x})) = 0$, ce qui est une des formes de la condition du 1er ordre nécessairement vérifiée par un minimum local $\overline{x}$ de f sur C.

***Exercice VI.22.** *Décomposition de Moreau*

Soient $(H, \langle\cdot,\cdot\rangle)$ un espace euclidien et K un cône convexe fermé non vide de H.

On désigne par K° le cône polaire de K, c'est-à-dire l'ensemble

$$K^\circ = \{x \in H \mid \langle x, y\rangle \leqslant 0 \text{ pour tout } y \in K\}.$$

Le théorème de décomposition de Moreau affirme que les propriétés (1) et (2) suivantes sont équivalentes :

(1) $\quad z = x + y,\ x \in K,\ y \in K^\circ$, et $\langle x, y\rangle = 0$;

(2) $\quad x = p_K(z),\ y = p_{K^\circ}(z)$.

La décomposition figurant en (1) s'appelle *la décomposition de Moreau de z suivant K et K°*.

Dans chacun des deux exemples qui suivent, déterminer le cône polaire de K et la décomposition de Moreau de l'élément de H proposé.

1°) $H = \mathbb{R}^n$ muni de son produit scalaire usuel ; K est l'orthant positif $(\mathbb{R}^+)^n$ de $\mathbb{R}^n$; l'élément à décomposer est un élément quelconque de $\mathbb{R}^n$.

2°) $H = \mathcal{S}_n(\mathbb{R})$ muni du produit scalaire $\langle\langle A, B\rangle\rangle :=$ trace de AB ; K est le cône des matrices (symétriques) semi-définies positives ; l'élément à décomposer est un élément quelconque de $\mathcal{S}_n(\mathbb{R})$.

Solution : 1°) $K^\circ = -K$; la décomposition de Moreau de $z = (z_1, \dots, z_n) \in \mathbb{R}^n$ suivant K et K° est $z = z^+ + z^-$, où $z^+ = (z_1^+, \dots, z_n^+)$ et $z^- = (z_1^-, \dots, z_n^-)$. Ici u^- signifie $\max(-u, 0)$.

2°) $K^\circ = -K$, c'est-à-dire le cône des matrices (symétriques) semi-définies négatives (revoir l'Exercice I.10, 2°), si nécessaire). Étant donné $Z \in H$, soit Ω une matrice orthogonale telle que

$$\Omega^\top Z \Omega = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$$

(matrice diagonale dont les éléments diagonaux sont les valeurs propres (réelles) de Z).

De façon naturelle, on pose

$$X := \Omega \operatorname{diag}(\lambda_1^+, \dots, \lambda_n^+)\Omega^\top,\ Y := \Omega \operatorname{diag}(\lambda_1^-, \dots, \lambda_n^-)\Omega^\top.$$

Alors $Z = X + Y$ est la décomposition de Moreau cherchée.

Commentaire : L'opération de projection sur un convexe fermé ainsi que la décomposition de Moreau, dont quelques propriétés ou exemples font l'objet des exercices précédents, prennent toute leur importance dans le contexte des espaces de Hilbert réels (espaces fonctionnels de référence dans l'étude de problèmes variationnels).

***Exercice VI.23.** Soit $C := [0,1]^n$.

1°) Quels sont les points extrémaux de C ?

2°) Soit $\overline{x} = (\overline{x}_1, \dots, \overline{x}_n) \in \{0,1\}^n$. On se propose de déterminer le cône normal $N(C, \overline{x})$ à C en $\overline{x}$. Pour cela on pose $I_0(\overline{x}) := \{i \mid \overline{x}_i = 0\}$ et $I_1(\overline{x}) := \{i \mid \overline{x}_i = 1\}$.

Démontrer l'équivalence des énoncés suivants :

$$u = (u_1, \dots, u_n) \in N(C, \overline{x}) \ ; \tag{6.28}$$

$$\sum_{i \in I_0(\overline{x})} u_i x_i + \sum_{i \in I_1(\overline{x})} u_i (x_i - 1) \leqslant 0 \text{ pour tout } x = (x_1, \dots, x_n) \in [0,1]^n \ ; \tag{6.29}$$

$$\sum_{i \in I_0(\overline{x})} u_i \tilde{x}_i + \sum_{i \in I_1(\overline{x})} u_i (\tilde{x}_i - 1) \leqslant 0 \text{ pour tout } \tilde{x} = (\tilde{x}_1, \dots, \tilde{x}_n) \in \{0,1\}^n \ ; \tag{6.30}$$

$$\sum_{i \in I_0(\overline{x})} u_i^+ + \sum_{i \in I_1(\overline{x})} u_i^- = 0 \ ; \tag{6.31}$$

$$u_i \leqslant 0 \text{ si } i \in I_0(\overline{x}) \text{ et } u_i \geqslant 0 \text{ si } i \in I_1(\overline{x}). \tag{6.32}$$

3°) Soit à minimiser sur C une fonction convexe différentiable f. Comment caractériser $\overline{x} = (\overline{x}_1, \dots, \overline{x}_n) \in \{0,1\}^n$ minimum de f sur C ?

Solution : 1°) L'ensemble $\operatorname{ext} C$ des points extrémaux de C est $\operatorname{ext} C = \{0,1\}^n$.

2°) – Par définition, $u \in N(C, \overline{x})$ signifie : $\langle u, x - \overline{x} \rangle \leqslant 0$ pour tout $x \in C$. Cela se traduit par :

$$\sum_{i \in I_0(\overline{x})} u_i x_i + \sum_{i \in I_1(\overline{x})} u_i (x_i - 1) \leqslant 0 \text{ pour tout } x \in C.$$

Donc (6.28) $\Leftrightarrow$ (6.29).

– Comme $\operatorname{ext} C = \{0,1\}^n$ et que $C = \operatorname{conv}(\operatorname{ext} C)$, avoir $\langle u, x - \overline{x} \rangle \leqslant 0$ pour tout $x \in C$ équivaut à avoir $\langle u, \tilde{x} - \overline{x} \rangle \leqslant 0$ pour tout $\tilde{x} \in \operatorname{ext} C$. D'où l'équivalence de (6.29) et (6.30).

– Puisque u_i^+ et u_i^- sont toujours positifs ou nuls, avoir (6.31) signifie : $u_i^+ = 0$ pour tout $i \in I_0(\overline{x})$ et $u_i^- = 0$ pour tout $i \in I_1(\overline{x})$; c'est-à-dire (6.32) .

– [(6.32) $\Rightarrow$ (6.30)] est évident ; démontrons [(6.30) $\Rightarrow$ (6.32)].

Si $i \in I_0(\overline{x})$, posons $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_n)$ défini par $\tilde{x}_j := \begin{cases} 1 \text{ si } j = i, \\ 1 \text{ si } j \in I_1(\overline{x}), \\ 0 \text{ sinon.} \end{cases}$

Alors $\sum_{i \in I_0(\overline{x})} u_i \tilde{x}_i + \sum_{i \in I_1(\overline{x})} u_i(\tilde{x}_i - 1) = u_i \leqslant 0$ d'après (6.30) .

De manière similaire, si $i \in I_1(\overline{x})$, posons $\tilde{x}$ défini par

$\tilde{x}_j := \begin{cases} 0 \text{ si } j \in \{i\} \cup I_0(\overline{x}), \\ 1 \text{ sinon.} \end{cases}$

D'après (6.30) , $-u_i \leqslant 0$.

Il est bon de visualiser ces résultats avec les points extrémaux de $[0,1]^2$ ou $[0,1]^3$.

3°) $\overline{x} \in C$ est un minimum de f sur C si, et seulement si, l'opposé de $\nabla f(\overline{x})$ est dans $N(C, \overline{x})$. Dans le cas où $\overline{x} \in \{0,1\}^n$, cela se traduit donc par :

$$\frac{\partial f}{\partial x_i}(\overline{x}) \geqslant 0 \text{ pour tout } i \in I_0(\overline{x})$$

et

$$\frac{\partial f}{\partial x_i}(\overline{x}) \leqslant 0 \text{ pour tout } i \in I_1(\overline{x}).$$

***Exercice VI.24.** Soient $f : \mathbb{R}^n \to \mathbb{R}$ une fonction strictement convexe et C un convexe compact non vide de $\mathbb{R}^n$. Montrer qu'un point maximisant f sur C est nécessairement un point extrémal de C.

Solution : La fonction f, puisque supposée convexe sur $\mathbb{R}^n$, y est continue ; C est supposé compact : il existe donc $\overline{x} \in C$ maximisant f sur C.

Supposons que $\overline{x}$ ne soit pas un point extrémal de C : il existe $\alpha \in]0,1[$, x_1 et x_2 dans C, différents, tels que $\overline{x} = \alpha x_1 + (1-\alpha)x_2$. Mais, f étant strictement convexe,

$$\begin{aligned} f(\overline{x}) &< \alpha f(x_1) + (1-\alpha) f(x_2) \\ &< \alpha f(\overline{x}) + (1-\alpha) f(\overline{x}) = f(\overline{x}), \end{aligned}$$

d'où la contradiction.

Commentaire :

– Un exemple typique est avec $f : x \longmapsto \| x - a \|^2$, où $\| \cdot \|$ désigne la norme euclidienne usuelle de $\mathbb{R}^n$ et $a \in \mathbb{R}^n$; alors le point de C le plus éloigné de a est nécessairement un point extrémal de C.

– Prolongement. Soit C un compact non vide de $\mathbb{R}^n$ (que l'on peut prendre convexe sans perte de généralité) ; on suppose que pour tout $a \in \mathbb{R}^n$, il n'y a qu'un point qui soit le plus éloigné de a dans C. Montrer que C est nécessairement réduit à un point. Hé hé...

***Exercice VI.25.** Soient C un convexe de $\mathbb{R}^n$ et $f : C \to \mathbb{R}$. On dit que f est quasi-convexe sur C lorsque

$$\begin{pmatrix} x_1, x_2 \in C \\ \alpha \in]0,1[\end{pmatrix} \Rightarrow (f(\alpha x_1 + (1-\alpha)x_2) \leqslant \max\{f(x_1), f(x_2)\}) . \tag{6.33}$$

1°) Montrer que f est quasi-convexe sur C si et seulement si :

$$\forall r \in \mathbb{R},\ \{x \in C \mid f(x) \leqslant r\} \text{ est convexe.}$$

2°) Soit f quasi-convexe sur C.

(i) Montrer que tout minimum local strict de f (sur C) est un minimum global.

(ii) Montrer que tout maximum local strict de f est nécessairement un point extrémal de C (et se trouve donc sur la frontière de C).

Solution : 1°) Soit f vérifiant (6.33) et considérons $S_r(f) := \{x \in C \mid f(x) \leqslant r\}$.

Si x_1 et x_2 sont pris dans $S_r(f)$ et α dans $]0,1[$, $f(\alpha x_1 + (1-\alpha)x_2) \leqslant r$ d'après (6.33) et donc $\alpha x_1 + (1-\alpha)x_2 \in S_r(f)$. L'ensemble $S_r(f)$ est bien convexe.

Réciproquement, étant donnés x_1 et x_2 dans C et α dans $]0,1[$, posons $r := \max\{f(x_1), f(x_2)\}$. Les points x_1 et x_2 sont dans $S_r(f)$ et la convexité de $S_r(f)$ fait alors que $\alpha x_1 + (1-\alpha)x_2$ est aussi dans $S_r(f)$. D'où $f(\alpha x_1 + (1-\alpha)x_2) \leqslant r$.

2°) (i) Soit $\overline{x}$ un minimum local strict de f sur C ; soit $\eta > 0$ tel que

$$\begin{pmatrix} \| x - \overline{x} \| \leqslant \eta \\ x \in C,\ x \neq \overline{x} \end{pmatrix} \Rightarrow (f(x) > f(\overline{x})) .$$

Prenons $x \in C$, $\| x - \overline{x} \| > \eta$, et posons $t := \frac{\eta}{\|x - \overline{x}\|}$, $x_t := (1-t)\overline{x} + tx$.

D'une manière claire, $x_t \in C$, $\| x_t - \overline{x} \| = \eta$ et $x_t \neq \overline{x}$. Si on avait $f(x) < f(\overline{x})$, on aurait

$$f(\overline{x}) < f(x_t) \leqslant \max\{f(\overline{x}), f(x)\} = f(\overline{x}),$$

d'où la contradiction. Donc on a bien $f(x) \geqslant f(\overline{x})$.

(*ii*) Soit $\overline{x}$ un maximum local strict de f sur C. Si $\overline{x}$ n'était pas un point extrémal de C, on pourrait trouver $u \neq 0$ tel que

$$\begin{array}{l} \overline{x} - u \in C \\ \overline{x} + u \in C \end{array} \quad \text{et} \quad \begin{array}{l} f(\overline{x} - u) < f(\overline{x}) \\ f(\overline{x} + u) < f(\overline{x}). \end{array}$$

Par suite, sachant que $\overline{x} = \frac{1}{2}[(\overline{x} - u) + (\overline{x} + u)]$, la quasi-convexité de f ferait que

$$f(\overline{x}) \leqslant \max\{f(\overline{x} - u), f(\overline{x} + u)\} < f(\overline{x}),$$

d'où la contradiction.

***Exercice VI.26.** Soit $\Omega := \{x = (x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_i > 0 \text{ pour tout } i\}$. On dit que $f : \Omega \to \mathbb{R}$ est de la forme $(\mathcal{G})$ lorsqu'elle s'exprime de la manière suivante :

$$x = (x_1, \ldots, x_n) \in \Omega \longmapsto f(x) = \sum_{k=1}^{p} c_k x_1^{a_{1k}} x_2^{a_{2k}} \ldots x_n^{a_{nk}},$$

où les c_k sont des réels positifs et les a_{ik} des réels quelconques (éventuellement < 0).

Des problèmes d'ingénierie, de génie chimique notamment, conduisent à des problèmes de minimisation du type :

$$(\mathcal{P}) \qquad \begin{cases} \text{Min } f_0(x) \\ x \in \Omega \qquad\qquad\qquad\qquad (\theta \text{ réel } > 0) \\ f_i(x) \leqslant \theta \text{ pour } i = 1, \ldots, m, \end{cases}$$

où les fonctions $f_0, f_1, \ldots, f_m$ sont du type $(\mathcal{G})$.

Montrer que, grâce à un changement de variables adéquat, $(\mathcal{P})$ peut être transformé en un problème de minimisation convexe qui lui est équivalent.

Solution : Posons $y_i = \ln\, x_i$ $(i = 1, \ldots, n)$, de manière qu'une fonction f de la forme $(\mathcal{G})$ se transforme en

$$y = (y_1, \ldots, y_n) \in \mathbb{R}^n \longmapsto g(y) := f(\mathrm{e}^{y_1}, \ldots, \mathrm{e}^{y_n}) = \sum_{k=1}^{p} c_k \mathrm{e}^{\langle a_k, y \rangle},$$

avec $a_k := (a_{1k}, a_{2k}, \ldots, a_{nk})$.

Les fonctions $y \longmapsto \mathrm{e}^{\langle a_k, y\rangle}$ sont convexes sur $\mathbb{R}^n$ (d'accord ?), les réels c_k sont positifs, donc la fonction g est convexe. $(\mathcal{P})$ est alors équivalent au problème de minimisation convexe suivant :

$$\left(\hat{\mathcal{P}}\right) \qquad \begin{cases} \operatorname{Min} g_0(y) \\ g_i(y) \leqslant \theta \quad \text{pour} \quad i = 1, \ldots, m. \end{cases}$$

***Exercice VI.27.** Soient S un compact non vide de $\mathbb{R}^n$ et $g, h : S \to \mathbb{R}$ deux fonctions continues. On suppose : $h(x) > 0$ pour tout $x \in S$, et on considère le problème d'optimisation (dit *fractionnaire*) suivant :

$$(\mathcal{P}) \qquad \begin{cases} \text{Minimiser} \quad f(x) := \dfrac{g(x)}{h(x)} \\ x \in S. \end{cases}$$

À $(\mathcal{P})$ on associe le problème d'optimisation suivant, paramétré par $\alpha \in \mathbb{R}$:

$$(\mathcal{P}_\alpha) \qquad \begin{cases} \text{Minimiser } g(x) - \alpha h(x) \\ x \in S. \end{cases}$$

1°) Montrer que la fonction $\alpha \longmapsto \theta(\alpha) := \min_{x\in S} \{g(x) - \alpha h(x)\}$ est concave, continue, strictement décroissante, et que l'équation $\theta(\alpha) = 0$ a une seule solution.

2°) Montrer :
– si $\overline{x}$ est solution de $(\mathcal{P})$, alors $\overline{\alpha} := f(\overline{x})$ vérifie $\theta(\overline{\alpha}) = 0$;
– si $\theta(\overline{\alpha}) = 0$, alors tout $\overline{x} \in S$ tel que $f(\overline{x}) = \overline{\alpha}$ est solution de $(\mathcal{P})$.

3°) On suppose : S convexe, g convexe positive sur S, h concave sur S. Quelles méthodes peut-on préconiser pour approcher d'une manière itérative la valeur optimale dans $(\mathcal{P})$?

Solution : 1°) La fonction θ est l'infimum de la famille de fonctions affines $a_x : \alpha \in \mathbb{R} \longmapsto g(x) - \alpha h(x)$, indexée par $x \in S$; elle est donc concave sur $\mathbb{R}$.

La fonction $(\alpha, x) \longmapsto \varphi(\alpha, x) := g(x) - \alpha h(x)$ étant continue et S compact, la fonction $\alpha \longmapsto \theta(\alpha) = \min_{x\in S} \varphi(\alpha, x)$ est continue (ceci est aisé à vérifier).

Prenons $\alpha_1 < \alpha_2$ et considérons une solution $\overline{x}_1$ de $(\mathcal{P}_{\alpha_1})$; alors

$$\begin{aligned} \theta(\alpha_1) = g(\overline{x}_1) - \alpha_1 h(\overline{x}_1) &> g(\overline{x}_1) - \alpha_2 h(\overline{x}_1) \quad (\text{car} \quad h(\overline{x}_1) > 0) \\ &\geqslant \min_{x\in S} \{g(x) - \alpha_2 h(x)\} =: \theta(\alpha_2). \end{aligned}$$

Soit $x_0 \in S$ et prenons les réels α_- et α_+ vérifiant : $\alpha_+ < \min_{x\in S} \dfrac{g(x)}{h(x)}$ et $\dfrac{g(x_0)}{h(x_0)} < \alpha_-$; il s'ensuit que $\theta(\alpha_-) < 0$ et $\theta(\alpha_+) > 0$.

Il existe donc un seul $\overline{\alpha}$ tel que $\theta(\overline{\alpha}) = 0$.

2°) Soit $\overline{x}$ une solution de $(\mathcal{P})$. On a :

$$\forall\, x \in S,\ \frac{g(x)}{h(x)} \geqslant \frac{g(\overline{x})}{h(\overline{x})} =: \overline{\alpha}\ ,$$

d'où

$$\min_{x\in S} \{g(x) - \overline{\alpha}h(x)\} \geqslant 0,\ g(\overline{x}) - \overline{\alpha}h(\overline{x}) = 0,$$

soit $\theta(\overline{\alpha}) = 0$.

Réciproquement, soit $\overline{\alpha} \in \mathbb{R}$ tel que $\theta(\overline{\alpha}) = \min_{x\in S} \{g(x) - \overline{\alpha}h(x)\} = 0$ et considérons $\overline{x} \in S$ tel que $g(\overline{x}) - \overline{\alpha}h(\overline{x}) = 0$, *i.e.* $\overline{\alpha} = f(\overline{x})$. Alors :

$$\forall x \in S,\ g(x) - \overline{\alpha}h(x) \geqslant g(\overline{x}) - \overline{\alpha}h(\overline{x}) = 0,$$

soit encore :

$$\forall x \in S,\ \frac{g(x)}{h(x)} \geqslant \overline{\alpha} = \frac{g(\overline{x})}{h(\overline{x})},$$

c'est-à-dire que $\overline{x}$ est solution de $(\mathcal{P})$.

3°) La première observation est que $f = g/h$ est à présent quasi-convexe sur S.

En effet :

$$\forall r \geqslant 0,\ \{x \in S \mid f(x) \leqslant r\} = \{x \in S \mid g(x) - rh(x) \leqslant 0\} \text{ est convexe.}$$

On peut donc penser à approcher la valeur optimale dans $(\mathcal{P})$ par une méthode de minimisation d'une fonction quasi-convexe sur un convexe.

Au vu du résultat de la 2$^{\text{e}}$ question, une deuxième méthode consiste à résoudre l'équation (dans $\mathbb{R}$) $\theta(\alpha) = 0$ par une méthode itérative ; l'évaluation $\theta(\alpha_k)$ résulte de la minimisation de la fonction $g - \alpha_k h$ qui est convexe sur S.

***Exercice VI.28.** Soit $\varphi : \mathbb{R}_*^+ \to \mathbb{R}$ convexe. On définit $I_\varphi : (\mathbb{R}_*^+)^n \times (\mathbb{R}_*^+)^n \to \mathbb{R}$ par

$$(p = (p_1, \ldots, p_n),\ q = (q_1, \ldots, q_n)) \longmapsto I_\varphi(p, q) := \sum_{i=1}^{n} q_i\ \varphi\left(\frac{p_i}{q_i}\right).$$

1°) Vérifier que I_φ est convexe. En déduire l'inégalité suivante :

$$I_\varphi(p,q) \geqslant \left(\sum_{i=1}^n q_i\right) \varphi\left(\frac{\sum_{i=1}^n p_i}{\sum_{i=1}^n q_i}\right).$$

2°) On définit $\varphi^\diamond : \mathbb{R}_*^+ \to \mathbb{R}$ par $\varphi^\diamond(x) := x\varphi(\frac{1}{x})$.

(i) $\varphi^\diamond$ est-elle convexe ?

(ii) Comment se comparent I_φ et $I_{\varphi^\diamond}$?

3°) Donner les expressions de $\varphi^\diamond$ et I_φ dans les cas suivants : $\varphi(t) = t \ln t$, $(1-\sqrt{t})^2$, t^α avec $\alpha > 1$, $(t-1)^2$, $|t-1|$.

Solution : 1°) Soit $f : (\mathbb{R}_*^+) \times (\mathbb{R}_*^+) \longmapsto \mathbb{R}$ définie par $f(x,y) := y\varphi\left(\frac{x}{y}\right)$. Il nous suffit de vérifier que f est convexe. À cet effet, prenons (x,y) et (u,v) dans $(\mathbb{R}_+^*) \times (\mathbb{R}_+^*)$ et $0 < \lambda < 1$. En notant $\overline{\lambda} = 1 - \lambda$, la convexité de φ implique

$$\begin{aligned}\varphi\left(\frac{\lambda x + \overline{\lambda} u}{\lambda y + \overline{\lambda} v}\right) &= \varphi\left(\frac{\lambda y}{\lambda y + \overline{\lambda} v}\frac{x}{y} + \frac{\overline{\lambda} v}{\lambda y + \overline{\lambda} v}\frac{u}{v}\right) \\ &\leqslant \frac{\lambda}{\lambda y + \overline{\lambda} v} y\varphi\left(\frac{x}{y}\right) + \frac{\overline{\lambda}}{\lambda y + \overline{\lambda} v} v\varphi\left(\frac{u}{v}\right).\end{aligned}$$

Après multiplication par $\lambda y + \overline{\lambda} v$, on obtient

$$(\lambda y + \overline{\lambda} v)\varphi\left(\frac{\lambda x + \overline{\lambda} u}{\lambda y + \overline{\lambda} v}\right) \leqslant \lambda y\varphi\left(\frac{x}{y}\right) + \overline{\lambda} v\varphi\left(\frac{u}{v}\right),$$

ce qui est l'inégalité de convexité (sur f) cherchée.

Dans l'inégalité de Jensen $\varphi\left(\sum_{i=1}^n t_i x_i\right) \leqslant \sum_{i=1}^n t_i \varphi(x_i)$, faisons $t_i = \frac{q_i}{\sum_{i=1}^n q_i}$, $x_i = \frac{p_i}{q_i}$. On en déduit immédiatement l'inégalité demandée.

2°) (i) Oui. C'est un résultat classique sur les fonctions convexes de la variable réelle (rappelé en début de chapitre), mais qui est aussi une conséquence de ce qui a été démontré au-dessus.

(ii) $I_{\varphi^\diamond}(p,q) = I_\varphi(q,p)$.

3°) $I_\varphi(p,q)$ est une sorte de mesure de proximité entre p et q ; elle joue un rôle important lorsque $p = (p_1, \ldots, p_n)$ et $q = (q_1, \ldots, q_n)$ sont des distributions de probabilité ($\sum_i p_i = \sum_i q_i = 1$) notamment.

– $\varphi(t) = t \ln\ t$. Alors $\quad \varphi^\diamond(t) = -\ln\ t,\ I_\varphi(p,q) = \sum_{i=1}^{n} p_i \ln\ \frac{p_i}{q_i}$.

– $\varphi(t) = (1-\sqrt{t})^2$. Ici $\quad \varphi^\diamond = \varphi,\ I_\varphi(p,q) = \sum_{i=1}^{n} (\sqrt{p_i} - \sqrt{q_i})^2$.

– $\varphi(t) = t^\alpha$. Alors $\quad \varphi^\diamond(t) = t^{1-\alpha},\ I_\varphi(p,q) = \sum_{i=1}^{n} (p_i)^\alpha (q_i)^{1-\alpha}$.

– $\varphi(t) = (t-1)^2$. Alors $\quad \varphi^\diamond(t) = t + \frac{1}{t} - 2,\ I_\varphi(p,q) = \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{q_i}$.

– $\varphi(t) = |\ t-1\ |$. Alors $\quad \varphi^\diamond = \varphi,\ I_\varphi(p,q) = \sum_{i=1}^{n} |\ p_i - q_i\ |$.

***Exercice VI.29.** *Sur l'espérance mathématique de l'inverse d'une matrice aléatoire*

Si X est une variable aléatoire strictement positive, $\mathbb{E}\left(\frac{1}{X}\right) \geqslant \frac{1}{\mathbb{E}(X)}$ pourvu que $\mathbb{E}(X)$ et $\mathbb{E}\left(\frac{1}{X}\right)$ existent. Dans cet exercice, on propose une généralisation de cette propriété au cas matriciel.

On désigne par matrice aléatoire une matrice A dont les coefficients a_{ij} sont des variables aléatoires. Lorsque les a_{ij} sont intégrables, on dira que A est intégrable et on posera $\mathbb{E}(A) := [\mathbb{E}(a_{ij})]_{1\leqslant i,j\leqslant n}$.

Soit A une matrice carrée aléatoire telle que, presque sûrement, $A(\omega) = [a_{ij}(\omega)]_{1\leqslant i,j\leqslant n}$ soit réelle, symétrique et définie positive ; on suppose que A et A^{-1} sont intégrables. On se propose de démontrer que $\mathbb{E}(A^{-1}) \succeq [\mathbb{E}(A)]^{-1}$.

1re approche

1°) Soient U et V deux matrices réelles symétriques définies positives de taille n, et c un élément de $\mathbb{R}^n$. On définit $f : [0,1] \to \mathbb{R}$ comme suit :

$$\forall t \in [0,1],\ f(t) := \left\langle [(1-t)U + tV]^{-1} c, c \right\rangle .$$

Montrer que f est continue sur [0,1] et deux fois dérivable sur]0,1[. En déduire que f est convexe.

2°) En déduire :

$$\forall t \in [0,1],\ [(1-t)U + tV]^{-1} \preceq (1-t)U^{-1} + tV^{-1}.$$

3°) Montrer alors, à l'aide de l'inégalité de Jensen,

$$\mathbb{E}\left(\left\langle A^{-1}c, c\right\rangle\right) \geqslant \left\langle [\mathbb{E}(A)]^{-1} c, c \right\rangle .$$

2^e approche

4°) On pose $\Delta(\omega) := A(\omega) - \mathbb{E}(A)$. Établir

$$\mathbb{E}(A^{-1}) = [\mathbb{E}(A)]^{-1} + [\mathbb{E}(A)]^{-1}\mathbb{E}(\Delta A^{-1}\Delta)[\mathbb{E}(A)]^{-1}.$$

5°) Vérifier que $[\mathbb{E}(A)]^{-1}\mathbb{E}(\Delta A^{-1}\Delta)[\mathbb{E}(A)]^{-1}$ est semi-définie positive. En déduire l'inégalité $\mathbb{E}(A^{-1}) \succeq [\mathbb{E}(A)]^{-1}$.

Solution : *1^re approche*

1°) On désigne par $\overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ l'ouvert (c'est en fait un cône convexe ouvert) des matrices réelles symétriques définies positives de taille n.

Puisque U et V sont dans $\overset{\circ}{\mathcal{P}}_n(\mathbb{R})$,

$$I := \left\{ t \in \mathbb{R} \mid (1-t)U + tV \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R}) \right\}$$

est un intervalle ouvert de $\mathbb{R}$ contenant [0,1].

On peut voir f comme la composée des applications suivantes :

$$\begin{aligned} t \in I &\overset{\varphi_1}{\longmapsto} \varphi_1(t) := (1-t)U + tV \\ M \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R}) &\overset{\varphi_2}{\longmapsto} \varphi_2(M) := M^{-1} \\ A \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R}) &\overset{\varphi_3}{\longmapsto} \varphi_3(A) := \langle Ac, c\rangle . \end{aligned}$$

φ_1 est une fonction affine, donc de classe C^∞, et $\varphi_1'(t) = V - U$ pour tout $t \in I$.

φ_2 est également une fonction de classe C^∞, et sa différentielle première en M est donnée par

$$H \longmapsto D\varphi_2(M)(H) = -M^{-1}HM^{-1}.$$

φ_3 est linéaire (continue), donc de classe C^∞, et $D\varphi_3(A) = \varphi_3$ en tout A.

Par suite, $f = \varphi_3 \circ \varphi_2 \circ \varphi_1$ est C^∞ sur $I \supset [0,1]$ et, pour tout $t \in I$,

$$\begin{aligned} f'(t) &= -\left\langle [(1-t)U + tV]^{-1}[V-U][(1-t)U + tV]^{-1}c, c\right\rangle , \\ f''(t) &= 2\left\langle [(1-t)U + tV]^{-1}[V-U][(1-t)U + tV]^{-1}[V-U] \right. \\ &\qquad \left. [(1-t)U + tV]^{-1}c, c\right\rangle . \end{aligned}$$

Posons $u := [(1-t)U+tV]^{-1}c$ et $v := [V-U]u$. D'après ce qui précède et en raison de la symétrie des matrices en jeu,

$$\begin{aligned} f''(t) &= 2\left\langle [V-U][(1-t)U+tV]^{-1}[V-U]u, u\right\rangle \\ &= 2\left\langle [(1-t)U+tV]^{-1}v, v\right\rangle . \end{aligned}$$

D'où $f''(t) \geqslant 0$ puisque $[(1-t)U+tV]^{-1}$ est symétrique définie positive. f est bien convexe sur I.

2°) L'inégalité de convexité (de base) relative à f conduit à :

$$\forall t \in [0,1],\ \left\langle [(1-t)U+tV]^{-1}c, c\right\rangle \leqslant \left\langle [(1-t)U^{-1}+tV^{-1}]c, c\right\rangle ,$$

et ce pour tout $c \in \mathbb{R}^n$. D'où l'inégalité annoncée.

3°) De par les résultats des questions précédentes, l'application $F : \overset{\circ}{\mathcal{P}}_n(\mathbb{R}) \to \mathbb{R}$ qui à A associe $F(A) := \left\langle A^{-1}c, c\right\rangle$ est convexe. Par application de l'inégalité de Jensen :

$$\text{soit} \quad \begin{array}{l} \mathbb{E}[F(A(\omega))] \geqslant F[\mathbb{E}(A)], \\ \mathbb{E}\left[\left\langle A^{-1}(\omega)c, c\right\rangle\right] \geqslant \left\langle [\mathbb{E}(A)]^{-1}c, c\right\rangle , \end{array}$$

d'où le résultat escompté puisque $\mathbb{E}\left[\left\langle A^{-1}(\omega)c, c\right\rangle\right] = \left\langle \mathbb{E}(A^{-1})c, c\right\rangle$.

2e approche

4°) $\Delta(\omega)A^{-1}(\omega)\triangle(\omega) = A(\omega) + \mathbb{E}(A)A^{-1}(\omega)\mathbb{E}(A) - 2\mathbb{E}(A)$,
d'où :

$$\mathbb{E}[\Delta(\omega)A^{-1}(\omega)\triangle(\omega)] = -\mathbb{E}(A) + \mathbb{E}(A)\mathbb{E}(A^{-1})\mathbb{E}(A)$$

et

$$[\mathbb{E}(A)]^{-1}\mathbb{E}[\Delta(\omega)A^{-1}(\omega)\triangle(\omega)][\mathbb{E}(A)]^{-1} = -[\mathbb{E}(A)]^{-1} + \mathbb{E}(A^{-1}).$$

5°) Comme la matrice $M := \mathbb{E}[\Delta(\omega)A^{-1}(\omega)\triangle(\omega)]$ est semi-définie positive, il en est de même de $[\mathbb{E}(A)]^{-1}M[\mathbb{E}(A)]^{-1}$. D'où l'inégalité annoncée.

****** Exercice VI.30.*** Soit $\overset{\circ}{\mathcal{P}}_n(\mathbb{R})$ le cône convexe ouvert des matrices symétriques définies positives de taille n. On définit $f : \mathbb{R}^n \times \overset{\circ}{\mathcal{P}}_n(\mathbb{R}) \to \mathbb{R}$ comme suit :

$$\forall\ x \in \mathbb{R}^n,\ \forall\ A \in \overset{\circ}{\mathcal{P}}_n(\mathbb{R}),\ f(x,A) := \left\langle A^{-1}x, x\right\rangle$$

($\langle \cdot, \cdot \rangle$ étant le produit scalaire usuel sur $\mathbb{R}^n$).

L'application f est-elle convexe ?

Solution : La réponse est oui. Alors que la convexité de $x \longmapsto \langle A^{-1}x, x\rangle$ est facile et celle de $A \longmapsto \langle A^{-1}x, x\rangle$ peut être aisément démontrée, c'est la convexité de $(x, A) \longmapsto \langle A^{-1}x, x\rangle$ comme fonction du *couple* (x, A) qui est prouvée ici.

– Soit tout d'abord $A := \mathrm{diag}(a_1, \ldots, a_n)$, $B := \mathrm{diag}(b_1, \ldots, b_n)$, avec les a_i et b_i strictement positifs et $\alpha \in]0, 1[$. Alors :

$$f(\alpha x + (1-\alpha)y, \alpha A + (1-\alpha)B) = \sum_{i=1}^{n} \frac{(\alpha x_i + (1-\alpha)y_i)^2}{\alpha a_i + (1-\alpha)b_i}$$
$$= \sum_{i=1}^{n} \left[\left(\alpha \frac{x_i^2}{a_i} + (1-\alpha)\frac{y_i^2}{b_i} \right) - \alpha(1-\alpha)(b_i x_i - a_i y_i)^2 \right]$$
$$\leqslant \sum_{i=1}^{n} \left(\alpha \frac{x_i^2}{a_i} + (1-\alpha)\frac{y_i^2}{b_i} \right) = \alpha f(x, A) + (1-\alpha) f(y, B).$$

– Soient à présent A et B quelconques dans $\overset{\circ}{\mathcal{P}}_n(\mathbb{R})$. On peut trouver Q inversible telle que $QAQ^\top$ et $QBQ^\top$ soient toutes deux diagonales (réduction simultanée des formes quadratiques associées à A et B, possible car A est définie positive). Or

$$f(x, A) = \langle A^{-1}x, x\rangle = \langle Q^{-T}A^{-1}Q^{-1}Qx, Qx\rangle$$
$$= \left\langle (QAQ^\top)^{-1}Qx, Qx \right\rangle = f(Qx, QAQ^\top) \ ;$$

donc
$\alpha f(x, A) + (1-\alpha) f(y, B) - f(\alpha x + (1-\alpha)y, \alpha A + (1-\alpha)B)$
$= \alpha f(Qx, QAQ^\top) + (1-\alpha) f(Qy, QBQ^\top)$
$= -f(\alpha Qx + (1-\alpha)Qy, \alpha QAQ^\top + (1-\alpha)QBQ^\top)$,
$\geqslant 0$ d'après le résultat du 1^{er} point.

***Exercice VI.31.** Soit $f : \mathbb{R}^n \to \mathbb{R}$ convexe et de classe C^2 sur $\mathbb{R}^n$, soit C un convexe fermé de $\mathbb{R}^n$.

On suppose qu'il existe $\overline{x} \in C$ minimisant f sur C.

1°) Montrer qu'un point $\tilde{x} \in C$ minimise f sur C si, et seulement si,

$$\langle \nabla f(\overline{x}), \tilde{x} - \overline{x}\rangle = 0 \quad \text{et} \quad \nabla f(\tilde{x}) = \nabla f(\overline{x}). \tag{6.34}$$

2°) On suppose que f est quadratique, *i.e.* $f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle + c$, avec A symétrique semi-définie positive, et que C est un polyèdre convexe fermé.

Montrer que l'ensemble des points minimisant f sur C est un polyèdre convexe fermé que l'on exprimera en fonction de C, A, b et x.

Solution : 1°) Soit $\tilde{x}$ un point de C vérifiant la condition (6.34). De part la convexité de f, on a

$$f(\overline{x}) \geqslant f(\tilde{x}) + \langle \nabla f(\tilde{x}), \overline{x} - \tilde{x}\rangle\,,$$

d'où $f(\overline{x}) \geqslant f(\tilde{x})$ à l'aide de (6.3.4). Ainsi, $\tilde{x}$ minimise aussi f sur C.

Réciproquement, soit $\tilde{x}$ un point de C minimisant f sur C. On a alors :

$$f(\overline{x}) = f(\tilde{x}) \geqslant f(\overline{x}) + \langle \nabla f(\overline{x}), \tilde{x} - \overline{x}\rangle\,, \text{ d'où } \langle \nabla f(\overline{x}), \tilde{x} - \overline{x}\rangle \leqslant 0,$$
$$\langle \nabla f(\overline{x}), \tilde{x} - \overline{x}\rangle \geqslant 0 \quad \text{(c'est la condition de minimalité)},$$

en définitive $\langle \nabla f(\overline{x}), \tilde{x} - \overline{x}\rangle = 0$.

En échangeant le rôle de $\overline{x}$ et $\tilde{x}$, on obtient de même $\langle \nabla f(\tilde{x}), \overline{x} - \tilde{x}\rangle = 0$.

Or

$$\nabla f(\overline{x}) - \nabla f(\tilde{x}) = \int_0^1 \nabla^2 f(\tilde{x} + t(\overline{x} - \tilde{x}))(\overline{x} - \tilde{x})dt \tag{6.35}$$
$$= H(\overline{x} - \tilde{x}), \text{ où } H := \int_0^1 \nabla^2 f(\tilde{x} + t(\overline{x} - \tilde{x}))dt.$$

Comme

$$0 = \langle \nabla f(\overline{x}) - \nabla f(\tilde{x}), \overline{x} - \tilde{x}\rangle = \langle H(\overline{x} - \tilde{x}), \overline{x} - \tilde{x}\rangle\,,$$

et que H est symétrique semi-définie positive, $H(\overline{x} - \tilde{x}) = 0$ nécessairement.

D'où $\nabla f(\overline{x}) - \nabla f(\tilde{x}) = 0$ d'après l'évaluation (6.35).

2°) L'ensemble des points minimisant f sur C est

$$C \cap \{\tilde{x} \in \mathbb{R}^n \mid \langle b, \tilde{x}\rangle = \langle b, \overline{x}\rangle \text{ et } A\tilde{x} = A\overline{x}\}\,.$$

This page intentionally left blank

VII

INITIATION AU CALCUL SOUS-DIFFÉRENTIEL ET DE TRANSFORMÉES DE LEGENDRE-FENCHEL

Rappels

Les fonctions $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ qui seront considérées, qu'elles soient convexes ou non, vérifieront au minimum les deux propriétés suivantes :

$$\left.\begin{array}{l} - \text{ Il y a un point au moins en lequel } f \text{ est finie ;} \\ - \text{ Il existe une fonction affine minorant } f \text{ sur } \mathbb{R}^n \end{array}\right\}. \tag{7.1}$$

Pour de telles fonctions f, on appelle *domaine* de f l'ensemble des x en lesquels f prend une valeur finie ($\operatorname{dom} f := \{x \in \mathbb{R}^n \mid f(x) < +\infty\}$), et *épigraphe* de f l'ensemble des $(x, r) \in \mathbb{R}^n \times \mathbb{R}$ « au-dessus du graphe de f » ($\operatorname{epi} f := \{(x, r) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) \leqslant r\}$). Si l'inégalité est stricte dans la définition de $\operatorname{epi} f$, on parlera de l'épigraphe strict de f, et on notera $\operatorname{epi}_s f$ cet ensemble.

VII.1. La transformation de Legendre-Fenchel

VII.1.1. Définitions

La *conjuguée* ou *transformée de Legendre-Fenchel* de f est la fonction $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ définie par :

$$\forall s \in \mathbb{R}^n,\ f^*(s) := \sup_{x \in \mathbb{R}^n} \{\langle s, x\rangle - f(x)\}. \tag{7.2}$$

L'application $f \longmapsto f^*$ est appelée conjugaison ou transformation de Legendre-Fenchel.

La définition s'étend sans difficulté aux f définies sur un espace euclidien $(E, \langle\cdot,\cdot\rangle)$, auquel cas f^* est définie sur E^* représenté par E *via* $\langle\cdot,\cdot\rangle$.

Comme conséquence de la définition (7.2), nous avons l'inégalité suivante :

$$f^*(s) + f(x) \geqslant \langle s, x\rangle \text{ pour tout } (s,x) \in \mathbb{R}^n \times \mathbb{R}^n,$$

appelée *inégalité de Fenchel.*

Par construction, f^* est *convexe et semi-continue inférieurement* ; cette propriété ajoutée au fait que f vérifie (7.1) entraîne que f^* vérifie également (7.1).

Exemples :

– Si S est une partie non vide de $\mathbb{R}^n$ et si $I_S : x \in \mathbb{R}^n \longmapsto I_S(x) := 0$ si $x \in S$, $+\infty$ sinon (I_S est la *fonction indicatrice* de S), alors

$$(I_S)^* : s \in \mathbb{R}^n \longmapsto (I_S)^*(s) = \sup_{x\in S} \langle s, x\rangle$$

est la *fonction d'appui* de S (notée σ_S également).

– Si S est une partie fermée non vide de $\mathbb{R}^n$, posons

$$f_S : x \in \mathbb{R}^n \longmapsto f_S(x) := \frac{1}{2} \parallel x \parallel^2 \text{ si } x \in S, +\infty, \text{ sinon}; \tag{7.3}$$

alors,

$$(f_S)^* : s \in \mathbb{R}^n \longmapsto (f_S)^*(s) = \frac{1}{2}\left[\parallel s \parallel^2 - d_S^2(s)\right],$$

où d_S désigne la fonction-distance à S.

Notation : $\Gamma_0(\mathbb{R}^n)$ désigne l'ensemble des $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ qui sont convexes semi-continues inférieurement sur $\mathbb{R}^n$ et finies en au moins un point de $\mathbb{R}^n$. On a déja signalé que $f^* \in \Gamma_0(\mathbb{R}^n)$ dès que f vérifie (7.1).

VII.1.2. Quelques propriétés et règles de calcul

– Si f est 1-coercive sur $\mathbb{R}^n$ (*i.e.* si $f(x)/\parallel x \parallel \to +\infty$ quand $\parallel x \parallel \to +\infty$), alors f^* est partout finie sur $\mathbb{R}^n$ (c'est-à-dire dom $f^* = \mathbb{R}^n$).

– Soit $f : \mathbb{R}^n \to \mathbb{R}$ strictement convexe, différentiable et 1-coercive sur $\mathbb{R}^n$; alors :

• f^* est également partout finie, strictement convexe, différentiable et 1-coercive sur $\mathbb{R}^n$;

• l'application $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ est bijective et

$$f^*(s) = \langle s, (\nabla f)^{-1}(s) \rangle - f((\nabla f)^{-1}(s)) \text{ pour tout } s \in \mathbb{R}^n. \tag{7.4}$$

Ceci est la formule de Legendre.

– Étant données deux fonctions f_1 et f_2 minorées par une fonction affine commune, *l'inf-convolution* (ou *somme épigraphique*) de f_1 et f_2 est la fonction notée $f_1 \square f_2$ définie comme suit :

$$\forall x \in \mathbb{R}^n,\ (f_1 \square f_2)(x) := \inf_{x_1 + x_2 = x} \{f_1(x_1) + f_2(x_2)\}.$$

Alors $\mathrm{dom}(f_1 \square f_2) = \mathrm{dom}(f_1) + \mathrm{dom}(f_2)$ et $\mathrm{epi}_s(f_1 \square f_2) = \mathrm{epi}_s(f_1) + \mathrm{epi}_s(f_2)$. La conjuguée de $f_1 \square f_2$ est $f_1^* + f_2^*$, et ce sans aucune hypothèse additionnelle sur f_1 et f_2.

– Soient deux fonctions f_1 et f_2 de $\Gamma_0(\mathbb{R}^n)$ telles que les intérieurs relatifs de leurs domaines se coupent (*i.e.*, $\mathrm{ir}(\mathrm{dom}\ f_1)$ et $\mathrm{ir}(\mathrm{dom}\ f_2)$ ont un point en commun) ; alors la conjuguée de $f_1 + f_2$ est $f_1^* \square f_2^*$ et cette inf-convolution est *exacte*, c'est-à-dire : si $s \in \mathrm{dom} f_1^* \square f_2^*$, il existe $(s_1, s_2) \in \mathrm{dom}\ f_1^* \times \mathrm{dom}\ f_2^*$ tel que $s = s_1 + s_2$ et $f_1^* \square f_2^*(s) = f_1^*(s_1) + f_2^*(s_2)$.

– Si $f \in \Gamma_0(\mathbb{R}^n)$, la fonction $f^{**} := (f^*)^*$ n'est autre que f. La transformation $f \longmapsto f^*$ est en fait une involution de $\Gamma_0(\mathbb{R}^n)$.

VII.2. Le sous-différentiel d'une fonction

VII.2.1. Définitions

Étant donnés une fonction f et $x \in \mathrm{dom}\ f$, un élément s de $\mathbb{R}^n$ est appelé *sous-gradient* de f en x lorsque

$$f(x') \geqslant f(x) + \langle s, x' - x \rangle \text{ pour tout } x' \in \mathbb{R}^n. \tag{7.5}$$

La définition s'étend sans difficulté aux f définies sur un espace euclidien $(E, \langle \cdot, \cdot \rangle)$.

L'ensemble des sous-gradients de f en x est appelé *sous-différentiel* de f en x et noté par le graphisme $\partial f(x)$.

Une caractérisation des $s \in \partial f(x)$, venant immédiatement de (7.5), est comme suit :

$$(s \in \partial f(x)) \Leftrightarrow (f^*(s) + f(x) - \langle s, x \rangle = 0)\ ;$$

il s'ensuit que le calcul sous-différentiel et la transformation de Legendre-Fenchel sont très étroitement liés.

VII.2.2. Quelques propriétés et règles de calcul

– Si $f \in \Gamma_0(\mathbb{R}^n)$, alors $\partial f(x)$ est assurément non vide dès que x est dans l'intérieur relatif de dom f.

– Si $f \in \Gamma_0(\mathbb{R}^n)$ et si $x \in \text{int}(\text{dom } f)$, $\partial f(x)$ est un convexe compact (non vide) dont la fonction d'appui est

$$d \longmapsto f'(x,d) := \lim_{t \to 0^+} \frac{f(x+td) - f(x)}{t},$$

appelée *dérivée directionnelle* de f en x. En conséquence, $f \in \Gamma_0(\mathbb{R}^n)$ est différentiable en $x \in \text{int}(\text{dom } f)$ si et seulement si $\partial f(x)$ est un singleton (c'est-à-dire qu'il n'y a qu'un sous-gradient de f en x) ; auquel cas $\partial f(x) = \{\nabla f(x)\}$.

– Soit $f \in \Gamma_0(\mathbb{R}^n)$; alors : $(s \in \partial f(x)) \Leftrightarrow (x \in \partial f^*(s))$.
Ceci est résumé symboliquement sous la forme $\partial f^* = (\partial f)^{-1}$.

– Soit $f : \mathbb{R}^n \to \mathbb{R}$ strictement convexe, deux fois différentiable et 1-coercive sur $\mathbb{R}^n$. Supposons de plus que $\nabla^2 f(x)$ soit définie positive pour tout $x \in \mathbb{R}^n$. Alors f^* jouit des mêmes propriétés et le calcul différentiel sur f^* à partir de celui sur f s'opère comme suit :

$$\nabla f^*(s) = (\nabla f)^{-1}(s), \quad \nabla^2 f^*(s) = \left[\nabla^2 f((\nabla f)^{-1}(s))\right]^{-1}. \tag{7.6}$$

– Soient deux fonctions f_1 et f_2 de $\Gamma_0(\mathbb{R}^n)$ telles que les intérieurs relatifs de leurs domaines se coupent ; alors en tout x en lequel f_1 et f_2 sont finies,

$$\partial(f_1 + f_2)(x) = \partial f_1(x) + \partial f_2(x). \tag{7.7}$$

(Dans le membre de droite il s'agit d'une somme vectorielle de deux convexes fermés ; cette règle généralise donc la règle de calcul différentiel usuelle $\nabla(f_1 + f_2)(x) = \nabla f_1(x) + \nabla f_2(x)$.)

– Soient deux fonctions f_1 et f_2 de $\Gamma_0(\mathbb{R}^n)$ et $x \in \text{dom}(f_1 \square f_2)$ pour lequel il existe $(x_1, x_2) \in \text{dom } f_1 \times \text{dom } f_2$ tels que $(f_1 \square f_2)(x) = f_1(x_1) + f_2(x_2)$ (c'est-à-dire l'inf-convolution de f_1 et f_2 est exacte en $x_1 + x_2$) ; alors :

$$\partial(f_1 \square f_2)(x) = \partial f_1(x_1) \cap \partial f_2(x_2).$$

– Soient des fonctions $f_1, \dots, f_k \in \Gamma_0(\mathbb{R}^n)$ et x un point intérieur à tous les $\operatorname{dom} f_i$, $i = 1, \dots, k$ (par exemple x est quelconque si les f_i sont partout finies sur $\mathbb{R}^n$). Alors :

$$\partial(\max_{i=1,\dots,k} f_i)(x) = \operatorname{conv}\left\{ \bigcup_{i \in I(x)} \partial f_i(x) \right\}, \tag{7.8}$$

où $I(x) := \{i \mid f_i(x) = f(x)\}$. En d'autres termes, un sous-gradient de $\max_{i=1,\dots,k} f_i$ en x est une combinaison convexe de sous-gradients de f_i en x, où l'on ne retient que les indices i pour lesquels $f_i(x) = f(x)$ (« là où ça touche »). La règle de calcul (7.8), sans équivalent dans le monde des fonctions différentiables, est sans doute la plus importante du royaume des fonctions convexes.

VII.3. La convexification d'une fonction

Étant donnée une fonction f (toujours vérifiant (7.1)), on se pose la question de construire l'enveloppe convexe de f, c'est-à-dire la plus grande fonction convexe minorant f sur $\mathbb{R}^n$; laquelle fonction est notée $\operatorname{conv} f$ (ou $\operatorname{co} f$ plus traditionnellement). La construction analytique de $\operatorname{conv} f$ est comme suit :

$$(\operatorname{conv} f)(x) = \inf\left\{ \sum_{i=1}^{n+1} \alpha_i f(x_i) \mid x = \sum_{i=1}^{n+1} \alpha_i x_i,\ \alpha_i \geqslant 0 \ \text{ pour tout } i \right.$$
$$\left. \text{et } \sum_{i=1}^{n+1} \alpha_i = 1 \right\};$$

parmi toutes les décompositions possibles de x comme combinaisons convexes de points x_i, on prend l'infimum des combinaisons convexes correspondantes des valeurs $f(x_i)$ de la fonction. (Mais attention ! l'épigraphe de $\operatorname{conv} f$ peut être légèrement plus gros que l'enveloppe convexe de $\operatorname{epi} f$.)

Plus utile que $\operatorname{conv} f$ est son enveloppe semi-continue inférieure, c'est-à-dire la fonction obtenue en fermant l'épigraphe de $\operatorname{conv} f$ (ce qui revient *in fine* à prendre l'enveloppe convexe fermée de $\operatorname{epi} f$) ; cette nouvelle fonction est notée $\overline{\operatorname{conv}} f$ (ou $\overline{\operatorname{co}} f$ plus traditionnellement). Outre la construction « interne » de $\overline{\operatorname{conv}} f$ évoquée précédemment, il y a une construction « externe » qui s'avère toute aussi parlante : La fonction $\overline{\operatorname{conv}} f$ est le supremum de toutes les fonctions affines minorant f sur $\mathbb{R}^n$.

Fort heureusement, il y a une situation, fréquente dans les applications, où $\operatorname{conv} f$ et $\overline{\operatorname{conv}} f$ ne diffèrent pas :

– Si f est semi-continue et 1-coercive sur $\mathbb{R}^n$, alors $\operatorname{conv} f = \overline{\operatorname{conv}} f$ et, pour tout x en lequel $\operatorname{conv} f$ est finie, il existe des $x_i \in \operatorname{dom} f_i$ et des α_i positifs de somme 1 tels que

$$x = \sum_{i=1}^{n+1} \alpha_i x_i \quad \text{et } (\operatorname{conv} f)(x) = \sum_{i=1}^{n+1} \alpha_i f(x_i).$$

Exemple de telle situation : Soient S un fermé non vide de $\mathbb{R}^n$ et $\varphi : \mathbb{R}^n \to \mathbb{R}$ une fonction continue telle que $\lim_{\substack{\| x \| \to +\infty \\ x \in S}} \varphi(x) = +\infty$; alors la fonction $f : x \in \mathbb{R}^n \longmapsto f(x) := \varphi(x)$ si $x \in S$, $+\infty$ sinon, vérifie les hypothèses du résultat énoncé au-dessus. Voir (7.3) pour un exemple particulier.

Le lien avec la transformation de Legendre-Fenchel s'exprime par la relation suivante : $\overline{\operatorname{conv}} f = f^{**}$.

Références. Chapitre VI de [12] et Chapitre X de [13].

***Exercice VII.1.** Soient k éléments $s_1, \dots, s_k$ de $\mathbb{R}^n$.

1°) (*Lemme de Gordan*). Montrer l'équivalence des deux propositions suivantes :

(P_1) $\max_{i=1,\dots,k} \langle s_i, x \rangle \geqslant 0$ pour tout $x \in \mathbb{R}^n$;

(P_2) Il existe $\alpha_1 \geqslant 0, \dots, \alpha_k \geqslant 0$, non tous nuls, tels que $\sum_{i=1}^{k} \alpha_i s_i = 0$.

2°) Comparer les propositions suivantes :

(P_3) $\max_{i=1,\dots,k} \langle s_i, x \rangle > 0$ pour tout $x \neq 0$ de $\mathbb{R}^n$;

(P_4) Il existe $\alpha_1 > 0, \dots, \alpha_k > 0$, tels que $\sum_{i=1}^{k} \alpha_i s_i = 0$, et $\operatorname{vect}\{s_1, \dots, s_k\} = \mathbb{R}^n$.

Commentaire : Pour « visualiser » ces conditions (P_i), il est conseillé de représenter graphiquement la fonction $x \longmapsto \max_{i=1,\dots,k} \langle s_i, x \rangle$ ainsi que le cône normal à son épigraphe en l'origine (de $\mathbb{R}^{n+1}$).

Solution : 1°) *Preuve 1.* Soit $f : x \longmapsto f(x) := \max_{i=1,\dots,k} \langle s_i, x\rangle$. Ce que dit (P_1) est que la fonction convexe (et même linéaire par morceaux) f est minimisée en 0. Une condition nécessaire et suffisante pour qu'il en soit ainsi est donc : $0 \in \partial f(0)$. Sachant que $\partial f(0) = \text{conv}\{s_1, \dots, s_k\}$, l'équivalence de (P_1) et (P_2) s'ensuit.

Preuve 2 (directe).

$[(P_2) \Rightarrow (P_1)]$. Il est clair que (P_2) implique l'égalité (pour tout x) $\sum_{i=1}^{k} \alpha_i \langle s_i, x\rangle = 0$. S'il existait $\overline{x}$ tel que $\max_{i=1,\dots,k} \langle s_i, \overline{x}\rangle < 0$, on aurait (sachant que l'un au moins des α_i est > 0) $\sum_{i=1}^{k} \alpha_i \langle s_i, \overline{x}\rangle < 0$. D'où contradiction.

$[(P_1) \Rightarrow (P_2)]$. Soit $L : \mathbb{R}^k \to \mathbb{R}^n$ définie par $L(\alpha_1, \dots, \alpha_k) = -\sum_{i=1}^{k} \alpha_i s_i$. Comme L est linéaire continue, elle transforme le simplexe-unité Δ_k de $\mathbb{R}^k$ en un (polyèdre) convexe compact $L(\Delta_k)$ de $\mathbb{R}^n$.

Supposons (P_2) fausse. Alors $0 \notin L(\Delta_k)$. Nous allons, sans le dire explicitement, séparer strictement $\{0\}$ et $L(\Delta_k)$. Désignons par p_0 la projection de 0 sur $L(\Delta_k)$. D'après la caractérisation de la projection d'un élément sur un convexe fermé,

$$\langle c - p_0, 0 - p_0\rangle = \langle p_0, p_0 - c\rangle \leqslant 0 \quad \text{pour tout } c \in L(\Delta_k).$$

C'est notamment le cas pour $c = -s_i,\ i = 1, \dots, k$. D'où

$$\langle s_i, p_0\rangle \leqslant -\langle p_0, p_0\rangle = -\parallel p_0 \parallel^2 < 0,$$

ce qui induit $\max_{i=1,\dots,k} \langle s_i, p_0\rangle < 0$. D'où contradiction avec (P_1).

2°) Les deux propositions (P_3) et (P_4) sont équivalentes.

$[(P_4) \Rightarrow (P_3)]$. Supposons qu'il existe $\overline{x} \neq 0$ tel que $\max_{i=1,\dots,k} \langle s_i, \overline{x}\rangle \leqslant 0$ et montrons qu'on arrive à une contradiction. Comme

$$0 = \left\langle \sum_{i=1}^{k} \alpha_i s_i, \overline{x} \right\rangle = \sum_{i=1}^{k} \alpha_i \langle s_i, \overline{x}\rangle$$

et que tous les α_i sont > 0, il s'ensuit :

$$\langle s_i, \overline{x}\rangle = 0 \text{ pour tout } i = 1, \dots, k.$$

Puisque $\overline{x} \neq 0$, il existe $s \in \mathbb{R}^n$ tel que $\langle s, \overline{x} \rangle \neq 0$. Or $\mathbb{R}^n = \text{vect}\{s_1, \ldots, s_k\}$ par hypothèse ; il existe donc des réels λ_i tels que $s = \sum_{i=1}^{k} \lambda_i s_i$. Par suite

$$\langle s, \overline{x} \rangle = \sum_{i=1}^{k} \lambda_i \langle s_i, \overline{x} \rangle$$

est impossible à tenir puisque $\langle s, \overline{x} \rangle \neq 0$ et $\langle s_i, \overline{x} \rangle = 0$ pour tout i.

[$(P_3) \Rightarrow (P_4)$]. Montrons tout d'abord que (P_3) implique que $\text{vect}\{s_1, \ldots, s_k\} = \mathbb{R}^n$. S'il n'en était pas ainsi, il existerait $\overline{x} \neq 0$ tel que $\langle s_i, \overline{x} \rangle = 0$ pour tout $i = 1, \ldots, k$ (d'accord ?). Mais ceci contredirait alors (P_3).

Au vu du résultat de la 1re question, (P_3) implique l'existence de coefficients $\alpha_i \geqslant 0$, non tous nuls, tels que $\sum_{i=1}^{k} \alpha_i s_i = 0$. Reste à prouver (ici) que l'on peut prendre les α_i tous strictement positifs. Désignons par K le cône convexe engendré par les vecteurs s_i, c'est-à-dire

$$K := \left\{ \sum_{i=1}^{k} t_i s_i \mid t_i \geqslant 0 \ \text{ pour tout } \ i = 1, \ldots, k \right\}.$$

Nous savons que K est fermé (*cf.* Exercice V.12). Considérons à présent la projection de $-\sum_{i=1}^{k} s_i$ sur (le cône convexe fermé) K ; cette projection est de la forme $\sum_{i=1}^{k} \overline{t}_i s_i$, avec $\overline{t}_i \geqslant 0$, pour tout i, et caractérisée par les relations suivantes (*cf.* rappels de la Section VI.2) :

$$\begin{cases} -\sum_{i=1}^{k} s_i - \sum_{i=1}^{k} \overline{t}_i s_i \in K^\circ \text{ (cône polaire de } K) \\ \text{et} \\ \left\langle -\sum_{i=1}^{k} s_i - \sum_{i=1}^{k} \overline{t}_i s_i, \sum_{i=1}^{k} \overline{t}_i s_i \right\rangle = 0. \end{cases}$$

Puisque K est engendré par les s_i, la 1re relation ci-dessus implique

$$\left\langle -\sum_{i=1}^{k} (1 + \overline{t}_i) s_i, s_j \right\rangle \leqslant 0 \ \text{ pour tout } \ j = 1, \ldots, k,$$

d'où, grâce à la 2e relation ci-dessus,

$$\left\langle \sum_{i=1}^{k} (1+\overline{t}_i) s_i, s_j \right\rangle = 0 \quad \text{pour tout} \quad j = 1, \dots, k.$$

Comme $\text{vect}\{s_1, \dots, s_k\} = \mathbb{R}^n$, on en déduit $\sum_{i=1}^{k} (1+\overline{t}_i) s_i = 0$. D'où le résultat escompté.

Commentaire : – Des exemples simples montrent qu'on ne peut se contenter de « $\alpha_i \geqslant 0$ pour tout i » dans (P_4), et qu'on ne peut se dispenser de l'hypothèse « $\text{vect}\{s_1, \dots, s_k\} = \mathbb{R}^n$ ».

– Prolongement de l'exercice : Montrer que (P_4) équivaut aussi à $\text{cône}\,\{s_1, \cdots, s_k\} = \mathbb{R}^n$.

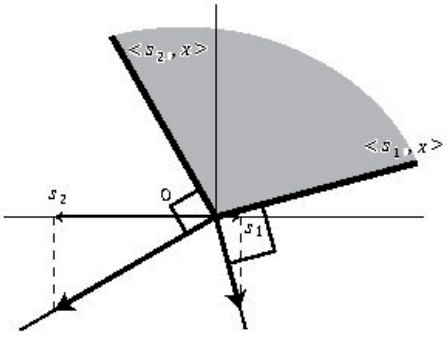


Figure 23. Illustration dans $\mathbb{R}^2$.

***Exercice VII.2.** Soit $f \in \Gamma_0(\mathbb{R})$ telle que $f(t) = f(-t)$ pour tout $t \in \mathbb{R}$. Rassembler toutes les propriétés possibles de f, notamment celles concernant f^* et ∂f.

Solution : • Le domaine de f est un intervalle symétrique :
si $0 \leqslant t \in \text{dom}\, f$, $[-t, +t] \subset \text{dom}\, f$.

• $0 \in \text{dom}\, f$ et minimise f sur $\mathbb{R}$. En effet, puisque f est paire et convexe

$$f(0) \leqslant \frac{1}{2}[f(t) + f(-t)] = f(t) \quad \text{pour tout } t \in \mathbb{R}.$$

• f^* est également paire.

- $\partial f(t) = -\partial f(-t)$ pour tout $t \in \text{dom } f$.
- Si $0 \in \partial f(t_0)$ pour un certain t_0, alors $f(t) = f(0)$ sur $[-t_0, +t_0]$.
- Si 0 est l'unique minimum de f et si f est dérivable en 0, il en est de même pour f^*.

***Exercice VII.3.** Soient $\theta \in \Gamma_0(\mathbb{R})$ paire et $f \in \Gamma_0(\mathbb{R}^n)$ définie par $f(x) := \theta(\| x \|)$.

Exprimer f^* en fonction de θ^*.

En déduire l'expression suivante de $\partial f(x)$:

$$\partial f(x) = \{s \in \mathbb{R}^n \mid \| s \| \in \partial\theta(\| x \|) \text{ et } \langle s, x\rangle = \| s \| \cdot \| x \|\}.$$

Solution :

- $f^*(s) = \theta^*(\| s \|)$.
- $s \in \partial f(x) \Leftrightarrow f(x) + f^*(s) = \langle s, x\rangle \Leftrightarrow \theta(\| x \|) + \theta^*(\| s \|) = \langle s, x\rangle$
  $\Leftrightarrow \theta(\| x \|) + \theta^*(\| s \|) = \| s \| \cdot \| x \| \text{ et } \langle s, x\rangle = \| s \| \cdot \| x \|$
  (utiliser l'inégalité de Schwarz, et celle de Fenchel pour θ).

Exercice VII.4. Construire graphiquement la conjuguée de la fonction suivante :

$$x \in \mathbb{R} \longmapsto f(x) := \frac{1}{2}x^2 + |x|$$

(graphiquement signifie en déterminant le graphe de ∂f, puis celui de $\partial f^* = (\partial f)^{-1}$, et en en déduisant f^*.)

Solution :

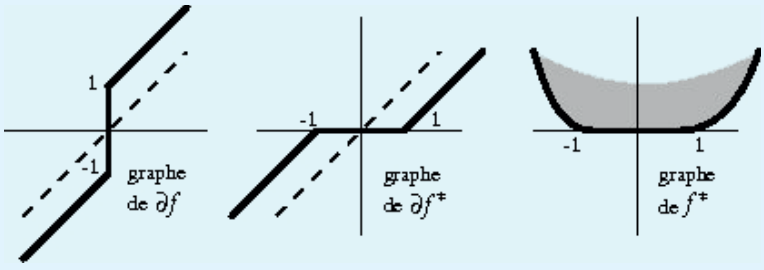


FIGURE 24.

D'où, par intégration de $s \longmapsto \partial f^*(s) = \{(f^*)'(s)\}$, et sachant que $f^*(0) = 0$:

$$f^*(s) = \begin{cases} 0 & \text{si } -1 \leqslant s \leqslant +1, \\ \frac{1}{2}(|\, s \,| - 1)^2 & \text{si } |\, s \,| > 1. \end{cases}$$

***Exercice VII.5.** Soit $A \in \mathcal{S}_n(\mathbb{R})$ définie positive. Montrer, à l'aide d'un simple calcul de conjuguée de fonction convexe, l'inégalité suivante

$$A + A^{-1} \succeq 2I_n.$$

Solution : Soit $f : x \in \mathbb{R}^n \longmapsto f(x) := \frac{1}{2}\langle Ax, x\rangle$. Alors $f^* : s \in \mathbb{R}^n \longmapsto f^*(s) = \frac{1}{2}\langle A^{-1}s, s\rangle$.

L'inégalité de Fenchel indique :

$$\forall x \in \mathbb{R}^n,\ \forall s \in \mathbb{R}^n,\ \langle Ax, x\rangle + \langle A^{-1}s, s\rangle \geqslant 2\langle x, s\rangle.$$

En faisant $x = s$, on trouve l'inégalité demandée.

***Exercice VII.6.** On se propose de trouver une primitive de la fonction $s \longmapsto$ Argsh s (notée aussi $(\mathrm{sh})^{-1}(s)$) par l'intermédiaire de la transformation de Legendre-Fenchel.

1°) Soit $f : x \in \mathbb{R} \longmapsto f(x) := \mathrm{ch}\, x$. Déterminer f^*.

2°) Indiquer pourquoi f^* est une primitive de la fonction Argsh.

Solution : 1°) Par définition, $f^*(s) = \sup_{x\in\mathbb{R}}\{sx - ch\, x\}$. Le supremum dans l'expression de $f^*(s)$ est atteint en un seul point $\overline{x}(s) =$ Argsh s. D'où

$$f^*(s) = s(\text{Argsh } s) - \mathrm{ch}(\text{Argsh } s) = s(\text{Argsh } s) - \sqrt{1+s^2}. \tag{7.9}$$

2°) On a $f' = sh$, d'où $(f^*)' = (f')^{-1} : s \in \mathbb{R} \longmapsto$ Argsh s. Donc f^* est la primitive de Argsh qui prend en 0 la valeur $f^*(0) = -\inf_{x\in\mathbb{R}} f(x) = -1$.

Le cas traité ici est une illustration de la formule de Legendre :

$$f^*(s) = s(f')^{-1}(s) - f[(f')^{-1}(s)].$$

***Exercice VII.7.** Pour $a > 0$ et $c \geqslant 0$ posons

$$f_{a,c} : x \in \mathbb{R} \longmapsto f_{a,c}(x) := -\sqrt{a^2 - (x-c)^2} \text{ si } |\, x - c \,| \leqslant a, \ +\infty \text{ sinon.}$$

1°) Calculer $(f_{a,c})^*$, puis $f_{a_1,c_1} \square f_{a_2,c_2}$.

2°) Vérifier sur cet exemple les formules reliant les dérivées première et seconde de $f_{a,c}$ à celles de $(f_{a,c})^*$.

Solution : 1°) $f^*_{a,c} : s \in \mathbb{R} \longmapsto f^*_{a,c}(s) = a\sqrt{1+s^2} + cs$. Ce calcul permet un va-et-vient entre les familles de fonctions $\{f_{a,c}\}$ et $\{f^*_{a,c}\}$. En particulier, les règles de calcul sur les conjuguées conduisent à :

$$f_{a_1,c_1} \square f_{a_2,c_2} = f_{a_1+a_2,c_1+c_2}.$$

2°) On a :

$$\begin{aligned} \forall x \in]c-a, c+a[, \quad & f'_{a,c}(x) = (x-c)[a^2-(x-c)^2]^{-1/2}, \\ & f''_{a,c}(x) = a^2[a^2-(x-c)^2]^{-3/2} \ ; \\ \forall s \in \mathbb{R}, \quad & (f^*_{a,c})'(s) = as(1+s^2)^{-1/2} + c, \\ & (f^*_{a,c})''(s) = a(1+s^2)^{-3/2}. \end{aligned}$$

Ainsi, l'application $f'_{a,c}$ est une bijection de $]c-a, c+a[$ sur $\mathbb{R}$ dont la bijection inverse est $(f^*_{a,c})'$. Pour ce qui est de la dérivée seconde, nous sommes dans les conditions où nous pouvons relier $f''_{a,c}$ et $(f^*_{a,c})''$ par la formule suivante :

$$\forall s \in \mathbb{R}, \ (f^*_{a,c})''(s) = \frac{1}{f''_{a,c}(x)}, \quad \text{où} \quad x = as(1+s^2)^{-1/2} + c.$$

Notons aussi sur cet exemple le comportement suivant :

Lorsque $a \downarrow 0$, $f_{a,c}(x) \uparrow I_{\{c\}}(x)$ pour tout x, et $f^*_{a,c}(s) \downarrow I^*_{\{c\}}(s) = cs$ pour tout s.

***Exercice VII.8.** Pour m et σ réels posons

$$q_{m,\sigma} : x \in \mathbb{R} \longmapsto q_{m,\sigma}(x) := \frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2 \text{ lorsque } \sigma \neq 0,$$

$$q_{m,0} := I_{\{m\}}.$$

Calculer $(q_{m,\sigma})^*$.

En déduire $q_{m,\sigma} \square q_{m',\sigma'}$.

Solution : $q^*_{m,\sigma} : s \in \mathbb{R} \longmapsto q^*_{m,\sigma}(s) = \frac{1}{2}\sigma^2 s^2 + ms$.

Par suite,

$$q_{m,\sigma} \square q_{m',\sigma'} = q_{m+m',\sqrt{\sigma^2+\sigma'^2}}.$$

Commentaire : Le résultat ci-dessus doit être rapproché de la formule correspondante en Probabilités : $\mathcal{N}(m,\sigma) * \mathcal{N}(m',\sigma') = \mathcal{N}(m+m', \sqrt{\sigma^2+\sigma'^2})$, où $*$ désigne l'opération de convolution usuelle en Analyse. On peut pousser le parallèle entre Probabilités et Analyse convexe et dresser le tableau de correspondance suivant :

Probabilités	*Analyse convexe*
addition de réels	infimum de réels
multiplication de réels	addition de réels
lois normales $\mathcal{N}(m,\sigma)$	fonctions quadratiques $q_{m,\sigma}$
convolution	inf-convolution
transformation de Fourier (fonctions caractéristiques)	transformation de Legendre-Fenchel (fonctions conjuguées)

***Exercice VII.9.** Soit $f \in \Gamma_0(\mathbb{R})$ définie de la façon suivante :

$$f(x) := \begin{cases} x \ln x + \frac{x^2}{2} - x & \text{si } x \geqslant 0 \text{ (avec } 0 \ln 0 = 0) \\ +\infty & \text{si } x < 0. \end{cases}$$

Ici $f^*(s)$ ne peut être exprimée analytiquement à l'aide des fonctions usuelles.

Considérons la fonction $l : \mathbb{R}^+ \to \mathbb{R}^+$, dite de Lambert, définie comme l'inverse de la fonction $x \longmapsto xe^x$ (c'est-à-dire : pour tout $u \geqslant 0$, $x = l(u)$ est l'unique solution de $xe^x = u$).

Exprimer f^* à l'aide de l et des fonctions usuelles.

Solution : Puisque $f(x)/\mid x \mid \to +\infty$ quand $\mid x \mid \to +\infty$, on sait déjà que f^* sera partout finie sur $\mathbb{R}$.

Soit $s \in \mathbb{R}$. Par définition

$$f^*(s) = \sup_{x \geqslant 0} \left\{ sx - x \ln x - \frac{x^2}{2} + x \right\}.$$

La fonction strictement concave $x \geqslant 0 \longmapsto \theta_s(x) := sx - x\ln x - \frac{x^2}{2} + x$ s'annule en 0, elle est dérivable sur $\mathbb{R}_*^+$ et $\theta_s'(x) = s - \ln x - x$ pour tout $x > 0$. Comme θ_s' s'annule en l'unique point $\overline{x}(s) = l(\mathrm{e}^s)$, le supremum de θ_s sur $\mathbb{R}^+$ est atteint en $\overline{x}(s)$. Ainsi :

$$\begin{aligned} f^*(s) &= sl(\mathrm{e}^s) - l(\mathrm{e}^s)[s - l(\mathrm{e}^s)] - \frac{[l(\mathrm{e}^s)]^2}{2} + l(\mathrm{e}^s) \\ &= \frac{1}{2}[l(\mathrm{e}^s)]^2 + l(\mathrm{e}^s). \end{aligned}$$

En définitive :

$$\forall s \in \mathbb{R}, \quad f^*(s) = \frac{1}{2}[l(\mathrm{e}^s)]^2 + l(\mathrm{e}^s).$$

***Exercice VII.10.** Soit $f \in \Gamma_0(\mathbb{R}^n)$ majorée sur $\mathbb{R}^n$. Que peut-on dire d'une telle fonction ?

Solution : Elle est nécessairement constante. En effet, de l'inégalité

$$f(x) \leqslant M \quad \text{pour tout } x \in \mathbb{R}^n,$$

on tire $f^* \geqslant I_{\{0\}} - M$ (fonction qui vaut partout $+\infty$ sauf en 0 où elle vaut $-M$).

Par conséquent, $f^* = I_{\{0\}} - C$ pour un certain $C \in \mathbb{R}$, d'où $f = f^{**} \equiv C$.

****Exercice VII.11.** Soient f et g dans $\Gamma_0(\mathbb{R}^n)$ telles que $f+g$ soit une fonction affine sur $\mathbb{R}^n$. Montrer que f et g sont nécessairement affines sur $\mathbb{R}^n$.

Indication. Calculer $(f+g)^*$.

Solution : Soient $s \in \mathbb{R}^n$ et $r \in \mathbb{R}$ tels que $f + g = \langle s, \cdot\rangle + r$. De cette égalité, il vient que f et g sont partout finies sur $\mathbb{R}^n$ et donc continues sur $\mathbb{R}^n$. Par conjugaison, il s'ensuit :

$$(f+g)^* = f^* \square g^* = I_{\{s\}} - r.$$

Comme $\mathrm{dom}\,(f^* \square g^*) = \mathrm{dom}\, f^* + \mathrm{dom}\, g^*$ et $\mathrm{epi}_s(f^* \square g^*) = \mathrm{epi}_s f^* + \mathrm{epi}_s g^*$, la seule possibilité pour vérifier la relation au-dessus est d'avoir :

$$f^* = I_{\{s_1\}} - r_1,\ g^* = I_{\{s_2\}} - r_2,\ s = s_1 + s_2,\ r = r_1 + r_2.$$

D'où $\quad f = (f^*)^* = \langle s_1, \cdot\rangle - r_1$ et $g = (g^*)^* = \langle s_2, \cdot\rangle - r_2$.

***Exercice VII.12.** Soit $\mathbb{R}^n$ muni du produit scalaire usuel noté $\langle \cdot, \cdot \rangle$ et de la norme euclidienne associée $\| \cdot \|$. La fonction $f : x \in \mathbb{R}^n \longmapsto f(x) := \frac{1}{2} \| x \|^2$ est sa propre conjuguée : $f = f^*$. Est-ce la seule fonction dans ce cas ?

Solution : Oui. Pour voir cela, considérons donc une fonction f sur $\mathbb{R}^n$ pour laquelle $f = f^*$. L'inégalité de Fenchel nous dit que

$$f(x) + f^*(s) \geqslant \langle s, x \rangle \quad \text{pour tout } (s, x) \in \mathbb{R}^n \times \mathbb{R}^n.$$

En faisant $s = x$ et sachant que $f = f^*$, il s'ensuit

$$f(x) \geqslant \frac{1}{2} \| x \|^2 \quad \text{pour tout } x \in \mathbb{R}^n.$$

Prenons la conjuguée membre à membre ci-dessus ; il vient $f^* \leqslant \frac{1}{2} \| \cdot \|^2$.

Donc $\frac{1}{2} \| \cdot \|^2$ est la seule solution de l'équation $f = f^*$.

Commentaire : Dans le même ordre d'idées, soit $f : \mathbb{R}^n \to \mathbb{R}$ convexe, de classe C^2, telle que $\det(\nabla^2 f(x)) = 1$ pour tout $x \in \mathbb{R}^n$. Que peut-on dire d'une telle fonction ? Réponse : elle est *quadratique*. Dur dur...

***Exercice VII.13.** Soit $f : \mathbb{R}^n \to \mathbb{R}$ telle que

$$\partial f(x) \neq \phi \quad \text{pour tout} \quad x \in \mathbb{R}^n.$$

Montrer qu'une telle fonction est nécessairement convexe sur $\mathbb{R}^n$.

Solution : *1re démonstration.* Soient x_0, x_1 dans $\mathbb{R}^n$ et $\alpha \in]0,1[$. Posons $x_\alpha := (1-\alpha)x_0 + \alpha x_1$. Prenons $s_\alpha \in \partial f(x_\alpha)$; alors :

$$f(x_0) \geqslant f(x_\alpha) + \langle s_\alpha, x_0 - x_\alpha \rangle = f(x_\alpha) + \alpha \langle s_\alpha, x_0 - x_1 \rangle ,$$
$$f(x_1) \geqslant f(x_\alpha) + \langle s_\alpha, x_1 - x_\alpha \rangle = f(x_\alpha) + (1-\alpha) \langle s_\alpha, x_1 - x_0 \rangle .$$

Multiplions la 1re inégalité par $1-\alpha$, la deuxième par α, et faisons la somme membre à membre ; il vient :

$$(1-\alpha) f(x_0) + \alpha f(x_1) \geqslant f(x_\alpha).$$

2e démonstration. Puisque $\partial f(a) \neq \phi$ pour au moins un $a \in \mathbb{R}^n$, f est minorée sur $\mathbb{R}^n$ par une fonction affine. On fait alors appel au résultat suivant :

$$(\partial f(x) \neq \phi) \Rightarrow ((\overline{\text{conv}}\, f)(x) = f(x)) .$$

On a donc $\overline{\text{conv}}\, f = f$ (dans le cas présent).

***Exercice VII.14.** Soient $f : \mathbb{R}^n \to \mathbb{R}$ convexe et C un convexe fermé sur lequel f est constante.

1°) Montrer que le sous-différentiel de f est constant sur l'intérieur relatif de C.

2°) On suppose f différentiable sur C. Montrer que le gradient de f est constant sur C.

Commentaire : On pourra illustrer cette propriété avec des fonctions $f : \mathbb{R}^2 \to \mathbb{R}$ et des segments $C = [A, B]$ de $\mathbb{R}^2$.

L'hypothèse « f constante sur C » peut être généralisée facilement à « f est affine sur C ».

Solution : 1°) a) Soient $x \in \text{ir } C$ et $d \in V :=$ sous-espace vectoriel direction de aff C. Alors nous disons que $\langle s, d\rangle = 0$ pour tout $s \in \partial f(x)$. Soit en effet $s \in \partial f(x)$. Par définition :

$$f(x + td) \geqslant f(x) + t\,\langle s, d\rangle \quad \text{pour tout} \quad t \in \mathbb{R}.$$

Sachant qu'il existe $\alpha > 0$ tel que $x + td \in C$ pour tout $t \in]-\alpha, +\alpha[$, nous avons $f(x + td) = f(x)$ pour de tels t. Divisons l'inégalité qui précède par $t > 0$ (resp. par $t < 0$) et faisons tendre t vers 0^+ (resp. 0^-) pour obtenir $\langle s, d\rangle \leqslant 0$ (resp. $\langle s, d\rangle \geqslant 0$).

b) Soit $x \in \text{ir } C$, $s \in \partial f(x)$ et $x' \in C$. Il vient tout d'abord :

$$f^*(s) + f(x) - \langle s, x\rangle = 0.$$

Or $x' - x \in V$, d'où $\langle s, x' - x\rangle = 0$ d'après le résultat du point précédent. Comme $f(x') = f(x)$, il vient

$$f^*(s) + f(x') - \langle s, x'\rangle = 0, \text{ soit } s \in \partial f(x').$$

On a donc démontré que $\partial f(x) \subset \partial f(x')$. Pour l'inclusion inverse, on considère $x' \in \text{ir } C$ et on opère de même.

2°) Nous avons $\nabla f(x) = v$ pour tout $x \in \text{ir } C$. Tout $x' \in C$ est limite d'une suite d'éléments x de ir C, et $\nabla f(x) \to \nabla f(x')$ quand $x \to x'$. D'où $\nabla f(x') = v$.

***Problème VII.15.** *Approximation et régularisation de Moreau-Yosida*

Soient $\mathbb{R}^n$ muni d'un produit scalaire noté $\langle \cdot, \cdot \rangle$, $\| \cdot \|$ la norme associée, et $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ une fonction convexe, semi-continue inférieurement (s.c.i.) et finie en au moins un point.

1°) Pour tout $r > 0$, on considère la fonction f_r définie sur $\mathbb{R}^n$ par :

$$x \in \mathbb{R}^n \longmapsto f_r(x) := \inf_{u \in \mathbb{R}^n} \left\{ f(u) + \frac{r}{2} \| x - u \|^2 \right\}. \tag{7.10}$$

(f_r est appelée régularisée de Moreau-Yosida de f.)

a) Vérifier que la fonction $u \longmapsto f(u) + \frac{r}{2} \| x - u \|^2$ est s.c.i. et 0-coercive sur $\mathbb{R}^n$.

En déduire que l'infimum est atteint dans la définition de $f_r(x)$.

Montrer que cet infimum est atteint en un point unique de $\mathbb{R}^n$, point que l'on notera x_r dans toute la suite.

b) Écrire f_r comme l'inf-convolution de deux fonctions convexes.

Vérifier que cette inf-convolution est exacte en tout point x de $\mathbb{R}^n$.

En déduire que f_r est différentiable en tout $x \in \mathbb{R}^n$ avec :

$$\nabla f_r(x) = r(x - x_r) \; ; \tag{7.11}$$

$$r(x - x_r) \in \partial f(x_r). \tag{7.12}$$

c) En écrivant les conditions de minimalité pour le problème de minimisation définissant $f_r(x)$, montrer que :

$$\begin{gathered} I + \tfrac{1}{r}\partial f \text{ est une multi-application surjective de } \mathbb{R}^n \text{ dans } \mathbb{R}^n \; ; \\ \forall x \in \mathbb{R}^n, \; \left(I + \tfrac{1}{r}\partial f\right)^{-1}(x) = x_r. \end{gathered} \tag{7.13}$$

(I désigne ici l'application identité de $\mathbb{R}^n$ dans $\mathbb{R}^n$).

2°) Déterminer $f_r(x)$ et x_r pour tout $x \in \mathbb{R}^n$ dans les cas suivants :

a) f est une forme affine sur $\mathbb{R}^n$, *i.e.*

$$u \in \mathbb{R}^n \longmapsto f(u) := \langle s, u \rangle + \alpha, \text{ où } \quad s \in \mathbb{R}^n \quad \text{et} \quad \alpha \in \mathbb{R}.$$

b) f est l'indicatrice d'un convexe fermé non vide C de $\mathbb{R}^n$.

c) $u \in \mathbb{R}^n \longmapsto f(u) := \frac{1}{2} \langle Au, u \rangle$, où $A : \mathbb{R}^n \to \mathbb{R}^n$ est linéaire autoadjoint.

3°) Montrer que x_r peut être caractérisé par l'une ou l'autre des conditions suivantes :

$$f(u) - f(x_r) + r \langle x_r - x, u - x_r \rangle \geqslant 0 \quad \text{pour tout} \quad u \in \mathbb{R}^n \; ; \tag{7.14}$$

$$f(u) - f(x_r) + r \langle u - x, u - x_r \rangle \geqslant 0 \quad \text{pour tout} \quad u \in \mathbb{R}^n. \tag{7.15}$$

Qu'expriment ces conditions dans le cas b) de la question précédente ?

4°) a) Montrer que l'application $x \longmapsto x_r$ est lipschitzienne de rapport 1.

b) Montrer que l'application $x \longmapsto \nabla f_r(x) = r(x - x_r)$ est lipschitzienne de rapport r.

c) Démontrer l'inégalité

$$0 \leqslant f_r(y) - f_r(x) - r\,\langle x - x_r, y - x\rangle \leqslant r \parallel x - y \parallel^2, \tag{7.16}$$

valable pour tout x, y dans $\mathbb{R}^n$.

5°) a) On suppose que f est bornée inférieurement sur $\mathbb{R}^n$. Indiquer pourquoi f_r est localement lipschitzienne et bornée inférieurement sur $\mathbb{R}^n$.

b) Quelle est la conjuguée (de Legendre-Fenchel) de la fonction $N_r : u \longmapsto \frac{r}{2} \parallel u \parallel^2$?

En déduire l'expression de la conjuguée f_r^* de f_r.

Comparer alors $\inf_{x\in\mathbb{R}^n} f(x)$ et $\inf_{x\in\mathbb{R}^n} f_r(x)$.

6°) a) Montrer que

$$f(x_r) \leqslant f_r(x) \leqslant f(x) \text{ pour tout } x \in \mathbb{R}^n. \tag{7.17}$$

b) Établir l'équivalence des assertions suivantes :

(i) x minimise f sur $\mathbb{R}^n$;
(ii) x minimise f_r sur $\mathbb{R}^n$;
(iii) $x = x_r$;
(iv) $f(x) = f(x_r)$;
(v) $f(x) = f_r(x)$.

7°) L'objet de cette question est l'étude du comportement de $f_r(x)$ quand $r \to +\infty$.

a) Soit $x \in \operatorname{dom} f := \{x \mid f(x) < +\infty\}$. Montrer tout d'abord que $x_r \to x$ quand $r \to +\infty$.

En déduire que $\{x \in \mathbb{R}^n \mid \partial f(x) \neq \phi\}$ est dense dans $\operatorname{dom} f$.

En déduire aussi que $f_r(x) \to f(x)$ quand $r \to +\infty$.

b) Soit $x \notin \operatorname{dom} f$. Montrer que $f_r(x) \to +\infty$ quand $r \to +\infty$. (On raisonnera par l'absurde en montrant que l'hypothèse « $\{f_r(x)\}_r$ est majorée » conduit à une contradiction.)

8°) On considère le problème de minimisation suivant :

$$(\mathcal{P}) \quad \text{Trouver } \overline{x} \in \mathbb{R}^n \text{ tel que } f(\overline{x}) = \overline{f} := \inf_{x\in\mathbb{R}^n} f(x),$$

où f vérifie en outre l'hypothèse suivante :

$$\forall \lambda \in \mathbb{R},\ \{x \in \mathbb{R}^n \mid f(x) \leqslant \lambda\} \text{ est borné.}$$

a) Indiquer pourquoi $\left\{x \in \mathbb{R}^n \mid f(x) = \overline{f}\right\}$ est un convexe compact non vide.

b) On considère la suite $\{x_k\}$ de $\mathbb{R}^n$ construite à partir de $x_0 \in \mathbb{R}^n$ de la manière suivante :

$$x_{k+1} := (I + \partial f)^{-1}(x_k),$$

c'est-à-dire x_{k+1} est l'unique point tel que

$$f(x_{k+1}) + \frac{1}{2} \parallel x_{k+1} - x_k \parallel^2 = \min_{u \in \mathbb{R}^n} \left\{ f(u) + \frac{1}{2} \parallel x_k - u \parallel^2 \right\}.$$

Montrer que la suite $\{f(x_k)\}$ est décroissante.
Montrer que la suite $\{x_k\}$ est bornée et que $\lim\limits_{k \to +\infty} \parallel x_{k+1} - x_k \parallel = 0$.
En déduire que $f(x_k) \to \overline{f}$ quand $k \to +\infty$.

Solution : 1°) a) La fonction $F_r : u \in \mathbb{R}^n \longmapsto F_r(u) := f(u) + \frac{r}{2} \parallel x - u \parallel^2$ est clairement s.c.i. Montrons qu'elle est 0-coercive sur $\mathbb{R}^n$. Comme f est convexe par hypothèse, elle possède une minorante affine : il existe $s_0 \in \mathbb{R}^n$ et $\alpha_0 \in \mathbb{R}$ tels que

$$f(u) \geqslant \langle s_0, u \rangle + \alpha_0 \text{ pour tout } u \in \mathbb{R}^n.$$

Par conséquent,

$$\begin{aligned} F_r(u) &\geqslant \langle s_0, u \rangle + \alpha_0 + \frac{r}{2} \parallel x - u \parallel^2 \\ &\geqslant \frac{r}{2} \parallel u + \left(\frac{s_0}{r} - x\right) \parallel^2 - \frac{r}{2} \parallel \frac{s_0}{r} - x \parallel^2 + \frac{r}{2} \parallel x \parallel^2 + \alpha_0, \end{aligned}$$

d'où on déduit : $\lim\limits_{\|u\| \to +\infty} F_r(u) = +\infty$.

Il en résulte qu'il existe bien un point minimisant F_r sur $\mathbb{R}^n$.

La fonction F_r étant strictement convexe, comme somme d'une fonction convexe et d'une fonction strictement convexe, il n'y a qu'un point minimisant F_r sur $\mathbb{R}^n$.

b) Soit $N_r := \frac{r}{2} \parallel \cdot \parallel^2$; il est clair que f_r est l'inf-convolution de f et de N_r : $f_r = f \square N_r$. De plus, cette inf-convolution est exacte en tout point de $\mathbb{R}^n$:

$$\forall x \in \mathbb{R}^n, \quad f_r(x) = (f \square N_r)(x) = f(x_r) + \frac{r}{2} \parallel x - x_r \parallel^2 . \tag{7.18}$$

On utilise alors la règle de calcul donnant le sous-différentiel d'une inf-convolution de fonctions pour obtenir

$$\forall x \in \mathbb{R}^n,\ \partial f_r(x) = \partial f(x_r) \cap \partial N_r(x - x_r).$$

Or $\partial N_r(x-x_r)=\{r(x-x_r)\}$, de sorte que

$$\partial f_r(x)=\{r(x-x_r)\} \text{ et } r(x-x_r)\in\partial f(x_r). \tag{7.19}$$

La fonction f_r est une fonction convexe de $\mathbb{R}^n$ dans $\mathbb{R}$, donc localement lipschitzienne sur $\mathbb{R}^n$; alors l'unique élément $r(x-x_r)$ de $\partial f_r(x)$ définit la différentielle (au sens de Fréchet) de f_r en x comme suit : $h\in\mathbb{R}^n\longmapsto\langle r(x-x_r),h\rangle$. Ainsi $\nabla f_r(x)=r(x-x_r)$.

c) Le point x_r est caractérisé par la condition suivante :

$$0\in\partial\left(f+\frac{r}{2}\parallel\cdot-x\parallel^2\right)(x_r),$$

ce qui revient à

$$0\in\partial f(x_r)+r(x_r-x) \text{ (condition déjà rencontrée en (7.19))},$$

ou encore

$$x\in\left(I+\frac{1}{r}\partial f\right)(x_r). \tag{7.20}$$

D'après ce qui a été vu plus haut, pour tout $x\in\mathbb{R}^n$, l'élément x_r, solution de l'équation multivoque (7.20), est défini de façon unique. En conséquence :

– la multi-application $I+\frac{1}{r}\partial f$ est surjective (*i.e.*, $\left(I+\frac{1}{r}\partial f\right)(\mathbb{R}^n)=\mathbb{R}^n$) ;

– l'inverse de la multi-application $I+\frac{1}{r}\partial f$ est en fait univoque ; il définit x_r précisément :

$$\forall x\in\mathbb{R}^n,\ x_r=\left(I+\frac{1}{r}\partial f\right)^{-1}(x). \tag{7.21}$$

2°) a) On a déjà utilisé la décomposition suivante dans 1°) a) :

$$\langle s,u\rangle+\alpha+\frac{r}{2}\parallel x-u\parallel^2=\frac{r}{2}\parallel u+\left(\frac{s}{r}-x\right)\parallel^2-\frac{r}{2}\parallel\frac{s}{r}-x\parallel^2+\frac{r}{2}\parallel x\parallel^2+\alpha.$$

Il s'ensuit :

$$x_r=x-\frac{s}{r},\ f_r(x)=-\frac{\parallel s\parallel^2}{r}+\langle s,x\rangle+\alpha.$$

On peut aussi calculer (immédiatement) x_r grâce à (7.21).

b) $f_r(x)=\inf_{u\in C}\left\{\frac{r}{2}\parallel x-u\parallel^2\right\}=\frac{r}{2}d_C^2(x)$, où d_C désigne la fonction-distance à C.

Quant à x_r, ce n'est autre que la projection de x sur C : $x_r=p_C(x)$.

c) Comme $\partial f(u)=\{Au\}$ pour tout u, il est aisé de déterminer x_r *via* (7.21) :

$$x_r=\left(I+\frac{A}{r}\right)^{-1}(x).$$

Pour ce qui est de $f_r(x)$,

$$f_r(x) = \frac{1}{2}\langle Ax_r, x_r\rangle + \frac{r}{2}\parallel x - x_r \parallel^2$$
$$= \frac{1}{2}\langle A_r x, x\rangle\,, \text{ où } A_r := A\left(I + \frac{A}{r}\right)^{-1} = r\left[I - \left(I + \frac{A}{r}\right)^{-1}\right].$$

3°) x_r est solution d'un problème de minimisation où la fonction-objectif est de la forme $f + g$, avec f convexe s.c.i. et g convexe différentiable sur $\mathbb{R}^n$. Les solutions $\overline{u}$ d'un tel problème sont caractérisées par l'une ou l'autre des conditions suivantes (*cf.* Exercice II.7) :

$$\forall u \in \mathbb{R}^n,\ \langle \nabla g(\overline{u}), u - \overline{u}\rangle + f(u) - f(\overline{u}) \geqslant 0\ ; \tag{7.22}$$

$$\forall u \in \mathbb{R}^n,\ \langle \nabla g(u), u - \overline{u}\rangle + f(u) - f(\overline{u}) \geqslant 0. \tag{7.23}$$

Avec $g : u \longmapsto g(u) = \frac{r}{2}\parallel u - x \parallel^2$, elles donnent précisément les conditions (7.14) et (7.15) attendues.

La première condition, c'est-à-dire (7.22), a déjà été vue puisqu'elle est équivalente à

$$-\nabla g(x_r) = r(x - x_r) \in \partial f(x_r).$$

Seule la condition (7.15), traduction de (7.23), a un caractère nouveau.

Si l'on considère l'exemple b) de la question précédente, on obtient :

$$\langle x - p_C(x), u - p_C(x)\rangle \leqslant 0 \text{ pour tout } u \in C$$
$$\text{(caractérisation variationnelle usuelle de } p_C(x))\ ; \tag{7.24}$$

$$\langle u - p_C(x), x - u\rangle \leqslant 0 \text{ pour tout } u \in C$$
$$(cf. \text{ Exercice 6.16}). \tag{7.25}$$

4°) a) L'inégalité (7.14) écrite successivement pour x_r et y_r donne :

$$\forall u \in \mathbb{R}^n,\quad f(u) - f(x_r) \geqslant r\langle x - x_r, u - x_r\rangle$$
$$\forall v \in \mathbb{R}^n,\quad f(v) - f(y_r) \geqslant r\langle y - y_r, v - y_r\rangle\,.$$

En faisant $u = y_r$ et $v = x_r$ et en additionnant, on obtient :

$$\parallel x_r - y_r \parallel^2 \leqslant \langle x_r - y_r, x - y\rangle\,, \tag{7.26}$$

d'où

$$\parallel x_r - y_r \parallel \leqslant \parallel x - y \parallel\,.$$

Donc l'application $x \longmapsto x_r$ est monotone et lipschitzienne de rapport 1.

b) On a :

$$\begin{aligned}\|(x-x_r)-(y-y_r)\|^2 &= \|\,(x-y)-(x_r-y_r)\,\|^2\\ &= \|\,x-y\,\|^2+\|\,x_r-y_r\,\|^2-2\,\langle x-y,x_r-y_r\rangle\\ &\leqslant \|\,x-y\,\|^2 \ \text{ d'après } (7.26)\,.\end{aligned}$$

L'application $x\longmapsto \nabla f_r(x)=r(x-x_r)$ est ainsi lipschitzienne de rapport r.

c) Puisque $r(y-y_r)\in\partial f_r(y)$, on a l'inégalité

$$f_r(x)-f_r(y)\geqslant r\,\langle y-y_r,x-y\rangle\,.$$

En retranchant $-r\,\langle x-x_r,x-y\rangle$ aux deux membres, on obtient :

$$\begin{aligned}f_r(x)-f_r(y)-r\,\langle x-x_r,x-y\rangle &\geqslant -r\,\|\,x-y\,\|^2+r\,\langle x-y,x_r-y_r\rangle\\ &\geqslant -r\,\|\,x-y\,\|^2,\end{aligned}$$

soit encore

$$f_r(y)-f_r(x)-r\,\langle x-x_r,y-x\rangle\leqslant r\,\|\,x-y\,\|^2\,.$$

Une autre méthode consisterait à utiliser la propriété de Lipschitz de ∇f_r dans un développement de Taylor-Lagrange (ou Taylor avec reste sous forme d'intégrale) du 1er ordre de f_r.

5°) a) Comme cela a déjà été dit, $f_r:\mathbb{R}^n\to\mathbb{R}$ est convexe, donc localement lipschitzienne sur $\mathbb{R}^n$. Comme

$$f_r(x)\geqslant f(x_r)\geqslant \inf_{x\in\mathbb{R}^n} f(x),$$

f_r est bornée inférieurement sur $\mathbb{R}^n$ dès que f l'est.

b) La conjuguée N_r^* de N_r est $s\in\mathbb{R}^n\longmapsto N_r^*(s)=\frac{1}{2r}\,\|\,s\,\|^2$.
Puisque $f_r=f\square N_r$, on a $f_r^*=f^*+N_r^*$, c'est-à-dire :

$$\forall s\in\mathbb{R}^n,\ \ f_r^*(s)=f^*(s)+\frac{1}{2r}\,\|\,s\,\|^2\,.$$

En particulier

$$\underset{(=-\inf_{x\in\mathbb{R}^n} f_r(x))}{f_r^*(0)} \qquad \text{coïncide avec} \qquad = \underset{(=-\inf_{x\in\mathbb{R}^n} f(x))}{f^*(0)}\,.$$

6°) a) De par les définitions mêmes :

$$f(x_r)+\frac{r}{2}\,\|\,x-x_r\,\|^2=f_r(x)\leqslant f(x)\ \text{ pour tout } x\in\mathbb{R}^n. \qquad (7.27)$$

b) $[(ii) \Leftrightarrow (iii)]$. On sait que f_r est convexe et différentiable, avec $\nabla f_r(x) = r(x - x_r)$ en tout x. Par conséquent, x minimise f_r sur $\mathbb{R}^n$ si et seulement si $\nabla f_r(x) = 0$.

$[(i) \Leftrightarrow (iii)]$. x_r est caractérisé comme l'unique élément vérifiant

$$x \in \left(I + \frac{1}{r}\partial f\right)(x_r).$$

Il est donc clair que $x = x_r$ si et seulement si $0 \in \partial f(x)$, ou, d'une manière équivalente, si x minimise f sur $\mathbb{R}^n$.

$[(iii) \Rightarrow (iv) \Rightarrow (v)]$. Immédiat à partir de (7.27) .

$[(v) \Rightarrow (iii)]$. Si $f_r(x) = f(x)$, cela signifie que la borne inférieure dans la définition de $f_r(x)$ est atteinte en x, donc $x = x_r$.

7°) a) Soit $x \longmapsto \langle s_0, x\rangle + \alpha_0$ une minorante affine de f. Avec (7.27) on obtient :

$$f(x) \geqslant \langle s_0, x_r\rangle + \alpha_0 + \frac{r}{2} \parallel x - x_r \parallel^2,$$

ce qui implique (*cf.* décomposition du 1°) a) :

$$+\infty > \frac{f(x)}{r} \geqslant \frac{1}{2} \parallel x_r - x + \frac{s_0}{r} \parallel^2 - \frac{1}{2} \parallel \frac{s_0}{r} - x \parallel^2 + \frac{1}{2} \parallel x \parallel^2 + \frac{\alpha_0}{r}.$$

De ce fait, $\parallel x_r - x + \frac{s_0}{r} \parallel$ tend vers 0 quand $r \to +\infty$, et donc $\parallel x_r - x \parallel$ aussi.

Comme $\partial f(x_r)$ n'est pas vide (car $r(x - x_r) \in \partial f(x_r)$), $x_r \in \operatorname{dom} \partial f := \{x \mid \partial f(x) \neq \phi\}$ pour tout $r > 0$; avec ce qui précède, on a donc démontré que $\operatorname{dom} f$ est contenu dans l'adhérence de $\operatorname{dom} \partial f$.

La convergence de x_r vers x, combinée avec la semi-continuité inférieure de f (en x) et l'inégalité $f(x_r) \leqslant f_r(x) \leqslant f(x)$, font que $f_r(x) \to f(x)$ quand $r \to +\infty$.

b) Soit $x \notin \operatorname{dom} f$, *i.e.* $f(x) = +\infty$. Il nous faut montrer que $f_r(x) \longrightarrow +\infty$ quand $r \to +\infty$. La suite $\{f_r(x)\}_r$ étant croissante, le contraire de l'énoncé est : il existe K tel que $f_r(x) \leqslant K$ pour tout $r > 0$.

Toujours grâce à la minorante $\langle s_0, \cdot\rangle + \alpha_0$ de f, l'inégalité

$$\begin{aligned}\frac{K}{r} &\geqslant \frac{1}{r} f_r(x) = \frac{1}{r}\left[f(x_r) + \frac{r}{2} \parallel x - x_r \parallel^2\right] \\ &\geqslant \frac{1}{2} \parallel x_r - x + \frac{s_0}{r} \parallel^2 - \frac{\parallel s_0 \parallel^2}{2r^2} + \frac{\langle s_0, x\rangle}{r} + \frac{\alpha_0}{r}\end{aligned}$$

conduit à nouveau à : $x_r \to x$ quand $r \to +\infty$.

Ensuite

$$f(x_r) \leqslant f_r(x) \leqslant K \text{ pour tout } r > 0,$$
$$x_r \to x \text{ quand } r \to +\infty,$$
$$f \text{ est semi-continue inférieurement en } x,$$

font que $f(x) \leqslant K$. D'où la contradiction.

8°) L'hypothèse faite sur f revient à sa 0-coercivité :

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

a) La convexité de f, sa semi-continuité et sa coercivité font que $\overline{f} > -\infty$ et que $\{x \mid f(x) = \overline{f}\}$ est un convexe compact non vide de $\mathbb{R}^n$.

b) Par définition même de x_{k+1}, on a :

$$f(x_{k+1}) + \frac{1}{2} \| x_{k+1} - x_k \|^2 \leqslant f(x_k).$$

La suite $\{f(x_k)\}$ est évidemment décroissante (et bornée inférieurement par $\overline{f}$).

Conséquences :

- $x_k \in \{x \in \mathbb{R}^n \mid f(x) \leqslant f(x_0)\}$ qui est borné par hypothèse ;
- $f(x_k) \downarrow \mu \geqslant \overline{f}$ quand $k \to +\infty$. (7.28)

Posons $F(x) := (I + \partial f)^{-1}(x)$. Cette application $F : \mathbb{R}^n \to \mathbb{R}^n$ (qui est à la base de la définition de la suite $\{x_k\}$ puisque $x_{k+1} = F(x_k)$) vérifie l'inégalité suivante :

$$\forall x, y \in \mathbb{R}^n, \ \| F(x) - F(y) \|^2 \leqslant \| x - y \|^2 - \| [F(x) - x] - [F(y) - y] \|^2 \quad (7.29)$$

qui n'est autre, après développement, que l'inégalité (7.26) pour $r = 1$.

Soit $\overline{x}$ minimisant f sur $\mathbb{R}^n$. Sachant que $F(\overline{x}) = \overline{x}$ et $x_{k+1} = F(x_k)$, il vient de (7.29) :

$$\| x_{k+1} - \overline{x} \|^2 \leqslant \| x_k - \overline{x} \|^2 - \| x_{k+1} - x_k \|^2 . \quad (7.30)$$

Conséquences :

• La suite $\{\| x_k - \overline{x} \|^2\}$ est décroissante, donc convergente ;

• $\| x_{k+1} - \overline{x} \|^2 - \| x_k - \overline{x} \|^2 \to_{k \to +\infty} 0$ (la suite $\{\| x_k - \overline{x} \|^2\}$ étant convergente) ;

• $\| x_{k+1} - x_k \| \to_{k \to +\infty} 0$ (cela résulte de (7.30)).

Revenons à présent à la suite $\{f(x_k)\}$.

Comme f est convexe, on a

$$\overline{f} = f(\overline{x}) \geqslant f(x_{k+1}) + f'(x_{k+1}, \overline{x} - x_{k+1}). \quad (7.31)$$

Puisque x_{k+1} minimise $u \longmapsto f(u) + \frac{1}{2} \| x_k - u \|^2$,

$$f'(x_{k+1}, \overline{x} - x_{k+1}) + \langle x_{k+1} - x_k, \overline{x} - x_{k+1} \rangle \geqslant 0. \tag{7.32}$$

Sachant que la suite $\{\| \overline{x} - x_{k+1} \|\}$ est bornée et que $\lim_{k \to +\infty} \| x_{k+1} - x_k \| = 0$, un passage à la limite dans (7.32) conduit à

$$\liminf_{k \to +\infty} f'(x_{k+1}, \overline{x} - x_{k+1}) \geqslant 0.$$

D'où on tire grâce à (7.31) : $\lim_{k \to +\infty} f(x_k) \leqslant \overline{\mu}$. Ce qui, avec (7.28), nous permet de conclure que $f(x_k) \to \overline{f}$ quand $k \to +\infty$.

On peut en fait démontrer que la suite (*toute entière*, pas seulement une sous-suite) $\{x_k\}$ converge vers *un* minimum de f sur $\mathbb{R}^n$.

***Exercice VII.16.** Soit $f : \mathbb{R}^n \to \mathbb{R}$ convexe et de classe C^2 sur $\mathbb{R}^n$. Montrer que l'application $c : \mathbb{R}^n \to \mathbb{R}^n$ définie par $c(p) := p + \nabla f(p)$ est un C^1-difféomorphisme de $\mathbb{R}^n$ sur lui-même.

Indication. Pour démontrer le caractère bijectif de c, on pourra minimiser la fonction $g : u \in \mathbb{R}^n \longmapsto g(u) := f(u) + \frac{1}{2} \| u - x \|^2$, où $x \in \mathbb{R}^n$ est donné.

Solution : *c est bijective.* Pour x donné, considérons

$$g : u \longmapsto g(u) = f(u) + \tfrac{1}{2} \| u - x \|^2 .$$

La fonction f étant convexe, elle est minorée par une fonction affine : il existe $s_0 \in \mathbb{R}^n$ et $\alpha_0 \in \mathbb{R}$ tels que $f(u) \geqslant \langle s_0, u \rangle + \alpha_0$ pour tout $u \in \mathbb{R}^n$.

Par conséquent,

$$\begin{aligned} g(u) &\geqslant \langle s_0, u \rangle + \alpha_0 + \frac{1}{2} \| u - x \|^2 \\ &\geqslant \frac{1}{2} \| u + s_0 - x \|^2 - \frac{1}{2} \| s_0 - x \|^2 + \frac{\| x \|^2}{2} + \alpha_0, \end{aligned}$$

d'où on déduit : $\lim_{\|u\| \to +\infty} g(u) = +\infty$. Par ailleurs, g est (continue et) strictement convexe sur $\mathbb{R}^n$. Il existe donc un unique élément de $\mathbb{R}^n$, noté $p(x)$, minimisant g sur $\mathbb{R}^n$; cet élément $p(x)$ est caractérisé par l'équation

$$\nabla f(p(x)) + p(x) - x = 0 \quad \text{(condition nécessaire et suffisante de minimalité)},$$

soit

$$c(p(x)) = x.$$

Caractère C^1 de c^{-1}. La différentielle de c en x est représentée par la matrice jacobienne $Jc(x)$ qui vaut ici $I_n+\nabla^2 f(x)$. Comme $\nabla^2 f(x)$ est symétrique semi-définie positive, $Jc(x)$ a un déterminant supérieur à dét $I_n = 1$. D'après le théorème d'inversion locale, $c^{-1} = p$ est C^1 sur $\mathbb{R}^n$ avec

$$Jp(x) = [Jc(p(x))]^{-1} \text{ pour tout } x \in \mathbb{R}^n.$$

Autrement dit :

$$Jp(x) = \left[I_n + \nabla^2 f(p(x))\right]^{-1} \text{ pour tout } x \in \mathbb{R}^n.$$

En résumé : c est une bijection de $\mathbb{R}^n$ sur lui-même, c et c^{-1} sont de classe C^1 sur $\mathbb{R}^n$; c est donc un C^1-difféomorphisme de $\mathbb{R}^n$ sur lui-même.

***Exercice VII.17.** Soit $f : \mathbb{R}^n \to \mathbb{R}$ convexe et de classe C^2 sur $\mathbb{R}^n$. Pour tout $r > 0$, on considère la fonction f_r définie sur $\mathbb{R}^n$ par :

$$x \in \mathbb{R}^n \longmapsto f_r(x) := \inf_{u\in\mathbb{R}^n} \left\{ f(u) + \frac{r}{2} \parallel x - u \parallel^2 \right\}.$$

(f_r est la régularisée de Moreau-Yosida de f.)

En utilisant les résultats des Exercices VII.15 (1re question) et 7.16, montrer que f_r est de classe C^2 sur $\mathbb{R}^n$ et exprimer $\nabla^2 f_r$ en fonction de $\nabla^2 f$ et de l'inverse de l'application $p \longmapsto p + \frac{1}{r}\nabla f(p)$.

Solution : Si $p_r(x)$ désigne l'unique élément de $\mathbb{R}^n$ minimisant la fonction $u \longmapsto f(u) + \frac{r}{2} \parallel x - u \parallel^2$ sur $\mathbb{R}^n$, on a (*cf.* Exercice VII.15, 1re question) :

$$\nabla f_r(x) = r(x - p_r(x)) = \nabla f(p_r(x)).$$

Par conséquent, f_r est deux fois (continûment) différentiable en x si, et seulement si, p_r est (continûment) différentiable en x ; dans ce cas

$$\nabla^2 f_r(x) = r(I_n - Jp_r(x)) = \nabla^2 f(p_r(x)) \cdot Jp_r(x)$$

Or, lorsque f est C^2 sur $\mathbb{R}^n$, l'application p_r est de classe C^1 sur $\mathbb{R}^n$, avec $Jp_r(x) = \left[I_n + \frac{\nabla^2 f(p_r(x))}{r}\right]^{-1}$ (*cf.* Exercice VII.16). En définitive, f_r est de classe C^2 sur $\mathbb{R}^n$ avec :

$$\begin{cases} \nabla^2 f_r(x) = r\left(I_n - \left[I_n + \frac{\nabla^2 f(p_r(x))}{r}\right]^{-1}\right) = \nabla^2 f(p_r(x)) \left[I_n + \frac{\nabla^2 f(p_r(x))}{r}\right]^{-1}, \\ p_r(x) = \left(I_n + \frac{\nabla f}{r}\right)^{-1}(x). \end{cases}$$

***Exercice VII.18.** Soit $f : \mathbb{R}^n \to \mathbb{R}$ convexe (mais pas nécessairement différentiable). On désigne par $\partial f(x)$ le sous-différentiel de f en x.

Soient $x \in \mathbb{R}^n$ et $y < f(x)$; on désigne par $(\overline{x}, \overline{y})$ la projection de (x, y) sur l'épigraphe de f.

– Vérifier que $\overline{y} = f(\overline{x})$ et $y < f(\overline{x})$.

– Montrer que

$$\frac{x - \overline{x}}{f(\overline{x}) - y} \in \partial f(\overline{x}).$$

Solution : $\mathbb{R}^n \times \mathbb{R}$ est structuré en espace euclidien grâce au produit scalaire

$$\big\langle (x, r), (x', r') \big\rangle_{\mathbb{R}^{n+1}} := \big\langle x, x' \big\rangle_{\mathbb{R}^n} + rr'.$$

L'épigraphe de f, epi $f := \{(x, r) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) \leqslant r\}$, est ici un convexe fermé de $\mathbb{R}^n \times \mathbb{R}$. Il peut être vu comme l'ensemble de sous-niveau (ou tranche) de la fonction $g : (x, r) \longmapsto g(x, r) := f(x) - r$ au niveau 0, *i.e.*,

$$\text{epi } f = \{(x, r) \in \mathbb{R}^n \times \mathbb{R} \mid g(x, r) \leqslant 0\}.$$

Comme g est une fonction convexe (continue) et qu'il existe $(x_0, r_0) \in \mathbb{R}^n \times \mathbb{R}$ tel que $g(x_0, r_0) < 0$ (hypothèse de Slater donc), on a :

$$\begin{aligned} \text{int}(\text{epi } f) &= \{(x, r) \mid g(x, r) < 0\} = \text{epi } f \setminus \text{gr } f, \\ \text{fr}(\text{epi } f) &= \{(x, r) \mid g(x, r) = 0\} = \text{gr } f. \end{aligned}$$

Par hypothèse, $(x, y) \notin \text{epi } f$; la projection $(\overline{x}, \overline{y})$ de (x, y) sur epi f se trouve donc sur la frontière de epi f, d'où $f(\overline{x}) = \overline{y}$.

L'élément $(\overline{x}, f(\overline{x}))$, projection de (x, y) sur epi f, est solution de l'inéquation variationnelle suivante :

$$\langle x - \overline{x}, u - \overline{x} \rangle + (y - f(\overline{x}))(v - f(\overline{x})) \leqslant 0 \text{ pour tout } (u, v) \in \text{epi } f. \quad (7.33)$$

Il s'ensuit :

• $y - f(\overline{x}) \leqslant 0$ (sinon on arrive à une contradiction dans (7.33) en prenant $(u, v + \rho) \in \text{epi } f$ et en faisant $\rho \to +\infty$) ;

• $y < f(\overline{x})$ en fait (sinon, avec $y = f(\overline{x})$, l'inégalité (7.33) indique que $\langle x - \overline{x}, u - \overline{x} \rangle \leqslant 0$ pour tout $u \in \mathbb{R}^n$ (car f est partout finie sur $\mathbb{R}^n$), et donc $x = \overline{x}$; on aurait alors $(x, y) = (\overline{x}, f(\overline{x})) \in \text{epi } f$, ce qui n'est pas le cas).

En divisant par $f(\overline{x}) - y$ dans (7.33), on obtient :

$$\left\langle \frac{x - \overline{x}}{f(\overline{x}) - y}, u - \overline{x} \right\rangle + f(\overline{x}) - v \leqslant 0 \quad \text{pour tout } (u, v) \in \text{epi } f.$$

Cette dernière inégalité, écrite pour tous les $(u, f(u))$, $u \in \mathbb{R}^n$, conduit à :

$$f(u) \geqslant f(\overline{x}) + \left\langle \frac{x-\overline{x}}{f(\overline{x})-y}, u-\overline{x} \right\rangle \quad \text{pour tout } u \in \mathbb{R}^n,$$

c'est-à-dire $\frac{x-\overline{x}}{f(\overline{x})-y} \in \partial f(\overline{x})$.

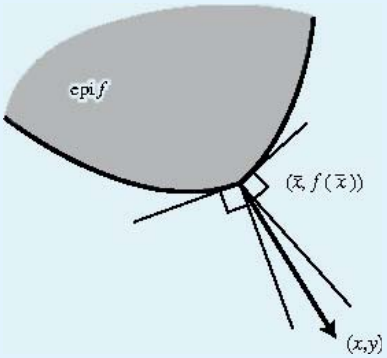


Figure 25.

Illustration du fait que $(x-\overline{x}, y-f(\overline{x}))$ est dans le cône normal à epi f en $(\overline{x}, f(\overline{x}))$, soit encore $\frac{x-\overline{x}}{f(\overline{x})-y} \in \partial f(\overline{x})$.

***Exercice VII.19.** Soit $E := \mathcal{S}_n(\mathbb{R})$ structuré en espace euclidien à l'aide du produit scalaire $\langle\langle \cdot, \cdot \rangle\rangle$ (rappel : $\langle\langle A, B\rangle\rangle := \mathrm{tr}(AB)$) ; soit $f : E \to \mathbb{R}$ définie comme suit :

$$\forall X \in E,\ f(X) := \frac{1}{n} \parallel X \parallel^2 - \left(\frac{\mathrm{tr}\, X}{n}\right)^2.$$

On souhaite calculer f^*, c'est-à-dire

$$S \in E \longmapsto f^*(S) := \sup_{X \in E} \{\langle\langle S, X\rangle\rangle - f(X)\}.$$

1°) Vérifier que $f(X) = \frac{1}{n} \parallel X - \frac{\mathrm{tr}\, X}{n} I_n \parallel^2$ pour tout $X \in E$. En déduire que f est convexe sur E.

2°) a) Montrer que le supremum de la fonction $X \longmapsto \langle\langle S, X\rangle\rangle - f(X)$ sur E est fini et atteint si, et seulement si, $\mathrm{tr}\ S = 0$.

En déduire $f^*(S)$ lorsque $\mathrm{tr}\ S = 0$.

b) Montrer que $f^*(S) = +\infty$ si $\mathrm{tr}\ S \neq 0$.

Indication. La décomposition $E = V \oplus V^\perp$, où $V = \{X \in E \mid \operatorname{tr} X = 0\}$ et $V^\perp = \mathbb{R} I_n$, peut être utile dans l'organisation des calculs.

Solution : 1°) En utilisant la règle de calcul

$$\| M - N \|^2 = \| M \|^2 \; + \; \| N \|^2 - 2 \langle\langle M, N \rangle\rangle ,$$

on obtient

$$\left\| X - \frac{\operatorname{tr} X}{n} I_n \right\|^2 = \| X \|^2 + \left(\frac{\operatorname{tr} X}{n} \right)^2 \| I_n \|^2 - \frac{2}{n} (\operatorname{tr} X)^2 .$$

Comme $\| I_n \|^2 = \langle\langle I_n, I_n \rangle\rangle = \operatorname{tr}(I_n^2) = n$, il vient

$$\left\| X - \frac{\operatorname{tr} X}{n} I_n \right\|^2 = \| X \|^2 - \left(\frac{\operatorname{tr} X}{n} \right)^2 .$$

D'où l'expression annoncée de $f(X)$.

L'application $X \in E \longmapsto X - \frac{\operatorname{tr} X}{n} I_n \in E$ est affine, la fonction $U \in E \longmapsto \frac{1}{n} \| U \|^2 \in \mathbb{R}$ est convexe ; par conséquent la composée des deux, qui n'est autre que f, est convexe.

2°) a) La fonction $g : X \in E \longmapsto g(X) := \langle\langle S, X \rangle\rangle - f(X)$ est concave et différentiable sur E, avec

$$\nabla g(X) = S - \frac{2}{n} \left(X - \frac{\operatorname{tr} X}{n} I_n \right) .$$

Le supremum de g sur E est fini et atteint si, et seulement si, l'équation $\nabla g(X) = 0$ est résoluble ; la valeur maximale est alors $g(X)$ où X est une des solutions de l'équation en question. Voyons pour quels S cela est possible.

Avoir $\nabla g(X) = 0$, *i.e.* $S = \frac{2}{n} \left(X - \frac{\operatorname{tr} X}{n} I_n \right)$, implique $\operatorname{tr} S = 0$. Réciproquement, si $\operatorname{tr} S = 0$, l'équation $\nabla g(X) = 0$ est résoluble avec $X = \frac{nS}{2}$. En conséquence, lorsque $\operatorname{tr} S = 0$,

$$f^*(S) = \langle\langle S, nS/2 \rangle\rangle - f\,(nS/2) = n/4 \operatorname{tr}(S^2).$$

b) Supposons $\operatorname{tr} S \neq 0$ et considérons $X_k := k(\operatorname{tr} S) I_n$, $k \in \mathbb{N}$. Alors

$$\langle\langle S, X_k \rangle\rangle - f(X_k) = k(\operatorname{tr} S)^2 \to +\infty \quad \text{quand} \quad k \to +\infty.$$

Il s'ensuit que $f^*(S) = +\infty$.

***Problème VII.20.** *Conjugaison et sous-différentiation de fonctions spectrales*

Soit $E := \mathcal{S}_n(\mathbb{R})$ structuré en espace euclidien à l'aide du produit scalaire $\langle\langle \cdot, \cdot \rangle\rangle$ (rappel : $\langle\langle A, B \rangle\rangle := \mathrm{tr}(AB)$). Étant donnée $f \in \Gamma_0(\mathbb{R}^n)$ symétrique (*i.e.* vérifiant $f(x_{\sigma(1)}, \ldots, x_{\sigma(n)}) = f(x_1, \ldots, x_n)$ pour tout $(x_1, \ldots, x_n) \in \mathbb{R}^n$ et toute permutation σ de $\{1, \ldots, n\}$), on définit $V_f : \mathcal{S}_n(\mathbb{R}) \to \mathbb{R} \cup \{+\infty\}$ de la manière suivante :

$$\forall M \in \mathcal{S}_n(\mathbb{R}),\ V_f(M) := f\left(\lambda_1(M), \ldots, \lambda_n(M)\right), \tag{7.34}$$

où $\lambda_1(M) \geqslant \lambda_2(M) \geqslant \ldots \geqslant \lambda_n(M)$ désignent les valeurs propres de M.

De telles fonctions V_f sont appelées *fonctions de valeurs propres* ou *fonctions spectrales*.

L'objet du problème est de déterminer la conjuguée (resp. le sous-différentiel) de V_f en fonction de la conjuguée (resp. du sous-différentiel) de f.

1°) a) Vérifier que $f^*(\in \Gamma_0(\mathbb{R}^n))$ est également symétrique.

b) Montrer :

$$V_f(M) = \sup_{S \in E} \left\{\mathrm{tr}(MS) - f^*\left(\lambda_1(S), \ldots, \lambda_n(S)\right)\right\}. \tag{7.35}$$

En déduire que $V_f \in \Gamma_0(E)$ et $(V_f)^* = V_{f^*}$.

2°) À l'aide de fonctions $f \in \Gamma_0(\mathbb{R}^n)$ symétriques, faire une liste d'exemples de fonctions de valeurs propres V_f qui soient convexes.

3°) a) Montrer que $S \in E$ est un sous-gradient de V_f en M (*i.e.*, $S \in \partial V_f(M)$) si et seulement si :

$$\begin{cases} (\lambda_1(S), \ldots, \lambda_n(S)) \in \partial f\left(\lambda_1(M), \ldots, \lambda_n(M)\right) \\ \text{et il existe} \\ U \text{ orthogonale telle que } \begin{cases} U^\top M U = \mathrm{diag}\left(\lambda_1(M), \ldots, \lambda_n(M)\right) \\ \text{et} \\ U^\top S U = \mathrm{diag}\left(\lambda_1(S), \ldots, \lambda_n(S)\right). \end{cases} \end{cases} \tag{7.36}$$

b) On suppose f différentiable en $(\lambda_1(M), \ldots, \lambda_n(M))$. Déduire de ce qui précède que V_f est différentiable en M, avec :

$$\nabla V_f(M) = U\left[\mathrm{diag}\, \nabla f\left(\lambda_1(M), \ldots, \lambda_n(M)\right)\right] U^\top, \tag{7.37}$$

où U est une matrice orthogonale (quelconque) telle que $U^\top M U = \mathrm{diag}(\lambda_1(M), \ldots, \lambda_n(M))$, et $\mathrm{diag}\, \nabla f(\lambda_1(M), \ldots, \lambda_n(M))$ désigne la matrice diagonale construite à partir du vecteur $\nabla f(\lambda_1(M), \ldots, \lambda_n(M))$.

4°) Illustrer les résultats obtenus dans le problème sur l'exemple suivant :

$$x=(x_1,\dots,x_n)\in\mathbb{R}^n \longmapsto f(x):=\begin{cases} -\sum_{i=1}^n \ln(x_i) & \text{si} \\ x_i>0 \ \text{ pour tout } \ i=1,\dots,n\ ; \\ +\infty \ \text{ sinon.} \end{cases}$$

Indication. Pour démontrer l'inégalité $\geqslant$ dans (7.35) et démontrer (7.36), on pourra utiliser le résultat suivant : pour M et S dans E,

$$\mathrm{tr}(SM)=\sum_{i=1}^n \lambda_i(SM)\leqslant \sum_{i=1}^n \lambda_i(S)\lambda_i(M),$$

avec égalité si, et seulement si, il existe U orthogonale telle que $U^\top MU = diag(\lambda_1(M),\dots,\lambda_n(M))$ et $U^\top SU = diag(\lambda_1(S),\dots,\lambda_n(S))$ (décomposition spectrale simultanée).

<u>*Solution :*</u> 1°) a) Par définition de f^*,

$$\forall s=(s_1,\dots,s_n)\in\mathbb{R}^n,\ f^*(s)=\sup_{(x_1,\dots,x_n)}\left\{\sum_{i=1}^n s_ix_i-f(x_1,\dots,x_n)\right\}. \quad (7.38)$$

Si σ est une permutation de $\{1,\dots,n\}$, $\sum_{i=1}^n s_ix_i=\sum_{i=1}^n s_{\sigma(i)}x_{\sigma(i)}$, $f(x_1,\dots,x_n)=f(x_{\sigma(1)},\dots,x_{\sigma(n)})$, et $\left\{(x_{\sigma(1)},\dots,x_{\sigma(n)})\,|\,(x_1,\dots,x_n)\in\mathbb{R}^n\right\}=\mathbb{R}^n$; il est ainsi clair, à partir de (7.38), que

$$f^*(s_{\sigma(1)},\dots,s_{\sigma(n)})=f^*(s_1,\dots,s_n).$$

Comme cela a été fait à partir de f pour V_f, il est donc possible de définir $V_{f^*}:S\in\mathcal{S}_n(\mathbb{R})\longmapsto V_{f^*}(S):=f^*(\lambda_1(S),\dots,\lambda_n(S))$.

b) *–Inégalité $\leqslant$ dans* (7.35)

Soit U orthogonale telle que $M=U^\top\mathrm{diag}(\lambda_1(M),\dots,\lambda_n(M))U$; ainsi

$$\mathrm{tr}(MS)=\mathrm{tr}[\mathrm{diag}(\lambda_1(M),\dots,\lambda_n(M))USU^\top].$$

L'application de E dans E qui à S associe $USU^\top$ étant bijective et préservant les valeurs propres, on a :

$$\sup_{S\in E}\{\mathrm{tr}(MS) - f^*(\lambda_1(S),\dots,\lambda_n(S))\} = \sup_{S\in E}\{\mathrm{tr}[\mathrm{diag}(\lambda_1(M),\dots,\lambda_n(M))S] - f^*(\lambda_1(S),\dots,\lambda_n(S))\}.$$

En se restreignant aux matrices diagonales $S = \mathrm{diag}(s_1,\dots,s_n)$, on en déduit

$$\sup_{S\in E}\{\mathrm{tr}(MS) - f^*(\lambda_1(S),\dots,\lambda_n(S))\} \geqslant \sup_{(s_1,\dots,s_n)\in\mathbb{R}^n}\left\{\sum_{i=1}^n \lambda_i(M)s_i - f^*(s_1,\dots,s_n)\right\}.$$

L'expression de droite n'est autre que $f^{**}(\lambda_1(M),\dots,\lambda_n(M))$, soit $f(\lambda_1(M),\dots,\lambda_n(M))$ puisque $f\in\Gamma_0(\mathbb{R}^n)$.

– *Inégalité* $\geqslant$ *dans* (7.35)

Pour tout $S\in E$,

$$\begin{aligned}\mathrm{tr}(MS) - f^*(\lambda_1(S),\dots,\lambda_n(S)) &\leqslant \sum_{i=1}^n \lambda_i(M)\lambda_i(S) - f^*(\lambda_1(S),\dots,\lambda_n(S))\\ &\leqslant f^{**}(\lambda_1(M),\dots,\lambda_n(M)) = V_f(M).\end{aligned}$$

– *Conséquences de* (7.35)

(7.35) exprime que V_f est la conjuguée de V_{f^*} ; donc $V_f\in\Gamma_0(E)$.

Par ailleurs, f^* étant à son tour une fonction de $\Gamma_0(\mathbb{R}^n)$ symétrique, $V_{f^*}\in\Gamma_0(E)$. Il s'ensuit $(V_f)^* = (V_{f^*})^{**} = V_{f^*}$.

2°)

Fonctions de $\Gamma_0(\mathbb{R}^n)$ symétriques	Fonctions de valeurs propres V_f correspondantes
$f(x_1,...,x_n) = \max\{x_1,...,x_n\}$.	$V_f(M) =$ plus grande valeur propre de M.
$f(x_1,...,x_n) =$ somme des m plus grandes valeurs dans $x_1,...,x_n$.	$V_f(M) =$ somme des m plus grandes valeurs propres de M.
$f(x_1,...,x_n) = \max\{x_1,...,x_n\} - \min\{x_1,...,x_n\}$.	$V_f(M) = \lambda_1(M) - \lambda_n(M) = \max_{(i,j)} \mid \lambda_i(M) - \lambda_j(M) \mid$ (largeur du spectre de$M$).
$f(x_1,...,x_n) = \begin{cases} -\sum_{i=1}^n \ln x_i \text{ si} \\ x_i > 0 \text{ pour tout } i, \\ +\infty \text{ sinon.} \end{cases}$	$V_f(M) = \begin{cases} -\ln(\det M) \text{ si } M \text{ est} \\ \text{définie positive,} \\ +\infty \text{ sinon.} \end{cases}$
$f(x_1,...,x_n) = \begin{cases} -1/\left(\sum_{i=1}^n 1/x_i\right) \text{ si} \\ x_i > 0 \text{ pour tout } i, \\ +\infty \text{ sinon.} \end{cases}$	$V_f(M) = \begin{cases} -1/\mathrm{tr}(M^{-1}) \text{ si } M \text{ est} \\ \text{définie positive,} \\ +\infty \text{ sinon.} \end{cases}$

3°) a) Nous avons :

$$S \in \partial V_f(M) \Leftrightarrow \langle\langle S, M\rangle\rangle = V_f(M) + (V_f)^*(S)$$
$$= V_f(M) + V_{f^*}(S) \quad \text{(d'après la 1re question).}$$

Or :

- l'inégalité (de Fenchel)

$$f(\lambda_1(M),\ldots,\lambda_n(M)) + f^*(\lambda_1(S),\ldots,\lambda_n(S)) \geqslant \sum_{i=1}^n \lambda_i(M)\lambda_i(S)$$

est toujours assurée, et

- $\langle\langle S, M\rangle\rangle = \mathrm{tr}(SM) \leqslant \sum_{i=1}^n \lambda_i(S)\lambda_i(M)$.

Il s'ensuit :

$$S \in \partial V_f(M) \Leftrightarrow \begin{cases} f(\lambda_1(M),\dots,\lambda_n(M)) + f^*(\lambda_1(S),\dots,\lambda_n(S)) \\ \qquad\qquad = \sum_{i=1}^{n} \lambda_i(M)\lambda_i(S) \\ \text{et} \\ \operatorname{tr}(SM) = \sum_{i=1}^{n} \lambda_i(M)\lambda_i(S). \end{cases}$$

La 1re partie de l'assertion de droite exprime que

$$(\lambda_1(S),\dots,\lambda_n(S)) \in \partial f(\lambda_1(M),\dots,\lambda_n(M)) \ ;$$

quant à la 2e, elle n'a lieu que s'il existe U orthogonale telle que

$$U^\top MU = \operatorname{diag}(\lambda_1(M),\dots,\lambda_n(M)) \text{ et } U^\top SU = \operatorname{diag}(\lambda_1(S),\dots,\lambda_n(S)).$$

b) Si f est différentiable en $(\lambda_1(M),\dots,\lambda_n(M))$, alors

$$\partial f(\lambda_1(M),\dots,\lambda_n(M)) = \{\nabla f(\lambda_1(M),\dots,\lambda_n(M))\} \ ;$$

il s'ensuit avec la règle de calcul établie en (7.36) :

$$\partial V_f(M) = \big\{U[\operatorname{diag}\nabla f(\lambda_1(M),\dots,\lambda_n(M))]U^\top \mid U \text{ orthogonale telle que } U^\top MU = \operatorname{diag}(\lambda_1(M),\dots,\lambda_n(M))\big\}.$$

Montrons que cet ensemble ne contient qu'un seul élément, ce qui assurera la différentiabilité de V_f en M et l'expression annoncée (7.37) de $\nabla V_f(M)$.

Tous les éléments du convexe $\partial V_f(M)$ ont même norme ($\| S \|^2 = \operatorname{tr}(S^2) = \| \nabla f(\lambda_1(M),\dots,\lambda_n(M)) \|^2$ pour tout $S \in \partial V_f(M)$), et comme cette norme $\|\cdot\|$ sur E (déduite du produit scalaire $\langle\langle\cdot,\cdot\rangle\rangle$) est strictement convexe, *i.e.*

$$\begin{aligned} &(\| S_1 \| = \| S_2 \| = 1,\ S_1 \neq S_2,\ \alpha \in]0,1[) \\ &\Rightarrow (\| \alpha S_1 + (1-\alpha)S_2 \| < \alpha \| S_1 \| + (1-\alpha) \| S_2 \|), \end{aligned}$$

$\partial V_f(M)$ ne contient qu'un seul élément.

4°) La fonction f proposée donne lieu à

$$V_f : M \in E \longmapsto V_f(M) = \begin{cases} -\sum_{i=1}^{n} \ln(\lambda_i(M)) = -\ln(\det M) & \text{si } \lambda_i(M) > 0 \text{ pour tout } i, \\ & \text{c'est-à-dire si } M \text{ est définie positive,} \\ +\infty & \text{sinon.} \end{cases}$$

– *Détermination de la conjuguée*

Le calcul de f^* est aisé, de par la structure « décomposée » de f :

la conjuguée de $\varphi : x \longmapsto \begin{cases} -\ln x \text{ si } x > 0 \\ +\infty \text{ sinon} \end{cases}$ est

$$\varphi^* : s \longmapsto \begin{cases} -1 + \ln\left(-\frac{1}{s}\right) & \text{si } s < 0, \\ +\infty \text{ sinon}\,; \end{cases}$$

d'où

$$f^* : s = (s_1, \ldots, s_n) \longmapsto f^*(s) = \begin{cases} -n + \displaystyle\sum_{i=1}^{n} \ln\left(-\frac{1}{s_i}\right) & \text{si } s_i < 0 \text{ pour tout } i, \\ +\infty \text{ sinon.} \end{cases}$$

Il s'ensuit :

$$(V_f)^*(S) = V_{f^*}(S) = f^*(\lambda_1(S), \ldots, \lambda_n(S)) = \begin{cases} -n - \ln(-\det S) \text{ si } S \\ \text{est définie négative,} \\ +\infty \text{ sinon.} \end{cases}$$

– *Calcul différentiel*

f est différentiable en $(\lambda_1(M), \ldots, \lambda_n(M))$ lorsque $\lambda_i(M) > 0$ pour tout i, c'est-à-dire lorsque M est définie positive. Alors, la règle de calcul établie en (7.37) conduit à retrouver le résultat

$$\nabla V_f(M) = -M^{-1}.$$

Commentaire : – Les fonctions de valeurs propres V_f vérifient

$$V_f(M) = V_f(U^\top M U) \text{ quelle que soit } U \text{ orthogonale}\,; \tag{7.39}$$

il est intéressant de savoir que la réciproque est vraie : toute fonction convexe V sur E (vérifiant (7.39) est de la forme V_f.

– Pour les fonctions de valeurs propres du type V_f, le calcul de la conjuguée et du sous-différentiel reviennent à ceux – plus simples – du calcul de la conjuguée et du sous-différentiel de f.

– Le résultat (7.37) est quelque peu curieux : l'expression de $\nabla V_f(M)$ ne dépend pas de la matrice orthogonale U ayant servi à diagonaliser M.

***Exercice VII.21.** Soit C un convexe fermé (non vide) symétrique de $\mathbb{R}^n$ (*i.e.* tel que $(x_{\sigma(1)},\dots,x_{\sigma(n)}) \in C$ pour tout $(x_1,\dots,x_n) \in C$ et toute permutation σ de $\{1,\dots,n\}$). On définit :

$$\lambda^{-1}(C) := \{M \in \mathcal{S}_n(\mathbb{R}) \mid (\lambda_1(M),\dots,\lambda_n(M)) \in C\}, \tag{7.40}$$

où $\lambda_1(M) \geqslant \lambda_2(M) \geqslant \dots \geqslant \lambda_n(M)$ désignent les valeurs propres de M.

1°) Montrer que $\lambda^{-1}(C)$ est un convexe fermé de $\mathcal{S}_n(\mathbb{R})$.

2°) *Illustrations.* Déterminer $\lambda^{-1}(C)$ dans les cas suivants :

$$C = \{0\},\ C = \left(\mathbb{R}^+\right)^n,\ C = \Lambda_n \ \text{ (simplexe-unité de } \mathbb{R}^n).$$

Indication. Pour répondre à la 1re question, utiliser le résultat de la 1re question du Problème VII.20.

Solution : 1°) Soit $f := I_C$ (fonction indicatrice de C) ; de par les hypothèses faites sur C, $f \in \Gamma_0(\mathbb{R}^n)$ et est symétrique. Qu'est-ce qu'alors la fonction

$$V_f : M \in \mathcal{S}_n(\mathbb{R}) \longmapsto V_f(M) := f(\lambda_1(M),\dots,\lambda_n(M)) \ ?$$

Il vient immédiatement : $V_f = I_{\lambda^{-1}(C)}$ (fonction indicatrice de $\lambda^{-1}(C)$).

D'après le résultat de la 1re question du Problème VII.20, V_f est convexe semi-continue inférieurement sur $\mathcal{S}_n(\mathbb{R})$; par conséquent, $\lambda^{-1}(C)$ est un convexe fermé de $\mathcal{S}_n(\mathbb{R})$ (assurément non vide).

2°) Si $C = \{0\}$, seule la matrice nulle est dans $\lambda^{-1}(C)$.

Si $C = (\mathbb{R}^+)^n$, $\lambda^{-1}(C)$ est clairement $\mathcal{P}_n(\mathbb{R})$ (l'ensemble des matrices M de $\mathcal{S}_n(\mathbb{R})$ qui sont semi-définies positives).

Si $C = \Lambda_n$, $\lambda^{-1}(C) = \{M \in \mathcal{P}_n(\mathbb{R}) \mid \operatorname{tr} M = 1\}$ (ensemble noté Ω_1 dans l'Exercice 6.12).

Commentaire : – Les exemples traités dans l'exercice montrent que $\lambda^{-1}(C)$ n'est pas nécessairement polyédral lorsque C l'est.

– Prolongement de l'exercice : Montrer que M est un point extrémal de $\lambda^{-1}(C)$ si, et seulement si, $(\lambda_1(M),\dots,\lambda_n(M))$ est un point extrémal de C.

***Exercice VII.22.** Soit $f : \mathbb{R} \to \mathbb{R}$ dérivable et minorée par une fonction affine. Montrer que $\overline{\text{conv}}\, f$ est dérivable sur $\mathbb{R}$.

Solution : Comme $\text{conv}\, f$ est convexe et partout finie, il n'y a pas lieu de distinguer ici $\text{conv}\, f$ et $\overline{\text{conv}}\, f$ $(= f^{**})$. Cela dit, $\overline{\text{conv}}(\text{epi}\, f)$ peut être différent de $\text{conv}(\text{epi}\, f)$ (prendre $f(x) = \mathrm{e}^{-x^2}$ par exemple).

Supposons qu'il existe un point x_0 en lequel la fonction $g := \overline{\text{conv}}\, f$ n'est pas dérivable, c'est-à-dire en lequel la dérivée à gauche de g diffère de la dérivée à droite.

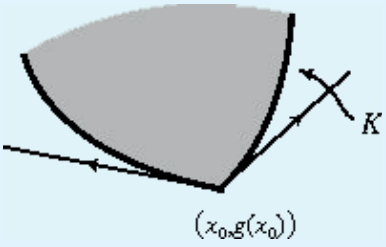


Figure 26.

Le cône K construit à partir des demi-dérivées de g en x_0 contient $\overline{\text{conv}}(\text{epi}\, f)$ et, de plus, $(x_0, g(x_0)) \in \overline{\text{conv}}(\text{epi}\, f)$. On peut donc approcher $(x_0, g(x_0))$ par une suite d'éléments de $\text{conv}(\text{epi}\, f)$, lesquels sont nécessairement dans K.

Il s'ensuit que $f(x_0) = g(x_0)$ (d'accord ?). Mais alors f ne serait pas dérivable en x_0, ce qui est contraire à l'hypothèse.

Commentaire : Ce résultat est particulier aux fonctions convexes de la variable réelle.

Considérons en effet $f : (x, y) \in \mathbb{R}^2 \longmapsto f(x, y) := \sqrt{x^2 + \mathrm{e}^{-y^2}}$; c'est une fonction de classe C^∞ sur $\mathbb{R}^2$ pour laquelle $(\overline{\text{conv}}\, f) : (x, y) \longmapsto (\overline{\text{conv}}\, f)(x, y) = |\, x\, |$.

***Exercice VII.23.** Soient f et $g : \mathbb{R} \to \mathbb{R}$ dérivables et minorées (chacune) par une fonction affine. On sait d'après l'exercice précédent que $\overline{\text{conv}}\, f$ et $\overline{\text{conv}}\, g$ sont également dérivables sur $\mathbb{R}$. La question à présent est : si $f' \leqslant g'$, a-t-on $(\overline{\text{conv}}\, f)' \leqslant (\overline{\text{conv}}\, g)'$?

Solution : La réponse est oui. Nous démontrons cela en plusieurs étapes.

Étape 1

Considérons $h : \mathbb{R} \to \mathbb{R}$ dérivable, minorée par une fonction affine, et $x_0 \in \mathbb{R}$. On suppose qu'il existe $a > (\overline{\text{conv}}\, h)'(x_0)$ et $b \in \mathbb{R}$ tels que

$$h(x) \geqslant a(x - x_0) + b \text{ pour } x \geqslant x_0.$$

Alors $b < (\overline{\text{conv}}\, h)(x_0)$ nécessairement.
En effet, si $b \geqslant (\overline{\text{conv}}\, h)(x_0)$, on aurait alors :

(i) $h(x) \geqslant a(x - x_0) + b \geqslant a(x - x_0) + (\overline{\text{conv}}\, h)(x_0)$ pour $x \geqslant x_0$ d'une part,

(ii) $h(x) \geqslant (\overline{\text{conv}}\, h)(x) \geqslant (\overline{\text{conv}}\, h)(x_0) + (\overline{\text{conv}}\, h)'(x_0)(x - x_0)$
$\geqslant (\overline{\text{conv}}\, h)(x_0) + a(x - x_0)$ si $x \leqslant x_0$ d'autre part.

En somme

$$h(x) \geqslant (\overline{\text{conv}}\, h)(x_0) + a(x - x_0) \text{ pour tout } x \in \mathbb{R},$$

et donc

$$(\overline{\text{conv}}\, h)(x) \geqslant (\overline{\text{conv}}\, h)(x_0) + a(x - x_0) \text{ pour tout } x \in \mathbb{R}.$$

Cette dernière inégalité implique $(\overline{\text{conv}}\, h)'(x_0) \geqslant a$ (en fait égalité), d'où contradiction avec l'hypothèse.

La même conclusion, à savoir $b < (\overline{\text{conv}}\, h)(x_0)$, est obtenue *mutatis mutandis* en supposant qu'il existe $a < (\overline{\text{conv}}\, h)'(x_0)$ et $b \in \mathbb{R}$ tels que

$$h(x) \geqslant a(x - x_0) + b \quad \text{pour} \quad x \leqslant x_0.$$

Étape 2

On ramène le problème à des fonctions coïncidant en un point $x_0 \in \mathbb{R}$.

Posons en effet $f_1(x) := f(x) - f(x_0)$ et $g_1(x) := g(x) - g(x_0)$. On a évidemment :

(1) $f_1' = f'$ et $g_1' = g'$;

(2) $(\overline{\text{conv}}\, f_1)' = (\overline{\text{conv}}\, f)'$ et $(\overline{\text{conv}}\, g_1)' = (\overline{\text{conv}}\ g)'$
(car $\overline{\text{conv}}\, f_1 = \overline{\text{conv}}\, f - f(x_0)$).

Étape 3

Sachant que $f_1' \leqslant g_1'$, montrons que $(\overline{\text{conv}}\, f_1)'(x_0) \leqslant (\overline{\text{conv}}\, g_1)'(x_0)$.

En premier lieu, notons :

– si $x \geqslant x_0$, $f_1(x) = \int_{x_0}^{x} f_1'(t)dt \leqslant \int_{x_0}^{x} g_1'(t)dt = g_1(x)$;

– si $x \leqslant x_0$, $g_1(x) \leqslant f_1(x)$.

Supposons que $(\overline{\text{conv}}\, f_1)'(x_0) > (\overline{\text{conv}}\, g_1)'(x_0)$ et arrivons à une contradiction. Nous avons :

– si $x \geqslant x_0$,

$$g_1(x) \geqslant f_1(x) \geqslant (\overline{\text{conv}}\, f_1)(x) \geqslant (\overline{\text{conv}}\, f_1)(x_0) + (\overline{\text{conv}}\, f_1)'(x_0)(x - x_0)$$

et, d'après le résultat de la 1re étape (appliqué à $h := g_1$, $a := (\overline{\text{conv}}\, f_1)'(x_0)$ et $b := (\overline{\text{conv}}\, f_1)(x_0)$), il vient que $(\overline{\text{conv}}\, f_1)(x_0) < (\overline{\text{conv}}\, g_1)(x_0)$.

– si $x \leqslant x_0$,

$$f_1(x) \geqslant g_1(x) \geqslant (\overline{\text{conv}}\, g_1)(x) \geqslant (\overline{\text{conv}}\, g_1)(x_0) + (\overline{\text{conv}}\, g_1)'(x_0)(x - x_0)$$

et, toujours d'après le résultat de la 1re étape (appliqué à $h := f_1$, $a := (\overline{\text{conv}}\, g_1)'(x_0)$ et $b := (\overline{\text{conv}}\, g_1)(x_0)$), il vient que $(\overline{\text{conv}}\, g_1)(x_0) < (\overline{\text{conv}}\, f_1)(x_0)$.

***Exercice VII.24.** Soient $f : \mathbb{R}^n \to \mathbb{R}$ convexe et $x_1, \ldots, x_k \in \mathbb{R}^n$. On suppose qu'il existe $\overline{\alpha}_1 > 0, \ldots, \overline{\alpha}_k > 0$, de somme 1, tels que $f\left(\sum_{i=1}^k \overline{\alpha}_i x_i\right) = \sum_{i=1}^k \overline{\alpha}_i f(x_i)$. Montrer que f est alors affine sur $\text{conv}\{x_1, \ldots, x_k\}$.

Solution : Soit $C := \text{conv}\{x_1, \ldots, x_k\}$. On veut montrer que s'il y a un point de l'intérieur relatif de C en lequel l'inégalité usuelle de convexité $f\left(\sum_{i=1}^k \overline{\alpha}_i x_i\right) \leqslant \sum_{i=1}^k \overline{\alpha}_i f(x_i)$ est une égalité, alors f est affine sur C.

Posons $\overline{x} := \sum_{i=1}^k \overline{\alpha}_i x_i$ et prenons $\overline{s} \in \partial f(\overline{x})$. La fonction $g : x \in \mathbb{R}^n \longmapsto g(x) := f(\overline{x}) + \langle s, x - \overline{x}\rangle$ est une minorante affine de f, coïncidant avec f en $\overline{x}$.

Nous disons que g coïncide avec f en tous les x_i, $i = 1, \ldots, k$. Si ce n'était pas le cas, nous aurions

$$g(\overline{x}) = \sum_{i=1}^k \overline{\alpha}_i g(x_i) < \sum_{i=1}^k \overline{\alpha}_i f(x_i) = f(\overline{x}),$$

d'où contradiction.

En conséquence, pour tout $x \in C$, disons $x = \sum_{i=1}^{k} \alpha_i x_i$,

$$
\begin{aligned}
f\left(\sum_{i=1}^{k} \alpha_i x_i\right) &\leqslant \sum_{i=1}^{k} \alpha_i f(x_i) \quad \text{(de part la convexité de } f\text{)} \\
&\leqslant \sum_{i=1}^{k} \alpha_i g(x_i) \quad \text{(puisque } f(x_i) = g(x_i) \text{ pour tout } i\text{)} \\
&\leqslant g\left(\sum_{i=1}^{k} \alpha_i x_i\right) \quad \text{(puisque } g \text{ est affine)} \\
&\leqslant f\left(\sum_{i=1}^{k} \alpha_i x_i\right) \quad \text{(puisque } g \text{ minore } f\text{)},
\end{aligned}
$$

d'où $f(x) = g(x)$.

****** Exercice VII.25.*** Soit C un convexe compact de $\mathbb{R}^n$, soit $f : C \to \mathbb{R}$ une fonction continue sur C et C^∞ sur $\overset{\circ}{C}$. On prolonge f à tout $\mathbb{R}^n$ en posant $f(x) = +\infty$ si $x \notin C$. La fonction $\overline{\text{conv}}\, f$ est-elle alors différentiable sur $\overset{\circ}{C}$?

Solution : Non. Voici un contre-exemple. Soit C le polyèdre convexe compact de $\mathbb{R}^2$ de sommets (1,0), (1,2), (0,3), (–1,2) et (–1,0) ; soit $f : (\xi_1, \xi_2) \in \mathbb{R}^2 \longmapsto f(\xi_1, \xi_2) := 1 - \xi_1^2$ si $(\xi_1, \xi_2) \in C$, $+\infty$ sinon. Alors $\overline{\text{conv}}\, f$ n'est pas différentiable sur le segment L joignant (1,2) à (–1,2).

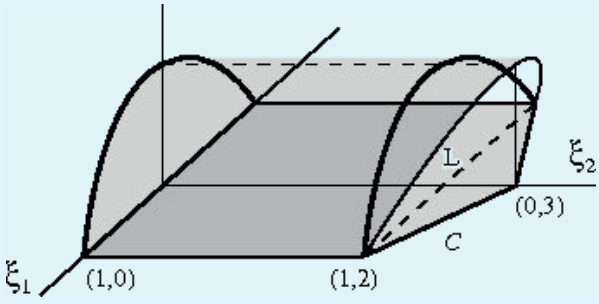


FIGURE 27.

Remarque : La fonction $\overline{\text{conv}}\, f$ construite à partir d'une fonction f comme dans l'exercice ci-dessus est continue sur $\overset{\circ}{C}$ (puisque $\overline{\text{conv}}\, f$ est convexe et C est le domaine de $\overline{\text{conv}}\, f$), mais il peut arriver qu'elle ne soit pas continue en des points de la frontière de C.

***Exercice VII.26.** S étant un fermé non vide de $\mathbb{R}^n$, on définit $f_S : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ par $f_S(x) := \frac{1}{2} \parallel x \parallel^2$ si $x \in S$, $+\infty$ sinon. On rappelle que la conjuguée f_S^* de f_S est $f_S^* : s \in \mathbb{R}^n \longmapsto f_S^*(s) = \frac{1}{2}\left[\parallel s \parallel^2 - d_S^2(s)\right]$.

Rappeler ce que sont $\partial f_S(x)$ et $(\text{conv}\, f_S)(x)$ pour $x \in S$, ainsi que $\partial f_S^*(s)$.

Illustrer les formules précédentes sur les exemples suivants :

$$S = \{x \in \mathbb{R} \mid 0 \leqslant x \leqslant 1 \text{ ou } 2 \leqslant x \leqslant 3\},$$

$$S = \big\{x = (\xi_1, \xi_2) \in \mathbb{R}^2 : \xi_1^2 + \xi_2^2 - 2 \mid \xi_1 \mid -2 \mid \xi_2 \mid +1 \geqslant 0 \text{ et } \mid \xi_1 \mid + \mid \xi_2 \mid \leqslant 1\big\}.$$

Solution : On note $P_S(x) := \{y \in S \mid d_S(x) = \parallel x - y \parallel\}$.

Considérons $x \in S$:

(i) $(s \in \partial f_S(x)) \Leftrightarrow (x \in P_S(s)) \Leftrightarrow (d_S(s) = \parallel s - x \parallel)$.

En particulier $x \in \partial f_S(x)$.

Les équivalences annoncées se voient grâce à la caractérisation de $\partial f_S(x)$ *via* f_S^* : $s \in \partial f_S(x)$ si, et seulement si, $f_S(x) + f_S^*(s) = \langle s, x\rangle$, ce qui revient ici à

$$I_S(x) + \frac{1}{2} \parallel x \parallel^2 + \frac{1}{2}\left[\parallel s \parallel^2 - d_S^2(s)\right] = \langle s, x\rangle.$$

Sachant que $I_S(x) = 0$, ceci équivaut à $d_S(s) = \parallel s - x \parallel$, *i.e.* $x \in P_S(s)$.

(ii) $\text{conv}\, f_S$ est une fonction convexe de domaine $\text{conv}\, S$.

Comme $\partial f_S(x) \neq \phi$ lorsque $x \in S$, $\text{conv}\, f_S$ et f_S coïncident en x ; de plus $\partial(\text{conv}\, f_S)(x) = \partial f_S(x)$.

(iii) $\partial f_S^*(s) = \text{conv}\, P_S(s)$.

En effet, $(\overline{\text{conv}}\, f_S =)\ \text{conv} f_S \in \Gamma_0(\mathbb{R}^n)$ de sorte que $x \in \partial f_S^*(s)$ équivaut à $s \in \partial(\ \text{conv}\, f_S)(x)$. Or, on est ici dans les conditions où on peut exprimer $\partial(\text{conv}\, f_S)(x)$ en fonction de ∂f_S (voir rappels de la Section VII.3) : il existe $x_1, \ldots, x_{n+1}$ dans S $(= \text{dom}\, f_S)$, $(\alpha_1, \ldots, \alpha_{n+1}) \in \Lambda_{n+1}$ tels que

$$x = \sum_{i=1}^{n+1} \alpha_i x_i \quad \text{et} \quad \partial(\text{conv}\, f_S)(x) = \bigcap_{\alpha_i > 0} \partial f_S(x_i).$$

Connaissant l'évaluation de $\partial f_S(x_i)$ lorsque $x_i \in S$ (*cf.* (*ii*)), on a bien traduit « $x \in \partial f_S^*(s)$ » en « $x \in \operatorname{conv} P_S(s)$ ».

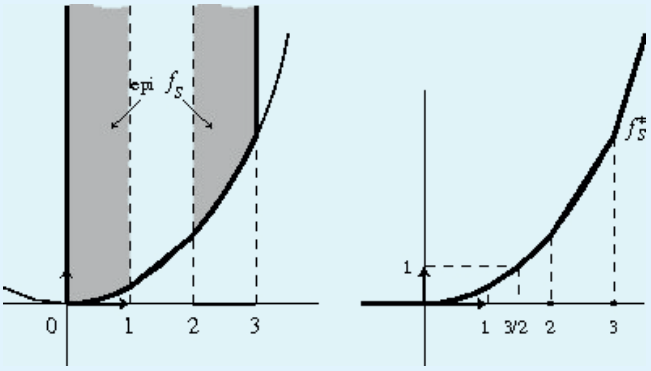


FIGURE 28.

$\partial f_S(1) = \partial(\operatorname{conv} f_s)(1) = \left[1, \frac{3}{2}\right] = \{s \text{ se projetant sur } S \text{ en } 1\}$.
En $x = \frac{3}{2}$, $P_S(x) = \{1, 2\}$ et $\partial f_S^*(x) = [1, 2]$.
Dans le 2e exemple, l'ensemble S de $\mathbb{R}^2$ est comme suit :

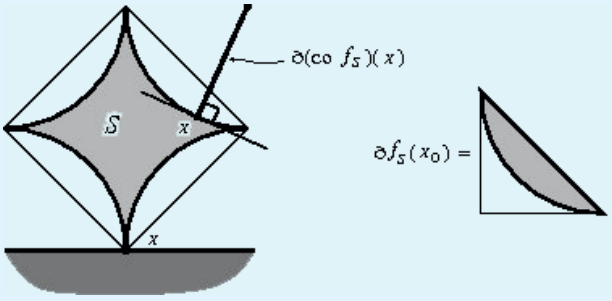


FIGURE 29.

f_S^* n'est pas différentiable en $x_0 = (1, 1)$.

Remarque : Dans le cadre plus général où S est un fermé non vide d'un espace de Hilbert, on montre facilement l'inclusion $P_S(x) \subset \partial f_S^*(x), x \in H$, de sorte que $\overline{\operatorname{conv}}\, P_S(x) \subset \partial f_S^*(x)$. Mais lorsque H est de dimension infinie, cette inclusion peut être stricte. Toutefois, bien des propriétés de la multi-application $x \longmapsto P_S(x)$, la monotonie par exemple, se déduisent des propriétés de $x \longmapsto \partial f_S^*(x)$ grâce à cette inclusion.

**** Problème VII.27.** *Minimisation de différences de fonctions convexes*

Soient g et h deux fonctions convexes de $\mathbb{R}^n$ dans $\mathbb{R}$ et on considère le problème d'optimisation suivant :

$$(\mathcal{P}) \qquad \text{Minimiser } f(x) := g(x) - h(x) \text{ sur } \mathbb{R}^n.$$

1°) Comment trouver une décomposition de f sous la forme $f = \tilde{g} - \tilde{h}$, où $\tilde{g}$ et $\tilde{h}$ sont toutes les deux fortement convexes ?

2°) Pour cette question et la suivante, on fait l'hypothèse $(\mathcal{H}_1)$ ci-dessous :

$$(\mathcal{H}_1) \qquad \{x \in \mathbb{R}^n : g(x) - h(x) \leqslant r\} \text{ est borné pour tout } r \in \mathbb{R}.$$

a) Indiquer brièvement pourquoi le problème $(\mathcal{P})$ a au moins une solution. Cette solution est-elle unique ?

b) On dit que $\overline{x}$ est un point T-critique de $(\mathcal{P})$ lorsque $\partial g(\overline{x}) \cap \partial h(\overline{x}) \neq \emptyset$. Montrer que toute solution de $(\mathcal{P})$ est un point T-critique de $(\mathcal{P})$.

3°) Dans cette question, on suppose de plus

$$(\mathcal{H}_2) \qquad g \text{ et } h \text{ sont fortement convexes sur } \mathbb{R}^n.$$

On considère alors l'algorithme décrit comme suit :

$\underline{k = 0}$: $x_0 \in \mathbb{R}^n$;

$\underline{k \to k+1}$: on prend un élément quelconque s_k de $\partial h(x_k)$ et on choisit x_{k+1} de sorte que $s_k \in \partial g(x_{k+1})$.

a) Montrer que la suite $\{f(x_k) = g(x_k) - h(x_k)\}_k$ est décroissante.

b) Montrer que les suites $\{x_k\}$ et $\{s_k\}$ sont bornées.

c) Montrer que $\displaystyle\sum_{k=0}^{+\infty} \| x_{k+1} - x_k \|^2 < +\infty$.

d) Montrer que si $\tilde{x}$ est limite d'une suite extraite de la suite $\{x_k\}$, alors $\tilde{x}$ est un point T-critique de $(\mathcal{P})$.

e) Soit $\theta : t \longmapsto \theta(t) := f[tx_{k+1} + (1-t)x_k]$. En supposant $x_{k+1} \neq x_k$, montrer que θ est strictement décroissante sur [0,1]. En déduire que $f(x_{k+1}) < f(x_k)$.

4°) On considère le problème $(\mathcal{Q})$ suivant (dans $\mathbb{R}^{n+1}$) :

$$(\mathcal{Q}) \qquad \begin{cases} \text{Minimiser } g(x) - y \\ \text{sous la contrainte : } (x,y) \in \mathbb{R}^n \times \mathbb{R},\ h(x) \geqslant y. \end{cases}$$

Établir les relations existant entre les solutions et valeurs optimales de $(\mathcal{P})$ et celles de $(\mathcal{Q})$.

Solution : 1°) La décomposition de f comme différence de fonctions convexes n'est évidemment pas unique ; il suffit d'ajouter à g et h une fonction convexe φ pour avoir une nouvelle décomposition : $f = (g + \varphi) - (h + \varphi)$. Si φ est fortement convexe, par exemple $\varphi : x \longmapsto \varphi(x) = \| x \|^2$, il s'ensuivra que $\tilde{g} := g + \varphi$ et $\tilde{h} := h + \varphi$ seront fortement convexes.

2°) a) La fonction f est continue (comme différence de fonctions convexes sur $\mathbb{R}^n$, donc continues). Soit $r_0 \in \mathbb{R}$ tel que

$$S_{r_0}(f) := \{x \in \mathbb{R}^n \mid f(x) \leqslant r_0\} \neq \emptyset.$$

$S_{r_0}(f)$ est compact et, bien entendu, minimiser f sur $\mathbb{R}^n$ revient à minimiser f sur $S_{r_0}(f)$. Le problème $(\mathcal{P})$ a donc une solution ; mais il n'y a aucune raison pour que cette solution soit unique.

b) Soient $\overline{x}$ une solution de $(\mathcal{P})$ et $s \in \partial h(\overline{x})$. On a :

$h(x) \geqslant h(\overline{x}) + \langle s, x - \overline{x}\rangle$ pour tout $x \in \mathbb{R}^n$ (puisque $s \in \partial h(\overline{x})$);

$g(x) - h(x) \geqslant g(\overline{x}) - h(\overline{x})$ pour tout $x \in \mathbb{R}^n$(puisque $\overline{x}$ minimise $f = g - h$ sur $\mathbb{R}^n$).

Par suite,

$$g(x) \geqslant g(\overline{x}) + \langle s, x - \overline{x}\rangle \text{ pour tout } x \in \mathbb{R}^n,$$

i.e., $s \in \partial g(\overline{x})$.

Ainsi on a montré : $(\phi \neq)\ \partial h(\overline{x}) \subset \partial g(\overline{x})$, et, *a fortiori*, $\partial g(\overline{x}) \cap \partial h(\overline{x}) \neq \phi$.

Remarques :

1. Il suffit que $\overline{x}$ soit minimum *local* de f pour avoir $\partial h(\overline{x}) \subset \partial g(\overline{x})$.

2. Si $\overline{x}$ avait été un maximum local de f, on aurait eu $\partial g(\overline{x}) \subset \partial h(\overline{x})$, soit encore $\partial g(\overline{x}) \cap \partial h(\overline{x}) \neq \phi$.

Autre démonstration de l'implication proposée. Supposons $\partial g(\overline{x}) \cap \partial h(\overline{x}) = \phi$ et montrons que cela conduit à une contradiction.

Avoir $\partial g(\overline{x}) \cap \partial h(\overline{x}) = \emptyset$ revient à avoir $0 \notin \partial g(\overline{x}) - \partial h(\overline{x})$. Ceci signifie exactement :

$$\exists\, d \in \mathbb{R}^n,\ \sigma_{\partial g(\overline{x})}(d) + \sigma_{-\partial h(\overline{x})}(d) < 0 \text{ (d'accord ?)}.$$

Or

$$\sigma_{-\partial h(\overline{x})}(d) = \sigma_{\partial h(\overline{x})}(-d) = h'(\overline{x}, -d) \geqslant -h'(\overline{x}, d).$$

Donc il existe $d \in \mathbb{R}^n$ tel que $g'(\overline{x}, d) - h'(\overline{x}, d) = f'(\overline{x}, d) < 0$, ce qui contredit la condition $(f'(\overline{x}, d) \geqslant 0$ pour tout $d \in \mathbb{R}^n)$ satisfaite en tout minimum (même local) de f.

3°) Choisir x_{k+1} de sorte que $s_k \in \partial g(x_{k+1})$ revient à choisir x_{k+1} minimisant $x \longmapsto g(x) - \langle s_k, x\rangle$ sur $\mathbb{R}^n$.

a) Puisque $s_k \in \partial g(x_{k+1})$ et que g est fortement convexe,

$$g(x_k) \geqslant g(x_{k+1}) + \langle s_k, x_k - x_{k+1}\rangle + \frac{c}{2} \parallel x_k - x_{k+1} \parallel^2 \qquad (7.41)$$

(c étant un module de forte convexité de g sur $\mathbb{R}^n$).

De même, puisque $s_k \in \partial h(x_k)$ et h est fortement convexe,

$$h(x_{k+1}) \geqslant h(x_k) + \langle s_k, x_{k+1} - x_k\rangle + \frac{d}{2} \parallel x_k - x_{k+1} \parallel^2 \qquad (7.42)$$

(d étant un module de forte convexité de h).

En additionnant (7.41) et (7.42), on obtient :

$$g(x_k) - h(x_k) \geqslant g(x_{k+1}) - h(x_{k+1}) + \frac{c+d}{2} \parallel x_k - x_{k+1} \parallel^2 . \qquad (7.43)$$

La suite $\{f(x_k) = g(x_k) - h(x_k)\}_k$ est bien décroissante. Elle est minorée par $\overline{f} := \inf_{x\in\mathbb{R}^n} f(x)$.

b) Puisque $f(x_k) \leqslant f(x_0)$ pour tout k et que $\{x \mid f(x) \leqslant f(x_0)\}$ est borné par hypothèse, la suite $\{x_k\}$ est bornée.

Comme l'image d'un borné par ∂h est un borné et que $s_k \in \partial h(x_k)$, la suite $\{s_k\}$ est également bornée.

c) On tire de (7.43) :

$$\left(\frac{c+d}{2}\right) \sum_{k=0}^{K-1} \parallel x_{k+1} - x_k \parallel^2 \leqslant f(x_0) - \overline{f} \text{ pour tout } K \in \mathbb{N}^*,$$

d'où $\sum\limits_{k=0}^{+\infty} \parallel x_{k+1} - x_k \parallel^2 < +\infty$.

d) Considérons $\{x_{k_l}\}_l$ telle que $x_{k_l} \to \tilde{x}$ quand $l \to +\infty$. Puisque la suite $\{s_{k_l}\}_l$ est bornée, on peut supposer – quitte à extraire une nouvelle sous-suite – que $s_{k_l} \to \tilde{s}$ quand $l \to +\infty$.

En raison du caractère fermé du graphe de la multi-application $x \longmapsto \partial h(x)$, un passage à la limite (sur l) de $s_{k_l} \in \partial h(x_{k_l})$ conduit à $\tilde{s} \in \partial h(\tilde{x})$.

Mais $x_{k_l+1} \underset{l\to+\infty}{\to} \tilde{x}$ également car $x_{k_l+1} - x_{k_l} \to 0$ quand $l \to +\infty$.

Donc, comme précédemment, un passage à la limite sur $s_{k_l} \in \partial g(x_{k_l+1})$ conduit à $\tilde{s} \in \partial g(\tilde{x})$.

En définitive, $\tilde{s} \in \partial h(\tilde{x}) \cap \partial g(\tilde{x})$; $\tilde{x}$ est bien un point T-critique de $(\mathcal{P})$.

e) $\theta(t) = (g-h)[tx_{k+1} + (1-t)x_k]$ par définition. Posons

$\tilde{x}_1 := t_1 x_{k+1} + (1-t_1)x_k$, $\tilde{x}_2 := t_2 x_{k+1} + (1-t_2)x_k$, où $t_1 < t_2$ et $t_1, t_2 \in [0,1]$, et choisissons $\tilde{s}_1 \in \partial h(\tilde{x}_1)$, $\tilde{s}_2 \in \partial h(\tilde{x}_2)$.

Il s'ensuit :

$$g(\tilde{x}_1) \geqslant g(\tilde{x}_2) + \langle \tilde{s}_2, \tilde{x}_1 - \tilde{x}_2 \rangle + \frac{c}{2} \parallel \tilde{x}_1 - \tilde{x}_2 \parallel^2,$$

$$h(\tilde{x}_2) \geqslant h(\tilde{x}_1) + \langle \tilde{s}_1, \tilde{x}_2 - \tilde{x}_1 \rangle + \frac{d}{2} \parallel \tilde{x}_1 - \tilde{x}_2 \parallel^2,$$

d'où, puisque $\parallel \tilde{x}_1 - \tilde{x}_2 \parallel = (t_2 - t_1) \parallel \tilde{x}_k - \tilde{x}_{k+1} \parallel > 0$,

$$\theta(t_1) - \theta(t_2) = g(\tilde{x}_1) - h(\tilde{x}_1) - g(\tilde{x}_2) + h(\tilde{x}_2) > \langle \tilde{s}_2 - \tilde{s}_1, \tilde{x}_1 - \tilde{x}_2 \rangle . \quad (7.44)$$

Mais

$$\langle \tilde{s}_2 - \tilde{s}_1, \tilde{x}_1 - \tilde{x}_2 \rangle = \langle s_k - \tilde{s}_2, \tilde{x}_2 - \tilde{x}_1 \rangle + \langle \tilde{s}_1 - s_k, \tilde{x}_2 - \tilde{x}_1 \rangle \quad (7.45)$$

et

$$\tilde{x}_2 - \tilde{x}_1 = (t_2 - t_1)(x_{k+1} - x_k),$$
$$x_{k+1} - \tilde{x}_2 = (1 - t_2)(x_{k+1} - x_k),$$

d'où :

$$\tilde{x}_2 - \tilde{x}_1 = \frac{t_2 - t_1}{1 - t_2}(x_{k+1} - \tilde{x}_2),$$

$$\tilde{x}_1 - x_k = t_1(x_{k+1} - x_k), \quad \text{d'où} \quad \tilde{x}_2 - \tilde{x}_1 = \frac{t_2 - t_1}{t_1}(\tilde{x}_1 - x_k).$$

Il vient alors de (7.45) :

$$\langle \tilde{s}_2 - \tilde{s}_1, \tilde{x}_1 - \tilde{x}_2 \rangle = \frac{t_2 - t_1}{1 - t_2} \langle s_k - \tilde{s}_2, x_{k+1} - \tilde{x}_2 \rangle + \frac{t_2 - t_1}{t_1} \langle \tilde{s}_1 - s_k, \tilde{x}_1 - x_k \rangle .$$

Or

$$(s_k \in \partial g(x_{k+1}),\ \tilde{s}_2 \in \partial g(\tilde{x}_2)) \Rightarrow (\langle s_k - \tilde{s}_2, x_{k+1} - \tilde{x}_2 \rangle \geqslant 0),$$
$$(s_k \in \partial h(x_k),\ \tilde{s}_1 \in \partial h(\tilde{x}_1)) \Rightarrow (\langle \tilde{s}_1 - s_k, \tilde{x}_1 - x_k \rangle \geqslant 0),$$

d'où $\langle \tilde{s}_2 - \tilde{s}_1, \tilde{x}_1 - \tilde{x}_2 \rangle \geqslant 0$, et (7.44) permet de conclure à $\theta(t_2) < \theta(t_1)$.

En particulier $\theta(1) < \theta(0)$, soit $f(x_{k+1}) < f(x_k)$.

4°) $(\mathcal{Q})$ est équivalent à $(\mathcal{P})$ au sens suivant :
– si $(\overline{x}, \overline{y})$ est solution de $(\mathcal{Q})$, alors $\overline{x}$ est solution de $(\mathcal{P})$ et $\overline{y} = h(\overline{x})$;
– si $\overline{x}$ est solution de $(\mathcal{P})$, alors $(\overline{x}, h(\overline{x}))$ est solution de $(\mathcal{Q})$.

Commentaire :

– Un problème analogue à $(\mathcal{Q})$ est le problème $(\mathcal{R})$ suivant (toujours dans $\mathbb{R}^{n+1}$) :

$$(\mathcal{R}) \qquad \begin{cases} \text{Maximiser } h(x) - y \\ \text{sous la contrainte : } (x,y) \in \mathbb{R}^n \times \mathbb{R},\ g(x) \leqslant y. \end{cases}$$

Alors, $(\mathcal{R})$ est équivalent à $(\mathcal{P})$ au sens suivant :

– si $(\overline{x},\overline{y})$ est solution de $(\mathcal{R})$, alors $\overline{x}$ est solution de $(\mathcal{P})$ et $\overline{y} = g(\overline{x})$;

– si $\overline{x}$ est solution de $(\mathcal{P})$, alors $(\overline{x}, g(\overline{x}))$ est solution de $(\mathcal{R})$.

$(\mathcal{Q})$ et $(\mathcal{R})$ ressemblent à des problèmes de minimisation convexe, mais l'une des convexités (dans la définition de la contrainte ou dans la fonction-objectif) est « à rebours ».

— L'association de $f^\diamondsuit := h^* - g^*$ à $f = g - h$ se trouve fort productive quant à la comparaison de points T-critiques, de valeurs optimales, etc. ; en plus, différentes décompositions de f en $g - h$ donnent lieu à différents $f^\diamondsuit$.

***** Exercice VII.28.** Soit $f \in \Gamma_0(\mathbb{R}^n)$, soit $\theta \in \Gamma_0(\mathbb{R}^n)$ vérifiant les hypothèses suivantes :

$$\theta \text{ est finie en } 0\,; \quad \lim_{\|x\|\to+\infty} \frac{\theta(x)}{\|x\|} = +\infty. \tag{7.46}$$

On se propose dans cet exercice de donner des formes explicites de solutions d'équations aux dérivées partielles (dites de *Hamilton-Jacobi*) suivantes :

$$\begin{cases} \dfrac{\partial F}{\partial t} + \theta^*(\nabla_x F) = 0 \text{ sur } \mathbb{R}^n \times\,]0,+\infty[, \\ F(x,0) = f(x) \text{ pour tout } x \in \mathbb{R}^n. \end{cases}$$

(Dans cette écriture, $\partial F/\partial t$ et $\nabla_x F$ désignent respectivement la dérivée partielle par rapport à t et le vecteur gradient par rapport à $x = (x_1,\dots,x_n)$ de la fonction $F : (x,t) \in \mathbb{R}^n \times\,]0,+\infty[\longmapsto F(x,t) \in \mathbb{R}$.)

1°) *Préliminaires.* a) Soit

$$H : (y,s) \in \mathbb{R}^n \times \mathbb{R} \longmapsto H(y,s) := \begin{cases} f^*(y) \text{ si } \theta^*(y) + s \leqslant 0, \\ +\infty \text{ sinon.} \end{cases}$$

Vérifier que $H \in \Gamma_0(\mathbb{R}^n \times \mathbb{R})$ et déterminer sa conjuguée H^*.

b) Posons

$$F : (x,t) \in \mathbb{R}^n \times \mathbb{R} \longmapsto F(x,t) := \begin{cases} \inf_{u\in\mathbb{R}^n} \left\{ f(u) + t\theta\left(\frac{x-u}{t}\right)\right\} \text{ si } t > 0, \\ f(x) \text{ si } t = 0, \\ +\infty \text{ si } t < 0. \end{cases}$$

Vérifier que, pour tout $x \in \mathbb{R}^n$,

$$F(x,t) \to f(x) \text{ quand } t \to 0^+. \tag{7.47}$$

2°) Outre les hypothèses de (7.46), on suppose :

$$\theta \text{ est différentiable en tout point où elle admet des sous-gradients.} \tag{7.48}$$

Montrer qu'alors F est différentiable en tout point (x,t) de $\operatorname{int}(\operatorname{dom} \mathrm{F})$ $(\subset \mathbb{R}^n \times]0,+\infty[)$ et y vérifie

$$\frac{\partial F}{\partial t}(x,t) + \theta^*(\nabla_x F(x,t)) = 0.$$

3°) *Exemples.* Donner des formes explicites de solutions d'équations de Hamilton-Jacobi suivantes :

$$\begin{cases} \dfrac{\partial F}{\partial t} + \dfrac{1}{2}\|\nabla_x F\|^2 = 0, \\ \lim\limits_{t\to 0^+} F(\cdot,t) = F(\cdot,0) = f. \end{cases} \tag{7.49}$$

$$\begin{cases} \dfrac{\partial F}{\partial t} + \sqrt{1 + \|\nabla_x F\|^2} = 0, \\ \lim_{t\to 0^+} F(\cdot,t) = F(\cdot,0) = f. \end{cases} \tag{7.50}$$

(Dans ces équations, $\|\cdot\|$ désigne la norme euclidienne usuelle de $\mathbb{R}^n$.)

Solution : 1°) a) Décomposons H en somme de deux fonctions H_1 et H_2 plus maniables ; soient

$$H_1 : (y,s) \in \mathbb{R}^n \times \mathbb{R} \longmapsto H_1(y,s) := f^*(y)$$
$$H_2 : (y,s) \in \mathbb{R}^n \times \mathbb{R} \longmapsto H_2(y,s) := I_{\mathbb{R}^-}(\theta^*(y) + s).$$

H_1 ne dépend que de y, $\operatorname{dom} H_1 = (\operatorname{dom} f^*) \times \mathbb{R}$, et puisque $f^* \in \Gamma_0(\mathbb{R}^n)$ il va de soi que $H_1 \in \Gamma_0(\mathbb{R}^n \times \mathbb{R})$.

H_2 est la composée de la fonction $r \longmapsto I_{\mathbb{R}^-}(r)$ qui est dans $\Gamma_0(\mathbb{R})$ et croissante, avec la fonction $(y,s) \longmapsto \theta^*(y)+s$ qui est visiblement dans $\Gamma_0(\mathbb{R}^n \times \mathbb{R})$; il s'ensuit aisément que $H_2 \in \Gamma_0(\mathbb{R}^n \times \mathbb{R})$. De plus, pour tout $y \in \operatorname{dom} f^*$, le couple $(y, -\theta^*(y))$ est à la fois dans le domaine de H_1 et dans celui de H_2. Pas de doute : $H = H_1 + H_2$ est bien une fonction de $\Gamma_0(\mathbb{R}^n \times \mathbb{R})$.

Soit $(x,t) \in \mathbb{R}^n \times \mathbb{R}$ et calculons

$$H^*(x,t) = \sup_{\begin{cases} y \in \mathbb{R}^n \\ \theta^*(y) + s \leqslant 0 \end{cases}} \{\langle y, x\rangle + st - f^*(y)\}.$$

Par « découplage » de l'opération qui consiste à prendre le supremum, nous avons

$$H^*(x,t) = \sup_{y \in \mathbb{R}^n} \sup_{\theta^*(y) \leqslant -s} \{\langle y, x\rangle + st - f^*(y)\}. \tag{7.51}$$

Il s'ensuit :

— Si $t < 0$, $H^*(x,t) = +\infty$ (pour voir cela, faire $s \to -\infty$ dans (7.51)).

— Si $t = 0$, $H^*(x,t) = \sup_{y \in \mathbb{R}^n} \{\langle y, x\rangle - f^*(y)\}$

$$= f^{**}(x) = f(x) \text{ (car } f \in \Gamma_0(\mathbb{R}^n)).$$

— Si $t > 0$, $H^*(x,t) = \sup_{y \in \mathbb{R}^n} \{\langle y, x\rangle - t\theta^*(y) - f^*(y)\} = (f^* + t\theta^*)^*(x)$.

Mais la fonction θ ayant été supposée 1-coercive sur $\mathbb{R}^n$ (c'est la deuxième partie de l'hypothèse (7.46)), sa conjuguée θ^* est partout finie. Nous sommes donc (largement) dans les hypothèses assurant que

$$(f^* + t\theta^*)^* = (f^*)^* \square\, (t\theta^*)^*. \tag{7.52}$$

Or $f^{**} = f$ et $(t\theta^*)^* : x \in \mathbb{R}^n \longmapsto t\theta^{**}\left(\frac{x}{t}\right) = t\theta\left(\frac{x}{t}\right)$. En conclusion :

— Si $t > 0$, $H^*(x,t) = [f \square\, t\theta(\frac{\cdot}{t})](x)$.

b) Comme cela a été observé en (7.52), on a pour tout $t > 0$:

$$F(\cdot, t) = f \square\, t\theta\left(\frac{\cdot}{t}\right) = (f^* + t\theta^*)^*,$$

c'est-à-dire :

$$F(x,t) = \sup_{y \in \mathbb{R}^n} \{\langle x, y\rangle - f^*(y) - t\theta^*(y)\} \text{ pour tout } x \in \mathbb{R}^n.$$

Il s'ensuit :

$$\begin{aligned} F(x,t) &\leqslant \sup_{y \in \mathbb{R}^n} \left\{\langle x, y\rangle - f^*(y) - t \inf_{y \in \mathbb{R}^n} \theta^*(y)\right\} \\ &\leqslant f(x) + t\theta(0) \text{ (car } f^{**} = f \text{ et } \inf_{y \in \mathbb{R}^n} \theta^*(y) = -\theta^{**}(0) = -\theta(0)) ; \end{aligned}$$

d'où : $\limsup_{t \to 0^+} F(x,t) \leqslant f(x)$.

D'autre part, on a pour un y quelconque dans $\mathbb{R}^n$:

$$F(x,t) \geqslant \langle x, y\rangle - f^*(y) - t\theta^*(y),$$

d'où : $\liminf_{t \to 0^+} F(x,t) \geqslant \langle x, y\rangle - f^*(y)$.

Par suite, $\liminf_{t\to 0^+} F(x,t) \geqslant \sup_{y\in\mathbb{R}^n}\{\langle x,y\rangle - f^*(y)\} = f(x)$.
En conclusion :

$$\lim_{t\to 0^+} F(t,x) = f(x) \text{ pour tout } x\in\mathbb{R}^n.$$

2°) Reprenons la décomposition $H = H_1 + H_2$ du 1° a) : $H_1 \in \Gamma_0(\mathbb{R}^n\times\mathbb{R})$ et $\operatorname{dom} H_1 = (\operatorname{dom} f^*)\times\mathbb{R}$, tandis que $H_2 \in \Gamma_0(\mathbb{R}^n\times\mathbb{R})$ et $\operatorname{dom} H_2 = \{(y,s)\in\mathbb{R}^n\times\mathbb{R} \mid \theta^*(y)+s\leqslant 0\}$. Puisque θ^* est partout finie sur $\mathbb{R}^n$, les intérieurs relatifs de $\operatorname{dom} H_1$ et de $\operatorname{dom} H_2$ se coupent ; cela est suffisant pour garantir $(H_1+H_2)^* = H_1^*\square H_2^*$ et l'exactitude de cette dernière inf-convolution. Ainsi, pour $(x,t)\in\operatorname{dom} F$, il existe $(z,u)\in\mathbb{R}^n\times\mathbb{R}$ tel que

$$F(x,t) = H_1^*(x-z,t-u) + H_2^*(z,u)\ ; \tag{7.53}$$

de plus

$$\partial F(x,t) = \partial H_1^*(x-z,t-u)\cap\partial H_2^*(z,u). \tag{7.54}$$

Explicitons ces résultats (7.53) et (7.54) . On a par de simples calculs :

$$H_1^*(w,r) = f(w) \ \text{ si } r=0,\ +\infty \ \text{ sinon}\,; \tag{7.55}$$

$$H_2^*(w,r) = r\theta\left(\frac{w}{r}\right) \ \text{ si } r>0. \tag{7.56}$$

On en déduit que $u=t$ dans (7.53) et

$$\partial F(x,t) = \partial H_1^*(x-z,0)\cap\partial H_2^*(z,t). \tag{7.57}$$

Précisons les choses en prenant $(x,t)\in\operatorname{int}(\operatorname{dom} F)$ $(\subset\mathbb{R}^n\times\,]0,+\infty[)$:

— La fonction convexe F est continue en (x,t), donc $\partial F(x,t)\neq\emptyset$; il vient alors avec (7.57) que $\partial H_2^*(z,t)\neq\emptyset$.

— De par la relation (7.56) liant H_2^* à θ, et sachant que θ est différentiable en tout point où elle admet des sous-gradients (c'est l'hypothèse (7.48)), on a :

$$\partial H_2^*(z,t) = \left\{\nabla\theta\left(\frac{z}{t}\right)\right\}\times\left\{\theta\left(\frac{z}{t}\right) - \frac{1}{t}\left\langle\nabla\theta\left(\frac{z}{t}\right),z\right\rangle\right\}.$$

— De par la relation (7.55), on a toujours :

$$\partial H_1^*(x-z,0) = \partial f(x-z)\times\mathbb{R}.$$

Ces informations permettent de conclure avec (7.57) :

— que $\partial F(x,t)$ est réduit à un seul élément, c'est-à-dire F est différentiable en (x,t) ;

— que $\nabla_x F(x,t) = \nabla\theta\left(\frac{z}{t}\right)$ et

$$\frac{\partial F}{\partial t}(x,t) = \theta\left(\frac{z}{t}\right) - \frac{1}{t}\left\langle\nabla\theta\left(\frac{z}{t}\right),z\right\rangle.$$

Dans cette dernière relation, observons que

$$\theta\left(\frac{z}{t}\right) - \frac{1}{t}\left\langle \nabla\theta\left(\frac{z}{t}\right), z\right\rangle = -\theta^*\left(\nabla\theta\left(\frac{z}{t}\right)\right),$$

d'où finalement : $\dfrac{\partial F}{\partial t}(x,t) = -\theta^*(\nabla_x F(x,t))$.

3°) Avec $\theta^* = \frac{1}{2}\|\cdot\|^2$, on a $\theta = \frac{1}{2}\|\cdot\|^2$: toutes les hypothèses (7.46) et (7.48) sont vérifiées ; des solutions de l'équation (7.49) correspondant à la condition initiale f sont données (sur un ouvert de $\mathbb{R}^n \times]0,+\infty[$) par :

$$(x,t) \longmapsto \inf_{u\in\mathbb{R}^n}\left\{ f(u) + \frac{\|x-u\|^2}{2t}\right\}. \tag{7.58}$$

Avec $\theta^* = \sqrt{1+\|\cdot\|^2}$, on a

$$\theta : x \in \mathbb{R}^n \longmapsto \theta(x) = -\sqrt{1-\|x\|^2} \text{ si } \|x\| \leqslant 1,\ +\infty \text{ sinon.}$$

(*cf.* Exercice VII.7 et Exercice VII.3 si nécessaire).

Toutes les hypothèses faites sur θ dans l'exercice, notamment (7.48), sont vérifiées ; des solutions de l'équation (7.50) correspondant à la condition initiale f sont données (sur un ouvert de $\mathbb{R}^n \times]0,+\infty[$) par :

$$(x,t) \longmapsto \inf_{\|u-x\|\leqslant t}\left\{ f(u) - \sqrt{t^2 - \|x-u\|^2}\right\}. \tag{7.59}$$

Commentaire : L'expression de $F(\cdot,t)$ comme inf-convolution de f et de $t\theta\left(\frac{\cdot}{t}\right)$ est connue sous le nom de *formule de Lax et Oleinik*.

This page intentionally left blank

SOURCES

« Certains auteurs, parlant de leurs ouvrages, disent : mon livre, mon commentaire, mon histoire... ils feraient mieux de dire : notre livre, notre commentaire, notre histoire, vu que d'ordinaire il y a plus du bien d'autrui que du leur ».
B. Pascal

Des écrits ou contributions orales de collègues nous ont conduit, partiellement et parfois indirectement, à la confection de certains exercices ; qu'ils en soient remerciés ici.

1.4 : G. Constans ; 1.5 : I. Glazman et Y. Liubitch ; 1.6 : H. Wolkowicz ; 1.7 : A. Nemirovsky et Yu. Nesterov ; 1.8 : W.W. Hager, L. Amodei (Application 4 de la 5^{e} question) ; 1.10 : R. Fletcher ; 1.11 : O. Mangasarian, J.-Y. Ranjeva ; 1.16 : P. Finsler et G. Debreu ; 1.18 : J.-M. Exbrayat ; 2.4 : R. Horst, P. Pardalos et Nguyen V. Thoai ; 2.5 : Yu. Ledyaev et E. Giner ; 2.6 : Revue de Mathématiques Spéciales (RMS) ; 2.7 : I. Ekeland et R. Temam ; 2.9 : J.-P. Dedieu et R. Janin ; 2.10 : C. Bès ; 2.11 : J. Hall ; 3.6 : Pham Dinh Tao ; 3.8 : V. Alexéev, E. Galéev et V. Tikhomirov ; 3.9 : C. Lemaréchal ; 3.12 : J.-P. Aubin et H. Frankowska ; 3.13 : RMS ; 3.17 : M. Kojima ; 3.18-3.19-3.20 : H. Moulin et F. Fogelman-Soulié ; 3.21-3.22 : J.-L. Goffin, S. Boyd et L. Vandenberghe ; 3.23 : J. Dennis et J. Moré ; 3.25 : R. Fletcher ; 3.27 : H. Moulin et F. Fogelman-Soulié ; 3.28 : J.D. Buys ; 3.30 : J. Lasserre et F. Bonnans ; 3.32 : RMS (1° et 2°), P. Huard (3°) ; 4.1 : A. Auslender ; 4.8 : N. Gaffke et R. Mathar ; 4.10-4.11-4.12 : S. Boyd et L. Vandenberghe ; 4.13 : R.T. Rockafellar ; 5.8-5.9 : R.L. Dykstra ; 5.11 : A. Badrikian ; 5.14 : S. Achmanov ; 5.15 (1°) : P. Thomas ; 5.17 : S. Robinson ; 5.19 : B.T. Polyak ; 5.21 : M. Sakarovitch ; 5.22, 5.24 : S. Achmanov ; 5.25 : J.-Ph. Vial ; 6.1 : S. Achmanov ; 6.8 : M. Volle ; 6.9 : L. Vanderberghe ; 6.13 : J.M. Borwein et R.C. O'Brien ; 6.14 : Z. Artstein ; 6.18 : S. Robinson ; 6.19 : F. Clarke et Ph. Loewen ; 6.25 : J.-P. Crouzeix et J.E.

Martinez-Legaz ; 6.27 : W. Dinkelbach ; 6.28 : I. Csiszár et A. Ben-Tal, A. Ben-Israel et M. Teboulle ; 6.29 (2°) : S. K. Mitra ; 6.30 : G. Letac ; 6.31 : O.L. Mangasarian ; 7.1 : RMS, M. Volle ; 7.3 : F. Flores ; 7.8 : J.-P. Quadrat ; 7.9 : J.M. Borwein ; 7.13 : L. Thibault ; 7.14 : O. Mangasarian et C. Gonzaga ; 7.15 : J.-J. Moreau et B. Martinet ; 7.19 : A. Seeger ; 7.20-7.21 : A. Lewis ; 7.23 : J.-C. Rochet et J. Benoist ; 7.27 : J. Toland et C. Michelot ; 7.28 : Ph. Plazanet.

Peuvent être considérés comme des « grands classiques » du domaine les exercices suivants : 1.4, 1.9, 1.13, 1.16, 1.18, 2.1, 2.7, 2.8, 2.10, 3.4, 3.16, 4.7, 4.9, 5.6, 5.12, 5.13, 5.18, 6.3, 6.4, 6.5, 6.10, 6.11, 6.17, 7.3, 7.12, 7.15.

RÉFÉRENCES GÉNÉRALES

[1] S. Achmanov, *Programmation linéaire,* Éditions Mir, Moscou (1984).

[2] V. Alexéev, E. Galéev et V. Tikhomirov, *Recueil de problèmes d'optimisation*, Éditions Mir (1987).

[3] D.P. Bertsekas, *Nonlinear programming*, Athena Scientific (1995).

[4] F. Bonnans, J.-C. Gilbert, C. Lemaréchal et C. Sagastizabal, *Optimisation numérique ; aspects théoriques et pratiques,* N° 27 de la série Mathématiques et Applications (1997). Version en anglais publiée en 2003.

[5] S. Boyd and L. Vanderberghe, *Convex optimization*, Cambridge University Press (2004).

[6] V. Chvatal, *Linear programming,* W.H. Freeman and Company, New-York (1983).

[7] Concis et dense. Très bon.

[8] J.-C. Culioli, *Introduction à l'Optimisation,* Ellipses (1994).

[9] F. Droesbeke, M. Hallin et C. Lefevre, *Programmation linéaire par l'exemple,* Ellipses (1986).

[10] J. Gauvin, *Leçons de programmation mathématique*, Éditions de l'École polytechnique de Montréal (1995).

[11] J.-B. Hiriart-Urruty, *L'Optimisation*, collection « Que sais-je ? », Presses Universitaires de France (1996).

[12] J.-B. Hiriart-Urruty et C. Lemaréchal, *Convex analysis and minimization algorithms, Vol. 1 : Fundamentals*, Grundlehren der mathematichen Wissenschaften **305**, Springer-Verlag (1993).

[13] J.-B. Hiriart-Urruty et C. Lemaréchal, *Convex analysis and minimization algorithms, Vol. 2 : Advanced theory and bundle methods*, Grundlehren der mathematichen Wissenschaften **305**, Springer-Verlag (1993).

[14] J.-B. Hiriart-Urruty et C. Lemaréchal, *Convex analysis andminimization algorithms*, Vol. I et II, Grundlehren der mathematichen Wissenschaften **305** & **306**, Springer-Verlag (1993).

[15] J.-B. Hiriart-Urruty et Y. Plusquellec, *Exercices d'Algèbre linéaire et bilinéaire*, CEPADUES-Éditions (1988).
[16] R.A. Horn and C.R. Johnson, *Matrix analysis*, Cambridge University Press (ré-impression de 1992).
[17] B. Lemaire et C. lemaire-Misonne, *Programmation linéaire sur micro-ordinateur,* Collection Méthode + Programmes, Masson (1988).
[18] D.G. Luenberger, *Linear and nonlinear programming*, Addison-Wesley, 2e édition (1984).
[19] M. Minoux, *Programmation mathématique* : *Théorie et Algorithmes*, Vol. I, Dunod (1983).
[20] C. Roos, T. Terlaky and J-Ph. Vial, *Theory and algorithms for linear optimization: an interior point approach,* J. Wiley (1997).
[21] Roseaux (nom collectif), *Exercices et problèmes résolus de Recherche Opérationnelle,* Tome III, Masson (1985).
[22] M. Sakarovitch, *Optimisation combinatoire. Graphes et Programmation linéaire,* Herman (1984).
[23] A. Schrijver, *Theory of linear and integer programming,* J. Wiley (1987).
[24] J. Teghem, *Programmation linéaire,* collection Statistique et Mathématiques Appliquées, Éditions de l'Université de Bruxelles et Éditions Ellipses (1996).
[25] S.J. Wright, *Primal-dual interior point methods,* SIAM Publications (1997).
[26] [21], [22] et [24] contiennent de nombreux exemples simples et des illustrations ; elles intègrent la Programmation linéaire dans un domaine plus vaste, répertorié sous le vocable de « Recherche Opérationnelle ».
[27] Description et analyse des principales méthodes de résolution numérique des problèmes de programmation linéaire reposant sur l'algorithme du simplexe, ainsi que les programmes nécessaires à leur mise en œuvre sur micro-ordinateur.
[28] [CS] Très complets sur la question ; de véritables « Bibles ». Longtemps dominée par les algorithmes du type « méthode du simplexe », la résolution numérique des programmes linéaires a subi un véritable révolution avec l'apport de N. Karmarkar (1984). Les techniques du type « points intérieurs » (*cf.* l'Exercice V.25 pour une idée) commencent à prendre place dans les formations du niveau 2e cycle : voir le chapitre 4 de [8] et les chapitres XIII et XIV de [24] par exemple. La 4e partie de [4] et les ouvrages [20] et [25] sont consacrés pour l'essentiel à ces nouvelles approches.

NOTICE HISTORIQUE

Bien des noms (Euler, Lagrange, Legendre, Lipschitz, etc.) ont été rencontrés dans d'autres contextes par l'étudiant-lecteur. Nous évoquons ici ceux cités dans ce recueil et/ou ayant marqué les développements modernes de l'Optimisation et de l'Analyse convexe.

K. ARROW, L. HURWICZ et **H. UZAWA.** Associés deux par deux ou tous les trois, ce sont guidés par des problèmes d'origine économique que ces économistes-mathématiciens ont publié ou édité les méthodes algorithmiques portant à présent leurs noms. L'Économie mathématique a beaucoup interagi avec l'Optimisation, et ce dès le début des années cinquante. Notons d'ailleurs que beaucoup de lauréats du Prix Nobel d'Économie (instauré en 1969) ont eu des activités très « mathématisées » : R. Frisch (1895–1973, norvégien, Prix Nobel en 1969) proposa en 1954 un type de pénalisation intérieure (ou de fonction-barrière) pour la programmation linéaire, K. Arrow (1921–, américain, Prix Nobel en 1972), L.V. Kantorovitch (1912–1986, russe, Prix Nobel en 1975), J.F. Nash (1928–, américain, Prix Nobel en 1994), sans oublier G. Debreu (1921–2004, d'origine française, Prix Nobel en 1983).

R. BAIRE (1874–1932) est un mathématicien français dont la période active ainsi que la carrière universitaire furent réduites à une douzaine d'années par des maladies nerveuses qui l'ont conduit à la mort. Il a enrichi l'Analyse de notions fondamentales, par exemple la semicontinuité qui s'avère essentielle dans les résultats d'existence dans un problème d'optimisation (*cf.* Exercice II.1).

G. BIRKHOFF (1911–1996). Garret de son prénom, il est souvent confondu avec George Birkhoff son père ; tous les deux sont des mathématiciens américains de renom qui furent longtemps professeurs à l'université de Harvard aux États-Unis. Un des résultats les plus connus de Garret Birkhoff est celui qui dit que « les sommets de l'ensemble des matrices bistochastiques sont les matrices de permutation » (1946), *cf.* Exercice V.11, dont les applications en géométrie convexe sont importantes.

C. CARATHÉODORY (1873–1950). Mathématicien allemand d'origine grecque. Le lecteur-étudiant a rencontré ou rencontrera son nom dans des domaines tels que la Théorie de la mesure et le Calcul des variations. Un de ses fameux théorèmes est : Si S est une

partie d'un espace vectoriel de dimension n, tout point de l'enveloppe convexe de S est combinaison d'une partie finie de S de cardinal au plus $n+1$ (*cf.* Chapitre VI).

A.L. CHOLESKY (1875–1918). Peu d'étudiants en Analyse numérique savent que Cholesky n'est pas un obscur polonais mais bel et bien un ingénieur militaire français originaire de la région Poitou-Charentes. C'est dans le but d'applications à la géodésie que Cholesky étudie la résolution de systèmes d'équations linéaires et introduit la factorisation qui porte son nom. Le « procédé du commandant Cholesky » fut publié de manière posthume en 1924 dans le Bulletin géodésique de Toulouse. N'importe quel cours d'Analyse numérique matricielle, n'importe où dans le monde, fait référence à la factorisation de Cholesky ; comme quoi, rien ne sert de... il faut publier à point. À quand le nom de Cholesky sur le site du Futuroscope ?

I. EKELAND (1944–). Mathématicien français, professeur retraité de l'Université de Paris IX – Dauphine. Ses travaux portent sur les problèmes variationnels, la commande optimale et l'économie mathématique. Sa condition nécessaire d'optimalité pour solutions approchées d'un problème d'optimisation, établie dans un contexte d'espace métrique complet (1974) (*cf.* Exercice II.3 pour une version simplifiée), est devenue un outil classique de l'Analyse variationnelle. I. Ekeland est également auteur de plusieurs ouvrages de popularisation des mathématiques.

(KY) FAN (1914–). Mathématicien américain d'origine chinoise. Arrivé comme boursier à Paris en 1939 avec tout viatique le plan du métro de Paris (comme il le raconte lui-même) et un grand désir de faire des mathématiques, Ky Fan obtient son doctorat en 1941 sous la direction de M. Fréchet. Après être resté en France jusqu'en 1945 puis occupé ensuite plusieurs postes aux États-Unis, il s'installe à Santa Barbara (Californie) en 1965 ; retraité, il y est toujours. Le champ d'activités de Ky Fan a été très vaste : théorie des opérateurs, analyse convexe et inégalités, analyse matricielle (avec des inégalités très fines sur les valeurs propres), topologie et théorèmes de point fixe. On lui doit un des tous premiers théorèmes de mini-maximisation (*cf.* Chapitre IV) et une formulation variationnelle de la somme des m plus grandes valeurs propres d'une matrice symétrique (*cf.* Exercice VI.12). Ky Fan est docteur *honoris causa* de l'université de Paris IX-Dauphine (1990).

G. FARKAS (1847–1930). Mathématicien hongrois qui a contribué à la théorie de la Physique ; dans ce domaine, les plus connus sont ses résultats en Mécanique et Thermodynamique. La première preuve (complète et correcte) de son théorème sur les inégalités linéaires homogènes (*cf.* Chapitre V) fut publiée en 1898.

W. FENCHEL (1905–1988), **J.-J. MOREAU** (1923–) et **R.T. ROCKAFELLAR** (1935–). Le développement de l'Analyse convexe moderne doit beaucoup à ces trois personnes. W. Fenchel, mathématicien danois d'origine allemande avait une approche très géométrique de la convexité ; J.-J. Moreau (à ne pas confondre avec l'actrice de cinéma de même nom) est un mathématicien-mécanicien de Montpellier qui, selon ses dires,« appliquait la Mécanique aux Mathématiques ». R.T. Rockafellar est un mathématicien américain dont les racines sont (huguenotes) françaises (« Rocquefeuille ») comme les financiers de noms voisins ; le concept de « problème dual » a été un des fils directeurs de l'approche

de cet auteur. R.T. Rockafellar est docteur *honoris causa* de l'université de Montpellier II (1995).

P. de FERMAT (1601–1665). Tout le monde en a entendu parler... y compris récemment. Ses célèbres questions d'Arithmétique ont fait un peu oublier son principe variationnel en Optique géométrique (le premier, historiquement, sans doute) et ses implications sur la loi de réfraction de la lumière ; une de ses citations est : « La nature agit toujours par les voies les plus courtes ». En 1629, soit treize ans avant la naissance de Newton, Fermat conçut sa méthode « De maximis et minimis » s'appliquant à la détermination des valeurs qui rendent maximum ou minimum une fonction ainsi que celles des tangentes aux courbes, ce qui revenait à poser les fondements du Calcul différentiel. À la Salle des Illustres du Capitole à Toulouse (Hôtel de ville) se trouve une statue de P. de Fermat avec la mention « inventeur du calcul différentiel », et le Guide du Routard d'ajouter « encore merci ! » (à Beaumont-de-Lomagne, village natal de Fermat). Alors, étudiants-lecteurs, si vous avez des problèmes en calcul différentiel, vous savez à qui vous plaindre...

J.W. GIBBS (1839–1903). Physicien américain (un « phénomène »...) auteur de travaux très profonds, en Thermodynamique entre autres. Le rôle de la convexification en Thermodynamique a été essentiellement étudié par lui. Dans ses travaux se trouvent aussi les racines des conditions de KKT (*cf.* Exercice IV.7). Une de ses citations : « Mathematics is the language of science ».

P. GORDAN (1837–1932). Mathématicien allemand connu pour ses travaux en Algèbre et Géométrie. *Cf.* Exercices VI.8 et VII.1 pour un lemme portant son nom.

L.V. KANTOROVICH (1912–1986). Mathématicien russe ayant contribué à l'Analyse appliquée, l'Analyse fonctionnelle et la Programmation linéaire. Bien que de formation entièrement mathématique, il montra une perception très pertinente des problèmes de nature économique auxquels il appliqua les techniques mathématiques. Considéré comme un des pères-fondateurs (avec G. Dantzig (1914–2005) aux États-Unis) de la Programmation linéaire, L. Kantorovitch partagea le prix Nobel d'Economie 1975 avec T. Koopmans. De nature robuste (des collègues l'ont vu avaler d'un trait une bouteille de vodka et plonger pour traverser un lac à la nage), il eût néanmoins à souffrir de sévices sous Staline.

W. KARUSH, H. KUHN et **A.W. TUCKER.** Non, non ce n'est pas la 3^{e} ligne de l'équipe de rugby d'Australie que doit rencontrer prochainement le Stade Toulousain... Alors que la règle de Lagrange (*i.e.* conditions d'optimalité dans des problèmes avec des contraintes du type égalité) était connue dès le début du XIX^{e} siècle, il faut attendre le milieu du XX^{e} siècle pour qu'on s'intéresse (par nécessité) aux conditions d'optimalité dans des problèmes avec des contraintes du type inégalité. A.W. Tucker (mathématicien américain d'origine canadienne, récemment décédé) et son élève H. Kuhn ont publié en 1951 in article fondamental à ce sujet. Il s'est avéré qu'ils avaient été précédés par W. Karush dans un travail publié en 1948 qui était pour l'essentiel son mémoire de Maîtrise de 1938 (étudiants-lecteurs, tous les espoirs vous sont permis !). Il est donc juste d'associer ces trois noms dans « les conditions de KKT » (*cf.* Chapitre III). Une citation : « The KKT theorem provides the single most important tool in modern economic analysis both from the theoretical and computational point of view » (D. Gale, 1967).

J. JENSEN (1859–1925). Mathématicien danois, autodidacte. Il n'occupa aucune position universitaire mais fut longtemps expert dans une compagnie téléphonique de Copenhague. L'inégalité de convexité qui porte son nom (*cf.* Chapitre VI) date de 1906. La tradition de convexité chez les mathématiciens scandinaves est commémorée par la flamme de la poste danoise ci-dessous.

F. JOHN (1910–1994) (on trouve souvent écrit Fritz John) est aussi un mathématicien de ce siècle qui a établi les conditions qui portent son nom (*cf.* Chapitre III) en 1948. Le travail de F. John était motivé par des questions de nature géométrique ; les ellipsoïdes pleins de volume extrémal mis en évidence dans les Exercices III.21 et III.22 sont d'ailleurs appelés ellipsoïdes de John (ou de Loewner-John). On doit aussi au mathématicien d'origine tchèque Ch. Loewner (ou K. Löwner) l'ordre partiel $\succeq$ dans $\mathcal{S}_n(\mathbb{R})$.

H. MINKOWSKI (1864–1909). Mathématicien allemand considéré comme le fondateur de la géométrie des nombres entiers ; il s'est intéressé aussi à la Physique mathématique (les fameux espaces de Minkowski de dimension 4) et a, le premier, procédé à une étude systématique de la convexité dans les espaces de dimension finie.

J. (VON) NEUMANN (1903–1957). Il y a plusieurs mathématiciens du nom de Neumann. Celui-ci, Von Neumann (John) est un mathématicien américain d'origine hongroise, mort relativement jeune, et qui fut pionnier dans ses idées et recherches en Mathématiques appliquées. Et quand il attaquait un problème, il n'enfonçait pas des portes ouvertes ! On lui doit notamment un théorème de mini-maximisation (*cf.* Chapitre IV).

INDEX