Institutional Repository - Research Portal
Dépôt Institutionnel - Portail de la Recherche
researchportal.unamur.be
University of Namur

THESIS / THÈSE

MASTER EN SCIENCES MATHÉMATIQUES

Sur la minimisation de la comme des q plus grandes valeurs propres et de la plus grande valeur propre en valeur absolue d'une matrice symétrique

WARTIQUE, Evelyne

Award date:
1996

Awarding institution:
Universite de Namur

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 25. Apr. 2024

FACULTÉS UNIVERSITAIRE NOTRE-DAME DE LA PAIX NAMUR
FACULTÉ DES SCIENCES

Sur la minimisation de la somme des q plus grandes valeurs propres et de la plus grande valeur propre en valeur absolue d'une matrice symétrique

Mémoire présenté pour l'obtention du grade
de Licencié en Sciences
mathématiques
par

EVELYNE WARTIQUE

Promoteur: *JEAN-JACQUES STRODIOT*

Année Académique 1995-1996

Sur la minimisation de la somme des q plus grandes valeurs propres et de la plus grande valeur propre en valeur absolue d'une matrice réelle symétrique

Résumé

Deux algorithmes implémentables et convergents sont proposés pour minimiser deux fonctions particulières convexes mais non nécessairement différentiables :

- la somme des q plus grandes valeurs propres d'une matrice réelle symétrique dont les éléments diagonaux constituent les seules variables, avec pour contrainte que la trace de la matrice soit constante ;
- la plus grande valeur propre en valeur absolue d'une matrice réelle symétrique paramétrisée A(x).

De telles fonctions trouvent leur origine respective en théorie des graphes et en ingénierie du contrôle.

Abstract

Two implementable algorithms for the minimization of particular convex but not everywhere differentiable functions are presented :

- the sum of the q largest eigenvalues of a real, symmetric matrix as a function of the diagonal entries of the matrix, constrained only by the requirement that the sum of these entries is constant ;
- the largest eigenvalue (in magnitude) of a real symmetric matrix function of x.

Such functions find there respective origin in graphs theory and in control engineering.

Mémoire de licence en sciences mathématiques.

Présenté par EVELYNE WARTIQUE.

Juin 1996.

Unité d'optimisation.
Promoteur : M. JEAN-JACQUES STRODIOT.

Nous tenons à exprimer toute notre gratitude à monsieur le professeur J.-J. STRODIOT qui a accepté de diriger ce mémoire et dont nous avons pu apprécier la compétence et l'amabilité.

Table des matières

Index

Partie I

Introduction

Nous nous proposons ici l'étude de deux algorithmes implémentables qui convergent pour minimiser deux fonctions appartenant à une classe particulière de fonctions convexes mais non nécessairement différentiables. En effet, chacune de ces fonctions est construite à partir des valeurs propres d'une matrice réelle symétrique. Le chapitre un nous fournira les outils nécessaires à l'étude de ces fonctions. Il constitue la partie théorique du mémoire de par l'analyse qu'il fait de la sensibilité de toutes les valeurs propres d'une matrice symétrique. Il ressortira entre autres de ce chapitre que la différentiabilité de nos fonctions "objectif" dépend de la multiplicité de certaines valeurs propres. De plus, les points minimisant ces fonctions sont souvent des points pour lesquels nous n'aurons pas la différentiabilité.

Le premier problème concerne la minimisation de la somme des q plus grandes valeurs propres d'une matrice réelle symétrique, où les variables sont les entrées diagonales de la matrice, avec pour seule contrainte que la somme de ces entrées soit constante. Il trouve son origine en théorie des graphes. Notamment dans les problèmes de partitionnement.

Les procédures existant pour minimiser une fonction convexe emploient soit une stratégie choisissant à chaque étape des directions au moyen du sous-différentiel, soit des caractérisations de certains élargissements du sous-différentiel facilement implémentables. Malheureusement, aucune d'entre elles n'est aisément applicable à notre fonction particulière. Notre démarche s'inspire de la méthode de plus grande descente. Sa principale caractéristique est d'utiliser une extension du sous-différentiel facilement implémentable. La procédure de minimisation que nous allons présenter au chapitre deux ne s'applique donc qu'à une classe spécifique de problèmes. Cependant, l'idée de base de notre algorithme, à savoir la mise à jour du sous-différentiel pour anticiper toute non-différentiabilité, a une application beaucoup plus générale. Notre étude se base essentiellement sur le rapport de J. Cullum, W.E. Donath et P. Wolfe [3].

Le second problème étudie la minimisation de la plus grande valeur propre en valeur absolue d'une matrice paramétrisée $A(x)$. Il provient entre autres de l'ingénierie du contrôle. Nous le réécrirons sous forme d'un problème avec des contraintes concernant les valeurs propres. Ce genre de problème est souvent rencontré en optimisation. Par exemple, en industrie de la structure, il arrive de minimiser le coût d'une structure pour laquelle la fréquence est soumise à des contraintes.

Nous présentons au chapitre trois un algorithme pour résoudre ce problème suivant la

méthode des contraintes actives. Sa convergence quadratique n'est pas démontrée ici. Il a pour caractéristique importante d'obtenir une direction de descente à partir de tout point qui n'est pas solution, en séparant, si nécessaire, des valeurs propres multiples. C'est pourquoi, l'algorithme apporte une amélioration aux méthodes du premier ordre décrites pour le même problème par Polak et Wardi [15] et Doyle [7]. Cette partie du travail est fortement influencée par deux ouvrages , l'un de Fletcher [9] et l'autre de Friedland, Nocedal et Overton [11].

N.B. : Pour faciliter l'accès aux références des théorèmes, lemmes, corollaires, exemples et propositions, nous avons inséré après la table des matières un index renvoyant aux pages correspondantes.

Problème de partitionnement.

Le problème de la minimisation de valeurs propres d'une matrice symétrique a de nombreuses applications notamment en théorie des graphes. Etant donné un multigraphe $G = (S, E)$, nous cherchons à partitionner ses sommets en q groupes, $S_1, ..., S_q$, de tailles respectives, $m_1, ..., m_q$, de sorte que le nombre d'arêtes dans E qui connectent les différents groupes (appelées connexions externes) soit minimal.

Par exemple, soit G_1 le graphe à quatre sommets $S = \{1, 2, 3, 4\}$ et quatre arêtes $E = \{(1,2)\,(2,3)\,(2,4)\,(3,4)\}$:

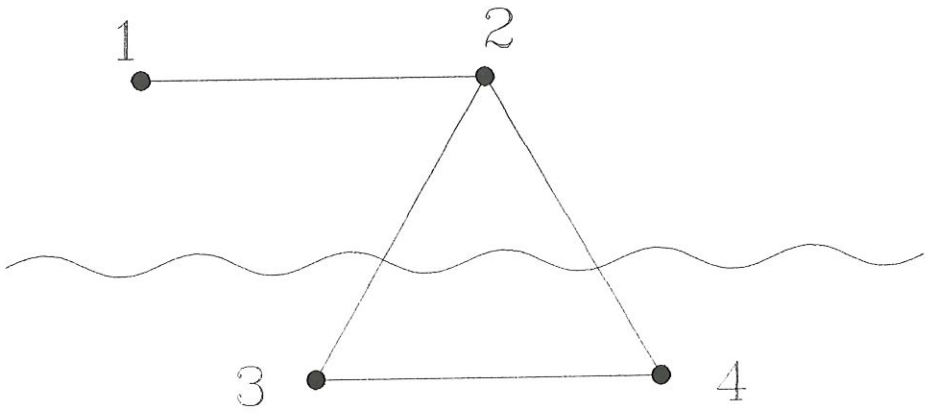


Figure 0-1:

Il y a trois partitions de G_1 en deux groupes de taille égale ($q = 2$, $m_1 = m_2 = 2$). Deux d'entre elles fournissent trois connexions entre les différents groupes. La partition optimale, comme on le voit sur la figure 1, est $S_1 = \{1, 2\}$ et $S_2 = \{3, 4\}$. Cette partition a deux arêtes externes $(2, 3)$ et $(2, 4)$.

En pratique, comme pour le schéma de montage des circuits intégrés d'un ordinateur, les graphes peuvent avoir de 100 à 2000 sommets, de nombreux arcs et le nombre de partitions peut aller jusqu'à 20. Pour de tels graphes, il n'est pas question de déterminer une partition optimale par une recherche exhaustive comme celle que nous avons employée pour G_1. L'objectif sera plutôt de rechercher une meilleure borne inférieure possible sur le nombre de connexions externes que toute partition peut avoir.

Afin de dériver une telle borne, Donath et Hoffman [6] utilisent les étapes suivantes. Notons

A la matrice d'incidence du graphe G ($A_{ij} = 1$ si (i,j) est une arête de G, et $A_{ij} = 0$ sinon), modifiée sur sa diagonale de sorte que $A_{ii} = -\sum_{j \neq i} A_{ij}$. Ce qui donne pour le graphe G_1 :

$$A = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 1 & -3 & 1 & 1 \\ 0 & 1 & -2 & 1 \\ 0 & 1 & 1 & -2 \end{bmatrix}.$$

Par définition de la trace de A, en tenant compte de la symétrie de A, nous avons :

$$tr(A) = \sum_i A_{ii} = \sum_i \left(-\sum_{j \neq i} A_{ij} \right) = -2 \sum_{j<i} A_{ij} \equiv -2 \, \#E,$$

où $\#E$ est le nombre d'arêtes dans G.

Pour toute partition de G, définissons P la matrice pour laquelle $P_{ij} = 1$ si i et j sont dans un même groupe de la partition, et $P_{ij} = 0$ sinon. P est la matrice d'incidence de l'union des sous graphes formés par la partition. Pour le graphe G_1, et la partition $P_1 = \{S_1, S_2\}$, où $S_1 = \{1,3\}$ et $S_2 = \{2,4\}$,

$$P = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}.$$

La matrice dont l'entrée i,j (pour $i \neq j$) est $A_{ij}P_{ij}$, est la matrice d'incidence du graphe dont les arêtes sont les arêtes communes à G et à un des sous graphes de la partition, ($A_{ij}P_{ij} = 1$ si $A_{ij} = 1$ et $P_{ij} = 1$). Le nombre $\#E_{NC}$ de ces arêtes est donné par

$$2\#E_{NC} = \sum_{\substack{i,j \\ i \neq j}} A_{ij}P_{ji} = \sum_i \sum_j A_{ij}P_{ji} - \sum_i A_{ii} = \sum_i \left[-A_{ii} + \sum_j A_{ij}P_{ji} \right]$$
$$= -trA + tr(AP) = 2\#E + tr(AP).$$

Pour notre exemple

$$[A_{ij}P_{ij}]_{i,j} = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & -3 & 0 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 1 & 0 & -2 \end{bmatrix},$$

$$AP = \begin{bmatrix} -1 & 1 & -1 & 1 \\ 2 & -2 & 2 & -2 \\ 2 & 2 & -2 & 2 \\ 1 & -1 & 1 & -1 \end{bmatrix} \text{ avec } \#E_{NC} = 1,$$

en effet, le seul arc commun à G_1 et P_1 est $(2,4)$.

Nous pouvons alors déduire le nombre $\#E_C$ de connexions externes, i.e. le nombre d'arêtes qui sont dans G et pas dans P :

$$2\#E_C = 2(\#E - \#E_{NC}) = -tr(AP). \tag{0.1}$$

(La partition P_1 produit donc trois connexions externes.)

Pour toute matrice diagonale D, telle que $tr(D) = 0$, l'équation (0.1) se généralise facilement à

$$tr((A+D)P) = -2\#E_C.$$

Pour toute matrice B, notons $\lambda_1(B) \geq \lambda_2(B) \geq ...$ ses valeurs propres rangées dans un ordre décroissant. Ainsi, $tr((A+D)P) = \sum_i^n \lambda_i\left[(A+D)P\right]$, et par l'inégalité de Hoffman-Wielandt [17] (car A, D, P sont symétriques),

$$-2\#E_C = tr((A+D)P) \leq \sum_i^n \lambda_i(P)\ \lambda_i(A+D). \tag{0.2}$$

Chaque matrice de partition a un rang q, le nombre de groupes de la partition. Ce qui veut dire que P a exactement q valeurs propres non nulles (car P est symétrique) : $\lambda_i(P)\ i = 1, ..., q$. De plus,

$$\lambda_i(P) = m_i \quad 1 \leq i \leq q, \tag{0.3}$$

où $m_1 \geq m_2 \geq ... \geq m_q$ sont les cardinaux des groupes de la partition. En effet, par des permutations identiques sur les lignes et les colonnes (i.e. si on permute les lignes i et j, on permute les colonnes i et j, et vice versa), P se ramène à une matrice formée de q blocs contenant uniquement des 1. Chacun de ces blocs donne une valeur propre égale à sa dimension, i.e. m_i. Les équations (0.2) et (0.3) font en sorte que le nombre $\#E_C$, des arêtes externes de toute partition de G en q groupes de taille $m_1 \geq m_2 \geq ... \geq m_q$, doit satisfaire l'inégalité suivante :

$$\#E_C \geq -\frac{1}{2}\sum_{i}^{q} m_i \, \lambda_i(A+D),$$

pour toute matrice diagonale D dont la trace est nulle.

Donc,

$$M \equiv - \min_{tr(D)=0} \sum_{i}^{q} m_i \, \lambda_i(A+D)/2$$

est la "meilleure" borne inférieure de $\#E_C$.

Nous développerons dans le chapitre deux un algorithme implémentable et convergent, pour déterminer un minimum de la fonction

$$f(D) = \sum_{i=1}^{q} \lambda_i(A+D),$$

soumise à la contrainte

$$tr(D) = 0.$$

Cette fonction particulière correspond au partitionnement d'un graphe en q groupes de tailles égales. La manipulation d'une partition en q groupes de tailles inégales implique de grandes modifications de l'algorithme proposé. Nous ne nous attarderons pas sur ce problème.

Partie II

Deux algorithmes implémentables

Chapitre 1

Analyse de sensibilité de toutes les valeurs propres d'une matrice symétrique

1.1 Contexte et notations

Nous allons étudier la sensibilité au premier ordre de toutes les valeurs propres d'une matrice A(x) réelle, symétrique, dépendant de façon différentielle d'un paramètre x. Nous contournerons les difficultés inhérentes à l'étude de chaque valeur propre λ_m, en considérant la somme f_m des m plus grandes valeurs propres de A(x). La raison en est que les formules de variation pour λ_m ne sont pas facilement utilisables pour l'analyse de sensibilité tandis qu'il en existe - elles sont dues à Ky Fan - pour les fonctions f_m directement liées aux techniques et résultats de l'analyse convexe et non différentiable. Nous ferons la distinction entre ce qui dépend de la variation du paramètre ($x \rightsquigarrow A(x)$) et ce qui est intrinsèque au problème de valeur propre ($A \rightsquigarrow \lambda(A)$).

Mettons au point quelques notations :

$M^{n,m}(R)$ désigne l'espace des matrices réelles à n lignes et m colonnes.

S^n désigne l'espace des matrices réelles et symétriques d'ordre n; on le munit du produit interne $\langle\langle A, B\rangle\rangle \equiv tr(AB)$ (trace de AB).

- $A(.) : R^p \rightarrow S^n \;\; x \rightsquigarrow A(x) \in C^1$

$\lambda_1(A(x)) \geq \lambda_2(A(x)) \geq ..... \geq \lambda_n(A(x))$ sont les n valeurs propres de A(x) rangées dans un ordre décroissant. Notons $\lambda_m(A(x)) = \lambda_m(x)$.

- $\sigma_m(.) : S^n \to R \;\; A \rightsquigarrow \sigma_m(A) \equiv \sum_{i=1}^{m} \lambda_i(A)$.

- $f_m(.) : R^p \to S^n \to R \;\; x \rightsquigarrow A(x) \rightsquigarrow f_m(x) \equiv \sigma_m(A(x)) = \sum_{i=1}^{m} \lambda_i(A(x))$.

  Nous avons donc $\lambda_m(x) = f_m(x) - f_{m-1}(x)$.

- $[B]_{ij}$ désigne l'élément i, j (i.e. situé sur la i-ème ligne et j-ème colonne) de la matrice $B = [b_1, b_2, ..., b_n]$ et sa k-ème colonne est notée $b_k = (b_{kl})_{l=1,...,n}$. Ainsi, b_{ji} désigne l'élément i, j de B ou encore la i-ème composante de b_j.

1.2 Propriétés des fonctions $\lambda_m(A)$, $\sigma_m(A)$ et $f_m(x)$

1.2.1 Propriétés des valeurs propres comme fonctions de A

Théorème 1 *(convexité)*

1. Les fonctions $\lambda_m : A \in S^n \rightsquigarrow \lambda_m(A)$ *avec m=1,.....,n sont positivement homogènes i.e.* $\lambda_m(\alpha A) = \alpha \lambda_m(A)$ *pour tout* $\alpha \geq 0$ *et* $A \in S^n$.

2. $\lambda_1(.)$ *est une fonction convexe et* $\lambda_n(.)$ *est une fonction concave.*

3. pour tout $m \in \{2, ..., n-1\}$, $\lambda_m(.)$ *est la différence de fonctions convexes positivement homogènes.*

Preuve. λ_m étant une valeur propre, la propriété 1 est automatiquement vérifiée.

La propriété 2 provient du caractère extrémal de λ_1(plus grande valeur propre) et de λ_n (plus petite valeur propre). $\lambda_1(A) = \sup_{\|y\|=1} \langle Ay, y \rangle$ et par la semi-linéarité du produit scalaire, on a pour tout $B \in S^n$ et $\mu \in [0, 1]$,

$$\lambda_1(\mu A + (1-\mu)B) = \sup_{\|y\|=1} \langle (\mu A + (1-\mu)B)\, y, y \rangle = \sup_{\|y\|=1} [\mu \langle Ay, y \rangle + (1-\mu) \langle By, y \rangle]$$
$$\leq \mu \sup_{\|y\|=1} \langle Ay, y \rangle + (1-\mu) \sup_{\|y\|=1} \langle By, y \rangle = \mu \lambda_1(A) + (1-\mu)\lambda_1(B).$$

La preuve est similaire pour $\lambda_n(A)$.

La propriété 3 deviendra claire lorsque nous aurons montré que $\sigma_m(A)$ est une fonction convexe positivement homogène pour tout $m \in \{2, ..., n-1\}$.

En effet $\lambda_m(A) = \sigma_m(A) - \sigma_{m-1}(A)$. ■

Remarque 1 *Un grand nombre de fonctions souvent rencontrées sont construites à partir des valeurs propres de A et préservent donc une structure de convexité. Par exemple :*

La largeur du spectre de A est une fonction convexe $A \in S^n \rightsquigarrow w(A) \equiv \lambda_1(A) - \lambda_n(A)$.

Le rayon spectral de A est une fonction convexe $A \in S^n \rightsquigarrow t(A) \equiv \max\{\lambda_1(A), -\lambda_n(A)\}$.

Le nombre de conditionnement de A est la différence de fonctions convexes sur le cône ouvert convexe P des matrices définies positives $A \in P \rightsquigarrow c(A) \equiv \lambda_1(A) / \lambda_n(A)$.

1.2.2 Formule variationnelle de Ky Fan pour $\sigma_m(A)$

$\sigma_m(A) \equiv \lambda_1(A) + ... + \lambda_m(A)$ est une fonction positivement homogène (car les λ_i le sont). Nous allons la réécrire comme un maximum de formes linéaires sur S^n. Nous aurons ainsi que $\sigma_m(A)$ est bien une fonction convexe. Pour comprendre la formule variationnelle de $\sigma_m(A)$, partons de celle de Rayleigh pour $\sigma_1(A)$.

Formule variationnelle pour $\sigma_1(A)$

La formule de Rayleigh donne

$$\lambda_1(A) = \sigma_1(A) = \max_{\|x\|=1} \langle Ax, x\rangle = \max_{\|x\|=1} \left\langle\left\langle A, xx^t\right\rangle\right\rangle \text{ pour tout } A \in S^n. \tag{1.1}$$

σ_1 est le maximum d'une collection de formes linéaires $A \in S^n \rightsquigarrow \langle\langle A, xx^t\rangle\rangle$ indexé par x. σ_1 est la fonction support de $R_1 \equiv \{xx^t : \|x\| = 1\}$ (ensemble des matrices normalisées symétriques définies positives de rang 1). L'enveloppe convexe fermée de R_1 est le plus grand ensemble dont la fonction support est σ_1. $\Omega_1 \equiv coR_1$ est bien fermé car R_1 est compact (fermé borné de S^n). On a donc

$$\sigma_1(A) = \max\{\langle\langle A, C\rangle\rangle : C \in \Omega_1\}.$$

Les points extrêmes de Ω_1 sont les matrices de R_1. Et maximiser une forme linéaire sur Ω_1 ou sur R_1 revient au même.

Proposition 1 *Réécriture de* Ω_1

$$\Omega_1 \equiv co\left\{xx^t : \|x\| = 1\right\} = \{C \geq 0 : trC = 1\}. \tag{1.2}$$

Preuve. Vérifions que $co\{xx^t : \|x\| = 1\} \subseteq \{C \geq 0 : trC = 1\}$.

Nous avons $\{xx^t : \|x\| = 1\} \subseteq \{C \geq 0 : trC = 1\}$.

Ce dernier ensemble est une partie convexe car pour tout $\mu \in [0,1]$ et pour tout $C_1 \geq 0 \;\; C_2 \geq 0$ vérifiant $trC_1 = trC_2 = 1$, nous avons

$\mu C_1 + (1-\mu)C_2 \geq 0$ et $tr\,(\,\mu C_1 + (1-\mu)C_2) = \mu\, tr\, C_1 + (1-\mu)\, tr\, C_2 = \mu + 1 - \mu = 1.$

Et comme coR_1 est le plus petit convexe contenant $\{xx^t : \|x\| = 1\}$, nous avons la première inclusion.

Il reste à montrer $co\{xx^t : \|x\| = 1\} \supseteq \{C \geq 0 : trC = 1\}$.

Soit $C \geq 0$ tel que $trC = 1$. Par la décomposition spectrale des matrices symétriques, nous construisons les matrices P et D contenant respectivement les vecteurs propres normalisés et les valeurs propres de C et nous écrivons $C = PDP^t = \sum_{i=1}^{n} \lambda_i\, p_i\, p_i^t$.

Montrons que C s'écrit comme combinaison convexe des éléments de $\{xx^t : \|x\| = 1\}$. Il suffit de prendre les λ_i pour coefficients de la combinaison. En effet $\|p_i\| = 1$, $\sum_{i=1}^{n} \lambda_i = trC = 1$ et $\lambda_i \geq 0$ car $C \geq 0$. ■

Formule variationnelle pour $\sigma_m(A)$

En généralisant la formule (1.1) de Rayleigh [2, Chap. 2] nous écrivons

$$\begin{aligned} \sigma_m(A) &= \max\left\{tr(AXX^t) : X \in M^{n,m}(R),\, X^tX = I_m\right\} \\ &= \max\{\langle\langle A, Y\rangle\rangle : Y \in R_m\} \end{aligned} \tag{1.3}$$

où $R_m = \{XX^t : X \in M^{n,m}(R),\, X^tX = I_m\}$.

Cela veut dire que σ_m est la fonction support du compact R_m. Notons Ω_m l'enveloppe fermée convexe de R_m. Tous les points extrèmes de Ω_m appartiennent à R_m. Donc

$$\sigma_m(A) = \max\{\langle\langle A, C\rangle\rangle : C \in \Omega_m\}.$$

Proposition 2 *Réécriture de* Ω_m

$$\Omega_m \equiv co\left\{XX^t : X \in M^{n,m}(R),\ X^tX = I_m\right\} = \{C \geq 0 : trC = m \text{ et } \lambda_1(C) \leq 1\} \tag{1.4}$$

Cela revient à dire que l'enveloppe convexe de l'ensemble des matrices de projection de rang m est l'ensemble des matrices symériques dont les valeurs propres appartiennent à $[0,1]$ *et dont la somme des valeurs propres vaut m.*

Avant d'entamer la preuve, définissons la *transformée de Legendre-Fenchel* pour la fonction support σ_m :

$$\sigma_m^* : C \in S^n \rightsquigarrow \sigma_m^*(C) = \sup_{A \in S^n} \{\langle\langle C, A\rangle\rangle - \sigma_m(A)\} \tag{1.5}$$

Preuve. *a*) Montrons que la fonction σ_m^* ne prend que deux valeurs: 0 et $+\infty$.

Si $C \in coR_m$ alors, pour tout $A \in S^n$, on a $\sigma_m(A) \geq \langle\langle A, C\rangle\rangle$, c'est-à-dire

$\langle\langle A, C\rangle\rangle - \sigma_m(A) \leq 0 \Rightarrow \sigma_m^*(C) = 0.$

Si $C \notin coR_m$ alors, $\sigma_m^*(C) = +\infty$.

- Pour le voir, remarquons d'abord qu'il existe $A \in S^n$ tel que $\langle\langle A, C\rangle\rangle > \sigma_m(A)$.

  -Si $tr(C) > m$, on prend $A = I^n$ et on a

  $\langle\langle I^n, C\rangle\rangle = tr(C) > \sigma_m(A) = \max\{tr(AX) : X \in coR_m\} = \max\{tr(X) : X \in coR_m\}$

  $= m.$

  -Si $tr(C) < m$, on prend $A = -I^n$ et on a

  $\langle\langle -I^n, C\rangle\rangle = tr(-C) > \sigma_m(A) = \max\{tr(AX) : X \in coR_m\}$

  $= \max\{tr(-X) : X \in coR_m\} = -m.$

  Le cas $tr(C) = m$ n'est pas à envisager car nous sommes dans le cas où $C \notin coR_m$.

- Le second point à remarquer est que pour tout $\mu > 0$, on a

  $$\mu A \in S^n \text{ et } \sigma_m^*(C) = \sup_{A \in S^n} \{\langle\langle C, A\rangle\rangle - \sigma_m(A)\} \geq \sup_{\mu > 0} \mu \underbrace{\{\langle\langle C, A\rangle\rangle - \sigma_m(A)\}}_{>0 \text{ pour } A \text{ bien choisi.}} = +\infty.$$

Nous avons donc que σ_m^* prend la valeur 0 précisément sur $coR_m = \Omega_m$.

b) Pour terminer notre preuve, il reste à déterminer les C pour lesquels σ_m^* est nul.

Par la décomposition spectrale de $C \in S^n$, nous avons $C = UDU^t$ avec $D = diag(\mu_1 ..... \mu_n)$.

Avant de calculer $\sigma_m^*(C)$, notons que $\langle\langle UDU^t, A\rangle\rangle = \langle\langle D, U^tAU\rangle\rangle \equiv \langle\langle D, B\rangle\rangle$. Un changement de variable dans (1.5) donne

$\sigma_m^*(C) = \sup_{B\in S^n}\left\{\sum_{i=1}^n \mu_i b_i - \sigma_m(B)\right\}$, où les b_i sont les éléments diagonaux de B.

Donc, $\sigma_m^*(C) = +\infty$ lorsque :

- un μ_j est <0 (prendre $B = diag(0, .., b_j ..., 0)$ avec $b_j \to -\infty$) ou

- un μ_j est >1 (prendre $B = diag(0, .., b_j ..., 0)$ avec $b_j \to +\infty$) ou

- $\sum_{i=1}^n \mu_i \neq m$ (prendre $B = diag(b, ..., b)$ avec $b \to -\infty$ ou $b \to +\infty$).

D'autre part un bon choix de X dans (1.3) donne pour tout $B \in S^n$,

$b_{i_1} + b_{i_2} + ... + b_{i_m} \leq \sigma_m(B)$ où $b_{i_1}, b_{i_2}, ..., b_{i_m}$ sont les m plus grands éléments parmi les b_i.

Ainsi, si les μ_i appartiennent à $[0, 1]$ et si leur somme vaut m, nous aurons

$\sum_{i=1}^n \mu_i b_i - \sigma_m(B) \leq \sum_{i=1}^n \mu_i b_i - (b_{i_1} + b_{i_2} + ... + b_{i_m}) \leq 0$

(on a l'égalité si $\mu_{i_1} = \mu_{i_2} = ... = \mu_{i_m} = 1$, $\mu_i = 0$ ailleurs).

Tout ceci revient à dire que $\sigma_m^*(C) = 0$ si et seulement si le spectre de C vérifie

$\mu_i \in [0, 1]$ pour tout $i = 1, ..., n$ et $\sum_{i=1}^n \mu_i = m$.

Nous avons donc prouvé

$\{C : \sigma_m^*(C) < +\infty\} = \{C : \sigma_m^*(C) = 0\} = \{C \geq 0 : trC = m \text{ et } \lambda_1(C) \leq 1\}$. ■

Théorème 2 *(convexité.)*

$\sigma_m(A) = \max\{\langle\langle A, C\rangle\rangle : C \in \Omega_m\}$ *est une fonction positivement homogène et convexe.*

Remarque 2 *La contrainte $\lambda_1(C) \leq 1$ de (1.4) n'apparaît pas pour m=1 (cfr 1.2) puisqu'elle est automatiquement vérifiée.*

1.2.3 Propriété de la somme des valeurs propres comme fonction de x

Théorème 3 *(convexité.)*

1. *$f_m(x)$ est localement Lipschitzienne.*

2. *Pour le cas particulier $A(x) = A_0 + diag\,(x_1, ..., x_n)$, on a que $f_m(x) = \sigma_m\,(A_0 + diag\,(x_i))$ est une fonction convexe où $diag(x_i) = diag\,(x_1, ..., x_n)$.*

Preuve. 1. $f_m(x)$ est localement Lipschitzienne car $f_m = \sigma_m \circ A$ où $A(.) : R^p \to S^n \in C^1$ et $\sigma_m(.) : S^n \to R$ est localement Lipschitzienne.

2. Pour tout $\mu \in [0,1]$ et $x, y \in \mathrm{R}^n$, nous avons

$f_m(\mu x + (1-\mu)y) = \sigma_m(A_0 + diag(\mu x_i + (1-\mu)y_i))$

$= \sigma_m(A_0 + \mu\, diag(x_i) + (1-\mu)\, diag(y_i)) = \sigma_m[\mu(A_0 + diag(x_i)) + (1-\mu)(A_0 + diag(y_i))]$

$\leq \mu\ \sigma_m(A_0 + diag(x_i)) + (1-\mu)\,\sigma_m(A_0 + diag(y_i)) = \mu\, f_m(x) + (1-\mu)\, f_m(y)$. ■

1.3 Eléments d'analyse convexe et non différentiable

Comme des fonctions non différentiables convexes ou localement Lipschitziennes interviennent naturellement dans notre étude, nous présentons dans cette section quelques résultats de l'analyse non différentiable.

1.3.1 Cas convexe non différentiable

Considérons $f : \Theta \subset X \to \mathrm{R}$ une fonction convexe définie sur un ouvert convexe Θ de X. Si $x \in \Theta$, alors la dérivée directionnelle $f'(x,.)$ de f en x existe :

$d \in X \rightsquigarrow f'(x,d) \equiv \lim\limits_{t \to 0+} \frac{f(x+td)-f(x)}{t} = \inf\limits_{t>0} \frac{f(x+td)-f(x)}{t}$ et le sous différentiel de f en x s'écrit

$$\begin{aligned} \partial f(x) &\equiv \{s \in X : f(x') \geq f(x) + \langle s, x'-x\rangle \text{ pour tout } x' \in X\} \\ &= \{s \in X : \langle s, d\rangle \leq f'(x,d) \text{ pour tout } d \in X\}. \end{aligned} \tag{1.6}$$

De plus $\partial f(x) = \partial[f'(x,.)](0)$.

$\partial f(x)$ joue le même rôle que le gradient d'une fonction convexe différentiable. En fait la fonction convexe est différentiable si et seulement si $\partial f(x) = \{\nabla f(x)\}$.

On peut réécrire le sous différentiel en utilisant la transformée de Legendre-Fenchel de f :

$$\partial f(x) = \{s \in X : f^*(s) + f(x) - \langle s, x\rangle \leq 0\}. \tag{1.7}$$

1.3.2 Cas non convexe non différentiable

Dans ce cas la dérivée directionelle n'existe pas nécessairement. Nous nous limitons à une fonction localement Lipschitzienne $f : \Theta \to \mathrm{R}$ définie sur un ouvert Θ. Pour $x \in \Theta$, définissons

la dérivée directionnelle généralisée de f en x par

$d \in X \rightsquigarrow f^\circ(x,d) \equiv \lim\limits_{\substack{x' \to x \\ t \to 0^+}} \sup \frac{f(x'+td)-f(x')}{t}$.

$f^\circ(x,.)$ est une fonction convexe et positivement homogène. Il est donc naturel de considérer son sous différentiel en 0 ou, ce qui est équivalent, l'ensemble convexe non vide compact de X dont la fonction support est $f^\circ(x,.)$:

$\partial f(x) = \partial\left[f^\circ(x,.)\right](0) \equiv \{s \in X : \langle s,d\rangle \leq f^\circ(x,d) \text{ pour tout } d \in X\}$.

Lorsque f est convexe, on a $f'(x,.) = f^\circ(x,.)$, de sorte que l'ensemble défini ci-dessus n'est rien d'autre que le sous différentiel de f en x (cfr 1.6). Pour cette raison, nous le noterons $\partial f(x)$ et l'appellerons *sous différentiel généralisé de f en x.*

Supposons que la fonction localement Lipschitzienne $f : \Theta \subset X \to$R admette une dérivée directionnelle $d \in X \rightsquigarrow f'(x',d)$ pour tout $x' \in \Theta$. Alors $f^\circ(x,d) = \lim\limits_{x' \to x} \sup f'(x',d)$ pour tout $d \in X$.

Mieux encore, supposons que

$$f^\circ(x,d) = f'(x',d) \text{ pour tout } d \in X, \tag{1.8}$$

alors le gradient généralisé de f en x est directement défini comme dans le cas convexe. Lorsque f vérifie la condition (1.8), f est dite *strictement tangentiellement convexe* en x ou régulière en x selon Clarke. Par exemple, les fonctions convexes ou les fonctions de classe C^1 sur Θ sont strictement tangentiellement convexes en $x \in \Theta$. De plus, les fonctions appartenant à C^1 sur Θ sont caractérisées parmi les fonctions localement Lipschitziennes sur Θ par :

$$\partial f(x) \text{ est un singleton pour tout } x \in \Theta \Longleftrightarrow f \text{ est de classe } C^1 \text{ sur } \Theta.$$

Proposition 3 *Soit X et Y des espaces Euclidiens, soit F: $\Theta \subset X \to Y$ de classe C^1 où Θ est un ensemble ouvert de X. Soit $\sigma : \Theta' \subset Y \to R$ une fonction localement Lipschitzienne sur un ouvert Θ' de Y contenant $F(\Theta)$.*

Alors $f \equiv \sigma_\circ F$ est une fonction localement Lipschitzienne sur Θ et, pour tout $x \in \Theta$,

$$\partial f(x) \subset [JF(x)]^* \, \partial\sigma(F(x)) \tag{1.9}$$

où $[JF(x)]^*$ est l'adjointe du Jacobien J de la fonction $F(x)$.

On a l'égalité dans (1.9) si σ est strictement tangentiellement convexe en $F(x)$ et dans ce cas, f est aussi strictement tangentiellement convexe en x. En particulier, cette dernière condition est vérifiée pour tout x lorsque σ est une fonction convexe.

1.4 Sous différentiel et dérivée directionnelle de $\sigma_m(A)$

Nous avons vu que $\sigma_m(A) : A \in S^n \rightsquigarrow \sigma_m(A) \equiv \lambda_1(A) + \lambda_2(A) + ... + \lambda_m(A)$ (somme des m plus grandes valeurs propres de la matrice symétrique A) est une fonction convexe positivement homogène sur S^n, plus précisément, elle est la fonction support de R_m ou de Ω_m. Ainsi, l'étude de sensibilité au premier ordre de la fonction σ_m revient à l'étude du sous différentiel ou de la dérivée directionnelle de σ_m. En effet, nous avons:

$$\sigma_m(A+H) = \sigma_m(A) + \sigma_m'(A,H) + \|H\|\,\varepsilon(H) \text{ pour tout } H \in S^n \tag{1.10}$$

où $\varepsilon(.)$ est une fonction de H convergeant vers 0 avec H.
Ou encore, il existe $\theta \in]0,1[$ tel que

$$\sigma_m(A+H) = \sigma_m(A) + \langle\langle C, H\rangle\rangle \text{ pour } C \in \partial\sigma_m(A+\theta H). \tag{1.11}$$

1.4.1 Expression implicite du sous différentiel de $\sigma_m(A)$

- Cas où $A = 0$

Par définition de σ_m comme fonction de support, c'est-à-dire $\sigma_m(D) = \max\{\langle\langle C, D\rangle\rangle : C \in \Omega_m\}$, nous avons : $\sigma_m(D) \geq \langle\langle C, D\rangle\rangle$ pour tout $C \in \Omega_m$. On réécrit alors Ω_m :

$$\Omega_m = \{C \in S^n : \langle\langle C, D\rangle\rangle \leq \sigma_m(D) \text{ pour tout } D \in S^n\}.$$

Ainsi, pour $A = 0$, la situation est assez simple :

Proposition 4 *(formule implicite pour $A = 0$)*

$$\partial\sigma_m(0) = \Omega_m \text{ et } \sigma_m^{'}(0,.) = \sigma_m(.).$$

Preuve. $\sigma_m^{'}(0,D) = \lim\limits_{t\to 0+} \frac{\sigma_m(0+tD)-\sigma_m(0)}{t} = \lim\limits_{t\to 0+} \frac{\sigma_m(tD)-0}{t} = \lim\limits_{t\to 0+} \frac{t\sigma_m(D)}{t} = \sigma_m(D).$

$\partial\sigma_m(0) = \left\{C \in S^n : \langle\langle C,D\rangle\rangle \leq \sigma_m^{'}(0,D) \text{ pour tout } D \in S^n\right\}$. ■

- Cas où $A \neq 0$

Théorème 4 *(formule implicite)*

Etant donné $A \in S^n$, nous avons:

$$\partial\sigma_m(A) = \{C \geq 0 : trC = m,\ \lambda_1(C) \leq 1 \text{ et } \langle\langle A,C\rangle\rangle = \sigma_m(A)\} \tag{1.12}$$

$$= co\left\{XX^t : X \in M^{n,m}(R),\ X^tX = I_m \text{ et } tr(AXX^t) = \sigma_m(A)\right\} \tag{1.13}$$

Preuve. Ecrivons l'expression du sous différentiel de $\sigma_m(A)$ en fonction de la transformée de Legendre Fenchel $C \rightsquigarrow \sigma_m^*(C)$ (cfr 1.7) : $\partial\sigma_m(A) = \{C \in X : \sigma_m^*(C) + \sigma_m(A) - \langle\langle A,C\rangle\rangle \leq 0\}$. Prenons $C \in \partial\sigma_m(A)$.

- Si $C \in \Omega_m$, la transformée de Legendre Fenchel de σ_m prend la valeur 0 (cfr la preuve de la proposition 2). Donc, $\sigma_m(A) \leq \langle\langle A,C\rangle\rangle$. Et comme $\sigma_m(A) \geq \langle\langle A,C\rangle\rangle$, nous avons $\sigma_m(A) = \langle\langle A,C\rangle\rangle$.
- Si $C \notin \Omega_m$, $\sigma_m^*(C) = +\infty$. Ce qui donne $\sigma_m(A) - \langle\langle A,C\rangle\rangle = -\infty$ c'est-à-dire $\sigma_m(A) = -\infty$, qui est impossible par définition de $\sigma_m(A)$ comme somme des m plus grandes valeurs propres de la matrice rélle A.

Nous écrivons dès lors :

$$C \in \partial\sigma_m(A) \Longleftrightarrow C \in \Omega_m \text{ et } \langle\langle A,C\rangle\rangle = \max\{\langle\langle A,D\rangle\rangle : D \in \Omega_m\} = \sigma_m(A)$$

Donc (1.12) est vérifié si on réécrit Ω_m en utilisant l'expression (1.4).

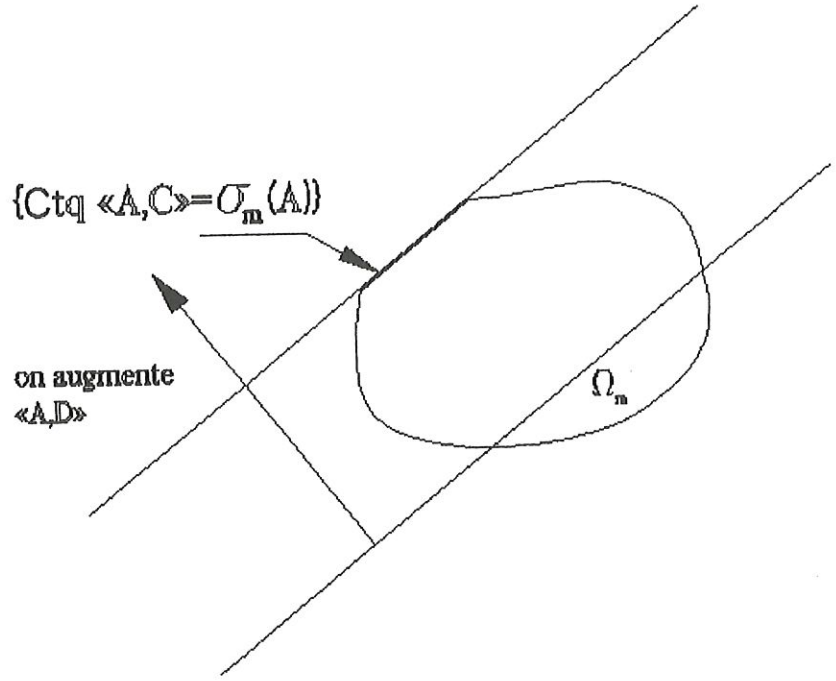
Afin de vérifier la seconde égalité de l'hypothèse, rapelons que $\Omega_m = coR_m$, et $\sigma_m(A) = \max\{\langle\langle A, C\rangle\rangle : C \in \Omega_m\} = \max\{\langle\langle A, P\rangle\rangle : P \in R_m\}$. De sorte que $C \in \partial\sigma_m(A)$ entraîne

$$C = \sum_{i=1}^{k} \alpha_i P_i \text{ avec } \sum_{i=1}^{k} \alpha_i = 1 \ , \ \alpha_i \geq 0 \ , \ P_i \in R_m \text{ pour tout } i \text{ avec } \langle\langle A, C\rangle\rangle = \sigma_m(A).$$

Ce qui implique que $P_i \in R_m$ et $\langle\langle A, P_i\rangle\rangle = \partial\sigma_m(A)$ pour tout i.
Donc $\partial\sigma_m(A) = co\{P \in R_m : \langle\langle A, P\rangle\rangle = \sigma_m(A)\}$, qui, par définition de R_m, est l'ensemble décrit par (1.13). ■

Signification géométrique du théorème 4 :

$\partial\sigma_m(A)$ est l'ensemble des matrices $C \in \Omega_m = \sigma_m(0)$ pour lesquelles la forme linéaire $D \in S^n \rightsquigarrow \langle\langle A, D\rangle\rangle$ atteint son maximum sur Ω_m; cet ensemble est donc la face de Ω_m exposée par A. Ou encore, pour construire cet ensemble, on rassemble tous les points extrêmes de Ω_m (i.e. les éléments de R_m) qui appartiennent à l'hyperplan affine $\{C \in S^n : \langle\langle A, C\rangle\rangle = \sigma_m(A)\}$, et on prend leur enveloppe convexe.



Exemple 1 *(cas extrêmes)*

- pour $m = n$, $\partial\sigma_m(A) = \{I_n\}$ pour tout $A \in S^n$ et nous avons que $\sigma_n = tr$ est une fonction linéaire sur S^n avec $\nabla\sigma_m(A) = I_n$

- pour $m = 1$, la formule du sous différentiel se simplifie un peu :

$$\partial\sigma_1(A) = \partial\lambda_1(A) = co\left\{uu^t : \|u\| = 1 \text{ et } Au = \lambda_1(A)u\right\} \tag{1.14}$$

Pour calculer un sous gradient de λ_1 en A, il suffit de calculer un vecteur propre normalisé u, associé à $\lambda_1(A)$, et de former le produit tensoriel uu^t.
La dérivée directionnelle de $\lambda_1(A)$ s'écrit :

$$\sigma_1'(A,H) = \lambda_1'(A,H) = \max\{\langle Hu, u\rangle : u \text{ est un vecteur propre normalisé, associé à } \lambda_1(A)\}$$

Exemple 2 *(forme particulière de A)*

Considérons une matrice ayant ses éléments fixes hors diagonale. Seuls les éléments diagonaux dépendent du paramètre x. On peut alors réécrire la matrice A sous la forme : $A = A_0 + \Lambda$, où Λ est une matrice n sur n , diagonale dépendant de x. Etudions $\sigma_1(A_0 + \Lambda) = \lambda_1(A_0 + \Lambda)$. Identifiant l'ensemble de ces matrices diagonales à l'ensemble R^n, nous obtenons la fonction

$$x = (\zeta_1, \zeta_2, ..., \zeta_n) \in R^n \rightsquigarrow f(x) \equiv \lambda_1(A_0 + diag(\zeta_1, \zeta_2, ..., \zeta_n) = \lambda_1(A_0 + \Lambda).$$

Elle est la composée de la fonction convexe λ_1 avec une fonction affine dont la partie linéaire est $\mathcal{A}_0$: $x = (\zeta_1, \zeta_2, ..., \zeta_n) \in R^n \rightsquigarrow \mathcal{A}_0(\zeta_1, \zeta_2, ..., \zeta_n) \equiv diag(\zeta_1, \zeta_2, ..., \zeta_n)$.
Par la proposition 3, avec $[J\mathcal{A}_0(x)]^* = \mathcal{A}_0^*$, f est une fonction convexe sur R^n pour laquelle

$$\partial f(x) = \mathcal{A}_0^*\, \partial\lambda_1\left(A_0 + \mathcal{A}_0(x)\right) \text{ pour tout } x \in R^n. \tag{1.15}$$

Déterminons $\mathcal{A}_0^*$. Par la définition de la trace et de l'adjointe, nous avons pour tout $x = (\zeta_1, \zeta_2, ..., \zeta_n) \in R^n$ et pour toute matrice $M \equiv [m_{ij}] \in S^n$:

$$\begin{aligned} \langle\langle \mathcal{A}_0(x), M \rangle\rangle &= \sum_{i=1}^{n} \zeta_i m_{ii} \\ &= \langle x, \mathcal{A}_0^*(M) \rangle . \end{aligned}$$

Donc $\mathcal{A}_0^*$ associe à $M \equiv [m_{ij}]$ le vecteur $(m_{11}, m_{22}, ...m_{nn})$. Pour $M = \partial\lambda_1 (A_0 + \mathcal{A}_0(x))$, $\mathcal{A}_0^*(M) = co(u_1^2, u_2^2, ..., u_n^2)$ (cfr 1.14). Nous réécrivons alors (1.15) sous la forme

$$\partial f(x) = co \left\{ \begin{array}{c} (u_1^2, u_2^2, ..., u_n^2) \text{ où } u = (u_1, u_2, ..., u_n) \\ \text{est un vecteur propre normalisé assosié à } \lambda_1 (A_0 + \mathcal{A}_0(x)) \text{ i.e. à } f(x) \end{array} \right\} .$$

Le cas plus général de la fonction $f_m : x \rightsquigarrow f_m(x) \equiv \sigma_m(A_0 + diag(\zeta_1, \zeta_2, ..., \zeta_n))$ est traité de la même manière : f_m est une fonction **convexe** *et $\partial f_m(x) = \mathcal{A}_0^* \, \partial\sigma_m (A_0 + \mathcal{A}_0(x))$ peut être écrit en utilisant le théorème 4.*

L'expression de $\sigma_m(A)$ du théorème 4 est dite *implicite* à cause de la contrainte affine $\langle\langle A, C \rangle\rangle = \sigma_m(A)$ ou $\langle\langle A, XX^t \rangle\rangle = \sigma_m(A)$ y apparaissant. En posant

$$M_m(A) \equiv \left\{ X \in M^{n,m}(\mathrm{R}) : X^t X = I_m \text{ et } \left\langle\left\langle A, XX^t \right\rangle\right\rangle = \sigma_m(A) \right\}, \tag{1.16}$$

(1.13) devient

$$\partial\sigma_m(A) = co \left\{ XX^t : X \in M_m(A) \right\} .$$

L'objectif de la prochaine section est de donner une expression plus *explicite* aux matrices X de $M_m(A)$ intervenant dans la formulation (1.13) de $\partial\sigma_m(A)$.

1.4.2 Expression plus explicite de $\sigma_m(A)$

Trois indices fondamentaux

Rappelons que nous notons λ_i les valeurs propres de A.

Définissons trois entiers i_m, j_m et r_m comme suit (leur dépendance en A est sous-entendue) :

$$\lambda_1 \geq ... \geq \lambda_{m-i_m} > \lambda_{m-i_m+1} = ... = \lambda_m = \lambda_{m+1} = ... = \lambda_{m+j_m} > \lambda_{m+j_m+1} \geq ... \geq \lambda_n \quad (1.17)$$

C'est-à-dire : r_m est la multiplicité de λ_m; j_m est le nombre de valeurs propres égales à λ_m, rangées strictement après λ_m; i_m est le nombre de valeurs propres égales à λ_m, rangées avant λ_m, λ_m compris.

Par exemple, pour $\lambda_1 \geq \lambda_2 > \lambda_3 = \lambda_4 = \lambda_5 = \lambda_6 = \lambda_7 > \lambda_8$,avec $m = 5$ nous obtenons : $i_m = 3$, $j_m = 2$ et $r_m = 5$.

Pour les valeurs propres extrêmes λ_1 et λ_n (i.e. pour m fixé à 1 et n), nous avons : $i_1 = 1$, $j_1 = r_1 - 1$ et $i_n = r_n$, $j_n = 0$.

Lorsque λ_m est une valeur propre simple (i.e. $r_m = 1$), $i_m = 1$, $j_m = 0$.

Il est facile de voir que pour tout $1 \leq m \leq n$, $i_m \geq 1$, $j_m \geq 0$ et $i_m + j_m = r_m$.

Nous commençons par démontrer deux lemmes techniques.

Lemme 1 *Soient α_1,...,α_n des réels vérifiant*

$$\alpha_i \in [0,1] \text{ pour tout } i = 1, ..., n \text{ et } \sum_{i=1}^{n} \alpha_i = m. \quad (1.18)$$

Alors,

$$\sum_{i=1}^{n} \alpha_i \lambda_i \leq \sum_{i=1}^{m} \lambda_i.$$

L'égalité est vérifiée si et seulement si :

$$\begin{aligned} &\alpha_i = 1 \text{ pour } i = 1, ..., m - i_m; \\ &\alpha_i = 0 \text{ pour } i > m + j_m; \\ &\sum_{k=1}^{r_m} \alpha_{m-i_m+k} = i_m. \end{aligned} \quad (1.19)$$

Preuve. Soit Γ_m le polyèdre convexe, compact de R^n décrit par les contraintes (1.18), et considérons le programme linéaire $\mathcal{P}(\lambda_m)$ suivant :

$$\begin{cases} \max \sum_{i=1}^{n} \alpha_i \lambda_i = \langle \alpha, \lambda \rangle \text{ où } \alpha = (\alpha_1, ..., \alpha_n) \text{ et } \lambda = (\lambda_1, ..., \lambda_n) \\ s.c.\ \alpha_i \in \Gamma_m \end{cases}$$

Par exemple pour $n = 3$ et $m = 2$, $\mathcal{P}(\lambda_2)$ s'écrit:

$$\begin{cases} \max \alpha_1\lambda_1 + \alpha_2\lambda_2 + \alpha_3\lambda_3 = \langle \alpha, \lambda \rangle \\ \text{s.c.} \quad \alpha_i \in [0,1] \text{ pour tout } i = 1,2,3 \text{ et } \sum_{i=1}^{3} \alpha_i = 2 \end{cases}$$

L'ensemble des points extrêmes (ou sommets) de Γ_m est formé de tous les vecteurs de R^n dont m coordonnées valent 1 et les autres 0. Il en découle que

$$\max_{\alpha_i \in \Gamma_m} \sum_{i=1}^{n} \alpha_i \lambda_i = \sum_{i=1}^{m} \lambda_i,$$

et $(1, ..., 1, 0, ..., 0)$ est *un* des sommets de Γ_m résolvant $\mathcal{P}(\lambda_m)$.

Il reste à déterminer l'ensemble optimal de $\mathcal{P}(\lambda_m)$, i.e. la face de Γ_m exposée par le vecteur $\lambda = (\lambda_1, ..., \lambda_n)$:

$$\left\{ (\alpha_1, ..., \alpha_n) \in \Gamma_m : \sum_{i=1}^{n} \alpha_i \lambda_i = \sum_{i=1}^{m} \lambda_i \right\}.$$

Par (1.17), nous obtenons de tels α_i optimaux en prenant pour les $m - i_m$ premières composantes de α la valeur 1, pour les $i \geq m + j_m + 1$, la valeur 0, et en distribuant le reste parmi les r_m coefficients rangés de $m - i_m + 1$ à $m + j_m$. Ceci est exactement ce qu'exprime (1.19). ■

Lemme 2 *Soit $X \in M^{n,m}(R)$ vérifiant $X^tX = I_m$, et notons par $x_i \equiv (x_{i1}, x_{i2}, ..., x_{in})^t$ la $i-$ème colonne de X. Alors*

$$\sum_{i=1}^{m} x_{ij}^2 \leq 1 \text{ pour tout } j = 1, ..., n.$$

Preuve. Comme $X^tX = I_m$, les m colonnes de X, $x_1, x_2, ..., x_m$ sont des vecteurs orthonormaux de R^n. Il existe donc $n-m$ vecteurs, $x_{m+1}, ..., x_n$ tels que la matrice complétée $\mathcal{X} = [x_1, ..., x_m, x_{m+1}, ..., x_n]$ soit orthogonale. Donc $\mathcal{X}^t\mathcal{X} = I_n$ et les vecteurs lignes de $\mathcal{X}$ sont normés. C'est-à-dire, avec $m \leq n$,

$$\sum_{i=1}^{m} x_{ij}^2 \leq \sum_{i=1}^{n} x_{ij}^2 = 1 \text{ pour tout } j = 1, ..., n.$$

■

A présent, nous sommes prêts à donner une forme plus explicite à l'ensemble $M_m(A)$.
A étant une matrice symétrique, peut être diagonalisée par une matrice orthogonale U :

$$U^tAU = diag(\lambda_1, ..., \lambda_n). \tag{1.20}$$

Proposition 5 *(réécriture de $M_m(A)$.)*

Posons

$$Q_m(A) \equiv \left\{ \begin{array}{c} H = [h_1, ..., h_m] \in M^{n,m}(R) : H^tH = I_m \text{ et } \sum_{i=1}^{m} h_{ij}^2 = 1 \text{ pour } j = 1, ..., m - i_m; \\ \sum_{j=m-i_m+1}^{m+j_m} \left(\sum_{i=1}^{m} h_{ij}^2 \right) = i_m; \ h_{ij} = 0 \text{ pour tout } 1 \leq i \leq m \text{ et } j \geq m + j_m + 1 \end{array} \right\}.$$

Nous avons : $X \in M_m(A)$ si et seulement si $X = UH$ pour $H \in Q_m(A)$.

Preuve. $\Leftarrow$: Soit $X = UH$ avec $H = [h_1, ..., h_m] \in Q_m(A)$.
Montrons que $X \in M_m(A) \equiv \{X \in M^{n,m}(R) : X^tX = I_m \text{ et } \langle\langle A, XX^t \rangle\rangle = \sigma_m(A)\}$.

- Comme U est une matrice orthogonale, $X^tX = H^tU^tUH = H^tH = I_m$. La première contrainte dans la définition de $M_m(A)$ est vérifiée.

- Comme $X = UH$, nous avons
  $\langle\langle A, XX^t \rangle\rangle = \langle\langle A, UHH^tU^t \rangle\rangle = \langle\langle U^tAU, HH^t \rangle\rangle = \langle\langle diag(\lambda_1, ..., \lambda_n), HH^t \rangle\rangle$ (cfr 1.20).
  Il nous reste à voir que

$$\left\langle\left\langle diag(\lambda_1, ..., \lambda_n), HH^t \right\rangle\right\rangle = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \lambda_j h_{ij}^2 \right) \tag{1.21}$$

$= \sum_{j=1}^{m+j_m} \lambda_j \left(\sum_{i=1}^{m} h_{ij}^2 \right)$ en tenant compte de $h_{ij} = 0$ pour tout $1 \leq i \leq m$ et $j \geq m + j_m + 1$;

$= \sum_{j=1}^{m-i_m} \lambda_j \left(\sum_{i=1}^{m} h_{ij}^2 \right) + \sum_{j=m-i_m+1}^{m+j_m} \lambda_m \left(\sum_{i=1}^{m} h_{ij}^2 \right)$ car $\lambda_j = \lambda_m$ pour $j = m - i_m + 1, ..., m + j_m$;

$= \sum_{j=1}^{m-i_m} \lambda_j .1 + \lambda_m . i_m = \sum_{j=1}^{m} \lambda_j$, après avoir utilisé le fait que $[h_1, ..., h_m] \in Q_m(A)$.

Nous avons prouvé que $\langle\langle A, XX^t \rangle\rangle = \sigma_m(A)$ et la première implication est ainsi vérifiée.

$\Rightarrow$: Prenons $X \in M_m(A)$. Démontrons que $H \equiv U^t X \in Q_m(A)$. Nous aurons bien $X = UH$ avec $H \in Q_m(A)$.

- $H^t H = X^t X = I_m$.
- Pour $X \equiv UH$, avec $\sigma_m(A) = \langle\langle A, XX^t \rangle\rangle$, nous écrivons

$$\sum_{j=1}^{m} \lambda_j = \left\langle\left\langle A, XX^t \right\rangle\right\rangle = \left\langle\left\langle U^t AU, HH^t \right\rangle\right\rangle = \left\langle\left\langle diag(\lambda_1, ..., \lambda_n), HH^t \right\rangle\right\rangle$$
$$= \sum_{j=1}^{n} \lambda_j \left(\sum_{i=1}^{m} h_{ij}^2 \right) \text{ (cfr 1.21)}$$

En vue d'appliquer le Lemme 1, posons $\alpha_j = \left(\sum_{i=1}^{m} h_{ij}^2 \right)$ pour tout $j = 1, ..., n$. En tenant compte de $H^t H = I_m$, remarquons que :

$$0 \leq \alpha_j \leq 1 \text{ pour tout } j \text{ (ceci résulte du Lemme 2);}$$
$$\sum_{j=1}^{n} \alpha_j = tr(HH^t) = tr(H^t H) = m$$

Nous sommes bien dans le contexte du Lemme 1.

Sachant que $\sum_{j=1}^{m} \lambda_j = \sum_{j=1}^{n} \lambda_j \alpha_j$, nous déduisons par la deuxième partie du Lemme 1 :

$$\alpha_j = 1 \text{ pour } j = 1, ..., m - i_m;$$
$$\alpha_j = 0 \text{ pour } j \geq m + j_m + 1 \text{ et } i = 1...m;$$
$$\sum_{k=1}^{r_m} \alpha_{m-i_m+k} = \sum_{j=m-i_m+1}^{m+j_m} \alpha_j = i_m.$$

Ainsi $H \in Q_m(A)$. ■

Le corollaire suivant met en évidence un sous ensemble de $M_m(A)$, facile à manipuler. Nous verrons qu'il joue un rôle essentiel pour clarifier l'expression de $\partial\sigma_m(A)$.

Corollaire 1 *Soit $Z_m(A)$ l'ensemble des matrices $Z \in M^{n,m}(R)$ de la forme suivante :*

$$Z = \begin{bmatrix} I_k & 0 \\ 0 & W \\ 0 & 0 \end{bmatrix}, \text{ avec } W \in M^{r_m,i_m}(R),\ W^tW = I_{i_m} \text{ et } k = m - i_m. \tag{1.22}$$

Alors

$$\{UZ : Z \in Z_m(A)\} \subset M_m(A).$$

Preuve. Par la proposition 5, il suffit de voir que $Z \equiv [z_{ij}] \in Q_m(A)$: $Z^tZ = I_m$ et $\sum_{i=1}^{m} z_{ij}^2 = 1$ pour $j = 1, ..., m - i_m$; $\sum_{j=m-i_m+1}^{m+j_m} \left(\sum_{i=1}^{m} z_{ij}^2 \right) = tr(W^tW) = i_m$; $z_{ij} = 0$ pour tout $1 \leq i \leq m$ et $j \geq m + j_m + 1$. ■

A présent, voici le principal résultat concernant la formule explicite de $\partial\sigma_m(A)$.

Théorème 5 *(Formule explicite)*

$$\partial\sigma_m(A) = co\left(U \begin{bmatrix} I_{m-i_m} & 0 & 0 \\ 0 & WW^t & 0 \\ 0 & 0 & 0 \end{bmatrix} U^t : W \in M^{r_m,i_m}(R),\ W^tW = I_{i_m} \right). \tag{1.23}$$

La matrice diagonale par blocs, $diag(I_{m-i_m}, WW^t, 0)$, apparaissant dans le membre de droite de (1.23) consiste en trois blocs : I_{m-i_m}, une matrice WW^t de dimension r_m sur r_m, et un bloc de dimension $n-(m+j_m)$ formé de zéros. Elle est donc de la forme ZZ^t avec Z vérifiant (1.22). Ainsi, (1.23) revient à

$$\partial\sigma_m(A) = co\left\{ UZZ^tU^t : Z \in Z_m(A) \right\}.$$

Preuve. $\supset$: par la définition de $M_m(A)$ (cfr 1.16), nous avons :

$$\partial\sigma_m(A) = co\left\{ XX^t : X \in M_m(A) \right\}.$$

En utilisant le corollaire 1, tout UZ avec $Z \in Z_m(A)$ appartient à $M_m(A)$. La première inclusion est alors vérifiée: $co\{XX^t : X \in M_m(A)\} \supset co\{UZZ^tU^t : Z \in Z_m(A)\}$.

$\subset$: par la proposition 5 nous écrivons

$$\partial\sigma_m(A) = co\left\{UHH^tU^t : H \in Q_m(A)\right\}.$$

Il nous faut donc prouver:

$$\left\{UHH^tU^t : H \in Q_m(A)\right\} \subset \left\{UZZ^tU^t : Z \in Z_m(A)\right\}.$$

Considérons une matrice HH^t avec $H \in Q_m(A)$. Notre objectif est de montrer qu'elle peut se réécrire sous la forme ZZ^t avec Z vérifiant (1.22).
Notant $p \equiv m + j_m$, rappelons les propriétés vérifiées par $H = [h_1, ..., h_m] \in Q_m(A)$ (cfr proposition 5) :

$$h_{ij} = 0 \text{ pour tout } 1 \leq i \leq m \text{ et } j > p; \tag{1.24}$$

$$\sum_{i=1}^{m} h_{ij}^2 = 1 \text{ pour } j = 1, ..., m - i_m; \tag{1.25}$$

$$H^tH = I_m. \tag{1.26}$$

- Nous définissons le "vecteur tronqué" de h_i, $h_i^0 \in \mathrm{R}^p$ comme suit :

$$h_{ij}^0 = h_{ij} \text{ pour } i = 1, ..., m \text{ et } j = 1, ..., p. \tag{1.27}$$

Par (1.24) et (1.26), $h_1^0, ..., h_m^0$ sont des vecteurs orthonormaux de R^p. Nous les complétons de j_m vecteurs $h_{m+1}^0, ..., h_p^0$ dans R^p de sorte que $\left\{h_1^0, ..., h_p^0\right\}$ forme une base orthonormale de R^p. D'où, la matrice $H^0 \equiv \left[h_1^0, ..., h_p^0\right]$ est une matrice orthogonale, et

$$\left(H^0(H^0)^t\right)_{jk} = \sum_{i=1}^{p} h_{ij}^0 h_{ik}^0 = \begin{cases} 1 \text{ si } j = k \\ 0 \text{ si } j \neq k \end{cases} \tag{1.28}$$

- Réécrivons le premier bloc diagonal de HH^t.
  L'équation (1.25) implique que $\sum_{i=1}^{m} \left(h_{ij}^0\right)^2 = 1$ pour $j = 1, ..., m - i_m$, donc en considérant l'orthogonalité de H^0 :

$$h_{ij}^0 = 0 \text{ pour } i = m+1, ..., p \equiv m + j_m \text{ et } j = 1, ..., m - i_m. \tag{1.29}$$

  En vue de (1.27) et (1.28), cette dernière égalité donne

$$\left(HH^t\right)_{jk} = \sum_{i=1}^{m} h_{ij}h_{ik} = \sum_{i=1}^{m} h_{ij}^0 h_{ik}^0 = \sum_{i=1}^{p} h_{ij}^0 h_{ik}^0 = (I_{m-i_m})_{jk} \text{ pour } \mathbf{j, k = 1, ..., m - i}_m.$$

- Avant de s'attaquer au second bloc de HH^t, définissons $s \equiv m - i_m$ et j_m vecteurs $v_1, ..., v_{j_m}$ dans R^{r_m} comme ceci :

$$(v_i)_j \equiv v_{ij} \equiv h_{m+i,s+j}^0 \text{ pour } i = 1, ..., j_m \text{ et } j = 1, ..., r_m.$$

  A nouveaux nous les complétons par i_m vecteurs $w_1, ..., w_{i_m}$ de sorte que l'ensemble $\{v_1, ..., v_{j_m}, w_1, ..., w_{i_m}\}$ forme une base orthonormale de R^{r_m}. Soient

$$V \equiv [v_1, ..., v_{j_m}] \text{ et } W \equiv [w_1, ..., w_{i_m}].$$

  La matrice $[V, W]$ est othogonale par construction.

- Occupons-nous du second bloc de HH^t.
  Calculons $\left(HH^t\right)_{jk}$ pour $s + 1 \leq j, k \leq p$ $(s = m - i_m)$. Dans ce cas, h_{ij}^0 est renoté v_{ij}.
  Nous déduisons tout d'abord de (1.27) et (1.28) la relation :

$$\sum_{i=1}^{m} h_{ij}h_{ik} = \begin{cases} 1 - \sum_{i=m+1}^{p} h_{ij}^0 h_{ik}^0 \text{ si } j = k \\ 0 - \sum_{i=m+1}^{p} h_{ij}^0 h_{ik}^0 \text{ si } j \neq k \end{cases} \tag{1.30}$$

Mais par l'orthogonalté de $[V, W]$ et par définition de V, nous avons :

$$\sum_{i=1}^{i_m} w_{ij}w_{ik} = \begin{cases} 1 - \sum_{i=1}^{j_m} v_{ij}v_{ik} = 1 - \sum_{i=m+1}^{p} h_{ij}^0 h_{ik}^0 \text{ si } j = k. \\ 0 - \sum_{i=1}^{j_m} v_{ij}v_{ik} = 0 - \sum_{i=m+1}^{p} h_{ij}^0 h_{ik}^0 \text{ si } j \neq k. \end{cases} \tag{1.31}$$

Il reste à comparer (1.30) et (1.31) pour obtenir :

$$\left(HH^t\right)_{jk} = \sum_{i=1}^{m} h_{ij}h_{ik} = \sum_{i=1}^{i_m} w_{ij}w_{ik} \text{ pour } \mathbf{j, k = s+1, ..., p.}$$

Le deuxième bloc diagonal de HH^t est bien la matrice de dimension $r_m = p - s$ sur $r_m = p - s$, WW^t.

- Pour tous les autres indices j et k, on vérifie facilement que $\left(HH^t\right)_{jk} = \left(HH^t\right)_{kj} = 0$.
  Par exemple, pour $j = 1, ..., m - i_m$ et $k = s + 1, ..., p$ nous avons, grâce aux équations (1.29) et (1.28) :

$$\left(HH^t\right)_{jk} = \sum_{i=1}^{m} h_{ij}h_{ik} = \sum_{i=1}^{p} h_{ij}^0 h_{ik}^0 = 0.$$

Nous sommes donc parvenus à réécrire HH^t sous la forme de ZZ^t avec :

$$Z = \begin{bmatrix} I_s & 0 \\ 0 & W \\ 0 & 0 \end{bmatrix}, \; W \in M^{r_m, i_m}(\mathrm{R}), W^tW = I_{i_m}.$$

■

Comme corollaires de l'expression explicite de $\partial\sigma_m(A)$, nous obtiendrons l'expression de la dérivée directionnelle $\sigma_m^{'}(A, .)$ de σ_m en A, et étudierons le cas où σ_m est différentiable. Pour cela, décomposons la matrice $U = [u_1, ..., u_n]$ qui diagonalise A (cfr 1.20), de façon à mettre en évidence deux sous matrices :

$$U_1 = [u_1, ..., u_{m-i_m}] \text{ et } U_2 = [u_{m-i_m+1}, ..., u_{m+j_m}]. \tag{1.32}$$

Nous déduisons alors du Théorème 5 :

Corollaire 2 *La dérivée directionnelle de σ_m en A dans la direction $H \in S^n$ est donnée par*

$$\sigma_m^{'}(A,H) = tr\left(U_1^t H U_1\right) + \sum_{k=1}^{i_m} \mu_k\left(U_2^t H U_2\right), \tag{1.33}$$

où $\mu_k\left(U_2^t H U_2\right)$ dénote la k-ème plus grande valeur propre de la matrice $U_2^t H U_2$.

La relation (1.33) contient donc deux parties :
- une partie *linéaire* : $H \rightsquigarrow$ la trace de la matrice $U_1^t H U_1$ de dimension $m - i_m$ sur $m - i_m$;
- une partie *convexe* : $H \rightsquigarrow$ la somme des i_m plus grandes valeurs propres de la matrice $U_2^t H U_2$ de dimension r_m sur r_m (appliquer le théorème 2 à la matrice $U_2^t H U_2$, avec $m = i_m$).

Preuve. Par définition

$$\begin{aligned} \sigma_m^{'}(A,H) &= \max\{\langle\langle C, H\rangle\rangle : C \in \partial\sigma_m(A)\} \\ &= \max\{\langle\langle C, H\rangle\rangle : C \in \Delta_m(A)\}, \end{aligned}$$

où $\Delta_m(A)$ est un sous ensemble de $M^n(\mathrm{R})$ tel que $co\left(\Delta_m(A)\right) = \partial\sigma_m(A)$. Utilisant l'expression (1.23) de $\partial\sigma_m(A)$ pour construire un tel $\Delta_m(A)$, il suit

$$\max\left\{ tr\left(U \begin{bmatrix} I_{m-i_m} & 0 & 0 \\ 0 & WW^t & 0 \\ 0 & 0 & 0 \end{bmatrix} U^t.H \right) : W \in M^{r_m,i_m}(\mathrm{R}),\ W^tW = I_{i_m} \right\}$$

$$= \max\left\{ tr\left(\left[U_1U_1^t + (U_2W)(U_2W)^t \right] H \right) : W \in M^{r_m,i_m}(\mathrm{R}),\ W^tW = I_{i_m} \right\}$$

$$= tr\left(U_1 H U_1^t\right) + \max\left\{ tr\left(U_2^t H U_2 W W^t\right) : W \in M^{r_m,i_m}(\mathrm{R}),\ W^tW = I_{i_m} \right\}.$$

Par (1.3), la seconde partie de l'expression ci -dessus vaut la somme des i_m plus grandes valeurs propres de la matrice $U_2^t H U_2$. ■

Remarque 3 *Observons que seuls les deux premiers blocs de la matrice U, diagonalisant A, apparaissent dans les expressions (1.23) et (1.33) de $\partial\sigma_m(A)$ et $\sigma_m^{'}(A,.)$: le troisième bloc $U_3 = [u_{m+j_m+1}, ..., u_n]$ n'intervient pas. Ceci s'explique en quelque sorte, par la manière dont*

sont rangées les valeurs propres de A. Ainsi la sensibilité de $\lambda_1 + ... + \lambda_m$ *fait appel aux valeurs propres de A ne dépassant pas le rang* $m + j_m$.

La différentiabilité de la fonction convexe σ_m en A peut être établie en utilisant les caractérisations suivantes :

- $\Delta_m(A)$ est un *singleton*, où $\Delta_m(A)$ est un sous ensemble de $M^n(\mathrm{R})$ tel que $co\,(\Delta_m(A)) = \partial\sigma_m(A)$;
- la fonction positivement homogène et convexe $\sigma'_m(A, .)$ est *linéaire.*

Il découle donc du Théorème 5 :

Corollaire 3 *(différentiabilité de* $\partial\sigma_m(A)$*)*
Si $j_m = 0$, *i.e. si m est la dernière place dans un groupe de valeurs propres égales, alors* σ_m *est différentiable en A et*

$$\Delta_m(A) = \nabla\sigma_m(A) = U_1U_1^t + U_2U_2^t.$$

Preuve. La condition $j_m = 0$ signifie que $\lambda_m > \lambda_{m+1}$ ou que $m = n$; donc $i_m = r_m$ de sorte que l'ensemble $\Delta_m(A)$, dont l'enveloppe convexe donne lieu à $\partial\sigma_m(A)$ dans le Théorème 5, est réduit au singleton $\{U_1U_1^t + U_2U_2^t\}$. ■

Conséquence du corollaire

Par définition, $m' \equiv m + j_m$ correspond à la dernière position dans un groupe de valeurs propres qui sont toutes égales à λ_m; et $m'' \equiv m - i_m$ numérote la dernière place dans le groupe juste précédent, contenant des valeurs propres égales. La conclusion du corollaire 3 donne

$$\sigma_{m'}(A) - \sigma_{m''}(A) = \lambda_{m-i_m+1} + ... + \lambda_m + ... + \lambda_{m+j_m} \text{ est une fonction différentiable en } A$$

dont le gradient est $\nabla\sigma_{m'}(A) - \nabla\sigma_{m''}(A)$.
Sans difficulté alors, nous pouvons établir le corollaire suivant.

Corollaire 4 *La fonction* $C \in S^n \rightsquigarrow \tau_m(C) \equiv (\lambda_{m-i_m+1} + ... + \lambda_m + ... + \lambda_{m+j_m})\,(C)$ *(i.e. la somme des* r_m *fonctions "valeurs propres" qui coïncident en A) est une fonction différentiable*

en A avec

$$\nabla\tau_m(A) = U_1U_1^t + U_2U_2^t - V_1V_1^t - V_2V_2^t$$

où U_1, U_2 (resp. V_1, V_2) sont les deux premières matrices dans la décomposition de U par blocs, correspondant à l'indice $m + j_m$ (resp. $m - i_m$).

1.4.3 Dérivée directionnelle de la m-ème valeur propre λ_m

Comme $\lambda_1 = \sigma_1$ et $\lambda_m = \sigma_m - \sigma_{m-1}$ pour $m \geq 2$, la dérivée directionnelle de λ_m est directement construite à partir de celle de σ_m :

$$\lambda_1'(A,.) = \sigma_1'(A,.) \text{ et } \lambda_m'(A,.) = \sigma_m'(A,.) - \sigma_{m-1}'(A,.) \text{ pour tout } m \geq 2;$$

et les développements du premier ordre, (1.10) et (1.11), s'appliquent aux fonctions $\lambda_m(.)$.

Nous allons dériver de l'expression explicite de $\sigma_m'(A,.)$, celle de $\lambda_m'(A,.)$.

Théorème 6 *(expression explicite de $\lambda_m'(A,.)$)*
la dérivée directionnelle de λ_m en A dans la direction $H \in S^n$ est la i_m-ème valeur propre de $U_2^tHU_2$:

$$\lambda_m'(A,H) = \mu_{i_m}(U_2^tHU_2).$$

Preuve.

- $m = 1$. Alors $i_1 = 1$, $j_1 = r_1 - 1$, U_1 est vide, $U_2 = [u_{1,\ldots,}u_{r_1}]$ et

$$\lambda_1'(A,H) = \sigma_m'(A,H) = \mu_1(U_2^tHU_2) \qquad \text{(cfr 1.33)}.$$

- $m \geq 2$.

  - Considérons dabord le cas où $i_m \geq 2$, i.e. $\lambda_{m-1} = \lambda_m$.

  Clairement, $i_{m-1} = i_m - 1$, $j_{m-1} = j_m + 1$ et $r_{m-1} = r_m$. La décomposition (1.32) de U en blocs $[U_1, U_2]$ est donc la même pour l'indice m ou $m-1$. C'est pourquoi

$$\sigma'_m(A,H)-\sigma'_{m-1}(A,H) = tr\,(U_1^tHU_1)+\sum_{k=1}^{i_m}\mu_k\,(U_2^tHU_2)-tr\,(U_1^tHU_1)-\sum_{k=1}^{i_{m-1}}\mu_k\,(U_2^tHU_2)$$
$$= \mu_{i_m}(U_2^tHU_2).$$

- A présent, si $i_m = 1$, i.e. $\lambda_{m-1} > \lambda_m$, nous avons $j_{m-1} = 0$. Cette fois les décompositions (1.32) relatives à m et $m-1$ font apparaître les blocs suivants :

$$U_1(m) = [u_{1,\ldots\ldots,}u_{m-1}] \quad \text{et} \quad U_2(m) = [u_{m,\ldots\ldots,}u_{m+j_m}],$$
$$U_1(m-1) = [u_{1,\ldots\ldots,}u_{m-1-i_{m-1}}] \quad \text{et} \quad U_2(m-1) = [u_{m-i_{m-1},\ldots\ldots,}u_{m-1}],$$

ce qui nous permet de noter que $[U_1(m-1), U_2(m-1)] = U_1(m)$.
Par le corollaire 3, avec $j_{m-1} = 0$ et $\Delta_{m-1}(A) = U_1(m-1)U_1^t(m-1)+U_2(m-1)U_2^t(m-1)$, σ_{m-1} est différentiable en A et

$$\begin{aligned}\sigma'_{m-1}(A,H) &= max\,\{\langle\langle C,H\rangle\rangle : C \in \Delta_{m-1}(A)\}\\ &= tr\left\{[U_1(m-1)]^t\,H\,[U_1(m-1)]\right\} + tr\left\{[U_2(m-1)]^t\,H\,[U_2(m-1)]\right\}\\ &= tr\left\{[U_1(m)]^t\,H\,[U_1(m)]\right\}.\end{aligned}$$

D'autre part, en se servant du corollaire 2 :

$$\sigma'_m(A,H) = tr\left\{[U_1(m)]^t\,H\,[U_1(m)]\right\} + \mu_1\left\{[U_2(m)]^t\,H\,[U_2(m)]\right\}.$$

Par conséquent, $\lambda'_m(A,H) = \sigma'_m(A,H)-\sigma'_{m-1}(A,H) = \mu_1\left\{[U_2(m)]^t\,H\,[U_2(m)]\right\}$, ce qu'il fallait démontrer. ■

Comportements de λ_m.

- Quand λ_m occupe la première place dans un groupe de valeurs propres égales, i.e. quand $\mathbf{i_m = 1}$, $\lambda'_m(A,H)$ est *la plus grande* valeur propre d'une matrice r_m sur $r_m\,U_2^tHU_2$ (cfr Théorème 6) on peut dire que la sensibilité au premier ordre de λ_m est modelée sur celle de la plus grande valeur propre λ_1 (puisque $\lambda'_1(A,H) = \mu_1\{U_2^tHU_2\}$). C'est donc une situation où $\lambda'_m(A,.)$ est *convexe.*

- Quand λ_m occupe la dernière place dans un groupe de valeurs propres égales, i.e. quand $\mathbf{j_m = 0}$, ou encore lorsque $i_m = r_m$, $\lambda'_m(A,H)$ est *la plus petite* valeur propre de $U_2^tHU_2$ (puisque la matrice $U_2^tHU_2$ ne possède pas plus de r_m valeurs propres rangées dans l'ordre décroissant) on

peut dire que λ_m imite la plus petite valeur propre λ_{r_m} (puisque $\lambda'_{r_m}(A,H) = \mu_{r_m}\{U_2^t H U_2\}$). C'est une situation où $\lambda'_m(A,.)$ est *concave.*

- Quand λ_m se situe à l'interieur d'un groupe de valeurs propres égales, i.e. quand $\mathbf{i}_m \geq \mathbf{2}$ et $\mathbf{j}_m \geq \mathbf{1}$, $\lambda'_m(A,H)$ est une valeur propre intermédiaire de $U_2^t H U_2$ (dont l'indice est le nombre de valeurs propres égales à λ_m et rangées avant m). Nous sommes dans le cas où $\lambda'_m(A,H)$ est la *différence de fonctions convexes.*

1.5 Gradient généralisé et dérivée directionnelle de $f_m(x)$

Dans cette section, nous tournons notre attention sur le calcul du gradient généralisé et de la dérivée directionnelle de $x \rightsquigarrow f_m(x) = \sigma_m[A(x)]$ où $A(x)$ est une matrice symétrique dépendant de façon différentiable d'un paramètre x. Plus précisément, soit Θ un ouvert de l'espace Euclidien R^p, et $A(.) : x \in \Theta \rightsquigarrow A(x) = [a_{ij}(x)] \in S^n$, une fonction de classe C^1; nous avons :

$$f_m = \sigma_m \circ A : \underbrace{\Theta \subset \mathrm{R}^p \xrightarrow{A(.)} S^n}_{(1)} \underbrace{\xrightarrow{\sigma_m} \mathrm{R}}_{(2)}$$

La composante (2) est intrinsèque au problème de valeur propre, indépendamment du fait que $A(x)$ varie en fonction du paramètre x; cette fonction σ_m a été étudiée en détail dans la section précédente. La composante (1) exprime la façon dont se comportent les entrées $a_{ij}(x)$ de $A(x)$ en fonction de $x \in \Theta$: elle concerne l'évaluation de la matrice $A(.)$, indépendamment des fonctions de $A \in S^n$ que nous allons étudier.

Comme indiqué dans la section 1.3, il n'y a pas de difficulté majeure à évaluer $\partial f_m(x)$ (resp. $f'_m(x,.)$) : une fois que nous connaissons $\partial\sigma_m[A(x)]$ (resp. $\sigma'_m[A(x)]$), il reste juste à appliquer la proposition 3.

Comme en (1.32), désignons par U_1 et U_2 les deux premières sous matrices de la matrice orthogonale U, diagonalisant $A(x)$ pour le vecteur $x \in \Theta$:

$$U = [u_1(x), ..., u_n(x)]\,,\ U^t A(x) U = diag\,(\lambda_1(x), ..., \lambda_n(x))\,;$$
$$U_1 = [u_1(x), ..., u_{m-i_m}(x)] \text{ et } U_2 = [u_{m-i_m+1}(x), ..., u_{m+j_m}(x)]\,.$$

(on comprend bien que U, i_m, j_m, r_m dépendent de x à travers $A(x)$. U_1 et U_2 dépendent de x

et m). Pour des raisons de facilité, notons pour tout $j = 1, ..., p$, la dérivée partielle de $A(.)$ par rapport à la j-ème variable x_j, par

$$A_j^{'}(x) \equiv \left[\frac{\partial a_{ik}}{\partial x_j}(x)\right]_{1\leq i,k\leq n} \in S^n.$$

1.5.1 Gradient généralisé de f_m

Théorème 7 *(expression explicite de $\partial f_m(x)$)*
Avec les hypothèses et notations ci-dessus, f_m est localement Lipschitzienne sur Θ et le gradient généralisé $\partial f_m(x)$ de f_m en $x \in \Theta$ est l'enveloppe convexe de l'ensemble suivant :

$$\Delta_m(x) \equiv \left\{(\eta_1, ..., \eta_p) \in R^p : \eta_j = tr\left(U_1^t A_j^{'}(x) U_1\right) + tr\left(\left[U_2 W\right]^t A_j^{'}(x)\left[U_2 W\right]\right)\right\}$$

où $W \in M^{r_m, i_m}(R), W^t W = I_{i_m}$.

Preuve. $f_m = \sigma_m \circ A$. Comme σ_m est une fonction convexe, elle est strictement tangentiellement convexe en $A(x)$, ce qui donne par la proposition 3 :

$$\partial f_m(x) = [\mathrm{J}A(x)]^* \, \partial\sigma_m(A(x)), \tag{1.34}$$

où $\mathrm{J}A(x) : \underset{(\zeta_1,...,\zeta_p)\rightsquigarrow \sum_{j=1}^{p} \zeta_j . A_j^{'}(x)}{R^p \longrightarrow S^n}$ est l'opérateur Jacobien de A en x.

- Par le Théorème 5, nous savons que $\partial\sigma_m\left(A(x)\right)$ est l'enveloppe convexe des matrices S de la forme

$$S = U \begin{bmatrix} I_{m-i_m} & 0 & 0 \\ 0 & WW^t & 0 \\ 0 & 0 & 0 \end{bmatrix} U^t = \begin{bmatrix} U_1 U_1^t & 0 & 0 \\ 0 & U_2 W \left[U_2 W\right]^t & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{1.35}$$

  avec $W \in M^{r_m, i_m}(\mathrm{R})$, $W^t W = I_{i_m}$.

- Pour obtenir $\partial f_m(x)$, il reste à évaluer $[\mathrm{J}A(x)]^* S$.
  Par définition de $\mathrm{J}A(x)$ et de son adjointe, nous avons pour tout $(\zeta_1, ..., \zeta_p) \in \mathrm{R}^p$ et

$S \in S^n$:

$$\mathrm{J}A(x)\left(\zeta_1, ..., \zeta_p\right) = \sum_{j=1}^{p} \zeta_j . A_j^{'}(x) \text{ et}$$

$$\begin{aligned}\left\langle\left(JA(x)\left(\zeta_1, ..., \zeta_p\right), S\right)\right\rangle &= \sum_{j=1}^{p} \zeta_j . \left\langle\left\langle A_j^{'}(x), S\right\rangle\right\rangle \\ &= \left\langle\left(\left(\zeta_1, ..., \zeta_p\right), \left[JA(x)\right]^* S\right)\right\rangle\end{aligned}$$

On en déduit que

$$[\mathrm{J}A(x)]^* \text{ est une fonction telle que : } S \in S^n \rightsquigarrow \left(..., tr\left(A_j^{'}(x).S\right), ...\right)_{j=1,...,p} \in R^p. \tag{1.36}$$

Lorsque S est de la forme décrite en (1.35), nous trouvons le résultat annoncé :

$$\eta_j = tr\left(A_j^{'}(x).S\right) = tr\left(U_1^t A_j^{'}(x) U_1\right) + tr\left(\left[U_2 W\right]^t A_j^{'}(x)\left[U_2 W\right]\right).$$

■

Exemple 3 *(forme particulière de A)*

Appliquons ce dernier théorème à la matrice $A = A_0 + diag(x_1, ..., x_n)$ *où (cfr exemple 2)* A_0 *est une matrice fixe, i.e. indépendante du paramètre* x.
Dans ce cas particulier, la matrice $A_j^{'}(x)$ *contient des zéros partout, sauf pour l'élément* $\left[A_j^{'}(x)\right]_{jj}$ *qui vaut* 1. *Dans le but de réécrire l'expression de l'ensemble* $\Delta_m(x)$, *calculons les éléments diagonaux de* $\left(U_1^t A_j^{'}(x) U_1\right)$ *et* $\left(\left[U_2 W\right]^t A_j^{'}(x)\left[U_2 W\right]\right)$, *où* $W = [w_1, ..., w_{i_m}]$:

$$\left(U_1^t A_j^{'}(x) U_1\right)_{ii} = \sum_{k=1}^{m-i_m} (U_1^t)_{ik} \underbrace{\left(A_j^{'}(x) U_1\right)_{ki}}_{= \begin{cases} 0 \ si \ k \neq j \\ u_{ij} \ si \ k = j \end{cases}} = u_{ij}^2 ,$$

$$\left([U_2W]^t A_j'(x)\,[U_2W]\right)_{ll} = [U_2W]^2_{jl} = \left(\sum_{k=1}^{m+j_m} [U_2]_{jk}\,[W]_{kl}\right)^2$$
$$= \left(\sum_{k=m-i_m+1}^{m+j_m} [U]_{jk}\,[W]_{k-m+i_m,\,l}\right)^2 = \left(\sum_{k=m-i_m+1}^{m+j_m} u_{kj} w_{l,\,k-m+i_m}\right)^2$$

pour tout $i = 1, ..., m - i_m$ *et* $l = 1, ..., i_m$.

Ainsi,

$$\eta_j = tr\left(U_1^t A_j'(x) U_1\right) + tr\left([U_2W]^t A_j'(x)\,[U_2W]\right) = \sum_{i=1}^{m-i_m} u_{ij}^2 + \sum_{l=1}^{i_m}\left(\sum_{k=m-i_m+1}^{m+j_m} u_{kj} w_{l\,k-m+i_m}\right)^2.$$

Ce qui nous donne :

$$\Delta_m(x) \equiv \left\{ \begin{array}{c} \eta \in R^p : \eta_j = \sum_{i=1}^{m-i_m} u_{ij}^2 + \sum_{l=1}^{i_m}\left(\sum_{k=m-i_m+1}^{m+j_m} u_{kj} w_{l\,k-m+i_m}\right)^2 \\ \text{où } W = [w_1, ..., w_{i_m}] \in M^{r_m, i_m},\ WW^t = I_{i_m}. \end{array} \right\}$$

Nous rencontrerons cette matrice particulière dans le prochain chapitre.

Les corollaires qui suivent sont les homologues des corollaires 3 et 4 traitant la différentiabilité de la fonction $\sigma_m(x)$.

Corollaire 5 *(différentiabilité de* $f_m(x)$*)*

Supposons que λ_m *occupe la dernière place d'un groupe de valeurs propres égales, alors* f_m *est différentiable en* x *et* $\nabla f_m(x) = \left(..., \dfrac{\partial f_m}{\partial x_j}(x), ...\right)$ *est donné par*

$$\frac{\partial f_m}{\partial x_j}(x) = tr\left(U_1^t A_j'(x) U_1\right) + tr\left(U_2^t A_j'(x) U_2\right) \text{ pour tout } j = 1, ..., p.$$

Preuve. Par le corollaire 3, σ_m est différentiable en $A(x)$ avec $\nabla\sigma_m\left(A(x)\right) = U_1^t U_1 + U_2^t U_2$. Le résultat suit alors de $\nabla f_m(x) = [\mathrm{J}A(x)]^*\,\nabla\sigma_m\left(A(x)\right)$ et (1.36). ■

Corollaire 6 *La fonction* $x' \in \Theta \rightsquigarrow t_m(x') \equiv (\lambda_{m-i_m+1} + ... + \lambda_m + ... + \lambda_{m+j_m})(x')$ *(c'est-à-dire la somme de toutes les valeurs propres qui coïncident en* x*) est différentiable en* x *avec*

$$\nabla t_m(x) = [JA(x)]^*\left(\nabla\sigma_{m+j_m}(x) - \nabla\sigma_{m-i_m}(x)\right).$$

Preuve. Immédiate par le corollaire 4. ■

1.5.2 Dérivée directionnelle de f_m

Par la proposition 3 , sous l'hypothèse de différentiabilité faite sur $A(.)$, la fonction composée $f_m = \sigma_m \circ A$ est strictement tangentiellement convexe en x avec

$$f_m'(x,d) = f_m^\circ(x,d) = \max\{\langle \xi, d\rangle : \xi \in \partial f_m(x)\} \text{ pour tout } d \in \mathrm{R}^p. \tag{1.37}$$

Pour obtenir $f_m'(x,d)$, il reste donc à rendre explicite l'expression du maximum.
En vue d'alléger les notations, nous posons pour $d = (d_1, ..., d_p) \in \mathrm{R}^p$,

$$E'(d) \equiv U_1^t \left(\sum_{j=1}^{p} d_j A_j'(x) \right) U_1 \qquad\qquad F'(d) \equiv U_2^t \left(\sum_{j=1}^{p} d_j A_j'(x) \right) U_2.$$

Théorème 8 *(Dérivée directionnelle de f_m)*

$$f_m'(x,d) = tr\left(E'(d)\right) + \sum_{k=1}^{i_m} \mu_k\left(F'(d)\right),$$

où $\mu_k\left(F'(d)\right)$ dénote la k-ème plus grande valeur propre de $F'(d)$.

Preuve. Une première façon d'atteindre le résultat est d'utiliser ∂f_m :
en effet, $\partial f_m = co\Delta_m(x)$ (cfr Théorème 7), et en utilisant (1.37), nous avons

$$f_m'(x,d) = \max\{\langle \xi, d\rangle : \xi \in \Delta_m(x)\} = \max_{\xi \in \Delta_m(x)} \sum_{j=1}^{p} \xi_j d_j.$$

L'expression détaillée de $\Delta_m(x)$ du théorème 7 permet d'écrire

$$\begin{aligned} f_m'(x,d) &= \max\left\{tr\left(E'(d)\right) + tr\left(W^t F'(d) W\right) : W^t W = I_{i_m}\right\} \\ &= tr\left(E'(d)\right) + \max\left\{tr\left(F'(d) W W^t\right) : W^t W = I_{i_m}\right\}. \end{aligned}$$

Nous reconnaissons dans l'expression du maximum la somme des i_m plus grandes valeurs propres de $F'(d)$ (cfr 1.3).

La seconde façon utilise la dérivée directionnelle de σ_m :

$$f_m^{'}(x,d) = \sigma_m^{'}\left(A(x), \mathrm{J}A(x)(d)\right) \text{ où } \mathrm{J}A(x)(d) = \sum_{j=1}^{p} d_j.A_j^{'}(x).$$

Le résultat découle du corollaire 2. ■

1.5.3 Dérivée directionnelle de $\lambda_m(x)$

Cette fois, la dérivée directionnelle de $\lambda_m(x)$ est construite à partir de celle de $f_m(x)$.

Théorème 9 *(dérivée directionnelle de λ_m).*

$$\lambda_m^{'}(x,d) = \mu_{i_m}(F^{'}(d)).$$

En particulier :

- *si $\lambda_m(x)$ occupe la première place dans un groupe de valeurs propres égales, alors $\lambda_m^{'}(x,.)$ est une fonction convexe,*
- *si $\lambda_m(x)$ occupe la dernière place dans un groupe de valeurs propres égales, alors $\lambda_m^{'}(x,.)$ est une fonction concave.*

Preuve. Connaissant les expressions de $f_m^{'}(x,d)$ et de $f_{m-1}^{'}(x,d)$ données par le théorème 8, la preuve suit le schéma de celle du théorème 6.

Une autre façon de le voir est d'observer que

$$\lambda_m^{'}(x,d) = \lambda_m^{'}\left(A(x), \mathrm{J}A(x)d\right).$$

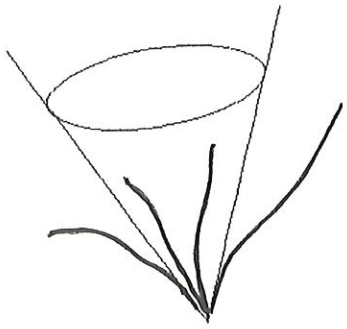
■

Résumons les trois comportements possibles de λ_m en fonction de x :

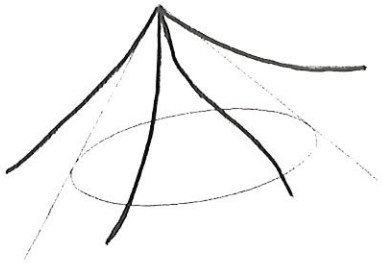
- quand $\lambda_m(x)$ occupe la première place dans un groupe de valeurs propres égales, l'approximation tangentielle $\lambda_m^{'}(x,.)$ de λ_m au point x est une fonction convexe positivement homogène (de plus, $\lambda_m^{'}(x,.) = \lambda_m^{\circ}(x,.)$, c'est-à-dire λ_m est *strictement tangentiellement convexe* en x).

- quand $\lambda_m(x)$ occupe la dernière place dans un groupe de valeurs propres égales, l'approximation tangentielle $\lambda'_m(x,.)$ est une fonction concave positivement homogène (en fait, $-\lambda'_m(x,.) = \lambda^\circ_m(x,.)$, ce qui revient à dire que λ_m est *strictement tangentiellement concave* en x).

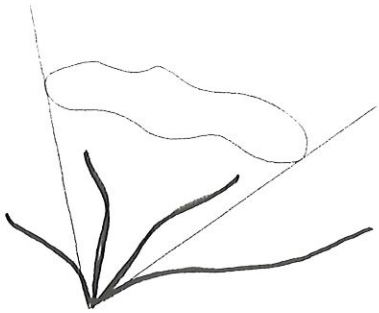
- quand $\lambda_m(x)$ est à l'intérieur d'un groupe de valeurs propres égales, l'approximation tangentielle $\lambda'_m(x,.)$ est la différence de fonctions convexes positivement homogènes.



Tangentiellement convexe.



Tangentiellement concave.



Chapitre 2

Minimiser certaines sommes non différentiables de valeurs propres de matrices symétriques

2.1 Problème et notations

Dans cette section, nous étudions le problème de minimiser la somme des q plus grandes valeurs propres d'une matrice réelle, symétrique, comme fonction de ses éléments diagonaux, avec pour seule contrainte que la somme de ces éléments diagonaux soit constante. Nous utiliserons dans ce chapitre des résultats de l'analyse des éléments propres et de la théorie de l'approximation. Le chapitre 1 (cfr exemple 2) nous a appris que cette fonction somme de valeurs propres est convexe mais pas nécessairement différentiable.

Introduisons quelques notations.

- Etant donné $x = (x_1, ..., x_n) \in \mathrm{R}^n$, construisons D, la matrice diagonale correspondante d'ordre n, i.e. $D_{ii} = x_i$ pour $1 \leq i \leq n$. A est une matrice réelle d'ordre n, symétrique et fixe.

- Notons par $\lambda_j(A + D)$ les n valeurs propres de $A + D$, rangées dans l'ordre décroissant : $\lambda_1(A + D) \geq \lambda_2(A + D) \geq ... \geq \lambda_n(A + D)$. Soit $U = \{u_1, ..., u_n\}$, une base orthonormale de vecteurs propres de $A + D$, correspondant à ces valeurs propres. Dans le chapitre 1,

ces vecteurs propres définissaient la caractérisation du sous différentiel. Nous verrons que leur calcul intervient dès qu'il s'agit d'employer les dérivées directionnelles. Ces calculs ne sont pas triviaux, car nous nous attendons à rencontrer des valeurs propres multiples. Il nous faudra donc nous référer à l'algorithme de Bloc de Lançzos [4].

- Définissons l'entier $i(\varepsilon)$ de sorte que

$$\lambda_1(A+D), ..., \lambda_{q-1-i(\varepsilon)}(A+D)$$

représentent toutes les valeurs propres plus grandes que $\lambda_q(A+D)+\varepsilon$, et notons l'ensemble des vecteurs propres correspondants par

$$U_1(x,\varepsilon) = \left\{u_1, ..., u_{q-1-i(\varepsilon)}\right\}.$$

Nous définissons alors l'entier $j(\varepsilon)$ tel que les valeurs propres

$$\lambda_{q-i(\varepsilon)}(A+D), ..., \lambda_{q-j(\varepsilon)}(A+D)$$

soient toutes celles appartenant à l'intervalle $[\lambda_q(A+D)-\varepsilon, \lambda_q(A+D)+\varepsilon]$, et

$$U_2(x,\varepsilon) = \left\{u_{q-i(\varepsilon)}, ..., u_{q+j(\varepsilon)}\right\}$$

désigne l'ensemble des vecteurs propres correspondants.

Pour toute base U de vecteurs propres de $A+D$, $U_1 = U_1(x,0)$ est l'ensemble des vecteurs propres de $A+D$ correspondant aux valeurs propres strictement plus grandes que λ_q (cfr 1.32) et $U_2 = U_2(x,0)$ est l'ensemble des vecteurs propres de $A+D$ correspondant à λ_q. Regardons à quoi correspondent ces entiers sur un exemple :
supposons que nous ayons $n = 11$, $q = 6$ et les valeurs propres rangées dans l'ordre décroissant comme suit :

$$\lambda_1 > \lambda_2 > \boldsymbol{\lambda_q+\varepsilon} > \lambda_3 > \lambda_4 > \lambda_{q-1} = \boxed{\lambda_q} = \lambda_{q+1} = \lambda_{q+2} > \lambda_9 > \lambda_{10} > \boldsymbol{\lambda_q-\varepsilon} > \lambda_{11}. \quad (2.1)$$

Comme $\lambda_2 = \lambda_{q-1-3}$ et $\lambda_{10} = \lambda_{q+4}$, nous écrivons $i(\varepsilon) = 3$ et $j(\varepsilon) = 4$. Il y a donc $i(\varepsilon)$ valeurs propres entre $\lambda_q + \varepsilon$ et λ_q non compris, et il y en a $j(\varepsilon)$ entre $\lambda_q - \varepsilon$ et λ_q non compris.
Pour $\varepsilon = 0$, $i(0) = 1$ (car il y a une valeur propre rangée entre $\lambda_q + \varepsilon = \lambda_{q-1}$ et λ_q non compris) et $j(0) = 2$ (car il y a deux valeurs propres entre $\lambda_q - \varepsilon = \lambda_{q+2}$ et λ_q non compris). La multiplicité de λ_q est donnée par $i(0) + j(0) + 1 = 4$

Les indices $i(\varepsilon)$ et$j(\varepsilon)$ sont respectivement appelés les multiplicités *intérieure* et *extérieure* à ε près de $\lambda_q(A + D)$. La multiplicité à ε près de $\lambda_q(A + D)$ est $i(\varepsilon) + j(\varepsilon) + 1$. Notons respectivement par i et j les multiplicités (exactes) intérieure et extérieure de $\lambda_q(A + D)$. Les nouveaux indices i,j sont liés aux indices i_q, j_q, r_q du chapitre 1, par : $i + 1 = i_q$, $j = j_q$ et $i + j + 1 = r_q$.

- Pour toute matrice $X = \{x_1, ..., x_m\}$, définissons le vecteur $T(X)$ par

$$T(X)_i = \sum_{j=1}^{m} x_{ji}^2$$

(x_{ji} est la i-ème composante de x_j)
H_a^b est l'ensemble de toutes les matrices de a lignes et b colonnes orthonormales.

- La fonction objectif est

$$f(x) = \sigma_q(A + D) = \sum_{j=1}^{q} \lambda_j(A + D). \tag{2.2}$$

Nous avons montré dans le chapitre 1 que cette fonction est convexe. Elle est différentiable, excepté aux points x pour lesquels $\lambda_q(A+D) = \lambda_{q+1}(A+D)$ (cfr le corollaire 5 du chapitre 1). Cela revient à dire que la multiplicité intérieure de λ_q et les multiplicités des valeurs propres distinctes de λ_q n'affectent pas la différentiabilité de f. L'exemple 2 du chapitre 1 nous donne l'expression explicite de son sous différentiel $\partial f(x)$. Ecrivons-le avec nos nouvelles notations :

$$\partial f(x) = co\,\Delta_q(x) = co\left\{\eta : \eta = T\left(U_1(x,0)\right) + T\left(U_2(x,0)H\right) \text{ où } H \in H_{i+j+1}^{i+1}\right\} \tag{2.3}$$

où rappelons-le, $U_1(x,0) = \{u_1, ..., u_{q-i-1}\}$ et $U_2(x,0) = \{u_{q-i}, ..., u_{q+j}\}$.

Nous cherchons donc à résoudre le problème d'optimisation suivant :

$$(\mathrm{P}) \begin{cases} \min \sum_{j=1}^{q} \lambda_j(A+D) \\ \text{s.c. } tr(D) = 0 \end{cases} \tag{2.4}$$

2.2 Algorithme général

2.2.1 Sous différentiel projeté et direction de plus grande descente

Dans la pratique, les points optimaux de (2.4) sont souvent des points en lesquels f n'est pas différentiable. Donc, les algorithmes conçus pour des fonctions différentiables ne peuvent être utilisés pour calculer ces points.

Comme le problème (P) contient la contrainte d'égalité $tr(D) = 0$, c'est-à-dire $\langle e, x\rangle = 0$ où $e = (1, 1, ..., 1)$, nous utiliserons la méthode du sous différentiel projeté. Notons $P = I - ee^t/n$ la projection orthogonale sur l'espace $\langle e, x\rangle = 0$. $P\partial f(x)$ est l'analogue pour une fonction convexe du gradient projeté d'une fonction différentiable.

L'existence de la dérivée directionnelle $f'(x,d)$ de f en x dans la direction d, nous invite à étudier la direction de plus grande descente. Cette notion est employée dans la suite. Cependant, elle ne peut être utilisés de façon naturelle car l'analogue pour des fonctions convexes de la méthode de plus grande descente pour des fonctions différentiables, ne garantit pas la convergence des itérés vers un point optimal.

En effet, dans le cas différentiable, la convergence de l'algorithme de plus grande descente est assurée par la continuité de l'opérateur ∇f :

$$\text{si } x_k \to x^* \text{ alors } -\nabla f(x_k) \to -\nabla f(x^*).$$

Dans le cas non différentiable, nous n'avons pas l'implication

$$x_k \to x^* \;\Rightarrow\; d_k \to d^*,$$

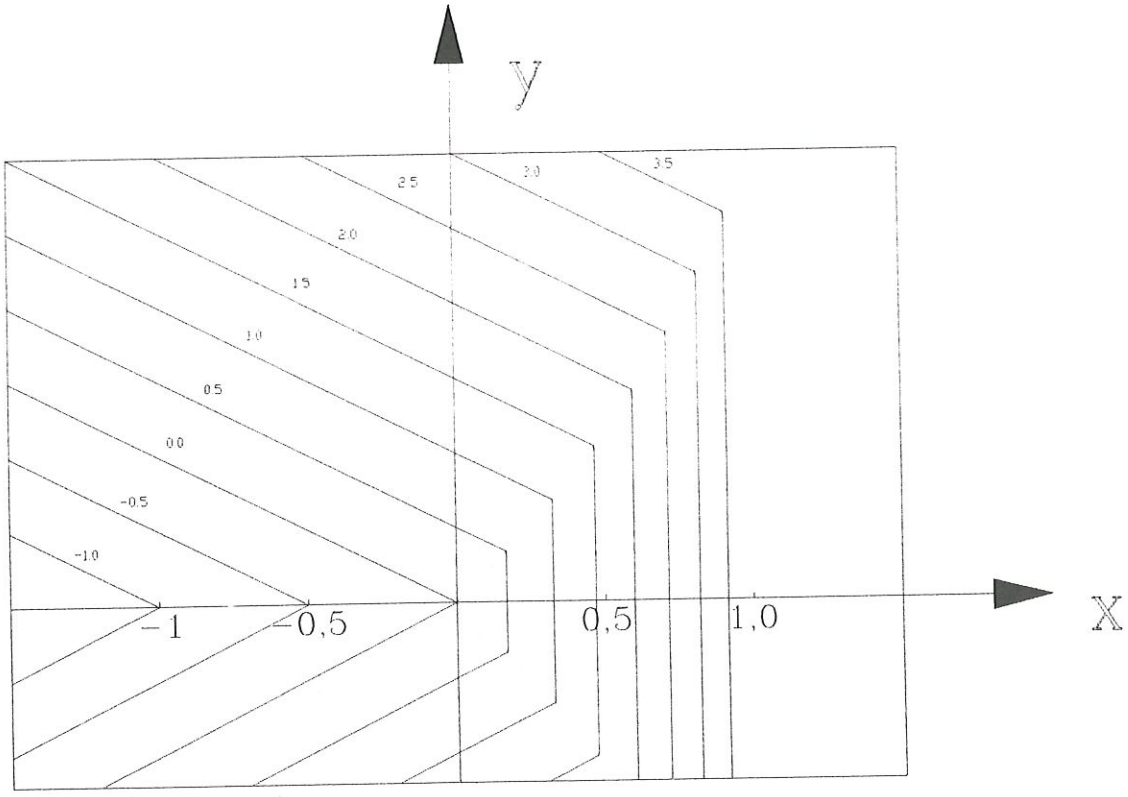
où $d_k = -\underset{\partial f(x_k)}{\mathrm{Pr}}(0)$ et $d^* = -\underset{\partial f(x^*)}{\mathrm{Pr}}(0)$ sont les directions de plus grande descente de f en x_k

et x^*, pour le cas non différentiable.

Exemple 4 *La foncton* $f(x,y) = \max\{x+2 \mid y \mid,\ 3x+x^3\}$ *est représentée dans la figure (2-1). Cette fonction n'est pas différentiable pour* $x \leq 0$ *,*$y = 0$ *et sur les courbes* $\pm y = x + 1/2x^3$, $x > 0$. *En chaque point d'une de ces courbes pour* $x < \frac{2}{3}$, *la direction de plus forte pente est tangente à cette courbe, et le point minimal dans cette directon se trouve sur l'autre courbe. Partant de tout point d'une de ces courbes avec* $x < \frac{2}{3}$, *la suite générée par la méthode de plus grande descente de Cauchy converge vers* $(0,0)$, *mais* $0 \notin \partial f(0,0)$: *en effet, si* $0 \in \partial f(0,0) = \left\{s : \langle s,d\rangle \leq f'((0,0);(d_1,d_2)) \text{ pour tout } d_1,d_2\right\}$, *alors* $f'((0,0);(d_1,d_2)) \geq 0$ *pour tout* d_1,d_2. *Or, pour* $d_1 = -1$ *et* $d_2 = 0$ *nous avons que*

$$\begin{aligned} f'((0,0);(d_1,d_2)) &= \lim_{t\searrow 0} \frac{\max\left\{td_1+2|td_2|\ ,\ 3td_1+(td_1)^3\right\}}{t} \\ &= \lim_{t\searrow 0} \max\{d_1+2 \mid d_2 \mid,\ 3d_1+t^2(d_1)^3\} = -1. \end{aligned}$$

Ce qui entraîne une contradiction.



Une façon de construire un algorithme de minimisation convergent pour une fonction convexe, serait de modifier la méthode de plus grande descente de manière à créer artificiellement ce manque de continuité.

2.2.2 Elargir le sous différentiel

Certaines procédures proposées pour minimiser une fonction convexe non différentiable simulent cette continuité en utilisant une approximation $S(x,\varepsilon)$ du sous différentiel de la fonction objectif. En particulier, Demyanov [5] élargit le sous différentiel $\partial f(x)$ en ajoutant à l'ensemble $\{XX^t : X \in M_q(A+D)\}$ (cfr (1.16)), toutes les matrices orthonormales Y, qui vérifient $\sigma_m(A) = \langle\langle A, Y\rangle\rangle \pm \varepsilon$. Si nous utilisions la procédure de Bertsekas et Mitter [1], le sous différentiel serait approximé par le sous différentiel à ε près, i.e.

$$S(x,\varepsilon) = \partial_\varepsilon f(x) = \left\{ s : f(x') \geq f(x) + \left\langle s, x' - x \right\rangle - \varepsilon \ \text{ pour tout } x' \right\}.$$

Au lieu de déterminer une direction de plus grande descente au point x_k, ils cherchent une direction de ε-plus grande descente, i.e. une direction d_k solution du problème

$$\min_{\|d\|\leq 1} \max_{s\in\partial_\varepsilon f(x_k)} \langle s, d\rangle .$$

Les deux procédures utilisent une information non-locale sur f, pour construire les ensembles approximants. Il est alors difficile de caractériser ces ensembles $S(x,\varepsilon)$.

Nous proposons ici une procédure n'ayant besoin que d'information locale, à savoir, le sous différentiel de f évalué à *l'itéré courant*. Cet algorithme suit aussi la méthode de ε-plus grande descente. Cependant, l'ensemble $S(x,\varepsilon)$ est complètement caractérisé par certains vecteurs propres de $(A+D)$, et peut être ainsi utilisé de façon implémentable. Nous utiliserons certaines idées des algorithmes de Bertsekas et Mitter, et Demyanov. La convergence d'une telle procédure est assurée par la construction d'une famille d'ensembles $S(x,\varepsilon)$ qui approximent $\partial f(x^*)$ dans le sens suivant : si $x_k \to x^*$, alors tout $s \in \partial f(x^*)$ est la limite d'une suite d'éléments $s_k \in S(x_k,\varepsilon)$. Inversément, la limite d'une suite de vecteurs $s_k \in S(x_k,\varepsilon)$ se trouve dans $\partial f(x^*)$. Nous vérifierons que l'ensemble $S(x,\varepsilon)$ que nous allons construire possède ces propriétés et nous verrons qu'il est complètement caractérisé par des quantités implémentables.

De plus, nous aurons pour tout x' et $\delta > 0$ que $S(x,\varepsilon) \subseteq \partial_\delta f(x')$ pour tout ε suffisamment petit et pour tout x suffisamment proche de x'.

Comme indiqué plus haut, un point minimisant la fonction objectif de (P) sera souvent un point pour lequel $\lambda_q(A+D) = \lambda_{q+1}(A+D)$ i.e. pour lequel f n'est pas différentiable. La procédure de descente développée dans ce chapitre a la faculté d'anticiper la multiplicité de λ_q en de tels points : si à l'itéré x_k la multiplicité de $\lambda_q(A+D_k)$ est 1, mais avec $\mid \lambda_{q+1}(A+D_k) - \lambda_q(A+D_k) \mid$ très petit, alors la procédure va considérer $\lambda_q(A+D_k)$ comme si c'était une valeur propre de multiplicité au moins 2. Ceci nous donne l'idée avec laquelle l'approximation du sous différentiel de f en x_k sera calculée.

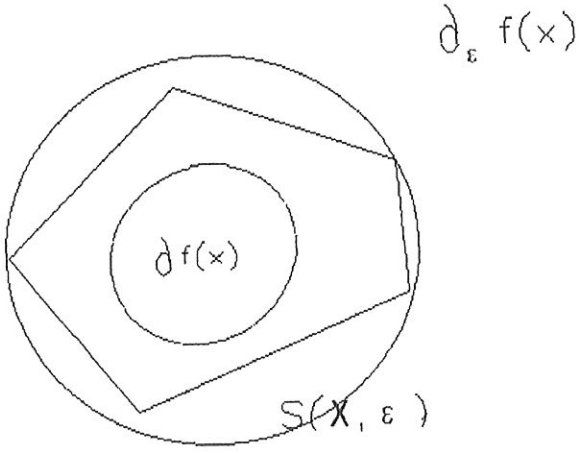
Un coup d'oeil sur l'équation (2.3) fournit une façon simple et réalisable d'élargir l'ensemble $\Delta_q(x)$: définissons

$$\Delta_q(x,\varepsilon) = \left\{\eta : \eta = T\left(U_1(x,\varepsilon)\right) + T\left(U_2(x,\varepsilon)H\right) \text{ où } H \in H^{i(\varepsilon)+1}_{i(\varepsilon)+j(\varepsilon)+1}\right\} \tag{2.5}$$

et

$$S(x,\varepsilon) = co\Delta_q(x,\varepsilon). \tag{2.6}$$

Clairement, $\Delta_q(x) \subseteq \Delta_q(x,\varepsilon)$. L'introduction du paramètre ε nous oblige à considérer les multiplicités à ε près, et fait en sorte que la construction de l'ensemble $\Delta_q(x) = \Delta_q(x,0)$ tienne compte de vecteurs propres supplémentaires (comparer (2.3) et (2.6)). Dans notre exemple, (cfr 2.1), cela revient à compléter les vecteurs propres associés à λ_5, λ_6, λ_7, λ_8, par les vecteurs propres qui correspondent à λ_3, λ_4, λ_9, et λ_{10}.



Afin d'appliquer la méthode du sous différentiel projeté, notons $PS(x,\varepsilon)$ et $P(\Delta_q(x,\varepsilon))$ les projections respectives de $S(x,\varepsilon)$ et $\Delta_q(x,\varepsilon)$ sur la contrainte $\langle e,x\rangle = 0$. Par le théorème de Carathéodory [8], $PS(x,\varepsilon) = coP(\Delta_q(x,\varepsilon))$ pour tout $\varepsilon \geq 0$.

2.2.3 Algorithme général de la somme des valeurs propres (S.V.P.)

A l'itération k, x_k, $\varepsilon_k > 0$, et $a > 1$ sont donnés.

Etape 1. Calculer $S(x_k, \varepsilon_k)$.

Etape 2. Calculer une direction de mouvement :

$d_k = -\frac{z_k}{\|z_k\|}$ où z_k est le vecteur de norme minimale de $PS(x_k, \varepsilon_k)$.

Etape 3. Mise à jour de ε_k et recherche linéaire :

si $d_k \neq 0$, alors $x_{k+1} = x_k + t_k d_k$ où $t_k = \arg\min_{t \geq 0} f(x_k + t d_k)$

et $\varepsilon_{k+1} = \varepsilon_k$. Aller à l'étape 1.

si $d_k = 0$, alors $0 \in PS(x_k, \varepsilon_k)$,

et si $\varepsilon_k \neq \varepsilon_{\min}$ alors $\varepsilon_{k-1} = \varepsilon_k / a$, $x_{k-1} = x_k$. Aller à l'étape 1.

sinon, aller à l'étape 4.

Etape 4. Critère d'arrêt :

si $\{0\} \in \partial f(x_k)$ alors x_k est solution optimale.

Afin que cet algorithme soit vraissemblable, nous devons vérifier au moins deux hypothèses.

1. Les itérés générés par l'algorithme S.V.P. "convergent" d'une certaine manière,
vers un point minimisant f pour $k \to \infty$.
2. L'algorithme S.V.P. est implémentable.

Nous étudierons la convergence dans la section 2.6. A présent, regardons de quelle façon ces 4 étapes s'implémentent.

2.3 Direction de mouvement et critère d'arrêt

2.3.1 Direction de mouvement sur $\langle e, x\rangle = 0$

Pour chaque itéré x_k, la direction de recherche d_k est solution du problème min-max suivant :

$$\varphi_k \equiv \varphi(x_k, \varepsilon_k) \equiv \min_{\|d\|\leq 1} \max_{s\in PS(x_k,\varepsilon_k)} \langle s, d\rangle . \tag{2.7}$$

Remarque 4 *Avec $\varepsilon_k = 0$, nous retrouvons le problème*

$$\min_{\|d\|\leq 1} \max_{s\in P\partial f(x_k)} \langle s, d\rangle \ \ i.e. \ \ \min_{\|d\|\leq 1} f'(x_k, d).$$

Dans le cas diffférentiable, l'équation (2.7) correspond au problème :

$$\min_{\|d\|\leq 1} \nabla f(x_k)^t d.$$

Bertsekas et Mitter ont démontré que la solution de (2.7) est

$$d_k = -\frac{z_{k_J}}{\|z_{k_J}\|}$$

où z_{k_J} est le vecteur de norme minimale de $PS(x_k, \varepsilon_k)$, avec

$$\varphi_k = -\|z_{k_J}\|.$$

Calculer d_k revient donc à chercher un vecteur de norme minimale d'un ensemble convexe, compact $PS(x_k, \varepsilon_k)$. Pour calculer un tel vecteur, Frank et Wolfe [10] proposent un algorithme utilisant des projections sur des segments de droites et le calcul de points de contact. Avant de le décrire, définissons la notion de point de contact. Par la suite, nous noterons $PS(x_k, \varepsilon_k)$ par $PS(k)$, $\Delta_q(x_k, \varepsilon_k)$ par $\Delta(k)$, $i(\varepsilon_k)$ par $i(k)$, et de même pour $j(\varepsilon_k)$.

Définition : $s^* \in PS(k)$ est un point de contact de $PS(k)$, correspondant à la direction d, si

$$\langle s^*, d\rangle = \max_{s\in PS(k)} \langle s, d\rangle . \tag{2.8}$$

Nous le désignons par $c(d)$.

L'algorithme de Frank-Wolfe procède comme suit :
Soit $g(s) \equiv \|s\|$ la fonction à minimiser sur $PS(k)$. Minimisons la linéarisation de $g(.)$ autour de $z_{k_j} \in PS(k)$. Ceci revient à résoudre le problème

$$\min_{s \in PS(k)} g(z_{k_j}) + \left\langle z_{k_j}, s - z_{k_j} \right\rangle \quad \text{i.e.} \quad \min_{s \in PS(k)} \left\langle z_{k_j}, s \right\rangle,$$

ou encore

$$\max_{s \in PS(k)} \left\langle -z_{k_j}, s \right\rangle,$$

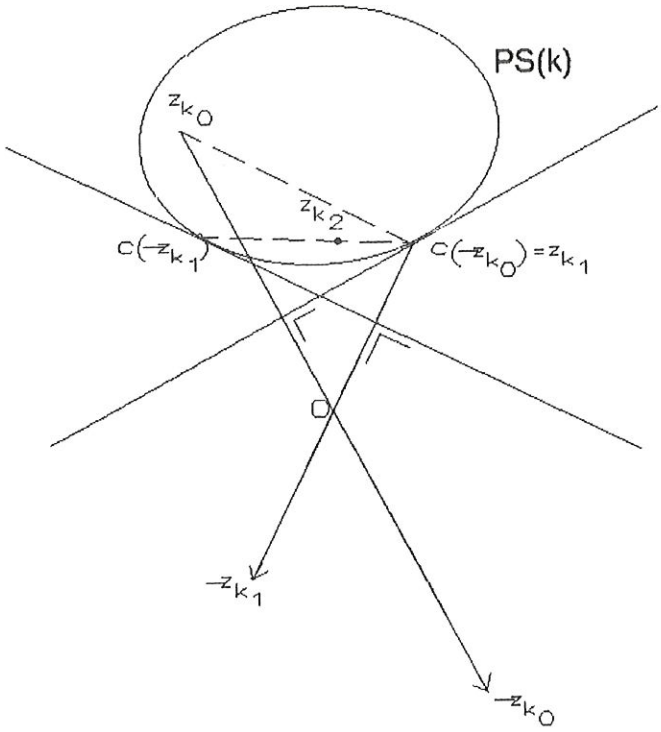
qui n'est rien d'autre que la recherche d'un point de contact de $PS(k)$, $c(-z_{k_j})$.

L'exécution des étapes qui suivent constitue l'algorithme de Frank-Wolfe.
Remarquons qu'à cette étape de l'algorithme général S.V.P., l'indice k est fixé. Supposons que nous disposons de $z_{k_0} \in PS(k)$ et posons $j = 0$.

Etape 1. Trouver $c(-z_{k_j})$.

Etape 2. Chercher $z_{k_{(j+1)}}$ défini comme étant le vecteur de norme minimale sur le segment $\left[z_{k_j}, c(-z_{k_j})\right]$, et poser $J = j + 1$.

Etape 3. Si $\max_{s \in PS(x_k, \varepsilon_k)} \langle s, z_{k_J} \rangle + \|z_{k_J}\| < \varepsilon_k$, alors $d_k = -\frac{z_{k_J}}{\|z_{k_J}\|}$, et on s'arrête.
Sinon, retourner à l'étape 1.



Ainsi d_k est déterminée si les points de contact de $PS(k)$ peuvent être calculés.

Comme pour tout d et x

$$f'(x,d) = \max_{s \in \partial f(x)} \langle s, d \rangle ,$$

la notion de point de contact reviendra pour calculer les dérivées directionnelles utilisées dans la recherche linéaire de l'algorithme général. La facilité avec laquelle un point de contact de $PS(k)$ peut être calculé est exprimée dans le prochain théorème. Pour le démontrer, nous aurons besoin du lemme suivant :

Lemme 3 $\sum_{j=1}^{q} \langle M\, x_j, x_j \rangle$ *atteint sa valeur maximale sur l'ensemble* $X = \{x_1, ..., x_q\}$ *de vecteurs orthonormaux, si et seulement si*

$$X \subseteq sp\{V_1, V_2\} \quad et \quad sp\{V_1\} \subseteq sp\{X\}$$

avec $V_1 = V_1(x,0) = \{v_1, ..., v_{q-i-1}\}$ *et* $V_2 = V_2(x,0) = \{v_{q-i}, ..., v_{q+j}\}$, *des matrices contenant des vecteurs propres de* M *associés aux valeurs propres* μ_k, *où* i *et* j *désignent cette fois les multiplicités intérieure et extérieure de* μ_q.

Preuve.

- $V \equiv [V_1, V_2]$ étant un ensemble de vecteurs orthonormaux, nous avons pour $1 \leq k \leq q$

$$x_k = \sum_{l=1}^{n} \langle x_k, v_l \rangle v_l. \tag{2.9}$$

v_k étant un vecteur propre de M, associé à μ_k, nous écrivons

$$Mv_k = \mu_k v_k. \tag{2.10}$$

Posons

$$T_l \equiv \sum_{k=1}^{q} \langle x_k, v_l \rangle^2 . \tag{2.11}$$

- Par la semi-linéarité du produit scalaire, en tenant compte de (2.9), (2.10) et (2.11), nous avons les égalités suivantes :

$$\begin{aligned} \sum_{k=1}^{q} \langle x_k, Mx_k \rangle &= \sum_{k=1}^{q} \sum_{l=1}^{n} \mu_l \langle x_k, v_l \rangle^2 + \mu_q \left[q - \sum_{k=1}^{q} \sum_{l=1}^{n} \langle x_k, v_l \rangle^2 \right] \\ &= \sum_{k=1}^{q} \mu_k + \sum_{l=1}^{q} \underbrace{(-\mu_l + \mu_q)}_{\leq 0} \underbrace{[1 - T_l]}_{\geq 0} + \sum_{l=q+1}^{n} \underbrace{(\mu_l - \mu_q)}_{\leq 0} \underbrace{T_l}_{\geq 0} \end{aligned} \tag{2.12}$$

La relation de Bessel Parceval nous donne,

$$\sum_{k=1}^{q} \langle x_k, v_l \rangle^2 \leq \|v_l\| = 1.$$

Cela veut dire que $1 - T_l \geq 0$. De plus, pour $l \leq q$, rappelons que $\mu_l \geq \mu_q$. Par contre, $\mu_l \leq \mu_q$ pour $l \geq q$. Tout ceci nous amène à dire que

$$\sum_{k=1}^{q} \langle x_k, Mx_k \rangle \leq \sum_{k=1}^{q} \mu_k. \tag{2.13}$$

- Chaque terme des sommes sur l de (2.12) étant négatif, nous avons l'égalité dans (2.13) si et seulement si chacun de ces termes s'annule. Sachant que $\mu_l - \mu_q = 0$ pour $l =$

$q-i, ..., q+j$, cette dernière condition revient à imposer :

1) $\langle x_k, v_l \rangle = 0$ pour $l > q+j$ et $1 \leq k \leq q$, et

2) $T_l \equiv \sum_{k=1}^{q} \langle x_k, v_l \rangle^2 = 1$ pour $l \leq q-i-1$.

La première relation que l'on impose implique que pour tout k, $x_k \in sp\{V_1, V_2\}$. En effet,

$$x_k = \sum_{l=1}^{n} \langle x_k, v_l \rangle v_l = \sum_{l=1}^{q+j} \langle x_k, v_l \rangle v_l + \sum_{l=q+j+1}^{n} \underbrace{\langle x_k, v_l \rangle}_{=0} v_l$$
$$= \sum_{l=1}^{q+j} \langle x_k, v_l \rangle v_l \in sp\{v_1, ..., v_{q+j}\}.$$

De la même façon, la deuxième relation exige que $v_l \in sp\{X\}$ pour $l \leq q-i-1$. ■

Théorème 10 *(point de contact)*
Pour toute direction d, un point de contact de $PS(x_k, \varepsilon_k)$ correspondant à d peut être calculé par la succession de quatre étapes :

1) Construire la matrice carrée M d'ordre $m \equiv i(k) + j(k) + 1$ telle que l'élément r, l est noté par

$$M_{rl} = \sum_{k=1}^{n} u_{(l-t)\,k}\; u_{(r-t)\,k}\; d_k$$

où $t \equiv q - i(k) - 1$, $1 \leq r, l \leq m$ et $U = \{u_1, ..., u_{q+j}\}$ est la matrice des vecteurs propres de $A + D$.

2) Calculer les vecteurs propres de M. Notons-les v_i.

3) Former la matrice $V = \left\{v_1, ..., v_{i(k)+1}\right\}$.

4) Le point de contact peut s'écrire comme :

$$s = T\left(U_1(x_k, \varepsilon_k)\right) + T\left(U_2(x_k, \varepsilon_k)V\right) - eq/n.$$

Preuve.

- Soit s un point de contact de $PS(k)$ correspondant à d. Par la linéarité de (2.8) et la convexité de $PS(k)$, nous avons :

$$\langle s, d\rangle = \max_{v \in PS(k)} \langle v, d\rangle = \max_{v \in P\Delta_q(k)} \langle v, d\rangle . \tag{2.14}$$

Par construction $v \in P\Delta(k)$, s'il existe, (cfr (2.5)), une matrice $H = \left[\eta_1, ..., \eta_{i(k)+1}\right]$ $\in H_m^{i(k)+1}$ telle que

$$v = T\left(U_1(x_k, \varepsilon_k)\right) + T\left(U_2(x_k, \varepsilon_k)H\right) - eq/n.$$

- Calculons $\langle v, d\rangle = \sum_{k=1}^{n} v_k d_k$.

$$\langle v, d\rangle = \sum_{k=1}^{n} \left[\sum_{l=1}^{t} u_{lk}^2 + \sum_{p=1}^{i(k)+1} \left(\sum_{l=q-i(k)}^{q-j(k)} u_{lk}\, \eta_{p(l-t)} \right)^2 \right] d_k \tag{2.15}$$

Les seules variables dans cette expression sont les vecteurs orthonormaux de H : η_l, $1 \leq l \leq i(k)+1$. D'où, seule la dernière somme de (2.15) doit être considérée lorsqu'on maximise (2.15). Réarrangeons cette somme :

$$\sum_{k=1}^{n} \left[\sum_{p=1}^{i(k)+1} \left(\sum_{l=q-i(k)}^{q+j(k)} u_{lk}\, \eta_{p(l-t)} \right)^2 \right] d_k = \sum_{k=1}^{n} \sum_{p=1}^{i(k)+1} \left[\sum_{\substack{l=1 \\ r=1}}^{m} \left(\eta_{pl}\, \eta_{pr}\, u_{(l+t)k}\, u_{(r+t)k} \right) \right] d_k.$$

$$= \sum_{p=1}^{i(k)+1} \left[\sum_{l,r=1}^{m} \eta_{pl} \left(\sum_{k=1}^{n} u_{(l+t)k}\, u_{(r+t)k}\, d_k \right) \eta_{pr} \right]. \tag{2.16}$$

- Définissons $M \equiv M(d, u)$, une matrice réelle symétrique d'ordre m telle que

$$M_{rl} = \sum_{k=1}^{n} u_{(l+t)k}\, u_{(r+t)k} d_k,$$

et notons ses vecteurs propres par v_k. L'équation (2.16) s'écrit alors :

$$\sum_{p=1}^{i(k)+1}\sum_{l=1}^{m}\eta_{pl}\left(\sum_{r=1}^{m}M_{rl}\,\eta_{pr}\right)=\sum_{p=1}^{i(k)+1}\sum_{l=1}^{m}\eta_{pl}\left[M\,\eta_p\right]_l$$

$$=\sum_{p=1}^{i(k)+1}\langle\eta_p,M\eta_p\rangle\,. \tag{2.17}$$

- Appliquons le lemme 3 pour maximiser (2.17). Le maximum de (2.17) est atteint sur tous les ensembles $Z=\left\{\eta_1,...,\eta_{i(k)+1}\right\}$ de vecteurs orthonormaux qui vérifient $Z\subseteq sp\left\{V_1,V_2\right\}$ et $sp\left\{V_1\right\}\subseteq sp\left\{Z\right\}$, où $V_1=\left\{v_1,...v_{q-i-1}\right\}$. Pour une telle matrice Z, un point de contact de $PS(k)$ pour d, s'exprime donc par :

$$s=T\left(U_1(x_k,\varepsilon_k)\right)+T\left(U_2(x_k,\varepsilon_k)Z\right)-eq/n.$$

Pour terminer la preuve il suffit de remarquer qu'en particulier, Z peut être un ensemble de vecteurs propres correspondant aux $i(k)+1$ plus grandes valeurs propres de M. ■

Nous retiendrons de ce théorème l'équivalence suivante :

$$\max_{v\in PS(k)}\langle v,d\rangle\Longleftrightarrow\max_{\eta_p}\underbrace{\sum_{p=1}^{i(k)+1}\langle\eta_p,M(d,u)\,\eta_p\rangle}$$

atteint sa valeur maximale sur les $i(k)+1$ vecteurs propres de $M(d,u)$, matrice dépendant des vecteurs propres $u_{q-i(k)},...,u_{q+j(k)}$.

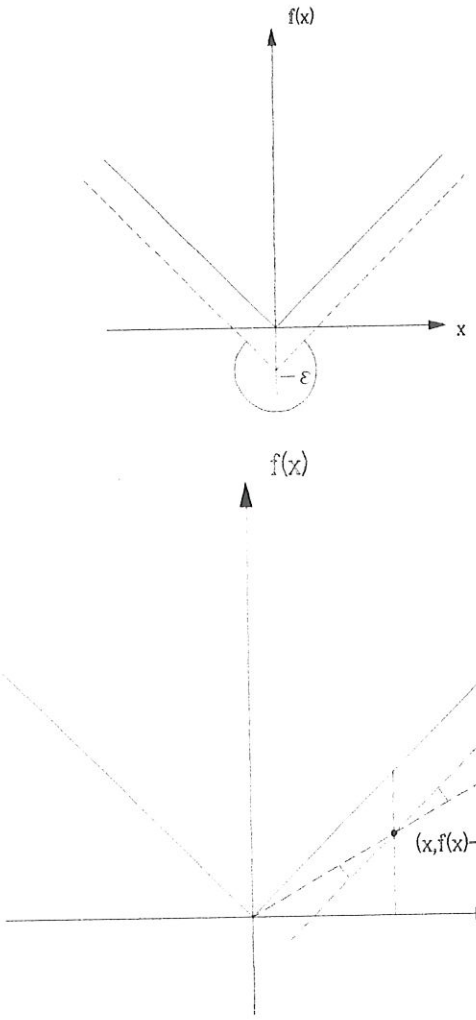
Le calcul d'un point de contact sur $PS(k)$, revient donc au calcul de vecteurs propres de $(A+D)$.

2.3.2 Critère d'arrêt

La conditon $\{0\}\in P\partial f(x)$ est nécessaire et suffisante pour que x soit un point minimisant f sur l'espace $\langle e,x\rangle=0$. Cependant, elle n'est pas réalisable puisque le sous différentiel n'est pas une fonction continue en x.

Considérons par exemple pour $x \in \mathrm{R}$, la fonction $f(x) = \mid x \mid = \max(x, -x)$. Pour $x^* = 0$, nous avons $f(x^*) = \min f(x)$ et $\{0\} \in \partial f(0)$. Mais $\partial f(x) = \{1\}$ pour tout $x > 0$. Par contre, si on élargit $\partial f(x)$ en constuisant $S(x, \varepsilon) = \partial_\varepsilon f(x)$, on obtient pour $\mid x \mid \leq \varepsilon$,

$$\{0\} \in S(x, \varepsilon) = \{-1 \leq s \leq 1\}.$$



Nous utilisons donc pour critère d'arrêt :

$$\{0\} \in PS(x_k, \varepsilon_{\min}),$$

où $\varepsilon_{\min}$ est déterminé par le degré de précision avec lequel on veut calculer les vecteurs propres utilisés dans l'expression de l'ensemble $S(x_k, \varepsilon_k)$.

2.4 Mise à jour de ε_k et recherche linéaire

2.4.1 Mise à jour de ε_k

$$\varphi_k \equiv \varphi(x_k, \varepsilon_k) \equiv \min_{\|d\| \leq 1} \max_{s \in PS(x_k, \varepsilon_k)} \langle s, d \rangle = \max_{s \in PS(x_k, \varepsilon_k)} \langle s, d_k \rangle .$$

Si $\varphi_k < -\varepsilon_k$, alors $\max_{s \in PS(x_k, \varepsilon_k)} \langle s, d_k \rangle < -\varepsilon_k < 0.$

Cette conditon, entraîne l'existence d'une direction de descente pour f en x_k, car l'inclusion $PS(k) \supseteq P\partial f(x_k)$ implique les inégalités suivantes :

$$0 > -\varepsilon_k > \max_{s \in PS(k)} \langle s, d_k \rangle \geq \max_{s \in P\partial f(x_k)} \langle s, d_k \rangle = f'(x_k, d_k).$$

De plus, nous avons la relation analogue à $0 \notin P\partial f(x_k)$, i.e.

$$0 \notin PS(k).$$

(En effet, si $0 \in PS(k)$, alors $\max_{s \in PS(x_k, \varepsilon_k)} \langle s, d_k \rangle \geq 0$.) Nous en concluons que $\|s\| > 0$ pour tout s appartenant à $PS(k)$. Par conséquent $\|z_{kJ}\| > 0$ et $d_k = \frac{-\|z_{kJ}\|}{\|z_{kJ}\|} < 0$.

Il reste alors à effectuer une recherche linéaire pour trouver le nouvel itéré x_{k+1}. On retourne à l'étape 1 en gardant l'ensemble $PS(k)$ pour approximation du sous différentiel de f, i.e. $\varepsilon_{k+1} = \varepsilon_k$.

Si $\varphi_k \geq -\varepsilon_k$, il n'est plus possible de suffisamment descendre avec ε_k car l'inclusion $PS(k) \subseteq P\partial f_{\varepsilon_k}(x_k)$
implique

$$f'_{\varepsilon_k}(x_k, d_k) = \max_{s \in P\partial f_{\varepsilon_k}(x_k)} \langle s, d_k \rangle \geq \max_{s \in PS(k)} \langle s, d_k \rangle \geq -\varepsilon_k.$$

Nous obtenons une sorte d'optimalité à ε_k près, de l'itéré x_k. Et cette fois,

$$0 \in PS(k).$$

Si $\varepsilon_k \neq \varepsilon_{\min}$, nous essayons d'améliorer l'approximation de $P\partial f(x_k)$ en diminuant ε_k. C'est-à-dire $\varepsilon_{k+1} = \varepsilon_k / a$. On retourne à l'étape 1 avec $x_{k+1} = x_k$ et $S(k+1) = S(x_k, \varepsilon_{k+1})$.

2.4.2 Recherche linéaire

La recherche linéaire utilise la convexité de f. La dérivée de la fonction $\phi_k(t) = f(x_k + td_k)$ est facile à évaluer pour tout $t \geq 0$, puisque le calcul de cette dérivée revient au calcul d'un point de contact $c(d_k)$ sur l'ensemble $\partial f(x_k + td_k)$:

$$\phi_k^{'}(t) = f^{'}(x_k + td_k, d_k) = \max_{s \in \partial f(x_k + td_k)} \langle s, d_k \rangle .$$

Comme démontré plus haut, ce calcul est réalisable une fois obtenus les vecteurs propres adéquats de la matrice $A + D$, évalués au point $x_k + td_k$, i.e. les vecteurs propres $u_{q-i(k)}, ..., u_{q+j(k)}$ de la matrice $A + D_k + t\ diag(d_{kl})$ où $d_k = [d_{kl}]_{1 \leq l \leq n}$. Donc en tout point de la direction de recherche, $\phi_k(t)$ et sa dérivée sont disponibles. Utilisant le fait que ϕ_k est convexe, nous pouvons construire une suite d'approximations successives dont les minima convergent vers un point minimum de ϕ_k. Comme pas initial de la recherche nous prenons $t_1 = \frac{-\varepsilon_k}{f^{'}(x_k, d_k)}$. Il conduit à une décroissance de f à ε_k près si $f(x_k + td_k) = f^{'}(x_k, d_k)t + f(x_k)$. Ayant posé $t_k = \arg\min \phi_k$, notons $x_{k+1} = x_k + t_k d_k$.

2.5 Algorithme S.V.P

A l'itération k, x_k, $\varepsilon_k > 0$, $\varepsilon_{\min} > 0$, $a > 1$, et $z_{k_0} \in PS(x_k, \varepsilon_k)$ sont donnés.

Etape 1. Calculer les valeurs et vecteurs propres de $(A + D_k)$ par l'algorithme de Lançzos.

$$\text{Calculer } \Delta(x_k, \varepsilon_k) = \left\{ \begin{array}{c} \eta : \eta = T\left(U_1(x_k, \varepsilon_k)\right) + T\left(U_2(x_k, \varepsilon_k)H\right) \\ \text{où } H \in H^{i(\varepsilon_k)+1}_{i(\varepsilon_k)+j(\varepsilon_k)+1} \end{array} \right\}$$

et former $PS(x_k, \varepsilon_k) = co\Delta(x_k, \varepsilon_k)$.

Etape 2. Calculer une direction de mouvement d_k sur $\langle e, x \rangle = 0$ par l'algorithme de Frank-Wolfe :

$d_k = -\frac{z_{k_J}}{\|z_{k_J}\|}$ où z_{k_J} est le vecteur de norme minimale à ε_k près de $PS(x_k, \varepsilon_k)$ et
$\varphi_k \equiv \varphi(x_k, \varepsilon_k) \equiv \min_{\|d\| \leq 1} \max_{s \in PS(x_k, \varepsilon_k)} \langle s, d \rangle < -\|z_{k_J}\| + \varepsilon_k$.

Etape 3. Mise à jour de ε_k et recherche linéaire :

si $\varphi_k < -\varepsilon_k$ alors $x_{k+1} = x_k + t_k d_k$ où $t_k = \arg\min_{t \geq 0} f(x_k + t d_k)$

et $\varepsilon_{k+1} = \varepsilon_k$. Aller à l'étape 1.

si $\varphi_k \geq -\varepsilon_k$ alors si $\varepsilon_k \neq \varepsilon_{\min}$, $\varepsilon_{k+1} = \varepsilon_k / a$, $x_{k+1} = x_k$. Aller à l'étape 1.

sinon, aller à l'étape 4.

Etape 4. Critère d'arrêt :

si $\{0\} \in PS(x_k, \varepsilon_{\min})$ alors x_k est solution optimale à $\varepsilon_{\min}$ près.

2.6 Convergence

Théorème 11 *1) f est bornée inférieurement sur l'ensemble $C = \{x : \langle x, e \rangle = 0\}$.*

2) L'intersection de C avec tout ensemble $\{x : f(x) \leq \alpha\}$ est bornée.

3) f atteint sa valeur minimale sur C.

Preuve. 1) Supposons par l'absurde que $f(x_k) \to -\infty$ pour une suite $\{x_k\} \subseteq C$. Alors, par définition de f comme somme des q plus grandes valeurs propres de $A + D_k$, et en tenant compte de l'ordre décroissant dans lequel elles sont rangées, nous obtenons $\lambda_l(A + D_k) \to -\infty$ pour $l > q$. Ce qui entraîne pour $x_k \in C$:

$$tr(A) = tr(A + D_k) = \sum_{l=1}^{n} \lambda_l(A + D_k) \to -\infty.$$

Ce qui est impossible vu que A est une matrice réelle d'ordre n, fixe.

2) Soit α une constante arbitraire. Par 1), f est bornée sur l'intersection en question. Cela veut dire que pour tout x appartenant à cette intersection,

$$\begin{aligned} c_1 &\leq \sum_{l=1}^{q} \lambda_l(A + D) \leq \alpha \text{ et} \\ c_2 &= tr(A + D) = \sum_{l=1}^{n} \lambda_l(A + D) = \sum_{l=1}^{q} \lambda_l(A + D) + \sum_{l=q+1}^{n} \lambda_l(A + D), \end{aligned} \tag{2.18}$$

où c_1, et c_2 désignent des constantes. Ainsi, $\sum_{l=q+1}^{n} \lambda_l(A+D)$ est bornée sur l'intersection de C avec $\{x : f(x) \leq \alpha\}$. Ceci combiné avec (2.18) fait en sorte que pour tout $l = 1, ..., n$, $\lambda_l(A+D)$ est borné sur l'intersection.

Posons $C_k = A + D_k$ et c_3 une constante. Par le théorème de Hoffman-Wielandt [17] nous écrivons :

$$\sum_{l=1}^{n} \Big(\underbrace{\lambda_l(C_k)}_{\text{borné pour tout } x_k \text{ dans l'intersecton}} - \lambda_l(D_k)\Big)^2 \leq \sum_{l=1}^{n} \lambda_l^2(A) = c_3.$$

Ce qui montre que pour $l = 1, ..., n$, $\lambda_l(D_k)$ est borné. Par définition de la matrice diagonale D_k, nous sommes parvenus à montrer que toutes les composantes du vecteur x_k sont bornées, pour $x_k \in C \bigcap \{x : f(x) \leq \alpha\}$.

3) Le dernier point est une conséquence du deuxième et de la continuité de f (en dimension finie, une fonction continue sur un ensemble fermé et borné atteint son minimum). ■
Le lemme suivant affirme qu'à la limite, la direction d_∞ ne peut être une directon de descente au point x_∞.

Lemme 4 *Soit f une fonction continue et Gâteaux différentiable. Soient les suites $\{x_k\}$ et $\{d_k\}$ vérifiant $f(x_{k+1}) \leq f(x_k + td_k)$, $0 \leq t \leq T$, $d_k \to d_\infty$, $x_k \to x_\infty$. Alors $f'(x_\infty, d_\infty) \geq 0$.*

Preuve. Par définition,

$$f'(x_\infty, d_\infty) = \lim_{t \searrow 0} \frac{f(x_\infty + td_\infty) - f(x_\infty)}{t}.$$

Or, pour tout $t \in [0, T]$, $f(x_\infty + td_\infty) = \lim_k f(x_k + td_k) \geq \lim_k f(x_{k+1}) = f(x_\infty)$, puisque f est continue. Nous obtenons alors le résultat du lemme. ■

Lemme 5 *Supposons que la suite $\{x_k\}, k \in K$ converge vers x_∞. Soit u_l^k un vecteur propre associé à $\lambda_l(A + D_k)$, $1 \leq l \leq n$, $k \in K$. Alors pour tout l,*

$$\lambda_l(A + D_k) \to \lambda_l(A + D_\infty),$$

et toute valeur d'adhérence de la suite $\left\{u_l^k\right\}$, $k \in K$ *est un vecteur propre associé à* $\lambda_l(A+D_\infty)$.

Théorème 12 *(continuité relative à* $PS(k)$*)*
Si $\{x_k\} \to x_\infty$ *pour* $k \in K$, *alors pour tout* $\varepsilon > 0$, *tout élément* $s \in P\partial f(x_\infty)$ *peut s'écrire comme la limite d'une suite d'éléments* $s_k \in PS(x_k, \varepsilon)$.

Preuve. Rappelons que

$$P(\partial f(x_\infty)) = coP\Delta(x_\infty) \text{ et } P(S(x_k, \varepsilon)) = coP\Delta(x_k, \varepsilon).$$

Définissons $i^k(\varepsilon)$ et $j^k(\varepsilon)$ les multiplicités intérieure et extérieure de $\lambda_q(A + D_k)$ à ε près et i, j les multiplicités intérieure et extérieure de $\lambda_q(A + D_\infty)$. Soit U^k l'ensemble des vecteurs propres de $A + D_k$ et

$$\begin{aligned} U_1^k(\infty) &= \{u_1, ..., u_{q-\mathrm{i}-1}\} \text{ contenant les } q - i - 1 \text{ premiers vecteurs propres de } U^k, \\ U_2^k(\infty) &= \{u_{q-\mathrm{i}}, ..., u_q, ..., u_{q-\mathrm{j}}\} \text{ contenant } i + j \text{ vecteurs propres de } U^k. \end{aligned}$$

Par le lemme 5, il existe $K' \subseteq K$ et un ensemble de vecteurs $\{U_1, U_2\}$ tels que pour tout $k \in K'$

$$U_1^k(\infty) \to U_1 \quad \text{et} \quad U_2^k(\infty) \to U_2.$$

($U_1 = U_1(x_\infty, 0)$ et $U_2 = U_2(x_\infty, 0)$).
Montrons que pour tout $\varepsilon > 0$ et pour tout $s \in coP\Delta(x_\infty)$, nous avons

$$s = \lim s_k \text{ où } s_k \in coP\Delta(x_k, \varepsilon).$$

- Soit $s \in coP\Delta(x_\infty)$. Cela veut dire qu'il existe des coefficients $\mu_l \geq 0$ tels que $\sum_{l=1}^{n+1} \mu_l = 1$ et

$$s = \sum_{l=1}^{n+1} \mu_l \, g_l,$$

où $g_l = T(U_1) + T(U_2 H) - eq/n$ pour $H \in H_{i+j+1}^{i+1}$ (car $g_l \in P\Delta(x_\infty)$).

Puisque $T(.)$ est continue, nous avons que

$$g_l = \lim_{k \in K'} g_{l_k},$$

où $g_{l_k} = T\left(U_1^k(\infty)\right) + T\left(U_2^k(\infty) H\right) - eq/n.$

- Par le lemme 5, pour tout ε et pour k assez grand, la multiplicité de λ_q^k se rapproche de celle de λ_q^∞ et nous obtenons :

  $U_1(x_k, \varepsilon) = \left\{u_1, ..., u_{q-1-i^k(\varepsilon)}\right\}$ contient au plus $q - i - 1$ vecteurs propres, et

  $U_2(x_k, \varepsilon) = \left\{u_{q-i^k(\varepsilon)}, ..., u_{q+j^k(\varepsilon)}\right\}$ contient au moins $i + j + 1$ vecteurs propres :

Etape $k = 1 \rightarrow$	$u_1^1 ... \ u_{q-i^1(\varepsilon)-1}^1 \ ...$	$...u_q^1...$	$..u_{q+j^1(\varepsilon)}^1...$
:	:	:	:
Etape $k \quad \rightarrow$	$u_1^k ... \ u_{q-i^k(\varepsilon)-1 \ ...}^k$	:	$..u_{q+j^k(\varepsilon)}^k...$
:	:	:	:
:	$u_1^{10}...$	:	:
	$\downarrow$	$\downarrow$	
A la limite $k = \infty \rightarrow$	$\underbrace{u_1 \ ..... \ u_{q-i-1}}_{\in U_1}$	$\underbrace{.....u_q.....u_{q+j}}_{\in U_2}$	...

Ainsi, pour k assez grand, nous notons

$$P\Delta_k(\infty) \equiv \left\{ \underset{\substack{\downarrow \\ \text{i.e. on prend à l'étape } k, \text{ les multiplicités } i \text{ et } j.}}{T\left(U_1^k(\infty)\right) + T\left(U_2^k(\infty) H\right)} - eq/n : H \in H_{i+j+1}^{i+1} \right\}$$

et

$$P\Delta_k(\infty) \subseteq P\Delta(x_k, \varepsilon).$$

- Nous concluons la preuve avec

  $s = \lim_{k \in K'} \sum_{l=1}^{n+1} \mu_l \, g_{l_k}$ où $g_{l_k} \in P\Delta_k(\infty) \subseteq P\Delta(x_k, \varepsilon)$, i.e.

  $s = \lim_{k \in K'} s_k$ où $s_k \in coP\Delta(x_k, \varepsilon)$. ■

Le Théorème 13 est la réciproque du Théorème 12. Ensemble, ils fournissent les arguments de continuité dont on a besoin pour prouver la convergence de l'algorithme.

Théorème 13 *(continuité relative à $PS(k)$)*
Pour $k \in K$, supposons que $\{x_k\} \to x_\infty$, $s_k \in PS(x_k, \varepsilon) = coP\Delta(x_k, \varepsilon)$ et $\{s_k\} \to s_\infty$. Si

$$\varepsilon < \varepsilon^* = \min \{ \mid \lambda_k - \lambda_l \mid : \lambda_k \neq \lambda_l, \text{ valeurs propres de } A + D_\infty \},$$

Alors $s_\infty \in P\partial f(x_\infty) = coP\Delta(x_\infty)$.

Preuve. Pour tout $k \in K$ il existe des coefficients $\mu_{k_l} \geq 0$ tels que $\sum_{l=1}^{n+1} \mu_{k_l} = 1$ et

$$s_k = \sum_{l=1}^{n+1} \mu_{k_l}\, g_{k_l}$$

où $g_{k_l} \in P\Delta(x_k, \varepsilon) \subseteq P\partial_\varepsilon f(x_k)$ pour tout $l = 1, ..., n+1$.
L'ensemble de ces coefficients μ_{k_l} est fermé et borné. Comme la suite $\{x_k\}_{k \in K}$ est bornée, l'ensemble $\bigcup_{k \in K} \partial_\varepsilon f(x_k)$ est aussi borné. Par suite, il existe $K' \subseteq K$ tel que g_{k_l} et μ_{k_l} convergent :

$$\exists\, K' \subseteq K \mid \text{pour } k \in K' \ \mu_{k_l} \to \mu_l \text{ et } g_{k_l} \to g_l.$$

Montrons que $s_\infty = \sum_{l=1}^{n+1} \mu_l\, g_l$ avec $g_l \in P\Delta(x_\infty)$.

- Soient i et j les multiplicités de $\lambda_q(A + D_\infty)$.
  Par le lemme 5, pour k assez grand, avec $\varepsilon < \varepsilon^*$ nous pouvons dire que :

  $U_1(x_k, \varepsilon) = \left\{u_1, ..., u_{q-1-i^k(\varepsilon)}\right\}$ contient $q - i - 1$ vecteurs propres,
  $U_2(x_k, \varepsilon) = \left\{u_{q-i^k(\varepsilon)}, ..., u_{q+j^k(\varepsilon)}\right\}$ contient $i + j + 1$ vecteurs propres,

  et

  $$\exists\, K'' \subseteq K' \mid \text{pour } k \in K'' \ U_1(x_k, \varepsilon) \to U_1 \text{ et } U_2(x_k, \varepsilon) \to U_2.$$

- Par définition de l'ensemble $P\Delta(x_k, \varepsilon)$, pour tout $l = 1, ..., n+1$ et $k \in K''$, il existe une

matrice orthonormale H_{k_l} permettant d'écrire g_{k_l} sous la forme :

$$g_{k_l} = T\left(U_1(x_k, \varepsilon)\right) + T\left(U_2(x_k, \varepsilon) H_{k_l}\right) - eq/n.$$

Par compacité de l'ensemble des matrices orthonormales, pour tout $l = 1, ..., n+1$,

$$\exists K^{'''} \subseteq K^{''} \mid \text{pour } k \in K^{'''} \; H_{k_l} \to H_l \in H_{i+j+1}^{i+1}.$$

La preuve est terminée puisque la continuité de $T(.)$ implique pour $1 \leq l \leq n+1$:

$$g_{k_l} \to g_l \equiv T\left(U_1\right) + T\left(U_2 H_l\right) - eq/n \in P\Delta(x_\infty).$$

■

Les Théorèmes 12 et 13 sont utilisés pour démontrer le Théorème de convergence qui suit.

Théorème 14 *(convergence)*

1) La suite $\{x_k\}$ des itérés engendrés par l'algorithme de la somme de valeurs propres est bornée.

2) $f(x_k)$, $k = 1, 2, ...$ converge vers la valeur minimale de f sur l'espace $\langle e, x \rangle = 0$.

3) Toute valeur d'adhérence x_∞ de la suite des itérés est un minimum de f sur $\langle e, x \rangle = 0$.

Preuve. 1) Conséquence du Théorème 11.

3) Par la deuxième affirmation du Théorème, $\lim\limits_{k \in K} f(x_k) = \min\limits_{x : \langle e, x \rangle = 0} f(x) \equiv f^*$. Par la continuité de f, nous avons le résultat :

$$f^* = \lim_{k \in K' \subseteq K} f(x_k) = f\left(\lim_{k \in K'} x_k\right) = f(x_\infty).$$

2) La preuve se fait en deux étapes :

- montrons que $\{\varepsilon_k\}$ converge vers 0.

Par l'absurde supposons que ε_k ne converge pas vers 0. Pour k assez grand, nous déduisons de l'algorithme que

$$\max_{s \in PS(x_k, \varepsilon)} \langle s, d_k \rangle = \varphi_k < -\varepsilon_k \equiv -\varepsilon \neq 0, \tag{2.19}$$

i.e. à partir d'un certain moment, ε_k ne diminue plus.

Soit x^* une valeur d'adhérence de la suite $\{x_k\}$. Il existe donc $K' \subseteq K$ tel que

$$x_k \rightarrow x^*, \text{ et} \tag{2.20}$$

$$d_k \rightarrow d^*, \tag{2.21}$$

puisque d_k est élément d'un ensemble compact.

Notons $\max_{s \in PS(x_k,\varepsilon)} \langle s, d^* \rangle \equiv \langle s^\circ, d^* \rangle$. L'ensemble $PS(x_k, \varepsilon)$ étant uniformément borné, il existe une constante R telle que

$$\begin{aligned} \max_{s \in PS(x_k,\varepsilon)} \langle s, d^* \rangle - \max_{s \in PS(x_k,\varepsilon)} \langle s, d_k \rangle &\leq \langle s^\circ, d^* \rangle - \langle s^\circ, d_k \rangle \\ = \langle s^\circ, d^* - d_k \rangle &\leq \|s^\circ\| \, \|d^* - d_k\| \leq R\|d^* - d_k\|. \end{aligned}$$

D'où, pour k suffisamment grand et par (2.19) et (2.21),

$$\max_{s \in PS(x_k,\varepsilon)} \langle s, d^* \rangle \leq \varphi_k + R\|d^* - d_k\| \leq \frac{-\varepsilon}{2}. \tag{2.22}$$

Soit $s \in P\partial f(x^*)$. En vue du Théorème 12 il existe un ensemble $K'' \subseteq K'$ et des éléments $s_k \in PS(x_k, \varepsilon)$ tels que pour $k \in K''$

$$s_k \rightarrow s.$$

Par construction et par (2.22) :

$$\langle s, d^* \rangle = \lim_{k \in K''} \langle s_k, d^* \rangle \leq \lim_{k \in K''} \max_{v \in PS(x_k,\varepsilon)} \langle v, d^* \rangle \leq -\tfrac{\varepsilon}{2}.$$

Comme s est arbitraire et $\langle e, d^* \rangle = 0$ (car d^* est une direction admissible), nous déduisons de la dernière ligne que

$$f'(x^*, d^*) \leq -\frac{\varepsilon}{2},$$

ce qui contredit le lemme 4 (f étant convexe, est bien Gâteaux différentiable).

- Montrons que $f'(x^*, d) \geq 0$ pour tout d vérifiant $\langle e, d \rangle = 0$.

Nous venons de montrer que $\varepsilon_k \to 0$. Nous pouvons alors trouver $K' \subseteq K$ tel que

$$\varphi_k > -\varepsilon_k.$$

C'est à dire, K' contient les indices pour lesquels ε_k diminue (cfr l'algorithme).

x^* étant une valeur d'adhérence de $\{x_k\}_{k\in K'}$, choisissons $K'' \subseteq K'$ de façon à avoir, $k \in K''$,

$$\varphi_k \to \delta \geq 0 \quad \text{et} \quad x_k \to x^*.$$

Par définition, pour tout d tel que $\|d\| = 1$, nous avons :

$$\varphi_k \leq \max_{s\in PS(x_k,\varepsilon_k)} \langle s, d\rangle .$$

On peut donc trouver $g_k \in PS(x_k, \varepsilon_k)$ satisfaisant l'inégalité

$$\varphi_k \leq \langle g_k, d\rangle . \tag{2.23}$$

Par le même argument utilisé dans la preuve précédente, l'ensemble $\bigcup_{k\in K} \partial_\varepsilon f(x_k)$ est borné et il existe $K''' \subseteq K''$ tel que

$$g_k \to g^*.$$

Pour k suffisamment grand et

$\varepsilon^* = \min\{\ |\ \lambda_k - \lambda_l\ |\ : \lambda_k \neq \lambda_l,$ valeurs propres de $A + D^*\}$, nous avons :

$$g_k \in PS(x_k, \varepsilon^*)$$

(car $\varepsilon_k \to 0$ et $PS(x_k, \varepsilon_k) \subseteq PS(x_k, \varepsilon^*)$ pour k assez grand). La conclusion du Théorème 13 nous amène à écrire que

$$g^* \in P\partial f(x^*). \tag{2.24}$$

Par définition du sous différentiel et en prenant la limite sur $k \in K'''$ dans la relation (2.23), (2.24) implique

$$f'(x^*, d) \geq \langle g^*, d\rangle \geq 0,$$

pour toute direction d vérifiant $\langle d, e\rangle = 0$. Par conséquent, x^* est un minimum de f sur $\langle d, e\rangle = 0$.

Comme f est continue et que $f(x_k)$ est monotone décroissante, la preuve se termine avec $f(x_k) \to f(x^*)$, la valeur minimale de f sur $\langle d, e\rangle = 0$ et toute valeur d'adhérence de $\{x_k\}$ est un point minimisant f. ■

Nous sommes parvenus à construire un algorithme implémentable et convergent, qui calcule "la meilleure" borne inférieure du nombre de connexions externes de toute partition des sommets d'un graphe en q groupes de taille égale.

Chapitre 3

Minimiser la plus grande valeur propre (en valeur absolue) d'une matrice symétrique

3.1 Problème et notations

Un problème souvent rencontré en ingénierie du contrôle est de minimiser $\phi(x)$, la plus grande valeur propre (en valeur absolue) d'une matrice symétrique dépendant du paramètre x. Si la matrice est une fonction affine, $\phi(x)$ est convexe. Cependant, $\phi(x)$ n'est pas différentiable, car les valeurs propres ne sont pas différentiables aux points où elles coïncident (cfr Chapitre 1). De plus, on s'attend à ce que la solution soit un point "non différentiable", puisque la minimisaton de $\phi(x)$ conduit, la plupart du temps, plusieurs valeurs propres à la même valeur minimale. Dans ce chapitre, nous construisons un algorithme qui converge de façon quadratique, vers le minimum de $\phi(x)$. Une caractéristique importante de cet algorithme est qu'il sépare si nécessaire, des valeurs propres égales, dans le but d'obtenir une direction de descente à partir de tout point qui n'est pas optimal.

Plus précisément, notons A, une matrice réelle, symétrique d'ordre $n \times n$, dont les entrées sont des fonctions affines de x, où $x \in \mathrm{R}^m$. Soient $\lambda_i(A(x))$ ses valeurs propres rangées dans un ordre décroissant, pour $i = 1, ..., n$. Le problème à résoudre est alors le problème convexe, non

différentiable suivant :

$$\min_{x \in \mathrm{R}^m} \phi(x) \tag{3.1}$$

où $\phi(x) = \max\limits_{1 \leq i \leq n} \mid \lambda_i(A(x)) \mid$.
Comme $A(x)$ est une fonction affine, nous pouvons l'écrire sous la forme

$$A(x) = A_0 + \sum_{k=1}^{m} x_k A_k.$$

Un cas intéressant est celui où

$$A_k = e_k e_k^t$$

avec e_k désignant la k-ème colonne de la matrice identité, de sorte que

$$A(x) = A_0 + diag(x).$$

Remarquons que le problème qui consiste à minimiser la plus grande valeur singulière (en valeur absolue) d'une matrice non symétrique $G(x)$, comme fonction affine de x peut être écrit sous la forme de (3.1) car les valeurs propres de

$$\begin{bmatrix} 0 & G(x) \\ G(x) & 0 \end{bmatrix}$$

sont plus ou moins les valeurs singulières de $G(x)$.

Le problème (3.1) peut être réécrit comme un problème d'optimisation non différentiable avec contraintes :

$$\begin{cases} \min\limits_{w \in \mathrm{R},\ x \in \mathrm{R}^m} w \\ \text{s.c. } -w \leq \lambda_i(A(x)) \leq w, \quad i = 1, ..., n, \end{cases} \tag{3.2}$$

ou de façon équivalente,

$$(\mathrm{P}) \begin{cases} \min\limits_{w \in \mathrm{R},\ x \in \mathrm{R}^m} w \\ \text{s.c. } wI - A(x) \geq 0 \\ \quad wI + A(x) \geq 0 \end{cases} \tag{3.3}$$

où "$\geq$" signifie "semi-définie positive" et I désigne la matrice identité. Comme une matrice

semi-définie positive a toutes ses valeurs propres positives ou nulles, les contraintes de (P) expriment bien les contraintes de (3.2). Résoudre le problème (P) revient à chercher le plus petit intervalle $[-w, w]$, contenant toutes les valeurs propres de $A(x)$.

3.2 Condition nécessaire et suffisante d'optimalité

Pour plus de facilité, considérons l'ensemble de toutes les matrices symétriques comme étant l'espace des variables, et notons K, le cône des matrices semi-définies i.e.

$$K = \{M \mid M \geq 0\}.$$

Définissons sur les matrices symétriques, le produit interne $\langle\langle A, B\rangle\rangle = trAB$. Le cône normal à l'ensemble convexe K en $M^{'}$ est défini par

$$\mathrm{N}K(M^{'}) = \left\{B \mid \left\langle\left\langle B, M - M^{'}\right\rangle\right\rangle \leq 0 \quad \forall M \in K\right\},$$

ou encore par

$$\mathrm{N}K(M^{'}) = \left\{B \mid \left\langle\left\langle B, M^{'}\right\rangle\right\rangle = \sup_{M \in K} tr\,(BM)\right\}.$$

Fletcher [9] a montré que ce cône normal peut encore s'écrire

$$\mathrm{N}K(M^{'}) = \left\{B \mid B = -ZUZ^{t},\ U = U^{t},\ U \geq 0\right\}, \tag{3.4}$$

où Z est une matrice dont les colonnes engendrent le noyau de $M^{'}$ (noté $\mathrm{Ker}(M^{'})$).

Notons ensuite

$$\begin{aligned}
\hat{K}_1 &= \{(w, x) \mid wI - A(x) \geq 0;\ w \in \mathrm{R};\ x \in \mathrm{R}^m\} \\
\hat{K}_2 &= \{(w, x) \mid wI + A(x) \geq 0;\ w \in \mathrm{R};\ x \in \mathrm{R}^m\},
\end{aligned}$$

les ensembles admissibles associés aux deux contraintes du problème (P). Par définition, et pour $i = 1, 2$,

$$\mathrm{N}\hat{K}_i(w^{'}, x^{'}) = \left\{(\delta, d) \mid \left(w^{'}, x^{'}\right)^{t}(\delta, d) = \sup_{(w,x) \in \hat{K}_i} (w, x)^{t}(\delta, d)\right\}.$$

Le lemme qui suit nous permettra d'établir la condition d'optimalité pour que x résolve (3.1). Pour la preuve, nous renvoyons le lecteur au théorème 4.1 de Fletcher [9].

Lemme 6

$$N\hat{K}_1(w^{'},x^{'}) = \left\{ \begin{array}{r} (\delta,d) \mid \delta = tr(B);\ d_k = -tr(BA_k),\ \ k=1,...,m, \\ B \in NK(w^{'}I - A(x^{'})) \end{array} \right\},$$

$$N\hat{K}_2(w^{'},x^{'}) = \left\{ \begin{array}{r} (\delta,d) \mid \delta = tr(B);\ d_k = -tr(BA_k),\ \ k=1,...,m, \\ B \in NK(w^{'}I + A(x^{'})) \end{array} \right\}.$$

Pour énoncer les conditions nécessaires et suffisantes d'optimalité, nous avons encore besoin de définir les notations suivantes :

Soit (w,x) un point admissible pour le problème (3.2). Nous désignerons par t (respectivement s) le nombre de contraintes $\lambda_i(A(x)) \leq w$ (respectivement $\lambda_i(A(x)) \geq -w$) actives en (w,x). Si aucune contrainte n'est active, nous poserons $t=0$ (respectivement $s=0$). Comme les valeurs propres sont classées par ordre décroissant, nous avons

$$\begin{aligned} \lambda_i(A(x)) &= w \quad i=1,...,t, \\ \lambda_i(A(x)) &= -w \quad i=n-s+1,...,n. \end{aligned}$$

Lorsque $t \neq 0$ ($s \neq 0$) nous noterons Q_1 (respectivement Q_2) la matrice dont les colonnes sont les vecteurs propres associés à λ_i, $i=1,...,t$ (respectivement $i=n-s+1,...,n$).

Théorème 15 *(C.N.S d'optimalité.)*
$\bar{x}$ est solution optimale de (3.1) si et seulement si il existe des matrices U et V de dimensions respectives $t \times t$ et $s \times s$, avec $U = U^t \geq 0$, $V = V^t \geq 0$, telles que

$$trU + trV = 1, \tag{3.5}$$

$$\left\langle \left\langle Q_1^t A_k Q_1, U \right\rangle \right\rangle - \left\langle \left\langle Q_2^t A_k Q_2, V \right\rangle \right\rangle = 0 \quad k=1,...,m, \tag{3.6}$$

où t, s, Q_1, et Q_2 ont été définis plus haut.

Preuve. Comme (3.1) est équivalent au problème convexe (3.3), la condition nécessaire et suffisante d'optimalité du point $(\bar{w}, \bar{x})$ est

$$-\nabla_{w,x} w \in \mathrm{N}C(\bar{w}, \bar{x}), \tag{3.7}$$

où C est l'ensemble décrit par les contrainte de (3.3), i.e. $C = \hat{K}_1 \bigcap \hat{K}_2$. La condition (3.7) devient alors :

$$\begin{bmatrix} 1 \\ 0 \\ : \\ 0 \end{bmatrix} + g_1 + g_2 = 0,$$

avec $g_1 \in \mathrm{N}\hat{K}_1(\bar{w}, \bar{x})$, et $g_2 \in \mathrm{N}\hat{K}_2(\bar{w}, \bar{x})$ [16, Chap. 5]. Par le Lemme 6, cette dernière expression peut se réécrire sous forme d'un système :

$$\begin{cases} 1 + tr(B_1) + tr(B_2) = 0, \\ \langle\langle B_1, A_k \rangle\rangle + \langle\langle B_2, A_k \rangle\rangle = 0 \quad k = 1, ..., m, \end{cases}$$

où $B_1 \in \mathrm{N}K(\bar{w}I - A(\bar{x}))$ et $B_2 \in \mathrm{N}K(\bar{w}I + A(\bar{x}))$. Par l'équation (3.4), nous avons

$$B_1 = -Q_1 U Q_1^t, \quad B_2 = -Q_2 V Q_2^t$$

pour $U = U^t \geq 0$, et $V = V^t \geq 0$, puisque Q_1 est une base de $\mathrm{Ker}(\bar{w}I - A(\bar{x}))$ et Q_2 est une base de $\mathrm{Ker}(\bar{w}I + A(\bar{x}))$.
Or, $tr(Q_1 U Q_1^t) = trU$, de même $tr(Q_2 V Q_2^t) = trV$, et $\langle\langle Q_1 U Q_1^t, A_k \rangle\rangle = tr(Q_1 U Q_1^t A_k)$
$= tr(U Q_1^t A_k Q_1) = \langle\langle Q_1^t A_k Q_1, U \rangle\rangle$, de même pour $\langle\langle Q_2 V Q_2^t, A_k \rangle\rangle$. Le théorème est ainsi prouvé. ■

Les matrices U et V de (3.5) et (3.6) jouent le rôle des multiplicateurs de Lagrange. Comme les conditions d'optimalité $U \geq 0$, $V \geq 0$ sont des conditions sur les matrices elles-mêmes et non sur leurs composantes, nous appellerons U et V les matrices de Lagrange.

3.3 Méthode des contraintes actives

3.3.1 Idée d'algorithme

Une stratégie de contraintes actives consiste à ramener la résolution de (P) à la résoluton d'un problème d'égalité plus simple :

$$(\text{P.E}) \begin{cases} \min\limits_{w\in\mathrm{R},\ x\in\mathrm{R}^m} w \\ \text{s.c. } \lambda_i(x) = w \quad i = 1, ..., t, \\ \quad \lambda_i(x) = -w \quad i = n-s+1, ..., n, \end{cases} \tag{3.8}$$

où t et s, les nombres de contraintes actives à l'optimum sont inconnues et donc à estimer. Notre méthode s'apparente à celles décrites par Friedland, Nocedal et Overton [11], pour résoudre le système

$$\lambda_i\left(A(x)\right) = w \quad i = 1, ..., t, \tag{3.9}$$

$$\lambda_i\left(A(x)\right) = \mu_i \quad i = t+1, ..., \hat{t}, \tag{3.10}$$

où $(w, \{\mu_i\})$ sont des valeurs données distinctes, et $t, \hat{t}$ sont "bien" choisis. Ils montrent aussi que la condition (3.9) impose sur le paramètre x, $t(t+1)/2$ contraintes linéaires indépendantes et non t, comme on pourrait le croire à première vue. Les méthodes numériques se basent sur cette considération. Notre algorithme peut être vu comme une généralisation des méthodes de Friedland, Nocedal et Overton, pour résoudre (P.E). Comme résultat de la minimisation, nous obtenons $w = \max(\lambda_1, \lambda_n)$ avec

$$w = \lambda_1 = ... = \lambda_t > \lambda_{t+1} \geq ... \geq \lambda_{n-s} > \lambda_{n-s+1} = ... = \lambda_n = -w. \tag{3.11}$$

Nous voyons ici que le fait de diminuer la multiplicité de $\lambda_1(x)$ ou $\lambda_n(x)$, entraîne la disparition de contraintes d'égalité. Nous utiliserons donc pour résoudre le problème (P), la stratégie des contraintes actives, reprise dans l'algorithme suivant.

Etant donné t, s, et x_1, un point admissible pour (P), poser $k = 1$.

Etape 1. Si x_k est solution du problème (P.E), aller à l'étape 2.
Sinon, calculer en x_k une direction de descente admissible d_k pour (P.E),
et aller à l'étape 3.

Etape 2. Evaluer les matrices de Lagrange U, V associées à (P.E).
Si $U \geq 0$ et $V \geq 0$ alors x_k est solution de (P).
Sinon, séparer une valeur propre multiple, et
obtenir une réduction de t et/ou de s,
calculer en x_k une direction de descente d_k, admissible pour le nouveau problème (P.E) et aller à l'étape 3.

Etape 3. Effectuer une recherche linéaire admissible pour (P), le long de d_k.
Poser $x_{k+1} = x_k + t_k d_k$.

Etape 4. Mise à jour des multiplicités t et s.

Etape 5. $k = k + 1$ et aller à l'étape 1.

Les prochaines sections ont pour objet d'éclaircir chacune de ces étapes.

3.3.2 Résolution du problème (PE)

Comme expliqué par Friedland, Nocedal et Overton, nous pouvons obtenir une méthode quadratiquement convergente pour résoudre le système non différentiable (3.9)(3.10) en appliquant une variante de la méthode de Newton au système non linéaire, mais différentiable

$$\begin{aligned} Q_1(x)^t A(x) Q_1(x) &= wI_t, \\ q_i(x)^t A(x) q_i(x) &= \mu_i \quad i = t+1, ..., \hat{t}, \end{aligned}$$

où les colonnes de $Q_1(x)$ forment un ensemble orthonormal de vecteurs propres de $A(x)$, correspondant à w (remarquons que $q_i(x)^t A(x) q_i(x)$ n'est rien d'autre que $\lambda_i(x)$). Ici, I_t dénote la matrice unité d'ordre t. En vue d'acquérir la convergence quadratique, nous avons besoin

que le nombre d'équations, $t(t+1)/2 + (\hat{t} - t)$, soit égal au nombre de variables. Le système d'équations à résoudre à chaque pas de la méthode de Newton est

$$\begin{aligned} Q_1(x)^t A(x+d) Q_1(x) &= wI_t, \\ q_i(x)^t A(x+d) q_i(x) &= \mu_i \quad i = t+1, ..., \hat{t}, \end{aligned}$$

(car la dérivée de $\lambda_i(x)$ par rapport à x_k s'exprime par $q_i(x)^t A_k q_i(x)$), où x est l'itéré courant et $x + d$ devient le nouvel itéré. La matrice $A(x + d)$ étant affine, ces équations forment un système linéaire en d. Une fois obtenu $x + d$, il faut calculer les valeurs et vecteurs propres de A au nouveau point, afin de commencer une nouvelle itération.

A présent, généralisons cette méthode pour résoudre (P.E). Pour l'instant, supposons que t et s sont connus. Appliquons la méthode de Newton au problème non linéaire mais différentiable,

$$\text{(P.E)} \begin{cases} \min\limits_{w \in \mathrm{R},\ x \in \mathrm{R}^m} w \\ \text{s.c. } wI_t - Q_1(x)^t A(x) Q_1(x) = 0, \\ \quad wI_s + Q_2(x)^t A(x) Q_2(x) = 0. \end{cases} \tag{3.12}$$

Le sous problème à résoudre à chaque itération est le problème quadratique suivant :

$$\text{(Q.P.E)} \begin{cases} \min\limits_{w \in \mathrm{R},\ d \in \mathrm{R}^m} w + \frac{1}{2} d^t W(x, U, V)\, d \\ \text{s.c. } wI_t - Q_1(x)^t A(x+d) Q_1(x) = 0, \\ \quad wI_s + Q_2(x)^t A(x+d) Q_2(x) = 0. \end{cases} \tag{3.13}$$

Les contraintes de (P.E) ont été linéarisées comme suit :

$$\begin{aligned} &[wI_t - Q_1(x)^t A(x) Q_1(x)] + \left[0 - \sum_{k=1}^{m} Q_1(x)^t A_k(x) Q_1(x)\, d_k\right] \\ &= wI_t - Q_1(x)^t \left[A(x) + \sum_{k=1}^{m} A_k d_k\right] Q_1(x) = wI_t - Q_1(x)^t A(x+d) Q_1(x). \end{aligned}$$

Le choix de la matrice $W(x, U, V)$ sera discuté plus tard.

Rappelons **l'algorithme (S.Q.P)** :

étant donné t, s, et x_1, U_1, V_1, poser $k = 1$.

Etape 1. Résoudre $(Q.P.E(x_k, U_k, V_k))$. On obtient pour solution $d_k$et les "multiplicateurs" U_{k+1}, V_{k+1}.

Etape 2. Si $d_k = 0$, alors (x_k, U_{k+1}, V_{k+1}) est un point de Kuhn-Tucker pour (P.E).
Sinon, $x_{k+1} = x_k + d_k$, $k = k + 1$ et retourner à l'étape 1.

Définissons la fonction lagrangienne associée à (3.12) par

$$L(w, x, U, V) = w - \left\langle\left\langle U, \left(wI_t - Q_1(x)^t A(x) Q_1(x)\right)\right\rangle\right\rangle - \left\langle\left\langle V, \left(wI_s + Q_2(x)^t A(x) Q_2(x)\right)\right\rangle\right\rangle \tag{3.14}$$

où $U = U^t$, $V = V^t$. La condition nécessaire du premier ordre pour que x soit solution de (P.E), à savoir $\nabla_{w,x} L = 0$, est qu'il existe des matrices symétriques U et V, vérifiant (3.5) et (3.6) (cette condition est aussi nécessaire pour le problème limite d'une séquence de problèmes (Q.P.E).). Cette équivalence avec les conditions d'optimalité (3.5), (3.6) est très importante, car elle veut dire que les matrices de Lagrange requises pour établir les conditions (3.5), (3.6) peuvent être obtenues en résolvant (P.E), ou plus précisément, en résolvant une suite de problèmes (Q.P.E). Le point clef est que (P.E) est plus facile à résoudre que le problème original. Un autre point à noter est que U et V n'ont pas besoin d'être semi-définies positives pour une solution optimale de (P.E), puisque les contraintes sont des égalités.

Remarquons que le problème (Q.P.E) n'admet pas toujours de solutions. En effet, si U ou V est indéfinie, alors t ou s est trop grand, et il est nécessaire de "séparer" une valeur propre multiple afin de diminuer le nombre de contraintes. Le problème (P.E) possède

$$\frac{t(t+1)}{2} + \frac{s(s+1)}{2} \tag{3.15}$$

contraintes.

- Si cette quantité est égale à $m+1$ (le nombre de variables de (P.E)), alors les contraintes seules suffisent à déterminer une solution unique de (P.E), et la méthode S.Q.P converge de façon locale et quadratique, sans se soucier de la matrice W.

- Si (3.15) est plus grand que $m+1$, alors, excepté dans les cas dégénérés, (P.E) est impossible à résoudre. En général, on espère que (3.15) soit plus petit ou égal à $m+1$. Malheureusement, on ne peut s'attendre à avoir l'égalité, par exemple pour $m = 4$, l'égalité est impossible.

- Si (3.15) est plus petit que $m+1$, alors, le choix de la matrice W est important pour assurer la convergence quadratique de la méthode S.Q.P. Un choix naturel pour cette matrice est de prendre le Hessien par rapport à x de la fonction lagrangienne (3.14). Il est possible de montrer que cette matrice est donnée par

$$W_{jk} = \left\langle\left\langle U, G_1^{j,k}\right\rangle\right\rangle - \left\langle\left\langle V, G_2^{j,k}\right\rangle\right\rangle, \tag{3.16}$$

où : 1.

$$G_l^{j,k} = 2Q_1(x)^t A_k \bar{Q}_l(x) \left(wJ_l - \bar{\Lambda}_l(x)\right)^{-1} \bar{Q}_1(x)^t A_j Q_l(x), \quad l = 1, 2,$$

2. les colonnes de $\bar{Q}_l(x)$ sont formées par les vecteurs $q_1, ..., q_n$, excepté ceux de $Q_l(x)$ $(l = 1, 2)$,

3. $\bar{\Lambda}_l(x)$ est une matrice diagonale dont les entrées sont les valeurs propres $\lambda_1, ..., \lambda_n$, excepté celles correspondant à $Q_l(x)$,

4. $J_1 = I_{n-t}$, $J_2 = I_{n-s}$.

Cependant, il faut être prudent lorsque nous utilisons U et V dans (3.16), car nous ne connaissons pas leur valeur au minimum de (P.E). Nous utiliserons donc des estimations de ces matrices lagrangiennes. La première idée est d'utiliser comme estimation, les valeurs obtenues au problème (Q.P.E) précédent. Malheureusement, ceci est inefficace, puisque les bases de vecteurs propres $Q_1(x)$ et $Q_2(x)$ évaluées à l'itéré courant n'ont, en général, pas de relations avec celles évaluées au point précédent. Ainsi, après avoir calculé $Q_1(x)$ et $Q_2(x)$, et avant de résoudre (Q.P.E), nous calculerons les estimations du premier ordre de U et V en minimisant la norme euclidienne du résidu provenant de la comparaison entre (3.5), (3.6) et la condition nécessaire d'optimisation pour le problème (Q.P.E) (pour plus de détails sur l'estimation des multiplicateurs de Lagrange pour des problèmes de *min-max*, voir Murray et Overton [13]).

3.3.3 Mise à jour des multiplicités t et s

Dans cette section, nous nous tournons sur l'importante question de la mise à jour des multiplicités supérieure et inférieure, t et s. Avant de commencer, supposons qu'initialement $t = 1$ et $s = 0$. Si nous résolvions (Q.P.E), la solution réduirait certainement $\lambda_1(x)$ à une valeur très inférieure aux autres valeurs propres. Ce qui ne correspondrait plus à notre problème (P). Pour éviter cela, il suffit d'incorporer au problème (Q.P.E), des contraintes d'inégalités qui forcent les valeurs propres à rester dans $[-w, w]$:

$$-w \leq q_i^t(x)A(x+d)q_i(x) \leq w, \quad t+1 \leq i \leq n-s. \tag{3.17}$$

A présent, nous pouvons obtenir la mise à jour de t et s en cherchant les contraintes qui sont actives en la solution de (Q.P.E), i.e. de (3.13) et (3.17). Une stratégie simple est d'augmenter t (respectivement s) par le nombre de contraintes qui atteignent la borne supérieure (respectivement inférieure) dans (3.17). A nouveau il nous faudra être prudent, car si t et s deviennent trop grands, le problème (P.E) devient impossible à résoudre. C'est pour cette raison que nous introduisons $\bar{t}$ et $\bar{s}$, des multiplicités supérieure et inférieure à une certaine tolérance près, qui seront définies en début de chaque itération par :

$$w - \lambda_i(x) \leq \text{TOL}, \quad i = 1, ..., \bar{t}, \tag{3.18}$$

$$w + \lambda_i(x) \leq \text{TOL}, \quad i = n - \bar{s} + 1, ..., n, \tag{3.19}$$

supposant que $\frac{\bar{t}(\bar{t}+1)}{2} + \frac{\bar{s}(\bar{s}+1)}{2} \leq m+1$, et que TOL est un nombre suffisamment petit, par exemple 10^{-2}. Comme $\bar{t} \geq t$, $\bar{s} \geq s$, si nous remplaçons t et s par $\bar{t}$ et $\bar{s}$ dans (3.17), nous constatons une diminution du nombre de contraintes. Cela revient à dire que lors de la mise à jour, nous compterons moins de bornes atteintes. Si nécessaire, t et s seront remplacés par $\bar{t}$ ou $\bar{s}$ comme expliqué plus loin. Mais, si t et s sont toujours initialisés à $\bar{t}$ et $\bar{s}$ au lieu d'utiliser l'information des contraintes actives évaluées en la solution du (Q.P.E) précédent, l'algorithme convergera beaucoup plus lentement; en fait, la convergence ne sera plus quadratique.

3.3.4 Recherche linéaire

Même si t et s ont les valeurs correctes définies en (3.11), il n'est pas garanti que la solution d de (3.13), (3.17) donnera

$$\phi(x+d) < \phi(x).$$

Fletcher suggère donc une stratégie de "région de confiance", à incorporer aux contraintes de Q.P.E :

$$\mid d_k \mid \leq \rho, \quad k = 1, ..., m, \tag{3.20}$$

où ρ est ajusté de façon dynamique. Il est clair que si ρ et TOL sont suffisamment petits, la solution d de (Q.P.E) i.e. de (3.13), (3.17) et (3.20), avec $t = \bar{t}$, $s = \bar{s}$, donnera $\phi(x+d) < \phi(x)$, à moins que $d = 0$.

3.3.5 Séparer des valeurs propres multiples

Si TOL$= 0$, $t = \bar{t}$, $s = \bar{s}$, et si la solution d de (Q.P.E) est nulle, alors le point x est minimum de (P.E). Et il résout (P), si les matrices lagrangiennes U et V sont semi-définies positives. Si U ou V est définie négative, il est possible et nécessaire pour progresser, de "séparer" une valeur propre multiple, comme expliqué ci-après.

Exemple 5 *Prenons $m = n = 2$, avec*

$$A_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \; A_1 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \; A_2 = \begin{bmatrix} 1 & k \\ k & 4 \end{bmatrix},$$

pour toute valeur réelle de k. Comme $A(x) = A_0 + A_1 x_1 + A_2 x_2$, le seul point pour lequel $A(x)$ a des valeurs propres multiples est $x = (0,0)^t$ avec $\lambda_1 = \lambda_2 = 1$. En effet l'équation en λ, $\det(A - \lambda I) = 0$, *a pour réalisant*

$$\rho = (3x_2)^2 + (2k\,x_2)^2 + (2x_1)^2 - 12x_1x_2,$$

et donc pour solution(s)

$$\lambda(x_1, x_2) = \frac{(2 + 5x_2) \pm \sqrt{\rho}}{2}.$$

$(0,0)^t$ *est donc une solution de (3.12) avec* $t = 2$ *et* $s = 0$. *Si* k *est assez grand, les valeurs propres, en valeur absolue s'éloignent de* 1, *et* $x = (0,0)^t$ *est bien un minimum de* $\phi(x)$. *Par contre, si* k *est suffisamment petit, il est possible de trouver une direction de descente en* $(0,0)$. *Il paraît donc nécessaire de distinguer ces situations et de trouver une direction de descente lorsqu'elle existe.*

Ecrivons les conditions d'optimalité. Le système (3.5) (3.6) donne avec le choix $Q_1 = I$,

$$\begin{bmatrix} 1 & 1 & 0 \\ -1 & 1 & 0 \\ -1 & -4 & -k \end{bmatrix} \begin{bmatrix} U_{11} \\ U_{22} \\ 2U_{12} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

Il a pour solution

$$U = \begin{bmatrix} \frac{1}{2} & \frac{-5}{4k} \\ \frac{-5}{4k} & \frac{1}{2} \end{bmatrix}.$$

La condition d'optimalité est $U \geq 0$, *i.e.*

$$| k | \geq \frac{5}{2}.$$

Nous allons voir comment obtenir une direction de descente si $| k | < \frac{5}{2}$. *L'astuce est de résoudre*

$$\delta I - \sum_{k=1}^{2} d_k A_k = -\mu u u^t,$$

où μ *est la valeur propre négative de* U *et* u *est le vecteur propre correspondant. Ceci donne pour* $k = 2.25$,

$$\begin{bmatrix} 1 & -1 & -1 \\ 1 & 1 & -4 \\ 0 & 0 & -k \end{bmatrix} \begin{bmatrix} \delta \\ d_1 \\ d_2 \end{bmatrix} = 2.78 \times 10^{-2} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},$$

i.e., $\delta = -3.09 \times 10^{-3}$, $d = (-1.85 \times 10^{-2}, -1.23 \times 10^{-2})^t$. *A présent* $\lambda(A(x+d)) \equiv \lambda(A(d_1, d_2)) = (0.941, 0.997)^t$, *de sorte que* $\phi(x+d) < \phi(x)$.

De façon plus générale, nous avons le théorème suivant :

Théorème 16 *Soient t et s définis par (3.11). Supposons que x est un minimum de (P.E) qui vérifie donc (3.5) et (3.6), pour des matrices symétriques $U \in R^{t\times t}$ et $V \in R^{s\times s}$. Si U est définie négative, avec μ une de ses valeurs propres négatives correspondant au vecteur propre u, et si (δ, d) résout*

$$\delta I_t - \sum_{k=1}^{m} d_k Q_1^t A_k Q_1 = -\mu u u^t, \tag{3.21}$$

$$\delta I_s + \sum_{k=1}^{m} d_k Q_2^t A_k Q_2 = 0, \tag{3.22}$$

alors, d est une direction de descente pour $\phi(x)$. De plus la multiplicité t est réduite au premier ordre d'exactement une unité le long de d, et le nouvel ensemble de vecteurs propres associés à $\lambda_i = w$ peut être pris au premier ordre, comme

$$\tilde{Q}_1 = Q_1 \bar{U},$$

où les colonnes de $\bar{U}$ sont les vecteurs propres de U exepté u.

Remarque 5 *Les équations (3.21) et (3.22) sont généralement solubles si (3.15) est inférieure ou égale à $m+1$. Les autres cas sont des situations dégénérées pour lesquelles il est difficile de trouver une direction de descente.*

Preuve. Prenons les produits internes de U avec (3.21) et de V avec (3.22). Additionnons-les. Comme $tr(I\,U) = trU$, et $Uuu^t = \mu$, nous obtenons

$$\delta\,(trU + trV) + \sum_{k=1}^{m} d_k\,(\langle\langle -U, Q_1^t A_k Q_1\rangle\rangle + \langle\langle V, Q_2^t A_k Q_2\rangle\rangle) = -\mu^2.$$

Il suit donc de (3.5) et (3.6) que

$$\delta = -\mu^2. \tag{3.23}$$

De plus, de la même façon que les contraintes de (Q.P.E) sont une linéarisation des contraintes de (P.E), (3.21) et (3.22) montrent que les contraintes de (P.E) sont vérifiées au premier ordre le long de la direction $x + \alpha d$, $\alpha \geq 0$ (car les membres de droite de (3.21) et (3.22) sont semi-définis positifs). Nous déduisons de (3.23) que d est une direction de descente. Finalement, nous vérifions la dernière hypothèse en multipliant (3.21) par $(u, \bar{U})^t$ à gauche et par $(u, \bar{U})$ à droite, de manière à obtenir :

$$\delta I_t - \sum_{k=1}^{m} d_k(u,\bar{U})^t Q_1^t A_k Q_1(u,\bar{U}) = \begin{bmatrix} -\mu & \\ & 0 \end{bmatrix}, \text{ c'est-à-dire,}$$

$$-\sum_{k=1}^{m} d_k(u,\bar{U})^t Q_1^t A_k Q_1(u,\bar{U}) = \begin{bmatrix} \mu^2 - \mu & \\ & diag(\mu^2) \end{bmatrix}. \blacksquare$$

En d'autres mots, toutes les valeurs propres, sauf une, sont réduites à μ^2 (au premier ordre), tandis que l'autre valeur propre est réduite à $\mu^2 - \mu$.

Notons que si U a plus d'une valeur propre négative (ou si U et V ont tous deux des valeurs propres négatives), alors nous pouvons réduire t de plus d'une unité (ou réduire t et s) en remplaçant le membre de droite de (3.21) (et (3.22)) par une somme de produits internes correspondant aux valeurs propres négatives.

Nous concluons ce chapitre avec un résumé de l'algorithme. Il nécessite des valeurs initiales pour TOL et ρ, et une estimation de la convergence donnée par ε.

3.4 Algorithme

Etape 0. Etant donné x, évaluer les valeurs et vecteurs propres $\{\lambda_i(x)\}$, et $\{q_i(x)\}$.
Définir $\bar{t}$, $\bar{s}$ par (3.18) (3.19). Poser $t \equiv \bar{t}$, $s \equiv \bar{s}$.

Etape 1. Résoudre le problème Q.P décrit par (3.13), (3.17), (3.20), en utilisant les estimations des matrices lagrangiennes pour définir W.
Si Q.P est insoluble, aller à l'étape 2.2.
Sinon, si $\|d\| \leq \varepsilon$, aller à l'étape 3,
sinon, aller à l'étape 2.

Etape 2. Evaluer $\{\lambda_i(x+d)\}$.

Si $\phi(x+d) < \phi(x)$, alors

2.1 Augmenter t et s, respectivement par le nombre de contraintes qui atteignent la borne supérieure et inférieure dans (3.17), de sorte que (3.15) soit plus petit ou égal à $m+1$. Remplacer x par $x+d$, évaluer $\{q_i(x)\}$ et définir $\bar{t}$, $\bar{s}$ par (3.18), (3.19). Doubler ρ, et aller à l'étape 1.

Sinon,

2.2 Remplacer t et s par $\bar{t}$ et $\bar{s}$. Diviser ρ par deux et aller à l'étape 1.

Etape 3. Si $U \geq 0$ et $V \geq 0$ alors

3.1 STOP, x est optimal.

Sinon,

3.2 Séparer une valeur propre multiple et obtenir une réduction de t et/ou de s. Ajuster $\bar{t}$, $\bar{s}$, t, s et aller à l'étape 1.

Partie III

Conclusion

Nous avons montré dans ce travail comment les résultats et les techniques de l'analyse convexe non différentiable permettent d'étudier la minimisation de fonctions de valeurs propres d'une matrice symétrique. L'analyse de sensibilité de ces valeurs propres repose sur des formulations variationnelles qui n'existent que pour le cas symétrique. Une première question se pose alors : que se passe-t-il si la matrice paramétrisée n'est plus symétrique? Nous ne parlons pas de l'analyse de sensibilité des valeurs singulières, analyse qui se ramène à celle des valeurs propres d'une matrice transformée symétrique.

Nous avons aussi obtenu une borne inférieure du nombre d'arêtes connectant les différents groupes de toute partition d'un graphe en q groupes de taille égale. Cette borne a été comparée dans l'article de W.E. Donath et A.J. Hoffman [6] au nombre de connexions externes provenant de partitions heuristiques de graphes aléatoires. Elle s'avère être une bonne estimation.

Bon nombre de problèmes peuvent être résolus par la méthode décrite pour minimiser la plus grande valeur propre en valeur absolue. Notamment, il est possible d'appliquer l'algorithme vu au chapitre trois pour résoudre

$$\min_x \max_{1\leq l\leq p} \max_{1\leq i\leq n} \mid \lambda_i(A^{(l)}(x)) \mid,$$

où $A^{(1)}(x), ..., A^{(p)}(x)$ sont des matrices symétriques dont les entrées sont des fonctions affines de x, en introduisant des contraintes supplémentaires au problème Q.P. L'algorithme permet aussi de résoudre des problèmes plus généraux d'optimisation comprenant des contraintes sur les valeurs propres d'une matrice paramétrisée. On pourrait encore inclure des contraintes sur les valeurs propres strictement comprises entre λ_1 et λ_n (malgré qu'elles ne soient plus des fonctions convexes).

Enfin, il est possible d'étendre l'algorithme à des matrices non linéaires $A(x)$ bien que le problème d'optimisation résultant ne soit plus nécessairement convexe. Les principales modifications à faire sont le remplacement de A_k par $\partial A(x)/\partial x_k$ et la vérification des conditions d'optimalité du second ordre.

Partie IV

Bibliographie

Bibliographie

[1] BERTSEKAS,D.P. and MITTER, S.K., Steepest descent for optimization problems with nondifferentiable cost functions, Proc 5th Annual Princeton Conf. on Information Sciences and Systems, 1971.

[2] BHATIA, R., Perturbation Bounds for Matrix Eigenvalues, Pitman Research Notes in Mathematics series 162, 1987.

[3] CULLUM, Jane, DONATH, W.E. and WOLFE, P., An algorithm for minimizing certain nondifferentiable convex functions, IBM RC 4611, 1973.

[4] CULLUM, Jane, DONATH, W.E. and KAHAN, W., A block Lançzos Algorithm for determining the q largets eigenvalues and eigenvectors of a real symmetric matrix, IBM Research Report, 1975.

[5] DENYANOV, V.F., Seeking a minimax on a bouded set, Soviet Math. Doklady, 11, pp. 517-521, trans. of Dokl. Akad. Nauk SSSR 191, 1970.

[6] DONATH, W.E. et HOFFMAN, A.J., Lower bounds for the partitioning of graphs, IBM J. Res. and Dev. 17 (5), (Septembre 1973), pp. 420-425.

[7] DOYLE, J., Analysis of feedback systems with structured uncertainties, IEEE Proc., 129, pp. 242-250, 1986, private communication.

[8] EGGLESTON, H.G., Convexity, Cambrige University Press, London, 1958.

[9] FLETCHER, R., Semi-definite matrix constraints in optimization, SIAM J. Control Optim., 23,pp. 493-513, 1985.

[10] FRANKE, M. and WOLFE, P., An algorithm for quadratic programming, Naval Research Logistics Q., 3 , pp. 95-110, 1956.

[11] FRIEDLAND, S. ,NOCEDAL, J. and OVERTON, M.L., The formulation and analysis of numerical methods for inverse eigenvalue problems, SIAM J. Numer. Anal., 24, pp. 634-667, 1987.

[12] HIRIART-URRUTY, J.-B. and YE, D., Sensitivity analysis of all eigenvalues of a symmetric matrix, Numerische Mathematik, 70, pp. 45-72, 1995.

[13] MURRAY, W. and OVERTON, M.L., A projected Lagrangian algorithm for nonlinear minimax optimization, SIAM J. Sci. Statist. Comput., 1, pp. 343-370, 1980.

[14] OVERTON, M.L., On minimizing the maximum eigenvalue of a symmetric matrix, SIAM J. Matrix Anal. Appl. 9, pp. 256-268, 1988.

[15] POLAK, E. and WARDI, Y., Nondifferentiable optimization algorithm for designing control systems having singular value inequalities, Automatica, 18, pp. 267-283, 1982.

[16] ROCKAFELLAR, R.T., The Theory of Subgradients and Its Applications to Problems of Optimization : Convex and Nonconvex Functions, Research and Education in Mathematics 1, Heldermann-Verlag, Berlin, 1981.

[17] WILKINSON, J.H., The algebraic eigenvelue problem, Clarendon Press, Oxford, 1965.