THÈSE
présentée pour le diplôme de
Doctorat de l'Université de Toulouse délivré par
l'Université Toulouse III – Paul Sabatier
en MATHÉMATIQUES APPLIQUÉES

par

Marc FUENTES

intitulée

Analyse et optimisation de problèmes sous contraintes d'autocorrélation

soutenue le 29 Octobre 2007 devant le jury composé de

M. Bergounioux	Professeur, Université d'Orléans	Rapporteur
D. Henrion	Chargé de Recherche, LAAS-CNRS	Examinateur
J.-B. Hiriart-Urruty	Professeur, Université Toulouse III	Directeur de Thèse
C. Lemaréchal	Directeur de recherches, INRIA Rhône-Alpes	Rapporteur
P. Mahey	Professeur, ISIMA, Clermont-Ferrand	Examinateur
P. Maréchal	Professeur, Université Toulouse III	Examinateur

Institut de Mathématiques de Toulouse, UMR 5219
Equipe Mathématiques pour l'Industrie et la Physique,
Université Paul Sabatier 31062 TOULOUSE Cédex 4

"Tu connais les premiers principes de la géométrie ? lui demanda-t-il. Un peu, monsieur Cyrus, répondit Harbert, qui ne voulait pas trop s'avancer. Tu te rappelles bien quelles sont les propriétés de deux triangles semblables ? Oui, répondit Harbert. Leurs côtés homologues sont proportionnels. Eh bien, mon enfant, je viens de construire deux triangles semblables, tous deux rectangles : le premier le plus petit, a pour côtés la perche perpendiculaire, la distance qui sépare le piquet du bas de la perche, et mon rayon visuel pour hypoténuse ; le second a pour côtés la muraille perpendiculaire, dont il s'agit de mesurer la hauteur, la distance qui sépare le piquet du bas de cette muraille, et mon rayon visuel formant également son hypoténuse qui se trouve être la prolongation de celle du premier triangle. Ah ! Monsieur Cyrus, j'ai compris ! s'écria Harbert. De même que la distance du piquet à la perche est proportionnelle à la distance du piquet à la base de la muraille, de même la hauteur de la perche est proportionnelle à la hauteur de cette muraille. C'est cela même, Harbert, répondit l'ingénieur, et quand nous aurons mesuré les deux premières distances, connaissant la hauteur de la perche, nous n'aurons plus qu'un calcul de proportions à faire, ce qui nous donnera la hauteur de la muraille et nous évitera de la mesurer directement."

Extrait de L'Île mystérieuse, (1874), Jules Verne

Remerciements

Ce travail de thèse n'aurait jamais vu le jour sans l'aide de nombreuses autres personnes, que je tiens à remercier chaleureusement ici :

Pour commencer, mon directeur de thèse Jean-Baptiste Hiriart-Urruty, d'une part pour m'avoir encadré durant ces trois années, mais aussi pour avoir suscité bien avant la thèse mon intérêt pour l'optimisation grâce à son livre original d'exercices [30].

Pour leurs discussions scientifiques motivantes, et leurs encouragements à poursuivre dans la recherche, je souhaiterais remercier particulièrement : Jérome Malick de l'INRIA de Grenoble pour toutes les discussions, qu'elles aient concerné l'optimisation ou non ; Didier Henrion, du LAAS qui m'a appris beaucoup sur les polynômes positifs et pour les échanges sur les algorithmes de points intérieurs concernants les LMI ; Claude Lémaréchal, pour son accueil à Grenoble en Janvier 2006, pour nos discussions sur M1QN3, mais également pour avoir accepté d'être rapporteur de ce travail ; Yves Lucet, de UBC Okanagan, pour l'accueil durant mon stage ATUPS en Colombie-Britannique. Je souhaiterais aussi remercier Maïtine Bergounioux d'avoir bien voulu être rapporteur de ma thèse , apportant ainsi son expertise en Traitement du Signal. Pour avoir accepté de faire partie de mon jury, je voudrais aussi exprimer mes remerciements à Philippe Mahey, dont j'ai suivi avec intérêt les cours d'optimisation à l'ISIMA, et Pierre Maréchal, qui par ses nombreuses remarques m'a permis d'améliorer le manuscrit.

Ma gratitude va aussi aux autres doctorants : les "contemporains" Raymond, Samy, Michaël, Jean-luc, Davuth, Abdelkader, Elie, Julien, Benjamin et Anne from Picard-ie, les anciens Jean-Pierre Bourgade, Nicolas Crouseilles, Mehdi, les jeunes, Domi, Sébastien, Mounir, et les jeunes jeunes, Benjamin, Xavier, Laurent, Salvador, Tiphaine, les post-docs Olivier, Aude, Sever, et certains permanents Marcel Mongeau, Sophie Jan, Francis Filbet (à Lyon mais souvent ici), Jérome Fehrenbach, Marcela Szopos. Je souhaiterais aussi remercier des permanents de l'INSA, particulièrement Alain Huard, Sandrine Scott, Sébastien Tordeux et Olivier Mazet, avec qui j'ai eu un bon contact lors des mes enseignements.

Ces quatre ans à Toulouse n'auraient pas été les mêmes sans les ami(e)s rencontrés ici, les grimpeurs (Le Poulpe, Max, Arnaud Tar-....e, Hélène, Nico (le petit et le grand), Martin, Loïc, Antêne, Sarah, Greg, Grand Stef, Emilie, Philou), les habitués de Goulier Claire, Elodie, Ludo, Hélène et Eglantine, sans oublier les hispanophones Amaia (Euskalduna ere), Jaime, Ricardo, Saul, Xiména, et les autres Sonia, Fermin, Lise.

Pour terminer, je souhaite remercier particulièrement mes amis d'ailleurs, Alexis et Sophie, Burz, Sioc, Fonf, Dédé et Agathe, Francois, Mirian eta Ilazkiñe, Andrea, Luca, Nordin, Manue qui fait des maths, et surtout ma famille (El padre y la madre, le brother Pierre et la sister-in-law Bernie, ainsi que mes tchottes frangines Annabelle et Sophie), qui malgré leur éloignement physique sont quand même très proches.

Table des matières

I Quelques Outils pour le Traitement du Signal **11**
I.1 Introduction au Filtrage Numérique 11
I.1.1 Systèmes Linéaires Discrets 11
I.1.2 Corrélation 14
I.1.3 Transformée de Fourier Discrète 16
I.2 Deux exemples de Problèmes en Traitement du Signal 19
I.2.1 Synthèse de filtre 19
I.2.2 Estimation de Densité Spectrale 21
I.2.3 La contrainte essentielle "$x \in \mathcal{C}_{n+1}$" 22

II Le cône $\mathcal{C}_{n+1}$ des vecteurs à composantes autocorrélées **23**
II.1 Historique 23
II.2 Définition de $\mathcal{C}_{n+1}$ 24
II.3 Propriétés de $\mathcal{C}_{n+1}$ 29
II.3.1 Ordre partiel induit par $\mathcal{C}_{n+1}$ 29
II.3.2 Géométrie de $\mathcal{C}_{n+1}$ 30
II.3.3 Représentation basée sur $\mathcal{S}^+_{n+1}(\mathbb{R})$ 50
II.4 Approche via le cône polaire de $\mathcal{C}_{n+1}$ 52
II.4.1 Opérateur adjoint et espace des matrices Toeplitz symétriques . . 53
II.4.2 Cône polaire de $\mathcal{C}_{n+1}$ 54
II.4.3 Isomorphisme vecteurs - matrices Toeplitz 57
II.4.4 Généralisations unidimensionelles 60

III Résolution numérique de problèmes d'optimisation avec contraintes d'autocorrélation **61**
III.1 Optimisation semi-infinie 61
III.2 Algorithmes de suivi de chemin 62
III.2.1 Introduction 63
III.2.2 Schéma algorithmique 65
III.2.3 Application au problème de projection sur $\mathcal{C}_{n+1}$ 66
III.3 Un algorithme basé sur des projections alternées 76

III.3.1 Algorithme de Boyle-Dykstra 76
III.3.2 Projection sur $\mathcal{T}_{n+1}(\mathbb{R})$ 78
III.3.3 Projection sur $\mathcal{N}(\mathcal{S}^{-}_{n+1}(\mathbb{R}))$ 78
III.3.4 Dans quels cas pourrait-on appliquer l'algorithme de Boyle-Dykstra ? 79
III.4 Algorithme de relaxation non-convexe 81
III.4.1 Mise en œuvre pratique de la relaxation non-convexe 84
III.4.2 Comparaison entre les algorithmes de suivi de chemin et celles traitant la relaxation non-convexe 86

IV Extensions Bidimensionnelles **89**
IV.1 Quelle généralisation choisir ? 90
IV.1.1 Corrélation de deux matrices 90
IV.1.2 Cône généralisé d'autocorrélation 91
IV.1.3 Polynômes Sommes de Carrés (SOS, pour Sums of Squares) . . . 91
IV.1.4 Polynômes trigonométriques positifs 94
IV.2 Propriétés et polarité sur les cônes introduits 95
IV.2.1 Quelques Propriétés de $\mathcal{C}_{m,n}$ 95
IV.2.2 Propriétés de $\mathcal{A}(\mathcal{S}^{+}_{mn}(\mathbb{R}))$ et de $\mathcal{P}^{+}_{m,n}(\mathbb{T})$ 97
IV.2.3 Cône polaire de $\mathcal{C}_{m,n}$ 98
IV.2.4 Cone polaire de $\mathcal{P}^{+}_{m,n}(\mathbb{T})$ 100
IV.3 Heuristique de projection sur $\mathcal{C}_{m,n}$ 102
IV.3.1 Résolution numérique du problème $(\mathcal{D})$ 104
IV.3.2 Une approche numérique pour $(\mathcal{NC})$ 106

Notations

- **Ensembles, éléments**
  - $\mathcal{C}_{n+1}$: cône des vecteurs à composantes autocorrélées
  - $\mathcal{M}_n(\mathbb{R})$: matrices carrés réelles d'ordre n
  - $\mathcal{S}_n(\mathbb{R})$: matrices réelles symétriques d'ordre n
  - $\mathcal{S}_n^+(\mathbb{R})$ (resp. $\mathcal{S}_n^-(\mathbb{R})$) : matrices réelles symétriques semi-définies positives (resp. négatives)
  - $\mathcal{T}_{n+1}(\mathbb{R})$: matrices réelles symétriques Toeplitz
  - $\mathcal{L}_n(\mathbb{R})$: cône dit "de Lorentz" ou épigraphe de la fonction norme euclidienne
  - P_1 : matrices "dyadiques" (du type $xx^\top$),
  - $\mathcal{C}_{m,n}$: ensemble des matrices autocorrélées
  - $\mathcal{P}_{m,n}^+(\mathbb{T})$: ensemble des polynômes trigonométriques bivariés positifs
  - $\mathbb{S}_n$: sphère euclidienne de $\mathbb{R}^{n+1}$
  - $\mathcal{O}(f(n))$: classe des fonctions majorées asymptotiquement par $f(n)$
  - $\mathbf{int}\, C$: intérieur de C
  - $\mathbf{cl}\, C$: fermeture topologique ou adhérence de C.
  - ∂C : frontière de C, i.e. $\mathbf{cl}\, C \backslash \mathbf{int}\, C$
  - $\mathtt{aff}(E)$: enveloppe affine de E
  - $\mathtt{cone}(E)$: enveloppe conique convexe de E
  - $\mathtt{conv}(E)$: enveloppe convexe de E
  - e_i : $i^{\text{ème}}$ vecteur de la base canonique de $\mathbb{R}^{n+1}$.
- **Opérations, Relations**
  - $A^\top$: transposée de A
  - $\mathcal{A}^\star$: adjoint de l'opérateur $\mathcal{A}$
  - $\mathtt{Tr}A$: trace de la matrice A
  - $\langle x, y\rangle$ ou $x^\top y$: produit scalaire canonique de $\mathbb{R}^n$,
  - $\langle\langle A, B\rangle\rangle$: produit scalaire de Frobenius dans $\mathcal{M}_n(\mathbb{R})$, i.e. $\mathtt{Tr}(A^\top B)$
  - $\mathtt{spec}(A)$: spectre de la matrice A
  - $P \succeq 0$: P est semi-définie positive
  - $P \succ 0$: P est définie positive
  - $x \succeq_K 0$: $x \in K$, où K est un cône convexe, fermé, solide et pointé.
  - $x \succ_K 0$: $x \in \mathbf{int}\, K$
  - K° : cône polaire négatif de K
  - σ_E : fonction de support de l'ensemble E

  - $N(K, x)$: cône normal à K en x
  - $\partial f(x)$: sous-différentiel de f en x
  - $\circ$: composition d'applications ou produit matriciel d'Hadamard
  - $\otimes$: produit de Kronecker (ou produit tensoriel) de matrices
  - $\mathtt{vec}$: isométrie d'identification entre $\mathcal{M}_{m,n}(\mathbb{R})$ et $\mathbb{R}^{mn}$
  - $\mathtt{svec}$: isométrie d'identification entre $\mathcal{S}_n(\mathbb{R})$ et $\mathbb{R}^{n(n+1)/2}$
  - $x \star y$: convolution discrète des signaux x et y
  - $\mathtt{corr}_a(x, y)$: corrélation acyclique des vecteurs x et y
  - $\mathtt{corr}_c(x, y)$: corrélation circulaire des vecteurs x et y
  - $a \mod b$: reste de la division euclidienne de a par b
  - $\mathcal{F}$: transformée de Fourier discrète
  - $[P(x)]$: Symbole d'Iverson pour le prédicat $P(x)$
  - $\overline{x}$: conjugaison complexe, espérance ou adhérence topologique de x (selon le contexte)
- **Abréviations**
  - BALC : Barrière Auto-concordante Logarithmiquement Convexe
  - BD : Boyle-Dykstra (Algorithme de)
  - BLAS : Basic Linear Algebra System
  - CPU : Central Processing Unit
  - FFT : Fast Fourier Transform
  - FIR : Finite Impulsional Response
  - LMI : Linear Matrix Inequality
  - LTI : Linear Time Invariant
  - SDP : semi-définie positive
  - TFD : Transformée de Fourier Discrète

Introduction

L'objet principal de cette thèse est l'étude, à l'aide d'outils fournis par l'Analyse Convexe et l'Optimisation, d'un cône convexe de $\mathbb{R}^n$ qui joue un rôle important en Traitement du signal, notamment en théorie du Filtrage Numérique et dans la modélisation de séries temporelles à l'aide de processus stochastiques. Il existe plusieurs définitions de ce cône, certaines utilisant la notion d'auto-corrélation (très présente en traitement du signal) et d'autres faisant appel à la positivité d'un polynôme sur un intervalle réel, qui n'est autre que la Transformée de Fourier de l'autocorrélation (que l'on désigne par Densité Spectrale d'Energie en traitement du signal). Même si le cadre de travail choisi ici se limite - pour des raisons d'application évidentes - au contexte de la dimension finie pour l'espace des variables, la prise en compte de ce type de contraintes peut faire apparaître des formulations dites semi-infinies (i.e., problèmes d'optimisation avec un nombre fini de variables mais un nombre infini de contraintes) qui rendent le problème difficile à traiter avec les outils et algorithmes traditionnels de l'optimisation. L'évolution rapide de l'optimisation ces deux dernières décennies nous a d'ailleurs amenés à utiliser des outils de la théorie "moderne" de l'Analyse convexe, comme la (dite) programmation semi-définie positive (SDP) et les Inégalités Linéaires Matricielles (LMI), ou encore les polynômes Sommes de Carrés (SOS). Dans ce cadre, d'ailleurs, ce cône convexe des vecteurs à composantes autocorrélées était absent de la "zoologie" classique des cônes convexes traditionnels, alors qu'il mériterait, à notre avis, lui aussi d'y occuper une bonne place ; et s'attacher à une étude plus en détail de ce cône nous paraissait être un moyen de pallier, à notre façon, ce manque. C'est ainsi qu'à partir des travaux précédents (et récents) sur le sujet, nous nous sommes efforcés d'élaborer une description plus précise en rassemblant un maximum de résultats concernant ce cône.

L'organisation générale de la thèse suit le plan suivant : la première partie comporte des rappels de Traitement du Signal utiles dans la suite, et des exemples afin de motiver notre étude ; la partie suivante contient le gros de la théorie concernant le cône des vecteurs à composantes autocorrélées, avec plusieurs résultats vraiment nouveaux, notamment le théorème sur les éléments propres des matrices qui "engendrent" le cône polaire, la démonstration de l'acuité du cône et un résultat concernant ses facettes ; il contient aussi des résultats sur le cône polaire et sur les identifications possibles entre matrices Toeplitz et vecteurs de $\mathbb{R}^n$. Dans la troisième partie sont présentées différentes approches algorithmiques qu'il est possible de mettre en œuvre lorsque ce cône apparaît comme contrainte spécifique dans un problème d'optimisation ; on en profite d'ailleurs pour rappeler au début quelques notions brèves sur les algorithmes (dits de) de suivi de chemin. Enfin la dernière partie aborde la gé-

néralisation au cas de signaux discrets bi-dimensionnels. Ces derniers présentant des difficultés aussi bien théoriques (définitions multiples envisageables) que pratiques (complexité calculatoire élevée) ; après une présentation des diverses généralisations possibles, nous proposons un algorithme heuristique pour la résolution du problème de projection sur ces cônes.

Chapitre I

Quelques Outils pour le Traitement du Signal

L'objet central de notre étude est un cône convexe solide de $\mathbb{R}^{n+1}$ que certains auteurs [1, 19] ont étudié pour son interêt pertinent dans la modélisation et l'étude des signaux numériques. Ce faisant, ils ont naturellement utilisé le langage et les outils classiques du Traitement du signal. Le langage principal de cette thèse étant mathématique, il paraît nécessaire d'introduire ou de re-préciser certains de ces outils de base avant d'entrer dans le vif du sujet. Ainsi, on ne rencontrera pas dans ce chapitre d'énoncés nouveaux - sauf peut être le *Lemme d'ajout des zéros*, qui n'est qu'une reformulation *ad hoc* d'un résultat très classique de Traitement du Signal - mais plutôt un rappel de certains résultats de base qui ne font pas forcément partie du bagage traditionnel de mathématicien. Nous introduirons ainsi, les outils fondamentaux de la Théorie du Filtrage Numérique, en particulier les différentes corrélations que l'on rencontre en Traitement du signal. Nous décrirons ensuite la Transformée de Fourier Discrète et ses mises en œuvre efficaces, qui sera dans la suite un outil de calcul fondamental. Enfin, pour des motivations pratiques de notre étude, nous présenterons deux problèmes de Traitement du Signal Numérique, l'un de synthèse et l'autre d'identification, qui font clairement apparaître la nécéssité d'étudier ce cône.

I.1 Introduction au Filtrage Numérique

I.1.1 Systèmes Linéaires Discrets

Considérons à cet effet, un système **discret**, **linéaire** et **invariant dans le temps** (LTI discret, en abrégé). On peut modéliser un tel système ou **filtre** comme un opérateur H de $\mathbb{C}^{\mathbb{Z}}$ dans lui même : il agit sur un **signal** d'entrée $\{x_n\}_{n\in\mathbb{Z}}$ pour produire un signal de sortie $\{y_n\}_{n\in\mathbb{Z}}$ - en principe différent de x - c'est la valeur ajoutée du filtre :

$$\begin{array}{rccc} H & : & \mathbb{C}^{\mathbb{Z}} & \to & \mathbb{C}^{\mathbb{Z}} \\ & & x & \mapsto & y = Hx. \end{array}$$

Définition I.1: *Le système modélisé par* H *sera un LTI si et seulement si*
- H *est un opérateur linéaire ;*
- H *est invariant dans le temps (stationnarité ou conservation du retard), c'est-à-dire* H *et* τ_a *(opérateur de translation) commutent pour tout* $a \in \mathbb{Z}$*, où*

$$\tau_a(x)_n = x_{n-a}\ ,\ \forall n \in \mathbb{Z}.$$

Exemple I.1: *Ainsi la* moyenne mobile *définie par*

$$H : \begin{cases} \mathbb{C}^{\mathbb{Z}} \to \mathbb{C}^{\mathbb{Z}} \\ x \mapsto \left(\frac{x_n + x_{n-1}}{2}\right)_{n\in\mathbb{Z}} \end{cases}$$

est un LTI, mais le filtre défini par

$$\begin{cases} \mathbb{C}^{\mathbb{Z}} \to \mathbb{C}^{\mathbb{Z}} \\ x \mapsto (ax_{-n})_{n\in\mathbb{Z}} \end{cases}$$

n'est pas un LTI, car même si H *est linéaire, il ne commute pourtant pas avec* τ_h

I.1.1.1 Convolution et Réponse Impulsionnelle

Pour calculer l'action d'un tel filtre sur un signal d'entrée x, l'outil mathématique est la convolution discrète : pour deux signaux x et y, on définit leur convolution discrète $x \star y$ comme un signal (lui aussi un élément de $\mathbb{C}^{\mathbb{Z}}$) dont les composantes valent

$$(y \star x)_n = \sum_{k\in\mathbb{Z}} y_{n-k}x_k = \sum_{k\in\mathbb{Z}} x_{n-k}y_k \text{ pour tout } n \in \mathbb{Z},$$

sous-réserve que les séries impliquées soient convergentes (c'est le cas dès qu'un des deux signaux est à support fini, mais on pourrait par exemple imposer que les séries soient de carré intégrable). Désignons alors par δ le signal impulsion unité représenté par la suite $\{\delta_n\}_{n\in\mathbb{Z}}$ suivante

$$\delta_n = \begin{cases} 1 \text{ si } n = 0 \\ 0 \text{ sinon.} \end{cases}$$

Alors si l'on désigne par h la *réponse impulsionnelle* $h = H\delta$, c'est-à-dire la sortie correspondante à une impulsion unité en entrée, on constate que, pour tout $n \in \mathbb{Z}$,

$$\begin{aligned} y_n &= (Hx)_n = \left(H\sum_{k\in\mathbb{Z}} x_k\tau_k \circ \delta\right)_n \\ &= \sum_{k\in\mathbb{Z}} x_k(H\circ\tau_k\circ\delta)_n = \sum_{k\in\mathbb{Z}} x_k\tau_k\circ(H\delta)_n \\ &= \sum_{k\in\mathbb{Z}} x_k\tau_k\circ h_n = \sum_{k\in\mathbb{Z}} x_k h_{n-k} = (h\star x)_n. \end{aligned}$$

Autrement dit, pour calculer la sortie $y = Hx$ d'un filtre, il suffit de convoler l'entrée x avec la réponse impulsionelle h. Un filtre est donc décrit de manière unique par sa réponse impulsionnelle.

I.1.1.2 Causalité et Stabilité

Parlant d'un filtre, on dit qu'il est **causal** si sa réponse impulsionnelle est nulle pour les indices négatifs :

$$\forall n \in \mathbb{Z},\ n < 0 \ \Rightarrow\ h_n = 0.$$

Ainsi, le calcul de la valeur de sortie y_n ne peut pas dépendre de la valeur de l'entrée à des instants futurs x_{n+k} avec $k > 0$.

Un filtre sera dit **stable**, s'il existe $M > 0$ tel que

$$\forall n \in \mathbb{Z},\ \ |y_n| \leqslant M \sup_{k\in\mathbb{Z}} |x_n|.$$

Enfin, on dit qu'un filtre est à **Réponse Impulsionnelle Finie**, (FIR en abrégé) si le support de h est fini, par exemple il existe $N > 0$ tel que

$$\forall n \in \mathbb{Z},\ |n| \geqslant N \Rightarrow h_n = 0.$$

On voit ainsi qu'un filtre FIR, est nécessairement *stable*. Dans la suite, nous considérerons uniquement des filtres FIR causaux réels, i.e. dont la réponse impulsionnelle est finie, et nulle pour les indices négatifs ; on pourra donc les représenter comme des vecteurs

$$h = (h_0, \cdots, h_M) \text{ de } \mathbb{R}^{M+1}.$$

I.1.1.3 Transformées en Z et de Fourier

Un outil fondamental en Traitement Numérique du Signal, est la *transformée en Z*, qui, à une réponse impulsionnelle h (ou tout autre signal d'ailleurs), associe la série de Laurent dans le plan complexe

$$H(z) = \sum_{n\in\mathbb{Z}} h_n z^{-n}.$$

Dans la suite, nous adopterons la convention suivante : pour un signal donné x (en minuscule) représenté par sa suite $\{x_n\}_{n\in\mathbb{Z}}$ on désignera avec la majuscule correspondante (ici $X(z)$) sa transformée en Z de l'argument z. La correspondance $x \in \mathbb{C}^{\mathbb{Z}} \leftrightarrow X(z)$ est bijective et possède la propriété intéressante par rapport à $\star$

$$y = x \star h \leftrightarrow Y(z) = X(z)H(z).$$

On définit la transformée de Fourier de h comme la restriction de $H(z)$ au cercle unité $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$:

$$H(\omega) := H(e^{i\omega}) = \sum_{n\in\mathbb{Z}} h_n e^{-in\omega}.$$

De même que précédemment, la correspondance minuscule-majuscule pour un signal et sa Transformée de Fourier sera utilisée dans la suite, à la différence près que l'argument sera toujours ω pour rappeler que c'est la restriction de $z \mapsto H(z)$ à $\mathbb{T}$. Dans notre cas, la transformée en Z est une fraction rationnelle de degré au plus M

$$H(z) = \sum_{n=0}^{M} h_n z^{-n},$$

et la transformée de Fourier est un polynôme trigonométrique

$$H(\omega) = \sum_{n=0}^{M} h_n e^{-in\omega}.$$

On définit la réponse fréquentielle comme le module de la transformée de Fourier, c'est-à-dire $|H(\omega)|$, ce qui mesure donc l'effet du filtre sur l'amplitude du signal en fonction de la pulsation ω.

I.1.2 Corrélation

On a vu précédemment que la convolution était un outil de calcul de base pour la réponse y d'un filtre h à une entrée donnée x. Une application bilinéaire assez semblable à la convolution, qui mesure la "ressemblance" d'un signal avec un autre décalé d'une translation donnée est la *corrélation*, appelée aussi parfois intercorrélation. *Grosso modo* la corrélation des signaux x et y revient à regarder le produit scalaire (dans $\mathbb{C}^{\mathbb{Z}}$ sous réserve de convergence des séries considérées) de x et $\tau_{-a}(y)$, ce qui s'écrit formellement

$$\mathtt{corr}(x,y)_n = \sum_{k\in\mathbb{Z}} \overline{x}_k y_{n+k} \text{ pour tout } n \in \mathbb{Z}.$$

Bien sûr, en se limitant à des signaux à support fini, les séries convergent et on peut définir naturellement deux corrélations sur $\mathbb{C}^n$ selon que l'on considère des signaux périodiques ou non.

I.1.2.1 Corrélation acyclique

La première que nous considèrons ici, c'est la corrélation précédente appliquée à des signaux dont le support est $\{0, \cdots, n\}$.

Définition I.2 (Corrélation acyclique): *Soit $x, y \in \mathbb{C}^{n+1}$, on appelle corrélation acyclique de x et de y, notée $\mathtt{corr}_a(x,y)$ le vecteur de $\mathbb{C}^{n+1}$ de coordonnées*

$$\mathtt{corr}_a(x,y)_k = \sum_{i=0}^{n-k} \overline{x}_i y_{i+k} \text{ pour } k \in \{0, \cdots, n\}.$$

On peut noter au passage, que $\mathtt{corr}_a(\cdot,\cdot)$ est bilinéaire, non-symétrique, et que $\mathtt{corr}_a(x,y)_0 = \langle x, y\rangle$ (produit hermitien usuel de x et y).

I.1.2.2 Corrélation cyclique

La seconde correspond en fait à la corrélation de deux signaux $n+1$-périodiques, et présente plusieurs propriétés intéressantes que l'on rencontre dans le cas de signaux continus.

Définition I.3 (Corrélation circulaire): *Soit $x, y \in \mathbb{C}^{n+1}$, on appelle corrélation cyclique ou circulaire de x et de y, noté $\mathtt{corr}_c(x,y)$ le vecteur de $\mathbb{C}^{n+1}$ de coordonnées*

$$\mathtt{corr}_c(x,y)_k = \sum_{i=0}^{n} \overline{x}_i y_{\{(i+k) \mod (n+1)\}} \text{ pour } k \in \{0, \cdots, n\},$$

où $a \mod b$ est le reste de la division euclidienne de a par b.

I.1.2.3 Ajout de zéros

La corrélation acyclique et la corrélation circulaire sont deux objets distincts mais possèdent néanmoins de nombreux points communs. En fait, il existe un lien direct entres les deux, et il est possible de calculer une corrélation acyclique à partir d'une corrélation circulaire, en ajoutant "suffisamment" de zéros (en anglais *zero padding*) aux vecteurs dont on veut calculer les corrélations acycliques. Plus précisemment, le lemme suivant expose les relations étroites entre $\mathtt{corr}_a$ et $\mathtt{corr}_c$:

Lemme I.1 (d'ajout des zéros): *Soit $n \in \mathbb{N}^*$, et $N \geqslant 2n+1$, $x, y \in \mathbb{C}^{n+1}$; désignons par $\widetilde{x} = (x, 0, \cdots, 0)$ et $\widetilde{y} = (y, 0, \cdots, 0)$ les vecteurs de $\mathbb{C}^N$ après avoir complété avec des zéros ; alors*

$$\mathtt{corr}_c(\widetilde{x},\widetilde{y})_k = \begin{cases} \mathtt{corr}_a(x,y)_k & \text{pour } k = 0, \cdots, n \\ \overline{\mathtt{corr}_a(y,x)}_{N-k} & \text{pour } k = N-n, \cdots, N-1 \\ 0 & \text{pour } n+1 \leqslant k \leqslant N-n-1 \end{cases}$$

Démonstration. Soit $k \leqslant n$, alors

$$\begin{aligned} \mathtt{corr}_c(\widetilde{x},\widetilde{y})_k &= \sum_{i=0}^{n-k} \overline{\widetilde{x}}_i \widetilde{y}_{\{(i+k) \mod N\}} + \sum_{i=n-k+1}^{n} \overline{\widetilde{x}}_i \underbrace{\widetilde{y}_{\{(i+k) \mod N\}}}_{0} + \sum_{i=n+1}^{N-1} \underbrace{\overline{\widetilde{x}}_i}_{0} \widetilde{y}_{\{(i+k) \mod N\}} \\ &= \sum_{i=0}^{n-k} \overline{x}_i y_{i+k} = \mathtt{corr}_a(x,y)_k. \end{aligned}$$

Si $k > N-n-1$, alors

$$\begin{aligned} \mathtt{corr}_c(\widetilde{x},\widetilde{y})_k &= \sum_{i=0}^{N-1-k} \overline{\widetilde{x}}_i \underbrace{\widetilde{y}_{\{(i+k) \mod N\}}}_{0} + \sum_{i=N-k}^{n} \overline{\widetilde{x}}_i \widetilde{y}_{\{(i+k) \mod N\}} + \sum_{i=n+1}^{N-1} \underbrace{\overline{\widetilde{x}}_i}_{0} \widetilde{y}_{\{(i+k) \mod N\}} \\ &= \sum_{i=N-k}^{n} \overline{x}_i y_{i+k-N} = \sum_{i=0}^{n-N-k} \overline{x}_{i+N-k} y_i = \overline{\mathtt{corr}_a(y,x)}_{N-k}. \end{aligned}$$

Si $n+1 \leqslant k \leqslant N-n-1$, alors

$$\mathtt{corr}_c(\widetilde{x},\widetilde{y})_k = \sum_{i=0}^{n} \overline{\widetilde{x}}_i \underbrace{\widetilde{y}_{\{(i+k) \mod N\}}}_{0} + \sum_{i=n+1}^{N} \underbrace{\overline{\widetilde{x}}_i}_{0} \widetilde{y}_{\{(i+k) \mod N\}} = 0.$$

□

On peut noter que si l'on applique une corrélation (classique, acyclique ou même circulaire) à un signal x et lui-même, on s'intéresse à la "ressemblance" de x avec des translations de lui-même : on parle alors *d'autocorrélation*.

I.1.3 Transformée de Fourier Discrète

La Transformée de Fourier étant un objet fonctionnel, on utilise une version discrétisée de celle-ci pour pouvoir "faire des calculs" : la *Transformée de Fourier Discrète* (TFD), c'est-à-dire un codage discret de la Transformée de Fourier, où *grosso modo*, on discrétise l'argument (ω) sur une grille à N éléments équirépartis dans $[0, 2\pi]$, et qui permet ainsi de travailler avec des vecteurs de $\mathbb{C}^N$ plutôt qu'avec une fonction. On rappellera brièvement ici la définition, ainsi que quelques propriétés qui nous seront utiles dans la suite. Les démonstrations de ces résultats n'étant pas difficiles, on pourra consulter pour plus de détail les références [8] (destiné à un public d'ingénieurs) ou [44] en français (plus mathématique, utilisant le formalisme de la théorie des groupes).

Définition I.4: *Soit $N \in \mathbb{N}^*$ et $x \in \mathbb{C}^N$; désignons par X (noté aussi parfois $\hat{x}$) le vecteur de $\mathbb{C}^N$ de coordonnées*

$$X_n = \sum_{k=0}^{N-1} x_k e^{\frac{2ikn\pi}{N}} \text{ pour } n \in \{0, \cdots, N-1\}.$$

Alors on appelle Transformée de Fourier Discrète d'ordre N, l'application linéaire définie par

$$\mathcal{F} : \begin{cases} \mathbb{C}^N \to \mathbb{C}^N \\ x \mapsto X \end{cases}.$$

Notons au passage, que $\mathcal{F}$ étant linéaire, on peut écrire sa matrice dans la base canonique de $\mathbb{C}^N$, W dont le terme général est $W_{kl} = \{e^{\frac{2ikl\pi}{N}}\}$ pour (k,l) dans $\{0, \cdots, N-1\}$. Cela donne matriciellement

$$W = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & e^{\frac{2i\pi}{N}} & \cdots & e^{\frac{2i(N-1)\pi}{N}} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{\frac{2i(N-1)\pi}{N}} & \cdots & e^{\frac{2i(N-1)^2\pi}{N}} \end{pmatrix}.$$

On voit que la matrice W est symétrique et l'on peut prouver assez facilement que $WW^\star = NI_N$, ce qui permet d'inverser très facilement $\mathcal{F}$ comme le signale la

Proposition I.1 (Transformée de Fourier Discrète inverse): *L'inverse $\mathcal{F}^{-1}$ de $\mathcal{F}$ est définie par*

$$x = \mathcal{F}^{-1}(X) = \frac{1}{N}\overline{W}X.$$

On a aussi l'équivalent discret du théorème de Parseval-Plancherel pour la transformée de Fourier continue :

Théorème I.1 (Parseval-Plancherel): *Soient $x, y \in \mathbb{C}^N$ et $\langle \cdot, \cdot \rangle$ le produit hermitien canonique de $\mathbb{C}^N$, alors*

$$\langle x, y \rangle = \frac{1}{N}\langle \mathcal{F}(x), \mathcal{F}(y) \rangle.$$

La TFD, en plus de ces nombreuses propriétés, présente un comportement "sympathique" par rapport à la corrélation circulaire : ainsi, on a le

Théorème I.2 (de corrélation discrète): *Soient $x, y \in \mathbb{C}^N$, alors*

$$\mathcal{F}(\mathtt{corr}_c(x, y)) = \overline{\mathcal{F}(x)} \circ \mathcal{F}(y),$$

où $\circ$ dénote le produit d'Hadamard (terme à terme).

I.1.3.1 Transformée de Fourier Rapide

La TFD, comme on l'a vu, est une application linéaire : avec une représentation matricielle, le coût de son calcul équivaut à un produit matrice-vecteur, autrement dit $\mathcal{O}(n^2)$[1]. Il existe depuis 1965 une classe d'algorithmes dits de Transformée de Fourier Rapide (en anglais Fast Fourier Transform, FFT) qui ramènent ce calcul à un coût de $\mathcal{O}(n \log n)$. Ces algorithmes dûs à Cooley et Tukey, sont basés sur le principe informatique de "diviser pour régner" : supposons que l'on doive traiter un problème de taille n ; si l'on s'aperçoit que pour résoudre ce problème, il "suffit" de traiter deux problèmes de taille $n/2$ et de fusionner leurs solutions, ceci à un coût au plus linéaire $L(n) \in \mathcal{O}(n)$, alors la complexité en temps du problème obéit à la récurrence suivante

$$T(n) = 2T(n/2) + L(n),$$

ce qui conduit en théorie de la complexité au fait que

$$T(n) \in \mathcal{O}(n \log_2 n).$$

On préfère écrire ici une relation d'appartenance plutôt qu'une égalité, car la la relation de domination asymptotique induit des classes de fonctions. Sans faire la théorie détaillée du calcul rapide d'une TFD, on peut montrer simplement comment apparaissent les deux problèmes sous-jacents de taille $n/2$ dans un problème de taille n.

Soit $x \in \mathbb{C}^n$ avec n une puissance de deux (n est donc pair). Alors sa TFD vaut pour chaque $k \in \{0, \cdots, n-1\}$

$$X_k = \mathcal{F}_n(X)_k = \sum_{p=0}^{n-1} x_p \omega_n^{-pk},$$

[1] Pour plus de détails sur les classes de fonctions de complexité, on pourra par exemple consulter [25]

avec $\omega_n = e^{\frac{2i\pi}{n}}$, et l'on indice la TFD par la dimension n des vecteurs considérés. On peut ensuite séparer la somme entre indices pairs et indices impaires :

$$X_k = \sum_{p=0}^{n/2-1} x_{2p}\omega_n^{-2pk} + \sum_{p=0}^{n/2-1} x_{2p+1}\omega_n^{-(2p+1)k} \quad (I.1)$$

$$= \sum_{p=0}^{n/2-1} x_{2p}\omega_n^{-2pk} + \omega_n^{-k}\sum_{p=0}^{n/2-1} x_{2p+1}\omega_n^{-2pk}. \quad (I.2)$$

Notons alors $x^\flat = (x_0, x_2, \cdots, x_n)$ et $x^\sharp = (x_1, x_3, \cdots, x_{n-1})$; avec cette écriture et en remarquant au préalable que

$$\omega^{-2pk} = \omega_{n/2}^{-pk},$$

alors X_k s'écrit

$$X_k = \begin{cases} \mathcal{F}_{n/2}(x^\flat)_k + \omega_n^{-k}\mathcal{F}_{n/2}(x^\sharp)_k \text{ si } k \in \{0, \cdots, n/2-1\} \\ \mathcal{F}_{n/2}(x^\flat)_{k-n/2} + \omega_n^{-k}\mathcal{F}_{n/2}(x^\sharp)_{k-n/2} \text{ si } k \in \{n/2, \cdots, n-1\}. \end{cases}$$

On voit bien alors que le calcul d'une FFT de dimension n se ramène à celui de deux FFT de taille $n/2$ et que l'assemblage du résultat ("l'opération de fusion") a un coût linéaire. On peut donc appliquer le principe "diviser pour régner". Faisons plusieurs remarques par rapport à la classe d'algorithmes que cette méthode engendre :

- On travaille avec des TFD qui sont des puissances de deux. Il existe des versions pour des tailles différentes, mais dans la suite on considérera uniquement des FFT pour des tailles étant des puissances de deux. Si l'on souhaite calculer une FFT d'une taille différente, il est possible d'ajouter des zéros pour atteindre la puissance de deux supérieure et l'on extrait ensuite un sous-vecteur du résultat.
- Un vrai algorithme de FFT, ne fait évidemment pas des appels récursifs qui, en nombre exponentiel, pourraient faire déborder rapidement la pile du processeur ; on préfère alors dé-récursifier l'algorithme afin de le rendre plus efficace.

On admettra alors le résultat de complexité suivant :

Proposition I.2: *Soit $N = 2^p$ et $x \in \mathbb{C}^N$; alors le calcul de de la Transformée de Fourier Discrète X (resp. Transformée de Fourier Discrète Inverse) peut se faire en au plus $\mathcal{O}(N \log_2 N)$ opérations arithmétiques de base.*

I.1.3.2 Corrélation Rapide

En utilisant le résultat précédent, ainsi que le lemme d'ajout des zéros et le théorème de corrélation discrète, on peut déduire un moyen rapide - sous-entendu à un coût au plus de $\mathcal{O}(n \log n)$ - de calculer une corrélation acylique :

Proposition I.3: *Soient x et y deux vecteurs de $\mathbb{C}^n$; alors le calcul de $\mathtt{corr}_a(x, y)$ peut se faire en $\mathcal{O}(n \log n)$.*

Démonstration. On pose $N = 2^{\lfloor \log_2(2n+1) \rfloor + 1}$, qui implique $N \geqslant 2n+1$ et $N \in \mathcal{O}(n)$. On forme alors $\widetilde{x} = (x, 0_{N-n})$ et $\widetilde{y} = (y, 0_{N-n})$ qui sont les "complétés" avec des zéros de x et y (coût $2N$) ; on calcule par FFT les TFD $\widetilde{X} = \mathcal{F}(\widetilde{x})$ et $\widetilde{Y} = \mathcal{F}(\widetilde{y})$ au coût de $2N \log N$, on calcule ensuite leur produit d'Hadamard conjugué $C = \overline{\widetilde{X}} \circ \widetilde{Y}$ (coût N) ; puis on compose par une TFD Inverse Rapide pour $c = \mathcal{F}^{-1}(C)$ (coût $N \log N$). Le tout a un coût de $3N + 3N \log N \in \mathcal{O}(n \log n)$. □

I.2 Deux exemples de Problèmes en Traitement du Signal

Les problèmes rencontrés en théorie du filtrage sont principalement de deux sortes. Les premiers sont des problèmes d'**identification**, où certains paramètres du filtre ne sont pas connus, mais on connaît par exemple des couples entrées/sorties du filtre et l'on souhaite estimer les paramètres du filtre correspondant. C'est un domaine du Traitement du signal qui a de fortes connexions avec la théorie des processus stochastiques et celle des problèmes inverses. L'autre catégorie de problèmes typiquement rencontrés, concerne les problèmes de **synthèse**, où l'on souhaite fabriquer un filtre (trouver sa réponse impulsionnelle) pour qu'il réalise certaines fonctions souhaitées. Nous présenterons ainsi un exemple de conception de filtre passe-bas tiré de [5] et un problème d'estimation de fonction d'autocorrélation qui provient de [1, 19].

I.2.1 Synthèse de filtre

Supposons que l'on souhaite réaliser un filtre qui atténue les basses fréquences : autrement dit, on voudrait que la réponse fréquentielle ait des valeurs faibles pour $\omega \approx 0$ et soit proche de 1 pour $\omega \approx \pi$. On peut se demander *a priori* pourquoi se limiter au segment $[0, \pi]$? Le filtre à concevoir étant linéaire, on peut décomposer chaque signal suivant ses différentes harmoniques, et se limiter ainsi à $[-\pi, \pi]$. On notera au passage que, puisque l'on considère uniquement des h réels, on a forcement $H(-\omega) = H^*(\omega)$ et donc plutôt que de considérer $[-\pi, \pi]$, on peut se limiter à $[0, \pi]$. Une manière d'obtenir une réponse fréquentielle répondant à la question peut se traduire par exemple sous la forme des contraintes suivantes :

$$|H(\omega)| \leqslant \varepsilon \,,\ \omega \in [0, \underline{\omega_c}] \qquad 1 - \varepsilon \leqslant |H(\omega)| \leqslant 1 + \varepsilon \,,\ \omega \in [\overline{\omega_c}, \pi],$$

où $\underline{\omega_c}$ et $\overline{\omega_c}$ représentent les fréquences de changement de régime.

Comme nous l'avons dit précédemment, les coefficients h_n sont les paramètres de conception de filtre, ce sont donc ces variables qui doivent être optimisées afin de réaliser le filtre selon le cahier des charges désiré. Imaginons que l'on souhaite minimiser l'énergie du filtre, alors on peut alors modéliser ce problème sous la forme

$$(\mathrm{Syn})\begin{cases} \min\limits_{h\in\mathbb{R}^{M+1}} & \displaystyle\int_{[0,\pi]} |H(\omega)|^2\,d\omega \\ \text{tel que} & |H(\omega)| \leqslant \varepsilon\,,\ \forall \omega \in [0,\underline{\omega_c}] \\ & 1-\varepsilon \leqslant |H(\omega)| \leqslant 1+\varepsilon\,,\ \forall \omega \in [\overline{\omega_c},\pi]. \end{cases} \tag{I.3}$$

Ce problème d'optimisation n'est malheureusement pas convexe et *a fortiori* n'est pas linéaire en les coefficients $(h_0,\cdots,h_M)$; ceci à cause du module complexe $|\cdot|^2$. Il y a pourtant un moyen de linéariser la fonction-objectif et d'une certaine manière les contraintes : posons $X(\omega)=|H(\omega)|^2$, alors

$$\begin{aligned} X(\omega) &= H(\omega)\overline{H}(\omega) \\ &= \left(\sum_{k=0}^{M} h_k e^{-ik\omega}\right)\left(\sum_{p=0}^{M} h_p e^{ip\omega}\right) \\ &= \sum_{l=-M}^{M} x_l e^{il\omega}. \end{aligned}$$

Il suffit, pour trouver les x_l, de transformer la somme double en somme simple et de déterminer le poids associé à $\{(p,k)\in\{0,\cdots,n\}^2 \,|\, p-k=l\}$. On trouve ainsi

$$x_l = \sum_{p=0}^{M-|l|} h_p h_{p+|l|}. \tag{I.4}$$

En particulier,

$$x_0 = \|(h_0,\cdots,h_M)\|^2, \tag{I.5}$$

ce qui a pour conséquence : h est nul si et seulement si $x_0=0$. Remarquons au passage que $x_{-l}=x_l$, donc

$$X(\omega) = x_0 + 2\sum_{l=1}^{M} x_l \cos(l\omega).$$

La fonction-objectif s'exprime sous la forme

$$x_0\pi + 2\sum_{k=1}^{M} x_k \int_{[0,\pi]} \cos(k\omega)\,d\omega,$$

qui est linéaire en x ! De même, élevons la première fonction-contrainte dans (I.3) au carré :

$$|H(\omega)|^2 = X(\omega) \leqslant \varepsilon^2,$$

ce qui équivaut à

$$\forall \omega \in [0,\underline{\omega_c}]\,,\ \varepsilon^2 - X(\omega) \geqslant 0.$$

Modulo une transformation linéaire basée sur les polynômes de Tchebitcheff comme décrit dans [1], on peut traiter cette contrainte sur n'importe quel intervalle $I \subset [0, \pi]$.

Ensuite, à un changement de variables affine près (qui revient à intégrer (ε^2) dans la constante et à changer de signe), ceci revient à s'intéresser à

$$X(\omega) = x_0 + 2\sum_{k=1}^{M} x_k \cos(k\omega) \geqslant 0\ ,\ \forall \omega \in [0, \pi]. \tag{I.6}$$

Or, grâce à un théorème que l'on verra plus tard, il y a équivalence entre la condition (I.6) sur les x_k et son expression sous forme (I.4).

Avec cette formulation, le problème devient linéaire et donc convexe en x, l'unique contrainte étant que les x doivent vérifier (I.4) ; Cette contrainte définit en réalité **un sous-ensemble convexe de $\mathbb{R}^{M+1}$ que nous allons étudier en détail par la suite et cela rend *a priori* le problème (I.3) plus simple à résoudre** ; désignons pour l'instant ce sous-ensemble par $\mathcal{C}_{M+1}$.

I.2.2 Estimation de Densité Spectrale

On considère un processus stochastique (complexe ou réel) $\{X_n\}_{n\in\mathbb{N}}$; on définit son espérance

$$\overline{X}_n = \mathcal{E}\{X_n\}$$

et sa fonction d'autocorrélation

$$(r_{XX})_{n_1,n_2} = \mathcal{E}\{X^*_{n_1} X_{n_2}\},$$

où $*$ désigne ici la conjugaison complexe pour éviter les confusions avec l'espérance. En réalité, cette définition d'autocorrélation est valable uniquement quand l'espérance est nulle. Dans la suite, nous ferons volontairement cet abus de langage.

On dit qu'un processus stochastique est *stationnaire au sens large* si

$$\begin{cases} \overline{X}_n &= \mu \\ (r_{XX})_{n_1,n_2} &= g(n_1 - n_2)\ ,\ \forall n_1, n_2 \in \mathbb{N}^2. \end{cases} \tag{I.7}$$

où g est une fonction quelconque : autrement dit, l'autocorrélation, ne dépend que du retard $n_1 - n_2$. Soit alors un processus stochastique $\{X_n\}_{n\in\mathbb{N}}$ stationnaire au sens large et supposons que l'on connaisse une suite x_k de réalisations des variables aléatoires X_k correspondantes.

Supposons que l'on recherche à estimer sa fonction d'autocorrélation pour des retards successifs ; il existe pour cela deux estimateurs classiques :

– le premier estimateur classique du k-ième coefficient d'autocorrélation

$$(\hat{r}_{XX})_k(x_0, \cdots, x_N) = \frac{1}{N-k+1}\sum_{i=0}^{N-k} x_i^* x_{i+k},$$

qui a l'avantage d'être sans biais $(\mathcal{E}\{(\hat{r}_{XX})_k(X_0, \cdots, X_N)\} = (r_{XX})_k)$, mais qui, par contre, présente une variance élevée pour des k grands.

– le second

$$(\check{r}_{XX})_k(x_0,\cdots,x_N) = \frac{1}{N}\sum_{i=0}^{N-k} x_i^* x_{i+k}$$

qui, lui, est biaisé, mais asymptotiquement sans biais, présente une variance moins élevée pour les k grands.

Si l'on choisit $\hat{r}_{XX}$, rien ne garantit que l'on obtienne bien une fonction d'autocorrélation (la *densité spectrale d'énergie*, autrement dit, la transformée de Fourier de la fonction d'autocorrélation peut être négative) ce qui peut être plutôt gênant dans les applications. Par contre, $\check{r}_{XX}$ est bien une fonction d'autocorrélation.

Une méthode possible pour forcer $\hat{r}_{XX}$ à vérifier une relation du type (I.4), consiste à projeter $\hat{r}_{XX}$ sur le cône $\mathcal{C}_{N+1}$, ce qui revient à résoudre

$$(\mathrm{Est}) \begin{cases} \min\limits_{x\in\mathbb{R}^{N+1}} & \|x-r\|_2^2 \\ & x \in \mathcal{C}_{N+1}, \end{cases} \tag{I.8}$$

où $r = \{(\hat{r}_{XX})_k\}_{k=0}^N$ est le vecteur des coefficients estimés, $N+1$ le nombre d'observations de $\{X_n\}_{n\in\mathbb{N}}$, et $\|\cdot\|_2$ la norme euclidienne dans $\mathbb{R}^{N+1}$.

I.2.3 La contrainte essentielle "$x \in \mathcal{C}_{n+1}$"

On voit dans ces deux exemples que la contrainte "$x \in \mathcal{C}_{n+1}$" surgit naturellement dans certains problèmes de Traitement du Signal. Il est donc logique de s'intéresser à la structure particulière qu'engendre cette contrainte dans des problèmes d'optimisation. En conséquence de quoi, on pourra privilégier une approche spécifique pour la résolution de ces problèmes, plutôt qu'une résolution au moyen d'algorithmes généraux, qui théoriquement fournissent des solutions, mais dont l'efficacité laisse fort à désirer lorsque la dimension augmente ou que l'on souhaite obtenir des certificats d'optimalité. Mais avant de parler plus précisement de ces méthodes, il convient d'étudier la structure, les différentes paramétrisations et la géométrie de ce cône.

Chapitre II

Le cône $\mathcal{C}_{n+1}$ des vecteurs à composantes autocorrélées

II.1 Historique

L'étude des cônes définis en tant que coefficients de polynômes positifs est fortement liée à la théorie des moments dont le développement principal correspond surtout au début du $20^{\text{ème}}$ siècle avec, entre autres, les recherches de Hilbert (*cf.* le $17^{\text{ème}}$ problème de Hilbert) et Schur par exemple. Parmi les différents travaux ultérieurs, on peut noter plus particulièrement l'ouvrage de Krein et Nudelman [33] qui fait le lien entre les deux théories et propose une étude systématique des cônes définis comme enveloppes convexes d'une courbe de $\mathbb{R}^n$, grâce aux notions de systèmes de fonction de Markov et de Tchebitcheff. Plus tard, les travaux de Shor [47] sont consacrés à la prise en compte algorithmique de contraintes définies par la positivité d'un polynôme. Puis, Nesterov dans son article [40] a démontré l'intérêt pour l'étude des polynômes qui sont exprimables comme sommes de carrés SOS (Sums Of Squares) : en effet, dans le cas d'une seule variable, cela est une condition nécessaire et suffisante de positivité. Le cas à plusieurs variables des polynômes sommes de carrés et leur lien fort avec les polynômes positifs, a par la suite été relativement bien étudié avec les travaux de Parillo [43], et Lasserre [34], ainsi que Megretski [37] dans le cas de polynômes trigonométriques.

De manière parallèle, dans le contexte du Traitement du Signal, le cas d'une seule variable trigonométrique a été bien étudié dans les travaux de Alkire *et al.* [1], Dumitrescu *et al.* [19], et ceux de Hachez [27], et ceux plus appliqués de Davidson *et al.* [15] qui ont présenté de plusieurs façons différentes le cône que nous allons étudier dans la suite. Les premiers en ont donné les formulations par autocorrélation et par positivité de la densité spectrale d'énergie, les seconds ont utilisé une formulation dite "par trace", et le troisième l'a présenté comme l'intersection avec $\mathbb{R}^n$ d'un cône de polynômes positifs de matrices. Il faut noter que depuis le début de cette thèse plusieurs articles intéressants [51, 18] sont venus compléter la littérature sur ce sujet et ses extensions au cas multivariables. Enfin, Lindquist et Byrnes [10], ont aussi étudié dans le cadre du traitement du signal, les relations étroites avec la théorie des moments, afin de résoudre des problèmes d'extension de covariance à l'aide de méthodes entropiques.

Notre objectif dans ce chapitre, qui présente plusieurs résultats nouveaux et reprend certaines démonstrations classiques tirées de [1], est de rendre l'objet mathématique $\mathcal{C}_{n+1}$ plus familier et plus intuitif pour les utilisateurs de l'analyse convexe : ainsi certains cônes tels que le cône des matrices semi-définies positives $\mathcal{S}_n(\mathbb{R})$, le cône de Lorentz $\mathcal{L}_n(\mathbb{R})$, ou encore le cône des matrices de distances $\mathbb{EDM}^N$ (voir par exemple, à ce sujet l'article [50]) sont des objets dont les propriétés géométriques, les faces ou le cône polaire commencent à être particulièrement bien connus dans la communauté mathématique. Certains résultats présentés ici ne semblent pas forcément très utiles dans les applications, mais il donnent assurément une meilleure intuition pour "penser" ou voir le cône $\mathcal{C}_{n+1}$.

II.2 Définition de $\mathcal{C}_{n+1}$

Il existe plusieurs définitions pour le cône $\mathcal{C}_{n+1}$ des *vecteurs à composantes autocorrélées*. On peut en donner une plutôt "algébrique" où les composantes d'un vecteur de $\mathbb{R}^{n+1}$, élément de ce cône, doivent vérifier une certaine relation : c'est notre

Définition II.1: *Dans l'espace euclidien* $F = \mathbb{R}^{n+1}$, *avec* $x = (x_0, \cdots, x_n)$ *et* $y = (y_0, \cdots, y_n)$, *on définit* $\mathcal{C}_{n+1}$ *par*

$$\mathcal{C}_{n+1} = \{\mathtt{corr}_a(y, y) \,|\, y \in F\} = \{x \in F \,|\, \exists y \in F\,,\ x_k = \sum_{i=0}^{n-k} y_i y_{i+k} \quad \forall k = 0, \cdots, n\},$$

où corr_a *est la corrélation acyclique définie au chapitre 1.*

Il est possible de reformuler cette définition d'une manière plus synthétique à l'aide des notations suivantes : posons E^k (la notation ^{k} est cohérente avec la puissance k-ième de la matrice définie par $k = 1$) la matrice de terme général

$$(E^k)_{ij} = [i = k + j] = \begin{cases} 1 \text{ si } i = k + j \\ 0 \text{ sinon,} \end{cases} \qquad \text{avec } i, j \in \{0, \cdots, n\}, \tag{II.1}$$

où $[i = k + j]$ désigne la convention d'Iverson utilisée en Informatique, définie par exemple dans [25], qui pour un prédicat p vaut

$$[p(x)] = \begin{cases} 1 \text{ si } p(x) \text{ est vrai,} \\ 0 \text{ sinon .} \end{cases}$$

Ainsi le delta de Kronecker s'écrirait $\delta_{i,j} = [i = j]$; l'avantage de cette notation réside principalement dans les situations où les prédicats sont compliqués, ce qui peut être gênant à écrire ou à lire en notation indicielle.

La matrice E^k alors introduite correspond au $k^{\text{ème}}$ décalage à droite

$$E^k \begin{cases} \mathbb{R}^{n+1} \to \mathbb{R}^{n+1} \\ (x_0, \cdots, x_n) \mapsto (0, \cdots, 0, x_0, \cdots, x_{n-k}) \end{cases} .$$

Si l'on considère sa partie symétrique définie par

$$A^{(k)} = \frac{1}{2}(E^k + (E^k)^\top) = \frac{1}{2}\begin{pmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 1 & \ddots & \ddots & \ddots & \ddots & \ddots & 1 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}, \tag{II.2}$$

on peut alors définir une application linéaire $\mathcal{A}$ de $\mathcal{M}_{n+1}(\mathbb{R})$ dans $\mathbb{R}^{n+1}$ par

$$\begin{array}{rccc} \mathcal{A}: & \mathcal{M}_{n+1}(\mathbb{R}) & \to & \mathbb{R}^{n+1} \\ & Q & \mapsto & \begin{pmatrix} \langle\langle A^{(0)}, Q\rangle\rangle \\ \vdots \\ \langle\langle A^{(n)}, Q\rangle\rangle \end{pmatrix}, \end{array}$$

où $\langle\langle A, B\rangle\rangle = \mathtt{Tr}(A^\top B) = \sum_{i,j} A_{ij}B_{ij}$ est le produit scalaire standard (dit de Frobenius) sur $\mathcal{M}_n(\mathbb{R})$. On a donc pour tout p dans $\{0, \cdots, n\}$ et $Q \in \mathcal{S}_{n+1}(\mathbb{R})$,

$$\mathcal{A}(Q)_p = \frac{1}{2}\sum_{k=0}^{n-p} Q_{k(k+p)} + Q_{(k+p)k} = \sum_{k=0}^{n-p} Q_{k(k+p)} = \sum_{\substack{0 \leqslant k,l \leqslant n \\ k-l=p}} Q_{kl}. \tag{II.3}$$

Par conséquent, pour une matrice symétrique positive de rang 1 (*i.e.* $Q = xx^\top$), on a

$$\mathcal{A}(xx^\top)_p = \sum_{k=0}^{n-p} x_k x_{k+p} = \mathtt{corr}_a(x, x)_p,$$

ce qui donne dans le cas non-symétrique (*i.e.* $Q = xy^\top$)

$$\mathcal{A}(xy^\top)_p = \frac{1}{2}\sum_{k=0}^{n-p} x_k y_{k+p} + x_{k+p}y_k = \frac{1}{2}(\mathtt{corr}_a(x, y)_p + \mathtt{corr}_a(y, x)_p). \tag{II.4}$$

Finalement, il en résulte que l'on peut exprimer $\mathcal{C}_{n+1}$ comme suit :

$$\mathcal{C}_{n+1} = \mathcal{A}(\{yy^\top \mid y \in \mathbb{R}^{n+1}\}). \tag{II.5}$$

Autrement dit, $\mathcal{C}_{n+1}$ **est l'image par une application linéaire de l'ensemble** P_1 **des matrices semi-définies positives (SDP) de rang inférieur ou égal à 1** . Cet ensemble P_1 - contenu intégralement dans la frontière $\partial\mathcal{S}^+_{n+1}(\mathbb{R})$ du cône des matrices SDP - est non-convexe ; cependant, comme on le verra plus tard, il est possible de re-paramétrer $\mathcal{C}_{n+1}$, en élargissant P_1 (on le remplace $\mathcal{S}^+_{n+1}(\mathbb{R})$), ce qui rend alors $\mathcal{C}_{n+1}$ évidemment convexe.

Remarque II.1: *Comme nous le verrons dans la suite, les cônes convexes (et plus généralement les objets convexes) admettent des formulations duales l'une de l'autre. Comme dans d'autres domaines mathématiques on peut* grosso modo *représenter un cône, soit comme généré par un ensemble d'objets (en l'occurrence des vecteurs), soit comme caractérisé par un ensemble de relations ou contraintes qu'il doit vérifier (des inégalités linéaires ou convexes).*

Pour faire un parallèle avec le "monde linéaire", on peut décrire un sous-espace vectoriel $\mathcal{V}$ de $\mathbb{R}^n$, globalement de deux façons : soit en en donnant une base $\mathcal{B} = (b^1, \cdots, b^p)$, et ainsi $\mathcal{V} = \mathtt{Vect}\,\mathcal{B}$, soit en fournissant une équation $Ax = 0$, et par conséquent $\mathcal{V} = \{x \in \mathbb{R}^n | Ax = 0\}$. La première formulation pourrait être qualifiée de "formulation par générateurs", tandis que la seconde quand à elle de "formulation par contraintes". La correspondance dans le monde des cônes convexes, sera pour une "formulation par générateurs", une expression du type

$$K = \mathtt{cone}(\{v^i\}_{i\in I}) = \left\{ \sum_{i\in I} \alpha_i v^i \,|\alpha \in (\mathbb{R}_+)^I \right\},$$

ou une formulation à l'aide d'inégalités ou contraintes

$$K = \{x \in \mathbb{R}^n \,|\, \langle a_i, x\rangle \leqslant 0 \ \forall i \in J\}.$$

Ainsi, la première définition que nous donnons semble se rapprocher d'une formulation par générateurs, même si l'enveloppe convexe conique n'apparaît pas explicitement. En fait, cette dernière surgira lorsque l'on montrera que l'on peut choisir justement $\mathcal{S}_{n+1}(\mathbb{R})$ au lieu de P_1 pour $\mathcal{C}_{n+1}$.

Exemple II.1: *Pour donner une idée de l'application linéaire $\mathcal{A}$ définie en* (II.3)*, calculons l'image d'une matrice symétrique Q de dimension 3 ; si*

$$Q = \begin{pmatrix} q_{11} & q_{12} & q_{13} \\ q_{12} & q_{22} & q_{23} \\ q_{13} & q_{23} & q_{33} \end{pmatrix} \text{ , alors } \mathcal{A}(Q) = \begin{pmatrix} q_{11} + q_{22} + q_{33} \\ q_{12} + q_{23} \\ q_{13} \end{pmatrix}.$$

Si la structure liée à l'autocorrélation n'apparaît pas forcément ici de manière explicite, en prenant $Q \in P_1$, sous la forme

$$Q = \begin{pmatrix} x_0^2 & x_0x_1 & x_0x_2 \\ x_0x_1 & x_1^2 & x_1x_2 \\ x_0x_2 & x_1x_2 & x_2^2 \end{pmatrix} \text{ , alors } \mathcal{A}(Q) = \begin{pmatrix} x_0^2 + x_1^2 + x_2^2 \\ x_0x_1 + x_1x_2 \\ x_0x_2 \end{pmatrix},$$

et l'on voit bien "surgir" dans les composantes de $\mathcal{A}(xx^\top)$ les coefficients d'autocorrélation de x.

On peut aussi adopter pour $\mathcal{C}_{n+1}$ une définition peut-être plus intuitive géométriquement, en tant qu'intersection d'une infinité de demi-espaces, qui consiste à voir $\mathcal{C}_{n+1}$ comme l'ensemble des polynômes trigonômetriques pair positifs.

Définition II.2: *Soit le vecteur* $\nu(\omega) = (1, 2\cos\omega, \cdots, 2\cos n\omega) \in \mathbb{R}^{n+1}$ *et le demi-espace* $H_\omega^+ = \{x \in \mathbb{R}^{n+1} \mid \langle \nu(\omega), x\rangle \geqslant 0\}$, *alors*

$$\mathcal{C}_{n+1} = \bigcap_{\omega\in[0,\pi]} H_\omega^+ = \{x \in \mathbb{R}^{n+1} \,:\, \forall\, \omega \in [0,\pi],\ \ x_0 + 2\sum_{k=1}^{n} x_k \cos k\omega \geqslant 0\}.$$

Par rapport à la remarque II.1, on voit ici clairement que c'est une formulation par contraintes. En toute rigueur, la Définition II.2 ne devrait pas avoir le statut de définition, mais plutôt de proposition, vu que l'on à déjà défini $\mathcal{C}_{n+1}$. Cependant, comme certains auteurs (c.f. Dumitrescu *et al.*, [19]) l'utilisent comme base de départ pour leur étude, nous préférons ici maintenir ce statut de définition, et nous allons démontrer juste après qu'en réalité, ces deux définitions sont totalement équivalentes, grâce à un corollaire du théorème de Riesz-Féjer, appelé aussi théorème de la factorisation spectrale.

Théorème II.1 (Riesz-Féjer): *Un polynôme trigonométrique pair, i.e*

$$R(\omega) = r_0 + 2\sum_{k=1}^{n} r_k \cos k\omega,$$

avec $r_0, \cdots, r_n$ *réels, peut se mettre sous la forme*

$$\left|\sum_{l=0}^{n} h_l e^{il\omega}\right|^2$$

où les $h_0, \cdots, h_n$ *sont réels si, et seulement si,* $R(\omega) \geqslant 0$ *pour tout* $\omega \in [-\pi, \pi]$ (*autrement dit sur* $[0, \pi]$).

Pour la démonstration du Théorème II.1 à caractère essentiellement algébrique mais sans difficultés majeures, on renvoie par exemple aux monographies [33, 5]. On peut donc démontrer maintenant que

Proposition II.1: *Les définitions II.1 et II.2, définissent le même sous-ensemble convexe de* $\mathbb{R}^{n+1}$.

Démonstration. Considérons un polynôme trigonométrique pair ($x_l = x_{-l}$)

$$X(\omega) = \sum_{l=-n}^{n} x_l e^{il\omega},$$

où les x_l sont réels. Le Théorème II.1 nous dit :

$$\exists H\ ,\ \forall \omega \in [0,\pi]\ ,\ X(\omega) = |H(\omega)|^2,$$

avec

$$H(\omega) = \sum_{k=0}^{n} h_k e^{ik\omega}$$

à coefficients réels si, et seulement si,

$$X(\omega) = x_0 + 2\sum_{k=1}^{n} x_k \cos k\omega \geqslant 0 \ , \ \forall \omega \in [0, \pi],$$

qui est exactement la Définition II.2.

Or si $X = |H|^2$, on a vu grâce à (I.4) que

$$x_k = \sum_{p=0}^{n-|k|} h_k h_{p+|k|}.$$

Cet élément x_k est défini pour $k \in \{-n, \cdots, n\}$, mais comme X est pair, il est défini de manière unique par $(x_0, \cdots, x_n)$ (car $x_{-k} = x_k$), et on peut étudier la formule en se limitant à $k \geqslant 0$; on obtient alors exactement l'expression de la Définition II.1. □

Le théorème de Riesz Féjèr, affirme donc l'équivalence entre le fait qu'un polynôme trigonométrique d'une variable soit un "carré" et sa positivité. Sa démonstration donne une idée pour construire les autres solutions y possibles du système $x = \mathcal{A}(yy^\top)$ grâce à la proposition suivante :

Proposition II.2: *Considérons le système polynômial suivant*

$$(S_x) \quad \mathcal{A}(yy^\top) = x,$$

et supposons qu'il existe $\check{y}$ *(dans* $\mathbb{R}^{n+1}$ *ou* $\mathbb{C}^{n+1}$*) vérifiant* S_x *: alors il existe au signe près au plus* $2^n - 1$ *autres solutions (eventuellement confondues) de* S_x *que l'on peut construire explicitement connaissant les racines complexes (éventuellement pour certaines réelles) du polynôme*

$$Y(z) = \sum_{k=0}^{n} \check{y}_k z^k.$$

Démonstration. En supposant pour la simplicité de la preuve que $\check{y}_n \neq 0$, si l'on pose $\widetilde{X}(z) = \sum_{k=-n}^{n} x_{|k|} z^k$ et $Y(z) = \sum_{k=0}^{n} \check{y}_k z^k$, alors le développement de $Y(z)Y(z^{-1})$ (on appelle cette écriture une **factorisation spectrale** de $\widetilde{X}(z)$) donne en regroupant la somme suivant les puissance de z

$$\sum_{k=-n}^{n} \left(\sum_{l=0}^{n-k} \check{y}_l \check{y}_{l+k} \right) z^k = \sum_{k=-n}^{n} \mathcal{A}_{|k|}(\check{y}\check{y}^\top) z^k.$$

Si l'on identifie alors les coefficients des puissances de z, on obtient l'équivalence

$$\widetilde{X}(z) = Y(z)Y(z^{-1}) \Leftrightarrow x = \mathcal{A}(\check{y}\check{y}^\top).$$

Il faut noter que l'on ne peut pas ici utiliser $\mathtt{corr}_a$ car $\check{y}$ peut être complexe, et l'identification $\mathtt{corr}_a(\check{y},\check{y}) = \mathcal{A}(\check{y}\check{y}^\top)$ n'est valable que dans le cas réel. A chaque solution y, on peut donc associer une factorisation spectrale de $\widetilde{X}$ sous la forme $\widetilde{X}(z) = Y(z)Y(z^{-1})$. Voyons maintenant comment déterminer les autres solutions : d'aprés le théorème de D'Alembert, $Y(z)$ admet au plus n racines $(\alpha_1, \cdots, \alpha_n)$. Alors

$$Y(z) = \lambda \prod_{i=1}^{n} (z - \alpha_i);$$

Soit $B \subset \{1, \cdots, n\}$ construisons de la façon suivante

$$Y_B(z) = \lambda \prod_{i \notin B} (z - \alpha_i) \prod_{i \in B} (1 - \alpha_i z);$$

Si l'on développe $Y_B(z)Y_B(z^{-1})$, on retrouve là encore $\widetilde{X}(z)$ et donc le vecteur y^B des coefficients dans la base canonique des z^k de Y_B est encore solution de (S_x). On a donc autant de solutions de (S_x) qu'il y a de parties dans $\{1, \cdots, n\}$. Pour ne construire que les solutions réelles, dans le cas où $Y(z)$ est réel, il faut appliquer la transformation $(z - \alpha_i) \mapsto (1 - \alpha_i z)$ (ou d'une autre manière $\alpha_i \mapsto 1/\alpha_i$) sur des couples de valeurs propres conjuguées. Comme on ne connaît pas *a priori* le nombre de valeurs propres conjuguées de $Y(z)$, la seule borne que l'on puisse donner sur le nombre de solutions réelles de (S_x), c'est 2^n. □

Le calcul de la factorisation spectrale est un problème important en Traitement du Signal particulièrement en grande dimension (les algorithmes de calcul de racines ne sont plus aussi performants qu'en petite dimension), et parmi toutes les factorisations possibles, on est en général interessé par celles dont toutes les racines sont à l'extérieur du cercle unité ($Y(z^{-1})$ aura alors elle toutes ses racines à l'intérieur) : en effet, une telle factorisation est dite stable ou *de phase minimum* , car le filtre associé est alors stable. Dans sa thèse [27], Hachez présente une méthode basée sur la programmation SDP qui peut se formuler en utilisant $\mathcal{C}_{n+1}$.

II.3 Propriétés de $\mathcal{C}_{n+1}$

II.3.1 Ordre partiel induit par $\mathcal{C}_{n+1}$

Proposition II.3: *$\mathcal{C}_{n+1}$ est un cône convexe, fermé, solide (i.e. d'interieur non vide) et pointé ; il induit par conséquent un ordre partiel sur $\mathbb{R}^{n+1}$ noté $\preceq_{\mathcal{C}_{n+1}}$ et défini par*

$$x \preceq_{\mathcal{C}_{n+1}} y \text{ si et seulement si } y - x \in \mathcal{C}_{n+1}.$$

Démonstration. – Le fait que $\mathcal{C}_{n+1}$ soit un cône est trivial.

– En considérant la définition II.2,

$$\mathcal{C}_{n+1} = \bigcap_{\omega\in[0,\pi]} H^+_\omega$$

$\mathcal{C}_{n+1}$ est l'intersection d'une infinité de convexes fermés (les demi-espaces H^+_ω). Par conséquent, il est convexe et fermé.

– $\mathcal{C}_{n+1}$ est *solide* ; en effet, $e_0 = (1, 0, \cdots, 0)$ vérifie strictement l'ensemble des inégalités de la Définition II.2, il existe donc une boule centrée en e_0 strictement incluse dans $\mathcal{C}_{n+1}$.

– $\mathcal{C}_{n+1}$ est *pointé* (ou *saillant*), c'est-à-dire que $\mathcal{C}_{n+1} \cap (-\mathcal{C}_{n+1}) = \{0\}$. En effet, soit $x \in \mathcal{C}_{n+1} \cap (-\mathcal{C}_{n+1})$; alors il existe $y \in \mathbb{R}^{n+1}$ tel que $x_0 = \|y\|^2$, mais $-x \in \mathcal{C}_{n+1}$ donc il existe $w \in \mathbb{R}^{n+1}$ tel que $-x_0 = \|w\|^2$, d'où

$$0 = x_0 + (-x_0) = \|y\|^2 + \|w\|^2,$$

et par conséquent, $y = w = 0$ et $x = 0$. Donc $\mathcal{C}_{n+1} \cap (-\mathcal{C}_{n+1}) = \{0\}$.

Puisque $\mathcal{C}_{n+1}$ est un cône convexe pointé, ceci nous assure que $\preceq_{\mathcal{C}_{n+1}}$ est bien un ordre partiel conformément à la définition donnée dans [5].

□

Le caractère fermé de $\mathcal{C}_{n+1}$ permet de passer à la limite dans les inégalités généralisées, et le fait que $\mathcal{C}_{n+1}$ soit d'intérieur non vide nous permet de définir une inégalité stricte par

$$x \prec_{\mathcal{C}_{n+1}} y \text{ si et seulement si } y - x \in \mathbf{int}\,\mathcal{C}_{n+1}.$$

Remarque II.2: *Prouver la convexité de $\mathcal{C}_{n+1}$ est particulièrement simple lorsqu'on utilise la formulation par contraintes. Par contre, démontrer la convexité avec Définition II.1, est à notre connaissance impossible. En réalité, le cas d'une seule variable nous permet de confondre comme un unique objet, deux objets mathématiques - les vecteurs à composantes autocorrélées et les coefficients de polynômes positifs (qui normalement seraient distincts, comme on le verra ultérieurement dans le cas de plusieurs variables). La convexité, qui est une propriété fondamentale du point de vue théorique comme pratique et qui nous sera très utile par la suite, semble donc être intrinsèquement liée aux polynômes positifs.*

II.3.2 Géométrie de $\mathcal{C}_{n+1}$

II.3.2.1 Base compacte de $\mathcal{C}_{n+1}$

Pour se faire une idée plus précise de la géométrie de $\mathcal{C}_{n+1}$, on peut rechercher ses directions extrêmes : pour cela, on peut chercher une **base** de $\mathcal{C}_{n+1}$, c'est à dire un ensemble $B \subset \mathbb{R}^{n+1}$ tel que pour tout $x \in \mathcal{C}_{n+1}$ il existe $t_x \geqslant 0$ et $b_x \in B$ tels que

$$x = t_x b_x.$$

Grosso modo, connaissant B, on peut engendrer tout $\mathcal{C}_{n+1}$ à l'aide des demi-droites issues de l'origine qui passent par les points de B.

Lemme II.1: *Soit* $\mathbb{S}_n = \{x \in \mathbb{R}^{n+1} \mid \|x\| = 1\})$ *et* $\Theta : \left\{ \begin{array}{ccc} \mathbb{R}^{n+1} & \to & \mathcal{C}_{n+1} \\ x & \mapsto & \mathcal{A}(xx^\top) \end{array} \right.$ *; alors l'ensemble*

$$\mathcal{U}_n := \Theta(\mathbb{S}_n),$$

est une base de $\mathcal{C}_{n+1}$.

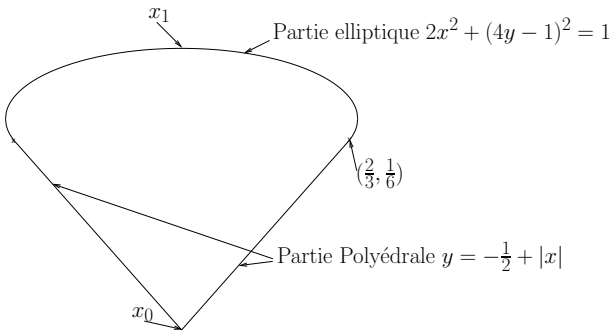
Démonstration. La fonction Θ est quadratique, homogène de degré 2, i.e. pour tout $y \in \mathbb{R}^{n+1}$ et $t \in \mathbb{R}$ alors $\Theta(ty) = t^2\Theta(y)$. Soit $x \in \mathcal{C}_{n+1}$, il existe $y \in \mathbb{R}^{n+1}$ tel que $x = \Theta(y)$, or $x_0 = \Theta(y)_0 = \|y\|^2$. Si $x_0 = 0$, ceci implique que y est le vecteur nul, et par conséquent x aussi ; alors x peut s'écrire comme $0 \cdot z$ où z est n'importe quel vecteur de $\mathcal{U}_n$. Sinon

$$y = \|y\| \cdot \tilde{y} \text{ avec } \tilde{y} = \frac{y}{\|y\|} \in \mathbb{S}_n,$$

et $x = \|y\|^2 \cdot \Theta(\tilde{y}) = x_0 \cdot u$ avec $u = \Theta(\tilde{y}) \in \mathcal{U}_n$. On obtient donc l'ecriture souhaitée. □

En utilisant la définition de $\mathcal{U}_n$, on obtient que pour $x \in \mathcal{U}_n$, $x_0 = \|s\|^2 = 1$, donc $\mathcal{U}_n$, est contenu intégralement dans le plan d'équation $x_0 = 1$, et peut être assimilé ainsi à un compact de $\mathbb{R}^n$; dans la suite, on considèrera à la fois $\mathcal{U}_n$ comme un sous-ensemble de $\mathbb{R}^n$, et comme un sous-ensemble plat dans $\mathbb{R}^{n+1}$ situé dans le plan $x_0 = 1$, cela dépendra des situations. Dans le cas, $n = 2$, on peut trouver facilement $\mathcal{U}_2$, par exemple en calculant la famille d'enveloppes de la courbe $\omega \mapsto (2\cos\omega, 2\cos 2\omega)$ sur $[0, \pi]$. On obtient alors la figure II.1.

FIG. II.1: Coupe par le plan $x_0 = 1$ de $\mathcal{C}_3$



Pour se faire une idée de $\mathcal{U}_n$ en dimension supérieure, on peut en chercher un encadrement. Si on cherche par exemple le plus petit pavé $P_{\alpha,\beta} = \prod_{i=1}^{n}[\alpha_i, \beta_i]$ contenant $\mathcal{U}_n$, ceci revient à chercher pour chaque composante i dans $\{1, \cdots, n\}$

$$\max_{x \in \mathbb{S}_n} \Theta_i(x) \text{ et } \min_{x \in \mathbb{S}_n} \Theta_i(x).$$

Or,

$$\max_{x\in\mathbb{S}_n} \Theta_i(x) = \max_{x\in\mathbb{S}_n} \langle\langle A^{(i)}, xx^\top\rangle\rangle = \max_{x\in\mathbb{S}_n} \langle A^{(i)}x, x\rangle = \lambda_{\max}(A^{(i)}),$$

la dernière égalité provenant des formulations variationnelles de Rayleigh pour les valeurs propres extrémales. Notons au passage, que l'on obtient la même formulation, *mutatis mutandis*, avec le minimum. Donc pour trouver le pavé $P_{\alpha,\beta}$, il suffit de connaître la plus grande et la plus petite des valeurs propres des $A^{(i)}$. Les $A^{(i)}$ ayant une structure bien particulière, il est possible de déterminer directement leurs valeurs propres et vecteurs propres associés.

II.3.2.2 Eléments propres de $A^{(i)}$ (cf. (II.2))

Si l'on considère le spectre des E^i, il est particulièrement simple, puisque réduit à $\{0\}$, du fait que ces matrices soient nilpotentes. Par contre, lorsque l'on considère les parties symétriques de ces matrices, la situation se complique de manière non négligeable. Le cas $i = 1$ est un exercice classique d'analyse matricielle faisant intervenir des $\sin k\pi/(n+1)$, et les cas où $i > \lfloor n/2\rfloor + 1$, s'analysent très facilement en utilisant des espaces stables qui permettent de conclure que le spectre est réduit à $\{-1/2, 0, 1/2\}$. Par contre, les cas où $i \in \{2, \cdots, \lfloor n/2\rfloor\}$, ne semblent pas être très connus. A notre connaissance, seul l'article de *Strang et al.* [24] y fait référence, et donne le spectre de ces matrices. De notre côté, après quelques expériences numériques faisant apparaître des regroupements de valeurs propres selon la valeur de i, nous avons déduit une formule inspirée du cas $i = 1$, puis nous avons démontré la validité de ces formules analytiques, en proposant un système de vecteurs propres qui n'était pas connu des auteurs de [24]. Ceci a donné lieu à l'article [22], dont nous proposons ici une version française. Le résultat principal concernant les éléments propres de ces matrices est le suivant :

Théorème II.2: *Soit* $i \in \{1, \cdots, n\}$ *et* p, m_1 *et* m_2 *définis par*

$$p = \left\lfloor \frac{n+1}{i} \right\rfloor \ , m_1 = i - n - 1 + ip \ , \ m_2 = n + 1 - ip = i - m_1.$$

Alors les valeurs propres des $A^{(i)}$ *sont de deux types :*

- *d'une part, pour tout* i *dans* $\{1, \cdots, n\}$*, on a une première classe*

$$\lambda^{1,l} = \cos\left(\frac{l\pi}{p+1}\right) \ \textit{pour}\ l \in \{1, \cdots, p\}$$

avec pour multiplicité m_1 *et comme vecteurs propres associés*

$$v^{1,l,q} = \sum_{j=1}^{p} \sin\left(\frac{lj\pi}{p+1}\right) e_{\{i(j-1)+m_2+q\}}$$

où $q \in \{1, \cdots, m_1\}$.

– *d'autre part, lorsque* i *ne divise pas* $n+1$*, il apparaît alors une classe supplémentaire,*

$$\lambda^{2,l} = \cos\left(\frac{l\pi}{p+2}\right) \text{ où } l \in \{1, \cdots, p+1\}$$

de multiplicité m_2 *et dont les vecteurs propres associés sont*

$$v^{2,l,q} = \sum_{j=1}^{p+1} \sin\left(\frac{lj\pi}{p+2}\right) e_{\{i(j-1)+q\}}$$

avec $q \in \{1, \cdots, m_2\}$.

Comme la démonstration est un peu longue et calculatoire, on va la décomposer en plusieurs résultats, et l'on va commencer par proposer un lemme essentiel pour prouver l'indépendance linéaire des vecteurs propres.

Lemme II.2: *Soit* $T^p \in \mathcal{S}_p(\mathbb{R})$ *la matrice symétrique de terme général*

$$T^p_{ij} = \sin\left(\frac{ij\pi}{p+1}\right) \text{ pour } i, j \in \{1, \cdots, p\}.$$

Alors T^p *vérifie la relation* $(T^p)^\top T^p = \frac{p+1}{2} I_p$.

Ainsi la matrice T^p est orthogonale à un facteur multiplicatif près, ce qui prouve son inversibilité. C'est la matrice d'une Transformée de Sinus Discrète (DST), autrement dit la partie imaginaire d'une Transformée de Fourier Discrète, privé de son noyau

Démonstration. Posons

$$u^l = \sum_{j=1}^{p} \sin\left(\frac{jl\pi}{p+1}\right) e_j.$$

Pour la suite, on notera que pour $x \neq 0 \mod 2\pi$

$$\sum_{j=1}^{p} \cos jx = \cos\left(\frac{(p+1)x}{2}\right) \frac{\sin\left(\frac{px}{2}\right)}{\sin\left(\frac{x}{2}\right)} = \cos\left(\frac{px}{2}\right) \frac{\sin\left(\frac{(p+1)x}{2}\right)}{\sin\left(\frac{x}{2}\right)} - 1. \tag{II.6}$$

Considérons u^l et u^k, on calcule

$$\langle u^k, u^l \rangle = \sum_{j=1}^{p} \sin\left(\frac{jk\pi}{p+1}\right) \sin\left(\frac{jl\pi}{p+1}\right) = \frac{1}{2} \sum_{j=1}^{p} \left[\cos\left(\frac{j(k-l)\pi}{p+1}\right) - \cos\left(\frac{j(k+l)\pi}{p+1}\right)\right]. \tag{II.7}$$

– Si $l = k$, alors $k - l = 0$ et $k + l = 2k$ et en utilisant (II.7) puis (II.6), on obtient

$$\langle u^k, u^l \rangle = \frac{p}{2} - \frac{1}{2}\left(\cos\left(\frac{2pk\pi}{2(p+1)}\right) \frac{\sin\left(\frac{2(p+1)k\pi}{2(p+1)}\right)}{\sin\left(\frac{2k\pi}{2(p+1)}\right)} - 1\right) = \frac{p+1}{2}.$$

– Si maintenant $k \neq l$, alors

$$\langle u^k, u^l \rangle = \frac{1}{2}\left(\cos\left(\frac{p(k-l)\pi}{2(p+1)}\right) \frac{\sin\left(\frac{(k-l)\pi}{2}\right)}{\sin\left(\frac{(k-l)\pi}{2(p+1)}\right)} - \cos\left(\frac{p(k+l)\pi}{2(p+1)}\right) \frac{\sin\left(\frac{(k+l)\pi}{2}\right)}{\sin\left(\frac{(k+l)\pi}{2(p+1)}\right)} \right).$$

Or

$$\sin\left(\frac{(k+l)\pi}{2}\right) = (-1)^l \sin\left(\frac{(k-l)\pi}{2}\right),$$

donc

$$\langle u^k, u^l \rangle = \frac{\sin\left(\frac{(k-l)\pi}{2}\right)}{2\sin\left(\frac{(k+l)\pi}{2(p+1)}\right)\sin\left(\frac{(k-l)\pi}{2(p+1)}\right)}$$
$$\left(\cos\left(\frac{p(k-l)\pi}{2(p+1)}\right) \sin\left(\frac{(k+l)\pi}{2(p+1)}\right) + (-1)^{l+1} \cos\left(\frac{p(k+l)\pi}{2(p+1)}\right) \sin\left(\frac{(k-l)\pi}{2(p+1)}\right)\right).$$

Si $k-l$ est pair, alors en raison du terme $\sin\left(\frac{(k-l)\pi}{2}\right)$

$$\langle u^k, u^l \rangle = 0.$$

Sinon $k-l$ est impair, et $k+l$ aussi. Comme

$$\cos\left(\frac{px\pi}{2(p+1)}\right) = (-1)^{\frac{x-1}{2}} \sin\left(\frac{x\pi}{2(p+1)}\right) \text{ si } x \in 2\mathbb{Z}+1,$$

en remarquant que $(-1)^{l+1} = (-1)^{l-1}$, on déduit

$$\langle u^k, u^l \rangle = \frac{\sin\left(\frac{(k-l)\pi}{2}\right)}{2}\left((-1)^{\frac{k-l-1}{2}} + (-1)^{-l+1}(-1)^{\frac{k+l-1}{2}}\right) = 0.$$

□

Avant de poursuivre dans la preuve, nous avons besoin d'introduire deux fonctions qui allégeront les notations dans la suite : soit

$$g_q : \begin{cases} \{1, \cdots, p\} \to \{1, \cdots, n+1\} \\ j \mapsto i(j-1) + m_2 + q \end{cases} \quad \text{pour } q \in \{1, \cdots, m_1\},$$

et

$$h_q : \begin{cases} \{1, \cdots, p+1\} \to \{1, \cdots, n+1\} \\ j \mapsto i(j-1) + q \end{cases} \quad \text{pour } q \in \{1, \cdots, m_2\}.$$

On remarque que, comme i est non nul, g_q et h_q sont strictement croissantes et donc injectives. Alors on va démontrer le lemme ci-dessous.

Lemme II.3: *Soient* p, m_1 *et* m_2 *définis comme dans le Théorème II.2 ; alors on peut partitionner l'ensemble* $\{1, \cdots, n+1\}$ *de la façon suivante*

$$\{1, \cdots, n+1\} = \bigcup_{q=1}^{m_1} g_q(\{1, \cdots, p\}) \cup \bigcup_{q=1}^{m_2} h_q(\{1, \cdots, p+1\}) \tag{II.8}$$

en remarquant que les ensembles images dans cette union sont disjoints deux à deux : pour tout $q, t \in \{1, \cdots, p\}$, $q \neq t$, *et* $(r, s) \in \{1, \cdots, p+1\}$, $r \neq s$, *on a*

$$\begin{cases} g_q(\{1, \cdots, p\}) \cap g_t(\{1, \cdots, p\}) = \emptyset, \\ h_r(\{1, \cdots, p+1\}) \cap h_s(\{1, \cdots, p+1\}) = \emptyset, \\ h_r(\{1, \cdots, p+1\}) \cap g_q(\{1, \cdots, p\}) = \emptyset. \end{cases}$$

Démonstration. – Si $x \in g_{q_1}(\{1, \cdots, p\}) \cap g_{q_2}(\{1, \cdots, p\})$, considérons d'abord le cas où $q_1, q_2 < m_1$, alors il existe $y_1, y_2 \in \{1, \cdots, p\}$ tels que

$$i(y_1) + \underbrace{m_2 + q_1}_{\leqslant i-1} = i(y_2 - 1) + \underbrace{m_2 + q_2}_{\leqslant i-1}$$

et par unicité de la division euclidienne par i

$$\begin{cases} y_1 - 1 = y_2 - 1, \\ m_2 + q_1 = m_2 + q_2. \end{cases}$$

Et finalement, on conclut que $q_1 = q_2$.
Si maintenant, $q_1 = m_1$ et $q_2 < m_1$ (ou $q_2 = m_1$ et $q_1 < m_1$), c'est impossible car i divise $n+1$ et $n+1 \mod i = m_2 + q_2 \neq 0$, donc il ne reste comme possibilité que $q_1 = q_2 = m_2$.

– si $x \in h_{q_1}(\{1, \cdots, p+1\}) \cap h_{q_2}(\{1, \cdots, p+1\})$ alors il existe $y_1, y_2 \in \{1, \cdots, p+1\}$ tels que $i(y_1 - 1) + q_1 = i(y_2 - 1) + q_2$. Comme $q_1 \leqslant m_2 < 1$ (et de même pour q_2), l'unicité de la division euclidienne nous permet de conclure que

$$\begin{cases} y_1 - 1 = y_2 - 1, \\ q_1 = q_2. \end{cases}$$

– Si $x \in g_{q_1}(\{1, \cdots, p\}) \cap h_{q_2}(\{1, \cdots, p+1\})$ alors il existe $(y_1, y_2) \in \{1, \cdots, p\} \times \{1, \cdots, p+1\}$ tels que

$$i(y_1 - 1) + m_2 + q_1 = i(y_2 - 1) + q_2.$$

Si $q_1 = m_1$ alors i divise $n+1$ et on obtient encore une contradiction, sinon $m_2 + q_1 < i$ et l'unicité de la division euclidienne par i donne

$$\begin{cases} y_1 - 1 = y_2 - 1, \\ m_2 + q_1 = q_2. \end{cases}$$

On suppose évidemment que $m_2 \neq 0$ sinon il n'y a rien à démontrer, et donc $m_2 < m_2 + q_1 = q_2 \leqslant m_2$ ce qui est contradictoire donc

$$g_{q_1}(\{1, \cdots, p\}) \cap h_{q_2}(\{1, \cdots, p+1\}) = \emptyset.$$

Ainsi, on a prouvé effectivement que tous les ensembles de la partition (II.8) sont deux à deux disjoints. Il reste donc à montrer que leur réunion recouvre $\{1, \cdots, n+1\}$. Soit $x \in \{1, \cdots, n+1\}$, alors il existe un unique couple (d, r) tel que

$$x = id + r \text{ avec } 0 \leqslant r \leqslant i - 1.$$

- si $r = 0$ alors $x = id = i(d-1) + m_1 + m_2 = g_{m_1}(d)$;
- si $0 \leqslant r \leqslant m_2$ alors $x = id + r = h_r(d+1)$;
- si $m_2 + 1 \leqslant r \leqslant i - 1$ alors $x = id + m_2 + r - m_2 = g_{r-m_2}(d+1)$.

□

On est maintenant à même de prouver la

Proposition II.4: *La famille*

$$\mathcal{V} = \{(v^{1,l,q})_{l=1,q=1}^{l=p,q=m_1}\} \cup \{(v^{2,l,q})_{l=1,q=1}^{l=p+1,q=m_2}\}$$

forme une base de $\mathbb{R}^{n+1}$.

Démonstration. On va montrer l'indépendance linéaire de tous ces vecteurs en une fois : supposons que pour $\{\alpha_{lq}\}_{q=1,l=1}^{q=m_1,l=p}$ et $\{\beta_{lq}\}_{q=1,l=1}^{q=m_2,l=p+1}$,

$$\sum_{q=1}^{m_1}\sum_{l=1}^{p} \alpha_{lq} v^{1,l,q} + \sum_{q=1}^{m_2}\sum_{l=1}^{p+1} \beta_{lq} v^{2,l,q} = 0. \tag{II.9}$$

Regardons l'équation vectorielle (II.9) pour chaque composante :

$$\sum_{q=1}^{m_1}\sum_{l=1}^{p} \alpha_{lq} v_k^{1,l,q} + \sum_{q=1}^{m_2}\sum_{l=1}^{p+1} \beta_{lq} v_k^{2,l,q} = 0,$$

où $k \in \{1, \cdots, n+1\}$.

Si on partitionne $\{1, \cdots, n+1\}$ selon les "paquets" du lemme II.8, alors choisissons $q_1 \in \{1, \cdots, m_1\}$

$$\sum_{q=1}^{m_1}\sum_{l=1}^{p} \alpha_{lq} v_{g_{q_1}(j)}^{1,l,q} + \sum_{q=1}^{m_2}\sum_{l=1}^{p+1} \beta_{lq} \underbrace{v_{g_{q_1}(j)}^{2,l,q}}_{0} = 0 \text{ pour } j = 1, \cdots, p, \tag{II.10}$$

ce qui équivaut à

$$\sum_{l=1}^{p} \alpha_{lq_1} \sin\left(\frac{lj\pi}{p+1}\right) = 0 \text{ pour } j = 1, \cdots, p, \tag{II.11}$$

où $(i,j) \in \{1,\cdots,p\}$, alors (II.11) est équivalent à

$$T^p \alpha_{\cdot q_1} = 0.$$

Le Lemme II.2 nous permet de conclure que $\alpha_{\cdot q_1} = 0$ pour tout $q_1 \in \{1,\cdots,m_1\}$, et donc $\alpha = 0$.

De même, si on regarde les paquets formés par $y = h_{q_2}(j)$ où $q_2 \in \{1,\cdots,m_2\}$, et $j \in \{1,\cdots,p+1\}$, alors

$$\sum_{q=1}^{m_1}\sum_{l=1}^{p} \alpha_{lq} \underbrace{\nu^{1,l,q}_{h_{q_2}(j)}}_{0} + \sum_{q=1}^{m_2}\sum_{l=1}^{p+1} \beta_{lq}\nu^{2,l,q}_{h_{q_2}(j)} = 0 \text{ pour } j = 1,\cdots,p+1, \tag{II.12}$$

ce qui se réécrit en

$$\sum_{l=1}^{p+1} \beta_{lq_2} \sin\left(\frac{ljp\pi}{p+1}\right) = 0 \text{ pour } j = 1,\cdots,p+1, \tag{II.13}$$

qui n'est rien d'autre que

$$T^{p+1}\beta_{\cdot q_2} = 0.$$

Là encore, on conclut que $\beta = 0$ grâce au Lemme II.2. La famille est donc libre, pour vérifier que c'est une base, il suffit que son cardinal soit égal à $n+1$, or par construction

$$n+1 = m_1 p + m_2(p+1).$$

□

Afin d'achever la preuve, il suffit de vérifier les relations entre les valeurs et les vecteurs propres suggérés, du type $Ax = \lambda x$ si x est un vecteur propre associé à λ. On commence ainsi avec la

Proposition II.5: *Pour tout* $(l,q) \in \{1,\cdots,p\} \times \{1,\cdots,m_1\}$, *la relation suivante*

$$A^{(i)}\nu^{1,l,q} = \lambda^{1,l}\nu^{1,l,q}$$

est vérifiée.

Démonstration. Pour alléger les notations, posons $A = A^{(i)}$, $\nu = \nu^{1,l,q}$ et $g = g_q$. Alors nous allons prouver que $A\nu = \lambda\nu$, ou de manière équivalente

$$(A\nu)_k = \lambda\nu_k \text{ pour } k \in \{1,\cdots,n+1\}.$$

Considérons deux cas, pour l'indice k.

– Si $k \notin g(\{1,\cdots,p\})$ alors $\lambda v_k = 0$, car

$$v = \sum_{l=1}^{p} \sin\left(\frac{lj\pi}{(p+1)}\right) e_{g(l)}$$

et

$$\begin{aligned}
(Av)_k &= \sum_{t=1}^{n+1} A_{kt} v_t = \sum_{t=1}^{n+1}\sum_{j=1}^{p} A_{kt} v_{g(j)}[t = g(j)] \\
&= \sum_{j=1}^{p} A_{kg(j)} v_{g(j)} = \frac{1}{2}\sum_{j=1}^{p}([g(j)+i=k]+[g(j)=k+i])v_{g(j)} \\
&= \frac{1}{2}\sum_{j=2}^{p}[g(j+1)=k]v_{g(j)} + [g(j-1)=k]v_{g(j)} \\
&+ [g(1)+i=k]v_{g(1)} + [g(p)=k+i]v_{g(p)} = 0,
\end{aligned}$$

parce que k n'appartient pas à $g(\{1,\cdots,p\})$.

– Si $k \in g(\{1,\cdots,p\})$, alors il existe $j_0 \in \{1,\cdots,p\}$ tel que $k = g(j_0)$ et donc

$$\lambda v_k = \cos\left(\frac{l\pi}{p+1}\right)\sin\left(\frac{lj_0\pi}{p+1}\right).$$

Pour comparer, si l'on calcule

$$\begin{aligned}
(Av)_k &= \sum_{j=1}^{p} A_{kg(j)} v_{g(j)} = \sum_{j=1}^{p} A_{g(j_0)g(j)} v_{g(j)} \\
&= \frac{1}{2}\left(\sum_{j=1}^{p}([g(j)+i=g(j_0)]v_{g(j)} + \sum_{j=1}^{p}[g(j)=g(j_0)+i]v_{g(j)}\right).
\end{aligned}$$

Nous devons distinguer trois cas selon la valeur de j_0.

– si $2 \leqslant j_0 \leqslant p-1$; alors en utilisant l'injectivité de g,

$$\begin{aligned}
(Av)_k &= \frac{1}{2}\left(\sum_{j=1}^{p}([g(j)=g(j_0-1)]v_{g(j)} + \sum_{j=1}^{p}[g(j)=g(j_0+1)]v_{g(j)}\right) \\
&= \frac{1}{2}\left(v_{g(j_0-1)} + v_{g(j_0+1)}\right) \\
&= \frac{1}{2}\left(\sin\left(\frac{l(j_0-1)\pi}{p+1}\right) + \sin\left(\frac{l(j_0+1)\pi}{p+1}\right)\right) \\
&= \sin\left(\frac{lj_0\pi}{p+1}\right)\cos\left(\frac{l\pi}{p+1}\right) = \lambda v_k.
\end{aligned}$$

– Si $j_0 = 1$, alors

$$(Av)_k = \frac{1}{2}\sin\left(\frac{2l\pi}{p+1}\right) = \sin\left(\frac{l\pi}{p+1}\right)\cos\left(\frac{l\pi}{p+1}\right) = \lambda v_k.$$

– Si $j_0 = p$, alors

$$\begin{aligned}(Av)_k &= \frac{1}{2}\sin\left(\frac{l(p-1)\pi}{p+1}\right)\\ &= \frac{1}{2}\left(\sin\left(\frac{lp\pi}{p+1}\right)\cos\left(\frac{l\pi}{p+1}\right) - \cos\left(\frac{lp\pi}{p+1}\right)\sin\left(\frac{l\pi}{p+1}\right)\right).\end{aligned}$$

Mais,

$$0 = \sin\left(\frac{l(p+1)\pi}{p+1}\right) = \left(\sin\left(\frac{lp\pi}{p+1}\right)\cos\left(\frac{l\pi}{p+1}\right) + \cos\left(\frac{lp\pi}{p+1}\right)\sin\left(\frac{l\pi}{p+1}\right)\right),$$

et donc

$$(Av)_k = \sin\left(\frac{lp\pi}{p+1}\right)\cos\left(\frac{l\pi}{p+1}\right) = \lambda v_k.$$

□

Puis pour vérifier les relations valeurs propres-vecteurs propres, de la seconde catégorie de vecteurs propres, il faut établir la

Proposition II.6: *Pour tout* $(l, q) \in \{1, \cdots, p+1\} \times \{1, \cdots, m_2\}$, *la relation suivante*

$$A^{(i)}v^{2,l,q} = \lambda^{2,l}v^{2,l,q}$$

est vérifiée.

Démonstration. On pose ici $v = v^{2,l,q}$ et $h = h_q$. Le cas où $k \notin h(\{1, \cdots, p+1\})$, se démontre de la même façon que précédemment, on peut donc légitimement supposer que $k \in h(\{1, \cdots, p+1\})$; il existe donc j_0 tel que $k = h(j_0)$. Ainsi

$$\lambda v_k = \cos\left(\frac{l\pi}{p+2}\right)\sin\left(\frac{lj_0\pi}{p+2}\right),$$

et

$$(Av)_k = \frac{1}{2}\left(\sum_{j=1}^{p+1}([h(j)+i=h(j_0)]v_{h(j)} + \sum_{j=1}^{p+1}[h(j)=h(j_0)+i]v_{h(j)}\right).$$

On doit encore faire une discussion selon la valeur de j_0

– si $2 \leqslant j_0 \leqslant p$, alors h étant injective,

$$\begin{aligned}(Av)_k &= \frac{1}{2}\left(v_{h(j_0-1)} + v_{h(j_0+1)}\right)\\ &= \frac{1}{2}\left(\sin\left(\frac{l(j_0-1)\pi}{p+2}\right) + \sin\left(\frac{l(j_0+1)\pi}{p+2}\right)\right)\\ &= \sin\left(\frac{lj_0\pi}{p+2}\right)\cos\left(\frac{l\pi}{p+2}\right) = \lambda v_k.\end{aligned}$$

– Si $j_0 = 1$, alors

$$(Av)_k = \frac{1}{2}\sin\left(\frac{2l\pi}{p+2}\right) = \sin\left(\frac{l\pi}{p+2}\right)\cos\left(\frac{l\pi}{p+2}\right) = \lambda v_k.$$

– Sinon $j_0 = p+1$, et

$$\begin{aligned}(Av)_k &= \frac{1}{2}\sin\left(\frac{l((p+1)-1)\pi}{p+2}\right)\\ &= \frac{1}{2}\left(\sin\left(\frac{l(p+1)\pi}{p+2}\right)\cos\left(\frac{l\pi}{p+2}\right) - \cos\left(\frac{l(p+1)\pi}{p+2}\right)\sin\left(\frac{l\pi}{p+2}\right)\right)\\ &= \sin\left(\frac{l(p+1)\pi}{p+2}\right)\cos\left(\frac{l\pi}{p+2}\right),\end{aligned}$$

pour les mêmes raisons que précédemment.

□

Finalement, on a trouvé une base de $\mathbb{R}^{n+1}$ qui diagonalise $A^{(i)}$, on est donc sûr d'avoir énuméré toutes les valeurs propres de $A^{(i)}$, et ainsi le Théorème II.2 est démontré. Puisque l'on connaît exactement les valeurs propres de $A^{(i)}$, on peut en déduire facilement le

Lemme II.4: *Soit $i \in \{1, \cdots, n+1\}$. Alors $A^{(i)}$ a un spectre symétrique par rapport à l'origine : c'est-à-dire que si λ est valeur propre de $A^{(i)}$ alors $-\lambda$ aussi.*

Démonstration. Soit

$$\lambda = \cos\left(\frac{j\pi}{p+1}\right) \text{ avec } j \in \{1, \cdots, p\};$$

alors

$$-\lambda = \cos\left(\frac{j\pi}{p+1} - \pi\right) = \cos\left(\frac{(j-(p+1))\pi}{p+1}\right) = \cos\left(\frac{k\pi}{p+1}\right),$$

avec $k = p+1-j$ qui appartient à $\{1, \cdots, p\}$ dès lors que $j \in \{1, \cdots, p\}$. Le cas où

$$\lambda = \cos\left(\frac{j\pi}{p+2}\right) \text{ avec } j \in \{1, \cdots, p+1\},$$

se traite de manière totalement similaire. □

II.3.2.3 Encadrement polyédrique de $\mathcal{U}_n$

En utilisant le Lemme II.4, on déduit que le pavé minimal contenant $P_{\alpha,\beta}$ s'écrit

$$\mathcal{O}_n = \prod_{i=1}^{n} [-\lambda_{max}(A^{(i)}), \lambda_{max}(A^{(i)})],$$

exprimé à l'aide du Théorème II.2 comme suit :

$$\mathcal{O}_n = \prod_{i|(n+1)} \left[-\cos\left(\frac{\pi}{\lfloor (n+1)/i \rfloor + 1}\right), \cos\left(\frac{\pi}{\lfloor (n+1)/i \rfloor + 1}\right)\right] \prod_{i\not|(n+1)} \left[-\cos\left(\frac{\pi}{\lfloor (n+1)/i \rfloor + 2}\right), \cos\left(\frac{\pi}{\lfloor (n+1)/i \rfloor + 2}\right)\right].$$

Ceci nous fournit une approximation polyédrique externe ; cependant, à l'aide de vecteurs propres, on peut aussi obtenir une approximation polyédrique interne. En effet, pour un vecteur propre $v^{j,l,q}$ (où $j \in \{1,2\}$, et l et q sont dans les ensembles adéquats), la valeur $\Theta(v^{j,l,q})$ est indépendante de q comme le montre la proposition suivante.

Proposition II.7: *Avec les notations précédemment définies,on a :*

Pour tout $(q,r) \in \{1,\cdots,m_1\}\times\{1,\cdots,m_2\}$, $\Theta(v^{1,l,q}) = \Theta(v^{1,l,1})$ *et* $\Theta(v^{2,l,r}) = \Theta(v^{2,l,1})$.

Démonstration. Calculons l'image de $v^{1,l,q}$ par Θ. Sa $k^{\text{ème}}$-composante vaut

$$\langle A^k v^{1,l,q}, v^{1,l,q}\rangle = \frac{1}{2}\sum_{w=1}^{p}\sum_{j=1}^{p} ([g_q(j)+k = g_q(w)] + [g_q(j) = k + g_q(w)]) v_{g_q(j)} v_{g_q(w)}.$$

Si on regarde alors le terme général de la somme on peut voir qu'en réalité, il ne dépend pas de q : en effet, par exemple l'assertion $g_q(j)+k = g_q(w)$ est équivalente à

$$i(j-1) + m_2 + q + k = i(w-1) + m_2 + q;$$

et la composante

$$v_{g_q(j)} = \sin\left(\frac{lj\pi}{p+1}\right),$$

ne dépend que de j. On peut faire le même raisonnement pour $[g_q(j) = k + g_q(w)]$ et $v_{g_q(w)}$; par conséquent nous avons besoin de calculer cette composante seulement pour $v^{1,l,1}$ pour connaître les images $\Theta(v^{1,l,q})$. La démonstration est exactement la même pour $v^{2,l,r}$, *mutatis mutandis* et en remarquant que

$$h_r(j) + k = h_r(w) \Leftrightarrow i(j-1) + r + k = i(w-1) + r,$$

qui est vrai quelque soit r et la composante

$$v_{h_r(j)} = \sin\left(\frac{lj\pi}{p+2}\right),$$

ne dépend pas en réalité de r. □

A l'aide de cette proposition, on peut considérer l'enveloppe convexe

$$\mathcal{I}_n = \mathrm{conv}\left(\left\{[i\,|\,(n+1)]\Theta(v^i) + (1-[i\,|\,(n+1)])\Theta(w^i) \;|\; i \in \{1,\cdots,n\}\right\}\right),$$

où

$$v^i = \sum_{j=1}^{\lfloor (n+1)/i \rfloor} \sin\left(\frac{j\pi}{\lfloor (n+1)/i \rfloor + 1}\right) \mathbf{e}_{i(j-1)+n+1-i\lfloor (n+1)/i \rfloor + 1}$$

et

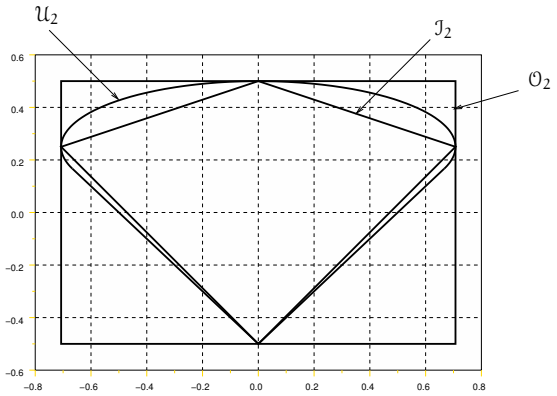
$$w^i = \sum_{j=1}^{\lfloor (n+1)/i \rfloor + 1} \sin\left(\frac{j\pi}{\lfloor (n+1)/i \rfloor + 2}\right) \mathbf{e}_{i(j-1)+1},$$

et déduire l'encadrement polyédrique de $\mathcal{U}_n$ suivant

$$\mathcal{I}_n \subset \mathcal{U}_n \subset \mathcal{O}_n,$$

qui est illustré dans le cas $n = 2$ sur la figure II.2

FIG. II.2: Encadrement de $\mathcal{U}_n$ pour $n = 2$



II.3.2.4 Estimation du volume de $\mathcal{U}_n$

Compte tenu de l'inclusion $\mathcal{U}_n \subset \mathcal{O}_n$, on peut majorer le volume de $\mathcal{U}_n$ comme suit :

$$\texttt{volume}(\mathcal{U}_n) \leqslant 2^n \prod_{i=1}^{n} \lambda_{\max}(A^{(i)}).$$

Or pour

$$\left\lfloor \frac{n+1}{2} \right\rfloor + 1 \leqslant i \leqslant n \text{ , on a } \lambda_{\max}(A^i) = 1/2,$$

et par conséquent,

$$\texttt{volume}(\mathcal{U}_n) \leqslant 2^{\lfloor (n+1)/2 \rfloor} \prod_{i=1}^{\lfloor (n+1)/2 \rfloor} \lambda_{\max}(A^{(i)}).$$

Puisque la fonction $\cos$ est décroissante sur $[0, \pi]$, on en déduit que

$$\texttt{volume}(\mathcal{U}_n) \leqslant 2^{\lfloor (n+1)/2 \rfloor} \prod_{i=1}^{\lfloor (n+1)/2 \rfloor} \cos\left(\frac{\pi}{\lfloor (n+1)/i \rfloor + 2}.\right)$$

Cette formule semble d'un intérêt limité pour avoir une idée de l'estimation de $\texttt{volume}(\mathcal{U}_n)$. Cependant, si on prend le logarithme du membre de droite et que l'on divise par n, on observe rapidement une convergence numérique vers une constante $c \approx 0.27$, ce qui laisse à penser que

$$\texttt{volume}(\mathcal{U}_n) \in \mathcal{O}(e^{cn}).$$

II.3.2.5 Positivité du produit scalaire sur $\mathcal{C}_{n+1}$

De manière classique, si l'on considère deux éléments de $\mathbb{R}^n_+$, et que l'on calcule leur produit scalaire, on obtient un réel positif. On observe le même comportement avec un couple de $\mathcal{S}_n^+(\mathbb{R})$ ou de $\mathcal{L}_n(\mathbb{R})$. Cela revient en fait à n'avoir que des angles aigus entre deux vecteurs quelconques de ces cônes. Le cône $\mathcal{C}_{n+1}$ jouit lui aussi de cette propriété.

Proposition II.8: *Le cône $\mathcal{C}_{n+1}$ est* aigu, i.e.

$$\forall (x, y) \in \mathcal{C}_{n+1}^2 \quad \langle x, y \rangle \geqslant 0.$$

Démonstration. Soient $(x, y) \in \mathcal{C}_{n+1}^2$; il existe $(z, t) \in (\mathbb{R}^{n+1})^2$ tels que $x = \texttt{corr}_a(z, z)$ et $y = \texttt{corr}_a(t, t)$; alors posons

$$\tilde{z} = \begin{pmatrix} z \\ 0_n \end{pmatrix}, \tilde{t} = \begin{pmatrix} t \\ 0_n \end{pmatrix} \text{ et } \tilde{x} = \texttt{corr}_c(\tilde{z}, \tilde{z}), \tilde{y} = \texttt{corr}_c(\tilde{t}, \tilde{t}).$$

En utilisant le théorème de Parseval-Plancherel, on a

$$\langle \tilde{x}, \tilde{y} \rangle = \sum_{i=0}^{2n} \widetilde{x}_i \widetilde{y}_i = \frac{1}{2n+1} \sum_{i=0}^{2n} \widetilde{X}_i \widetilde{Y}_i,$$

et en utilisant le fait que

$$\widetilde{X}_i = (\bar{Z} \circ Z)_i = \bar{Z}_i Z_i = |Z_i|^2,$$

qui est aussi vrai pour $\widetilde{Y}$, on en déduit $\langle \tilde{x}, \tilde{y} \rangle \geqslant 0$. Si on décompose ce produit scalaire, on obtient

$$\begin{aligned}
\langle \tilde{x}, \tilde{y} \rangle &= \tilde{x}_0 \tilde{y}_0 + \sum_{i=1}^{n} \tilde{x}_i \tilde{y}_i + \sum_{i=n+1}^{2n} \tilde{x}_i \tilde{y}_i \\
&= x_0 y_0 + \sum_{i=1}^{n} x_i y_i + \sum_{i=n+1}^{2n} \mathtt{corr}_c(\tilde{y}, \tilde{y})_i \mathtt{corr}_c(\tilde{t}, \tilde{t})_i \\
&= x_0 y_0 + \sum_{i=1}^{n} x_i y_i + \sum_{i=n+1}^{2n} \mathtt{corr}_a(y, y)_{2n+1-i} \mathtt{corr}_a(t, t)_{2n+1-i} \\
&= x_0 y_0 + 2 \sum_{i=1}^{n} x_i y_i,
\end{aligned}$$

où l'on a utilisé le Lemme d'ajout de zéros du premier chapitre. Posons $A = x_0 y_0 = \|t\|^2 \|z\|^2 \geqslant 0$ et $B = \sum_{i=1}^{n} x_i y_i$, alors

$$A + 2B \geqslant 0.$$

On en déduit

$$\langle x, y \rangle = A + B \geqslant A - A/2 = A/2 \geqslant 0.$$

□

On verra dans la suite que ceci a une incidence, sur la convexité dans certains cas d'une fonction coût intervenant dans un algorithme de projection sur $\mathcal{C}_{n+1}$.

II.3.2.6 Imbrications successives

Il existe aussi une relation "gigogne" qui lie $\mathcal{C}_n$ et $\mathcal{C}_{n+1}$, donnée par la

Proposition II.9: *Soit $\mathfrak{I} : \mathbb{R}^n \to \mathbb{R}^{n+1}$ l'injection linéaire définie par $\mathfrak{I}(x) = (x, 0)$; alors*

$$\mathfrak{I}(\mathcal{C}_n) = \mathcal{C}_{n+1} \cap \{x_n = 0\}.$$

Démonstration. Pour démontrer l'inclusion ($\subset$), il suffit de prendre $y_n = 0$ dans les équations de la Définition II.1 de $\mathcal{C}_{n+1}$ et vérifier que cela correspond à un élément de $\mathcal{C}_{n+1}$ avec la dernière coordonnée nulle. Démontrons l'autre inclusion : si $x \in \mathcal{C}_{n+1} \cap \{x_n = 0\}$ alors il existe $(y_0, \cdots, y_n) \in \mathbb{R}^{n+1}$ tels que

$$x_k = \sum_{i=0}^{n-k} y_i y_{i+k} \text{ pour } k = 0, \cdots, n,$$

et pour $k = n$, $x_n = y_0 y_n = 0$. Ceci implique que $y_0 = 0$ ou $y_n = 0$. Dans le second cas, on voit que $x = \mathfrak{I}(\mathcal{A}(\check{y}\check{y}^\top))$ avec $\check{y}_i = y_i$ pour $i \in \{0, \cdots, n-1\}$, et dans le premier $x = \mathfrak{I}(\mathcal{A}(\check{y}\check{y}^\top))$ où $\check{y}_i = y_{i+1}$ pour $i \in \{0, \cdots, n-1\}$. □

II.3.2.7 Fonction d'appui et frontière de $\mathcal{C}_{n+1}$

On peut obtenir une explicitation de $\mathcal{C}_{n+1}$ à l'aide de la fonction d'appui d'un certain ensemble ; on rappelle à cet effet, que la fonction d'appui σ_E de l'ensemble $E \subset \mathbb{R}^n$ est définie comme

$$\begin{array}{rccl} \sigma_E : & \mathbb{R}^n & \to & \mathbb{R} \cap \{+\infty\} \\ & x & \mapsto & \sup\limits_{s \in E} \langle x, s \rangle, \end{array}$$

et qu'il existe une correspondance bi-univoque entre les ensembles convexes fermés et les fonctions d'appui associées ; ainsi

$$\sigma_E = \sigma_{\overline{\mathrm{conv}}(E)}.$$

On résume parfois cela, en disant de manière imagée que la fonction d'appui d'un ensemble ne "voit" que son enveloppe convexe. Considérons

$$g(x) = \max_{\omega \in [0,\pi]} \langle -v(\omega), x \rangle,$$

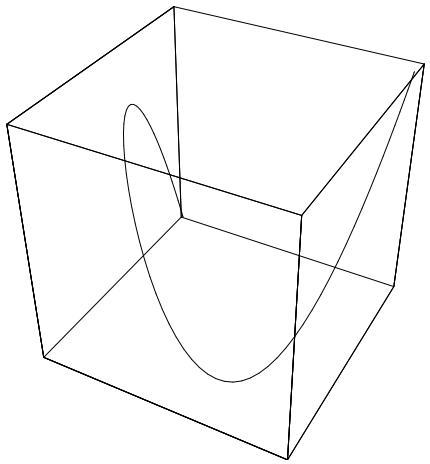
où $v(\omega) = (1, 2\cos\omega, \cdots, 2\cos n\omega) \in \mathbb{R}^{n+1}$; le maximum étant atteint car $[0, \pi]$ est compact dans $\mathbb{R}$ et $\omega \mapsto \langle v(\omega), x \rangle$ est continue. En définissant

$$S = \{-v(\omega) \mid \omega \in [0, \pi]\},$$

on peut voir alors que g est la fonction d'appui σ_S de l'ensemble S.

Comme on peut le voir sur la Figure II.3, l'ensemble S est une courbe paramétrée contenue dans l'hyperplan $x_0 = -1$.

FIG. II.3: la coupe de l'ensemble S par l'hyperplan $x_0 = -1$ dans $\mathbb{R}^4$



Si l'on était donc en mesure de connaître précisément l'enveloppe convexe de S, on pourrait alors facilement prendre en compte la contrainte "$x \in \mathcal{C}_{n+1}$" grâce à la proposition suivante.

Proposition II.10: *L'appartenance à $\mathcal{C}_{n+1}$ est déterminée par le signe de* g *:*

$$\begin{array}{c}(i)\ x \in \mathcal{C}_{n+1} \text{ si et seulement si } g(x) \leqslant 0\\ (ii)\ x \in \operatorname{int} \mathcal{C}_{n+1} \text{ si et seulement si } g(x) < 0\\ (iii)\ x \in \partial\mathcal{C}_{n+1} \text{ si et seulement si } g(x) = 0\\ (iv)\ x \notin \mathcal{C}_{n+1} \text{ si et seulement si } g(x) > 0.\end{array}$$

Démonstration. Le (i) est une conséquence de la Définition II.2, la négativité de $g(x)$ assure la positivité du polynôme trigonométrique pair $P(\omega) = \langle \nu(\omega), x\rangle$ ce qui est exactement la Définition II.2. (iv) n'est autre que (i) sous forme de négation. Quant aux assertions (ii) et (iii), elles proviennent de l'écriture sous la forme $\{x \mid g(x) \leqslant 0\}$ de $\mathcal{C}_{n+1}$, du fait que g soit convexe, et qu'il existe $\bar{x}$ tel que $g(\bar{x}) < 0$: en effet, sous ces hypothèses, la description à l'aide d'inégalités utilisant g de l'intérieur et de la frontière de l'ensemble $\{x \mid g(x) \leqslant 0\}$ sont des résultats classiques d'optimisation convexe (voir par exemple [31]). □

II.3.2.8 Cône tangent et cône normal à $\mathcal{C}_{n+1}$ en un point de $\mathcal{C}_{n+1}$

Pour décrire un ensemble convexe au voisinage d'un de ces points, il est possible d'en donner des approximations coniques : la première est le cône tangent $T(C, x)$ à un ensemble convexe C en $x \in C$, qui est défini comme

$$T(C, x) = \mathtt{cl}\{d \in \mathbb{R}^n \mid d = \alpha(y - x),\ y \in C,\ \alpha \in \mathbb{R}^+\},$$

où $\mathtt{cl}(\cdot)$ dénote la fermeture topologique. C'est l'équivalent en Analyse convexe du plan tangent pour les variétés différentielles. Dans notre cas, on peut en donner une écriture plus simple : comme

$$\mathcal{C}_{n+1} = \{x \in \mathbb{R}^{n+1} | - \langle \nu(\omega), x\rangle \leqslant 0,\ \forall \omega \in [0, \pi]\},$$

alors

$$T(\mathcal{C}_{n+1}, x) = \{d \in \mathbb{R}^{n+1} \mid \langle s, d\rangle \leqslant 0,\ \forall s \in J(x)\},$$

où

$$J(x) = \{-\nu(\omega) |\ \omega \in [0, \pi],\ \langle \nu(\omega), x\rangle = 0\},$$

ce qui conduit finalement à :

$$T(\mathcal{C}_{n+1}, x) = \{d \in \mathbb{R}^{n+1} \,|\langle \nu(\omega), d\rangle \geqslant 0, \text{ pour tous les } \omega \text{ tels que } \langle \nu(\omega), x\rangle = 0\}.$$

Un autre ensemble conique assez utile lié à un ensemble C convexe est le cône normal en x ; il est défini par

$$N(C, x) = \{d \in E \mid \forall y \in C\ ,\ \langle y - x, d\rangle \leqslant 0\} = [T(C, x)]^\circ,$$

où $(\cdot)^\circ$ désigne le polaire d'un ensemble (nous reverrons cette notion plus tard). Mais là encore, la définition de $\mathcal{C}_{n+1}$ nous permet de décrire simplement (cf. [31]) $N(C, x)$ par

$$N(\mathcal{C}_{n+1}, x) = \text{cone}\,\{-\nu(\omega) \mid \exists \omega \in [0, \pi]\,, \langle \nu(\omega), x\rangle = 0\}, \tag{II.14}$$

où $\text{cone}(\cdot)$ désigne l'enveloppe convexe conique, c'est-à-dire

$$\text{cone}(E) = \left\{ \sum_{i=1}^{k} \alpha_i x_i \,:\, x_i \in E\,,\ \alpha_i \geqslant 0,\ k \in \mathbb{N}^* \right\}.$$

La détermination du cône normal à $\mathcal{C}_{n+1}$ présente plusieurs applications intéressantes que nous allons voir maintenant.

II.3.2.9 La frontière de $\mathcal{C}_{n+1}$ n'est pas lisse

Le cône normal correspond à la généralisation dans le monde convexe, de l'orthogonal du plan tangent dans le cas des variétés différentiables. Par analogie, on comprendra aisement que si le cône tangent est un demi-espace (sa frontière étant le plan tangent) - cas que l'on qualifiera de *lisse* - alors le cône normal sera quant à lui réduit à une demi-droite orthogonale au plan tangent. Par conséquent, si la frontière de $\mathcal{C}_{n+1}$ était *lisse*, alors pour chaque point x de cette frontière, $N(\mathcal{C}_{n+1}, x)$ serait réduit à une demi-droite. Pour voir que ceci n'est pas vrai, considérons le point $x_0 = e_0 - \frac{1}{2}e_2$; alors

$$\langle x_0, \nu(\omega)\rangle = 1 - \cos(2\omega) \geqslant 0,$$

et $\langle x_0, \nu(\omega)\rangle = 0$ pour $\omega \in \{0, \pi\}$. Donc $g(x) = 0$ et x_0 est sur la frontière d'après la Proposition II.10. Alors

$$N(\mathcal{C}_{n+1}, x) = \text{cone}\{(-1, -2, -2, 0, \cdots, 0), (-1, 2, -2, 0, \cdots, 0)\},$$

qui n'est clairement pas une demi-droite. Par conséquent, le point x_0 est un point de non-lissité, pour tout $n \geqslant 2$!

II.3.2.10 La frontière de $\mathcal{C}_{n+1}$ n'est pas polyédrale

Un autre point pertinent pour l'étude de la frontière de $\mathcal{C}_{n+1}$ est le point $x_1 = e_0 + \frac{1}{2}e_2$, qui est en quelque sorte le point "opposé" à x_0 dans le compact $\mathcal{U}_n$. On peut remarquer d'abord que

$$\langle x_1, \nu(\omega)\rangle = 1 + \cos(2\omega) \geqslant 0,$$

expression qui s'annule pour $\omega = \frac{\pi}{2}$; donc $x_1 \in \partial\mathcal{C}_{n+1}$ et

$$N(C, x_1) = \text{cone}\{(-1, 0, 2, 0, \cdots, 0)\},$$

ce qui assure que x_1 est un point "régulier", car le cône normal étant en ce point une demi-droite.

Pour montrer que $\partial\mathcal{C}_{n+1}$ n'est pas polyédrale, nous allons montrer que localement autour de x_1, la courbure de $\partial\mathcal{C}_{n+1}$ n'est pas nulle. Pour cela, considérons la direction suivante $d = (0, 1, 0, \cdots, 0)$; on remarque ainsi que d est orthogonale à x_1. Alors, intéressons-nous à

$$g(x_1 + \varepsilon d) = \max_{\omega\in[0,\pi]} \{-1 - 2\varepsilon\cos\omega - \cos(2\omega)\},$$

où le maximum est atteint pour $\omega_0 = -\arccos(\varepsilon/2)$; par suite

$$\nabla g(x_1 + \varepsilon d) = \begin{pmatrix} -1 \\ \varepsilon \\ 2 - \varepsilon^2 \end{pmatrix},$$

ce qui implique

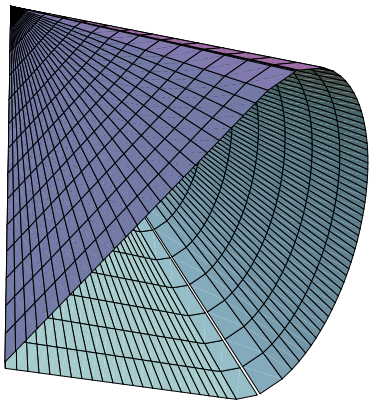
$$\frac{\nabla g(x_1 + \varepsilon d) - \nabla g(x_1)}{\varepsilon} = \begin{pmatrix} 0 \\ 1 \\ -\varepsilon \end{pmatrix}.$$

La variation de ∇g dans la direction d étant non-linéaire, la courbure n'est pas nulle localement autour de x_1 et, par conséquent, on est assuré de ne pas être sur une face de dimension supérieure à 1 ; donc $\partial\mathcal{C}_{n+1}$ n'est pas polyédrale.

On est donc en présence d'un cône dont la frontière ne présente pas de régularité uniforme comme $\mathcal{L}_n(\mathbb{R})$, mais possédant des parties lisses et des parties polyédrales à la fois, un peu comme $\mathcal{S}_n^+(\mathbb{R})$ qui présente des faces de dimension $\frac{p(p+1)}{2}$ et possède pourtant une infinité non-dénombrable de rayons extrêmaux.

Pour avoir une idée de la frontière de $\mathcal{C}_{n+1}$, on peut la tracer en dimension 3, comme sur la figure II.4.

FIG. II.4: le cône $\mathcal{C}_3$



II.3.2.11 Faces et facettes exposées de $\mathcal{C}_{n+1}$

Dire que $\mathcal{C}_{n+1}$ n'a pas une frontière lisse est certes une information en soi, mais peu précise. On peut en dire beaucoup plus, en s'intéressant aux faces de $\mathcal{C}_{n+1}$. Avant de les décrire, rappelons brièvement quelque définitions :

Définition II.3: *Soit* C *un sous-ensemble convexe de* $\mathbb{R}^n$. *On appelle* face exposée *tout ensemble* $F \subset C$ *tel qu'il existe* $a \in \mathbb{R}^n$ *et* $b \in \mathbb{R}$ *tels que*

$$\forall x \in C\ ,\ \langle a, x\rangle \leqslant b \quad \textit{et}\ F = C \cap \{x \in \mathbb{R}^n \mid \langle a, x\rangle = b\}.$$

La première condition signifie que $H_{a,b} = \{x \in \mathbb{R}^n | \langle a, x\rangle = b\}$ est un hyperplan d'appui de C, autrement dit : C est entièrement contenu dans le demi-espace $H^-_{a,b} = \{x \in \mathbb{R}^n \mid \langle a, x\rangle \leqslant b\}$ et $H_{a,b}$ est un sous-espace affine contenant l'enveloppe affine de la face F. Dans la suite, pour simplifier, nous appelerons faces[1] des faces exposées. On appelera facette toute face dont l'enveloppe affine est de dimension $n-1$. On va donc déterminer les faces de $\mathcal{C}_{n+1}$. On peut déjà remarquer que pour chercher les faces, on peut se limiter à caractériser les intersections de $\mathcal{C}_{n+1}$ avec des plans contenant l'origine - du type $H_s = \{x \in \mathbb{R}^{n+1} \mid \langle s, x\rangle = 0\}$ - grâce au Lemme suivant

Lemme II.5: *Soit* K *un cône et* $H_{a,b} = \{x \in \mathbb{R}^n \mid \langle a, x\rangle = b\}$ *un hyperplan d'appui de* K *tel que* $K \cap H_{a,b} \neq \emptyset$ *; alors* $b = 0$.

Démonstration. Soit $x \in K \cap H_{a,b}$. Alors $\langle a, x\rangle = b$ et $H^-_{a,b}$ étant un hyperplan d'appui de K, pour tout $z \in K$

$$\langle a, z\rangle \leqslant b.$$

Comme K est un cône, quel que soit $t > 0$, $tx \in K$ de sorte que

$$\langle a, tx\rangle \leqslant b \text{ pour tout } t > 0. \tag{II.15}$$

Si $b = 0$, il n'y a rien à démontrer ; si $b < 0$ alors pour $t = \frac{1}{2}$

$$\langle a, tx\rangle = \frac{1}{2}b > b$$

qui contredit (II.15) ; si $b > 0$, on prend $t = 2$ et on aboutit là encore à une contradiction. □

Les hyperplans à considérer sont donc les $H_s = H_{s,0}$, et une face F pourra ainsi être paramétrée par le vecteur normal s à l'hyperplan H_s ; la forme en est $F_s = \mathcal{C}_{n+1} \cap H_s$. Pour caractériser les faces exposées de $\mathcal{C}_{n+1}$, on a seulement réussi qu'à émettre la conjecture suivante :

Conjecture II.1: *Soit* F_s *une face exposée de* $\mathcal{C}_{n+1}$ *telle que* $\dim \mathtt{aff}\ F_s = k$ *; alors*

$$\mathtt{rg}\, \mathcal{A}^\star(s) \leqslant n + 1 - k$$

.

[1] La notion de face est plus complexe, nous nous limitons ici volontairement.

Remarque II.3: *Si l'on utilise seulement la linéarité de $\mathcal{A}$, on ne peut pas démontrer la conjecture car on peut avoir $(x_1x_1^\top, \cdots, x_px_p^\top)$ formant un système libre avec pourtant les $(x_1, \cdots, x_p)$ formant un système lié. Extraire un système libre de x_i doit donc sans doute s'appuyer sur des propriétés caractéristiques de $\mathcal{A}$.*

Si l'on utilise cependant la conjecture précédente dans le cas $k = n$, on peut essayer de trouver explicitement les directions exposant les facettes de $\mathcal{C}_{n+1}$. Elles correspondent au cas où $\mathcal{A}^\star(s)$ est une matrice de rang 1. Or il n'existe que deux solutions (à un facteur multiplicatif prés) d'une telle condition ; on peut le démontrer assez facilement ou consulter le livre [13] p. 252. On peut ensuite prouver que les candidats que l'on a trouvés sont effectivement des directions exposant des facettes.

Proposition II.11: *$\mathcal{C}_{n+1}$ a (au moins) deux faces de dimension n (i.e. des facettes) dont les normales sont dirigées par $s_1 = (1, 2, \cdots, 2)$ et $s_2 = (1, -2, \cdots, 2(-1)^n)$.*

Démonstration. Commençons par s_1. Pour démontrer que F_{s_1} est une face, considérons le système de points $(0, p_1, \cdots, p_n)$ avec

$$p_i = 2e_0 - e_i \text{ pour } i = 1, \cdots, n;$$

alors, suivant la Proposition II.10,

$$g(p_k) = -\min_{\omega\in[0,\pi]} \langle p_k, v(\omega)\rangle = -\min_{\omega\in[0,\pi]} 2(1 - \cos k\omega) = 0,$$

donc $p_k \in \partial\mathcal{C}_{n+1} \subset \mathcal{C}_{n+1}$. Or, on constate que $\langle p_k, s_1\rangle = 0$ pour tout k, et les p_k étant clairement indépendants, il existe $n + 1$ points affinement indépendants et contenus dans $F_{s_1} = \mathcal{C}_{n+1} \cap H_{s_1}$. Donc $\dim \mathtt{aff} F_{s_1} = n$. Pour s_2, on considère 0 et les éléments q_i, pour $i = 1, \cdots, n$, définis par

$$q_i = 2e_0 + (-1)^i e_i,$$

et l'on conclut de manière analogue. □

II.3.3 Représentation basée sur $\mathcal{S}^+_{n+1}(\mathbb{R})$

La Définition II.1 de $\mathcal{C}_{n+1}$ comme l'image par une application linéaire de l'ensemble P_1 des matrices dyadiques peut être satisfaisante sur le plan de l'intuition géométrique ; par contre, elle ne nous donne pas une paramétrisation utilisable facilement : si on considère le problème suivant où f est supposée convexe,

$$\begin{aligned} \min_{x\in\mathbb{R}^{n+1}} \quad & f(x) \\ & x \in \mathcal{C}_{n+1}, \end{aligned}$$

on peut le réécrire en

$$\begin{aligned} \min_{X\in\mathcal{S}_{n+1}(\mathbb{R})} \quad & f(\mathcal{A}(X)) \\ & X \succeq 0 \\ & \mathtt{rg}\, X = 1. \end{aligned}$$

Malheureusement, la dernière contrainte est loin d'être convexe, ce qui rend le problème difficile à résoudre. L'idée serait de pouvoir relaxer cette contrainte pour en faire un problème plus simple à résoudre. On va voir, dès maintenant, que cette démarche est légitime, car on peut donner une autre caractérisation de $\mathcal{C}_{n+1}$ basée sur le cône $\mathcal{S}^+_{n+1}(\mathbb{R})$ des matrices symétriques semi-définies positives.

II.3.3.1 Relaxation de la contrainte de rang

Dans la formulation (II.5) de $\mathcal{C}_{n+1}$, supposons que l'on prenne $\mathcal{S}^+_{n+1}(\mathbb{R})$ tout entier au lieu de P_1. On désigne cela en Optimisation par une "relaxation de contrainte", car on agrandit le domaine admissible. Alors, on peut montrer comme dans [1] que l'image par $\mathcal{A}$ de $\mathcal{S}^+_{n+1}(\mathbb{R})$ reste $\mathcal{C}_{n+1}$. En d'autres termes

$$\mathcal{C}_{n+1} = \mathcal{A}(P_1) = \mathcal{A}(\mathcal{S}^+_{n+1}(\mathbb{R})). \tag{II.16}$$

Démonstration. Montrons l'inclusion $\mathcal{A}(\mathcal{S}^+_{n+1}(\mathbb{R})) \subset \mathcal{C}_{n+1}$, l'autre étant directe. Nous présentions ici une variante de la démonstration de Alkire *et al.* : Soit $x = \mathcal{A}(Y)$ pour un certain $Y \in \mathcal{S}^+_{n+1}(\mathbb{R})$. Considérons alors

$$z_\omega = (1, e^{i\omega}, \cdots, e^{in\omega}) \in \mathbb{R}^{n+1};$$

si on s'intéresse à

$$\langle Yz_\omega, z^*_\omega \rangle = \sum_{0 \leqslant k,l \leqslant n} y_{kl} e^{i(k-l)\omega} = \sum_{p=-n}^{n} \left(\sum_{k-l=p} y_{kl} \right) e^{ip\omega},$$

qui est positif puisque $Y \in \mathcal{S}^+_{n+1}(\mathbb{R})$, alors en utilisant l'identité (II.3) et le fait que Y est symétrique

$$\sum_{k-l=-p} y_{kl} = \mathcal{A}(Y)_p = \sum_{l-k=p} y_{kl},$$

on déduit que

$$\langle Yz_\omega, z^*_\omega \rangle = \mathcal{A}(Y)_0 + 2\sum_{p=0}^{n} A(Y)_p \cos p\omega \geqslant 0$$

car $Y \succeq 0$. Ainsi (II.16) est démontrée. □

Il faut noter que ce type d'égalité est assez exceptionnel. On sait en effet que $\mathcal{S}^+_n(\mathbb{R}) = \mathrm{cone}(P_1)$ grâce au théorème de décomposition spectrale des matrices symétriques réelles. Pour un ensemble G et une application A, on a toujours

$$A(G) \subset A(\mathrm{cone}(G)),$$

mais ici on a

$$\mathcal{A}(\mathrm{cone}(P_1)) \subset \mathcal{A}(P_1),$$

ce qui nettement plus fort : **il suffit simplement de connaître l'image par $\mathcal{A}$ de tous les rayons extrémaux de $\mathcal{S}_n^+(\mathbb{R})$ pour connaître toute l'image de $\mathcal{S}_n^+(\mathbb{R})$ par $\mathcal{A}$.**

En posant $g(X) = f(\mathcal{A}(X))$, on peut donc utiliser la formulation

$$\begin{array}{ll} \min\limits_{X \in \mathcal{S}_{n+1}(\mathbb{R})} & g(X) \\ & X \succeq 0, \end{array}$$

qui ne présente qu'une contrainte du type Inégalités Linéaires Matricielles (LMI), type de problème pour lequel il existe des codes de calcul assez efficaces (jusqu'à $n \approx 500$). Par contre, il faut tempérer notre engouement pour cette formulation, car d'un ensemble à $(n+1)$ paramètres, on passe à une formulation LMI dans un espace de matrices à $(n+1)(n+2)/2$ paramètres, ce qui, en grande dimension, peut être prohibitif (prendre par exemple, $n+1 = 1000$, $X \in \mathbb{R}^{500500}$!). Pour travailler avec le même nombre de paramètres, on va s'intéresser à une autre approche, l'approche duale.

II.4 Approche via le cône polaire de $\mathcal{C}_{n+1}$

La dualité conique est en Analyse convexe la généralisation de la dualité linéaire où, *grosso modo*, on a remplacé des égalités par des inégalités. Pour comprendre le sens de l'affirmation précédente, considérons la décomposition orthogonale dans un espace de dimension finie E, relativement à un produit scalaire $\langle \cdot, \cdot \rangle_E$; on sait que pour tout sous-espace vectoriel $F \subset E$, on a

$$E = F \oplus F^\perp,$$

où $F^\perp$ désigne l'orthogonal de F pour $\langle \cdot, \cdot \rangle_E$, avec

$$\forall x \in F\ ,\ \forall y \in F^\perp\ ,\ \langle x, y \rangle_E = 0.$$

De plus, pour calculer effectivement cette décomposition, on utilise l'opérateur p_F de projection orthogonale sur F, défini par

$$p_F(x) = \arg\min_{y \in F} \|x - y\|^2,$$

alors

$$\forall x \in E\ ,\ x = \underbrace{p_F(x)}_{\in F} + \underbrace{x - p_F(x)}_{\in F^\perp}\ ,\ p_{F^\perp}(x) = x - p_F(x).$$

L'analogue de cette décomposition en Analyse convexe est la décomposition dite de Moreau : supposons que l'on remplace $\langle x, y \rangle_E = 0$ par $\langle x, y \rangle_E \leqslant 0$, et que F soit un cône convexe fermé ; désignons par F° l'ensemble

$$F^\circ = \{x \in E \mid \forall y \in F\ ,\ \langle x, y \rangle_E \leqslant 0\},$$

appelé cône polaire de F(et déjà mentionné précédemment). Alors la décomposition de Moreau précise :

Théorème II.3 (Moreau): *Soit* F *un cône convexe fermé et* $(x, x_F, x_{F^\circ}) \in E^3$ *; alors les assertions suivantes sont équivalentes :*

$$\begin{gathered}(i)\ x = x_F + x_{F^\circ},\ x_F \in F,\ x_{F^\circ} \in F^\circ,\ \textit{et}\ \langle x_F, x_{F^\circ}\rangle_E = 0,\\ (ii)\ x_F = p_F(x)\ \textit{et}\ x_{F^\circ} = p_{F^\circ}(x).\end{gathered}$$

Pour une démonstration de ce résultat et de toutes les bonnes propriétés de p_F, on pourra consulter [31] par exemple.

L'ensemble F°, polaire de F, est un cône qui possède plusieurs propriétés intéressantes, et qui permet notamment de dualiser les contraintes de type conique. Une des conclusions pertinentes de cette décomposition provient du fait que *lorsque l'on cherche la projection* x_F *d'un élément* x *sur un cône convexe fermé* F*, il est parfois plus simple de calculer sa projection* x_{F° *sur* F°*, et d'obtenir simplement* x_F *par* $x - x_{F^\circ}$. Avant de déterminer le cône polaire de $\mathcal{C}_{n+1}$, il nous faut définir un sous-espace vectoriel de $\mathcal{S}_{n+1}(\mathbb{R})$ qui interviendra dans ce calcul.

II.4.1 Opérateur adjoint et espace des matrices Toeplitz symétriques

Désignons par $\mathcal{A}^\star$ l'application linéaire suivante :

$$\begin{array}{rccc}\mathcal{A}^\star : & \mathbb{R}^{n+1} & \to & \mathcal{S}_{n+1}(\mathbb{R})\\ & (x_0, \cdots, x_n) & \mapsto & \displaystyle\sum_{i=0}^{n} x_i A^{(i)},\end{array}$$

qui à pour écriture matricielle

$$\mathcal{A}^\star(x) = \frac{1}{2}\begin{pmatrix} 2x_0 & x_1 & \cdots & x_n\\ x_1 & 2x_0 & \ddots & \vdots\\ \vdots & \ddots & \ddots & x_1\\ x_n & \cdots & x_1 & 2x_0\end{pmatrix}.$$

L'opérateur $\mathcal{A}^\star$ est bien entendu construit pour être l'adjoint par rapport à $\langle\cdot,\cdot\rangle$ et $\langle\langle\cdot,\cdot\rangle\rangle$ de $\mathcal{A}$; la notation choisie est donc justifiée par la proposition suivante

Proposition II.12: *L'opérateur* $\mathcal{A}^\star$ *est l'opérateur adjoint de* $\mathcal{A}$*, i.e.*

$$\forall (X, y) \in \mathcal{S}_{n+1}(\mathbb{R}) \times \mathbb{R}^{n+1},\quad \langle\langle X, \mathcal{A}^\star(y)\rangle\rangle = \langle \mathcal{A}(X), y\rangle.$$

Démonstration.

$\forall M \in \mathcal{S}_{n+1}(\mathbb{R})$, $\forall x \in \mathbb{R}^{n+1}$,

$$\langle\langle M, \mathcal{A}(x)\rangle\rangle = \langle\langle M, \sum_{k=0}^{n} x_k A^{(k)}\rangle\rangle = \sum_{k=0}^{n} x_k \underbrace{\langle\langle M, A^{(k)}\rangle\rangle}_{\mathcal{A}_k(M)} = \langle \mathcal{A}(M), x\rangle.$$

□

Il est facile de voir que l'opérateur $\mathcal{A}^\star$ est injectif, et par conséquent il réalise un isomorphisme de $\mathbb{R}^{n+1}$ sur son image $\mathcal{A}^\star(\mathbb{R}^{n+1})$, ensemble que précise mieux la

Définition II.4: *On appelle ensemble des matrices Toeplitz symétriques, noté $\mathcal{T}_{n+1}(\mathbb{R})$, le sous-espace vectoriel de $\mathcal{S}_{n+1}(\mathbb{R})$ défini par*

$$\mathcal{A}^\star(\mathbb{R}^{n+1}) = \mathtt{Vect}\,\{A^{(0)}, A^{(1)}, \cdots, A^{(n)}\},$$

ou de manière équivalente

$$\mathcal{T}_{n+1}(\mathbb{R}) = \{M \in \mathcal{S}_{n+1}(\mathbb{R}) \mid \exists x \in \mathbb{R}^{n+1} \text{ tel que } M_{ij} = x_{|i-j|} \text{ pour tout } i, j = 1, \cdots, n+1\}.$$

$\mathcal{T}_{n+1}(\mathbb{R})$ est, de fait, l'ensemble des matrices symétriques constantes le long de leurs diagonales. Ces matrices interviennent dans plusieurs champs des mathématiques et ont fait l'objet de nombreuses études : concernant leur inversion, le calcul rapide de leurs valeurs propres, etc. Une référence classique sur les matrices Toeplitz souvent citée est [26], qui fournit des résultats asymptotiques sur le spectre. Pour les cas en dimension finie, on pourra consulter par exemple [9, 16]. On remarquera que $\mathcal{T}_{n+1}(\mathbb{R})$ est de dimension $n+1$ et que $A^{(0)}, A^{(1)}, \cdots, A^{(n)}$ en constitue une base, puisque l'on voit facilement que c'est une famille libre de $n+1$ vecteurs. L'isomorphisme $\mathcal{A}^\star$ entre $\mathcal{T}_{n+1}(\mathbb{R})$ et $\mathbb{R}^{n+1}$ permettra ainsi de faire une identification - un tant soit peu abusive - lors de la définition de $\mathcal{C}^\circ_{n+1}$.

II.4.2 Cône polaire de $\mathcal{C}_{n+1}$

II.4.2.1 Formulation par contraintes

Proposition II.13: *Le cône polaire de $\mathcal{C}_{n+1}$ est, à un isomorphisme près, l'ensemble des matrices Toeplitz semi-définies négatives, c'est-à-dire*

$$\mathcal{C}^\circ_{n+1} = \{x \in \mathbb{R}^{n+1} \mid \mathcal{A}^\star(x) \preceq 0\}.$$

Cette démonstration est déjà présente dans [1, 19, 33] sous diverses formes ; nous la détaillons quand même ici en raison de sa simplicité.

Démonstration. On cherche $s \in \mathbb{R}^{n+1}$ tel que, pour tout $x \in \mathcal{C}_{n+1}$, $\langle s, x\rangle \leqslant 0$; alors :

$$\forall y \in \mathbb{R}^{n+1} : \ \langle s, \mathcal{A}(yy^\top)\rangle \leqslant 0 \Leftrightarrow \langle\langle \mathcal{A}^\star(s), yy^\top\rangle\rangle \leqslant 0 \Leftrightarrow \langle \mathcal{A}^\star(s)y, y\rangle \leqslant 0.$$

Autrement dit, $\mathcal{A}^\star(s)$ est une matrice symétrique semi-définie négative. □

Cette formulation permet de faire une passerelle avec la théorie des moments et le genre de méthodes qui sont utilisées dans le cas multivariable comme dans [34]. La formulation $\mathcal{A}^\star(x) \preceq 0$ nous dit que la forme Toeplitz

$$\sum_{k=0}^{n}\sum_{l=0}^{n} y_{k-l} x_k \bar{x}_l$$

est non-négative en ayant posé au préalable

$$y_k = \begin{cases} y_0 = x_0 \text{ si } k = 0 \\ x_k/2 \text{ si } 1 \leqslant |k| \leqslant n \\ 0 \text{ sinon.} \end{cases}$$

Ce qui est équivalent d'après [33] (p. 65) à l'existence d'une mesure **positive** σ sur $[0, 2\pi]$ telle que

$$-\frac{1}{2}x_{|k|} = \int_0^{2\pi} e^{-ikt} d\sigma(t) \text{ pour } k \in \mathbb{Z}$$

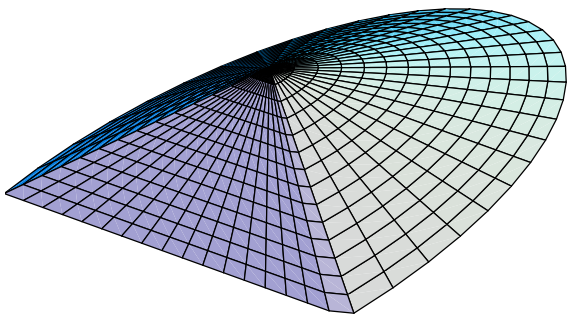
Vu que x_k est réel on obtient

$$x_k = \begin{cases} x_0 = -\int_{[0,2\pi]} d\sigma \\ \int_0^{2\pi} (-2\cos kt) d\sigma(t) \text{ pour } k = 1, \cdots, n, \end{cases}$$

ce qui, précisément, revient à écrire que x est dans l'enveloppe conique convexe de l'ensemble $\{-\nu(\omega) | \omega \in [0, \pi]\}$. D'après [33] (p. 15), ceci n'est autre que la formulation par générateurs de $\mathcal{C}_{n+1}^\circ$, que nous allons voir juste après.

On peut déjà noter une différence de taille avec les cônes convexes classiques de l'Analyse convexe $((\mathbb{R}_+)^n, \mathcal{L}_n(\mathbb{R}), \mathcal{S}_n^+(\mathbb{R}))$. Le cône $\mathcal{C}_{n+1}$ n'est pas auto-polaire (au signe près), c'est-à-dire $\mathcal{C}_{n+1}^\circ \neq -\mathcal{C}_{n+1}$, comme c'est le cas des trois cônes précités. C'est pour cela que l'approche des problèmes concernant $\mathcal{C}_{n+1}$ via la considération de son polaire, présente une complexité moindre (n paramètres contre $n(n+1)/2$), ce dont nous allons tirer profit par la suite.

FIG. II.5: Le cône polaire $\mathcal{C}_3^\circ$



Si l'on utilise la formulation $\mathcal{A}^\star(x) \preceq 0$, on peut tirer une nouvelle expression pour le cône normal à $\mathcal{C}_{n+1}$ en un de ses points x. En effet, on sait (cf. par exemple [31]) que pour un cône convexe fermé K

$$N(K, x) = \{y \in K^\circ \mid \langle x, y \rangle = 0\}.$$

On en déduit donc ici :

$$N(\mathcal{C}_{n+1}, x) = \{y \in \mathbb{R}^{n+1} \mid \mathcal{A}^\star(y) \preceq 0, \langle x, y\rangle = 0\}.$$

II.4.2.2 Conditions d'optimalité

Une des principales utilisations du cône normal à un ensemble convexe est l'écriture des conditions d'optimalité du premier ordre pour un problème d'optimisation dont cet ensemble est l'ensemble-contrainte. Ainsi, si l'on veut minimiser une fonction convexe f sur un ensemble C convexe, alors $\bar{x}$ sera une solution du problème, i.e.

$$\forall x \in C\ ,\ f(x) \geqslant f(\bar{x}), \text{ si et seulement si } \bar{x} \in C \text{ et } 0 \in \partial f(\bar{x}) + N(C, \bar{x}).$$

Ici $\partial f(\bar{x})$ désigne le sous-différentiel de f en $\bar{x}$:

$$\partial f(\bar{x}) = \{s \in E \mid \forall y \in E\ , \langle s, y - \bar{x}\rangle + f(\bar{x}) \leqslant f(y)\}.$$

Pour plus de détails concernant le sous-différentiel on se reportera à [31], ou autre livre d'Analyse convexe moderne. Ainsi, par exemple, dans le cas du problème de projection d'un point c sur C, i.e. avec $f(x) = \frac{1}{2}\|c - x\|_2^2$, on obtient comme condition :

$$\bar{x} \in \mathcal{C}_{n+1} \text{ et } 0 = \bar{x} - c + s, \text{ où } \mathcal{A}^\star(s) \preceq 0 \text{ et } \langle s, \bar{x}\rangle = 0,$$

ce qui se résume en

$$(\mathrm{Opt}) \begin{cases} \bar{x} \in \mathcal{C}_{n+1} \\ \mathcal{A}^\star(c - \bar{x}) \preceq 0 \ \ (c - \bar{x} \in \mathcal{C}_{n+1}^\circ) \\ \langle \bar{x}, c - \bar{x}\rangle = 0. \end{cases} \tag{II.17}$$

On retrouve comme cela les trois conditions du Théorème II.3 dans le cas où $F = \mathcal{C}_{n+1}$.

II.4.2.3 Expression (ou formulation) à l'aide de générateurs

On a donné une formulation du cône polaire sous la forme d'une inégalité linéaire matricielle que doit vérifier un élément de ce cône $\mathcal{C}_{n+1}^\circ$. On peut aussi donner une formulation par ses générateurs coniques, c'est-à-dire exprimer tout élément de ce cône comme une combinaison linéaire à coefficients positifs d'un certain nombre d'éléments. On utilise pour cela l'ensemble S défini précédemment :

Proposition II.14: *Soit* $S = \{-v(\omega) : \omega \in [0, \pi]\}$*; alors*

$$\mathcal{C}_{n+1}^\circ = \mathrm{cone}(S).$$

Cette formulation exprime $\mathcal{C}_{n+1}^\circ$ comme l'enveloppe convexe conique d'une courbe paramétrée de $\mathbb{R}^{n+1}$. C'est cette approche qui est choisie comme point de départ dans l'ouvrage classique de Krein et Nudelman [33]. On pourra noter qu'on retrouve le cas particulier du résultat (II.14) pour $x = 0$ (le cône normal en 0 d'un ensemble est son cône polaire).

Démonstration. On a

$$\mathcal{C}_{n+1} = \{x \in \mathbb{R}^{n+1} | \langle x, -\nu(\omega)\rangle \leqslant 0 \text{ pour } \omega \in [0,\pi]\}.$$

D'après le lemme de Farkas généralisé (voir [31]), si

$$\langle y, x\rangle \leqslant 0 \text{ pour tout } x \in \mathcal{C}_{n+1},$$

alors

$$\begin{pmatrix} y \\ 0 \end{pmatrix} \in \text{cone}\left(\left\{\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right\} \cup \left\{\begin{pmatrix} -\nu(\omega) \\ 0 \end{pmatrix}\right\}_{\omega\in[0,\pi]}\right);$$

en déduit donc que

$$\mathcal{C}_{n+1}^{\circ} \subset \text{cone}(S).$$

Réciproquement, si $y \in \text{cone}(S)$, il existe alors $(\alpha_0, \cdots, \alpha_n) \in (\mathbb{R}_+)^{n+1}$ tels que

$$y = -\sum_{k=0}^{n} \alpha_k \nu(\omega_k) \text{ avec } \omega_k \in [0,\pi];$$

alors, pour $x \in \mathcal{C}_{n+1}$,

$$\langle y, x\rangle = -\sum_{k=0}^{n} \alpha_k \langle \nu(\omega_k), x\rangle \leqslant 0,$$

ce qui implique $\text{cone}(S) \subset \mathcal{C}_{n+1}^{\circ}$. □

II.4.3 Isomorphisme vecteurs - matrices Toeplitz

L'identification possible entre $\mathbb{R}^{n+1}$ et $\mathcal{T}_{n+1}(\mathbb{R})$ à l'aide de $\mathcal{A}^\star$ semble *a priori* idéale. Pourtant, pour nos applications, elle présente un défaut majeur : elle ne transfère pas la structure euclidienne de $(\mathbb{R}^{n+1}, \langle\cdot,\cdot\rangle)$ sur $(\mathcal{T}_{n+1}(\mathbb{R}), \langle\langle\cdot,\cdot\rangle\rangle)$.

II.4.3.1 Conservation du produit scalaire

Si l'isomorphisme $\mathcal{A}^\star$ ne transporte pas le produit scalaire, il a néanmoins le mérite de montrer clairement que les $A^{(i)}$ constituent une base de $\mathcal{T}_{n+1}(\mathbb{R})$. Mais la famille des $A^{(i)}$ a une propriété plus intéressante, elle est orthogonale pour le produit scalaire de Frobenius $\langle\langle\cdot,\cdot\rangle\rangle$ dans $\mathcal{S}_{n+1}(\mathbb{R})$, ce qui va nous permettre, modulo une normalisation, de transférer la structure euclidienne de $\mathbb{R}^{n+1}$. On peut donc démontrer la

Proposition II.15: *Les $\{A^{(i)}\}_{i=0}^{n}$ forment une base orthogonale de $\mathcal{T}_{n+1}(\mathbb{R})$; plus spécifiquement*

$$\langle\langle A^{(i)}, A^{(j)}\rangle\rangle = [i=j](n-i+1).$$

Démonstration. Le produit scalaire de Frobenius n'est rien d'autre qu'un produit scalaire sur $\mathbb{R}^{(n+1)^2}$; or pour $i \neq j$, $A^{(i)}$ et $A^{(j)}$ n'ont aucun coefficient non nul en commun, donc $\langle\langle A^{(i)}, A^{(j)}\rangle\rangle = 0$. Le cas $i = j$ résulte d'un simple calcul. □

Puisque l'on connaît une base *orthogonale* de $\mathcal{T}_{n+1}(\mathbb{R})$, on peut facilement en faire une base *orthonormale* : pour cela, considérons l'endomorphisme $\mathcal{N}$ défini comme suit :

$$\begin{array}{rccc} \mathcal{N}: & \mathcal{T}_{n+1}(\mathbb{R}) & \to & \mathcal{T}_{n+1}(\mathbb{R}) \\ & \sum_{i=0}^{n} x_i A^{(i)} & \mapsto & \sum_{i=0}^{n} \frac{x_i}{\sqrt{n-i+1}} A^{(i)}. \end{array}$$

Alors, on a la proposition suivante qui établit une isométrie linéaire entre $\mathbb{R}^{n+1}$ et $\mathcal{T}_{n+1}(\mathbb{R})$:

Proposition II.16: *L'application $\varphi = \mathcal{N} \circ \mathcal{A}^\star$ est une isométrie de $(\mathbb{R}^{n+1}, \langle\cdot,\cdot\rangle)$ sur $(\mathcal{T}_{n+1}(\mathbb{R}), \langle\langle\cdot,\cdot\rangle\rangle)$, i.e. :*

$$\textit{pour tout } x, y \in \mathbb{R}^{n+1} \ \langle\langle \varphi(x), \varphi(y)\rangle\rangle = \langle x, y\rangle.$$

Démonstration. L'application φ est bijective, car $\mathcal{A}^\star$ est un isomorphisme et $\mathcal{N}$ est un opérateur diagonal sans valeur propre nulle ; de plus, φ conserve le produit scalaire, car elle correspond à la normalisation d'une base orthogonale :

$$\langle\langle \varphi(x), \varphi(y)\rangle\rangle = \langle\langle \sum_{k=0}^{n} \frac{x_k}{\|A^{(k)}\|} A^{(k)}, \sum_{l=0}^{n} \frac{y_l}{\|A^{(l)}\|} A^{(l)}\rangle\rangle = \sum_{0\leqslant k,l\leqslant n} x_k y_l [k=l] = \langle x, y\rangle.$$

□

Remarque II.4: *On a défini précédemment l'application $\mathcal{N}$ de $\mathcal{T}_{n+1}(\mathbb{R})$ dans lui-même. Pourtant, pour des matrices symétriques qui ne sont pas Toeplitz, il est encore possible de diviser chaque diagonale par un coefficient constant selon la diagonale. On peut par conséquent prolonger $\mathcal{N}$ à $\mathcal{S}_{n+1}(\mathbb{R})$ sans aucun problème, et c'est pourquoi, dans la suite, ce prolongement sera implicitement sous-entendu lorsque qu'on appliquera $\mathcal{N}$ à des matrices symétriques qui ne sont pas Toeplitz.*

II.4.3.2 Application à la projection sur le cône polaire $\mathcal{C}_{n+1}^\circ$

Considérons le problème de projection suivant :

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^{n+1}} & \|x - r\|^2 \\ & x \in \mathcal{C}_{n+1}^\circ. \end{array} \tag{II.18}$$

Comme $\|\varphi(u)\|_F^2 = \|u\|^2$ pour tout $u \in \mathbb{R}^{n+1}$, on reformule ce problème comme suit :

$$\begin{array}{ll} \min\limits_{X\in\mathcal{S}_{n+1}(\mathbb{R})} & \|X - R\|_F^2 \\ & \varphi^{-1}(X) \in \mathcal{C}_{n+1}^\circ, \end{array}$$

où $R = \varphi(r)$ et $\|\cdot\|_F$ est la norme de Frobenius (derivée de $\langle\langle\cdot,\cdot\rangle\rangle$).

Or $\varphi^{-1} = [\mathcal{A}^\star]^{-1} \circ \mathcal{N}^{-1}$, donc

$$\varphi^{-1}(X) \in \mathcal{C}_{n+1}^\circ \Leftrightarrow \mathcal{N}^{-1}(X) \in \mathcal{A}^\star(\mathcal{C}_{n+1}^\circ).$$

On a vu précédemment (Proposition II.13) que

$$\mathcal{A}^\star(\mathcal{C}_{n+1}^\circ) = \mathcal{T}_{n+1}(\mathbb{R}) \cap \mathcal{S}_{n+1}^-(\mathbb{R}),$$

et, comme $\mathcal{N}$ est injective,

$$\mathcal{N}(\mathcal{T}_{n+1}(\mathbb{R}) \cap \mathcal{S}_{n+1}^-(\mathbb{R})) = \mathcal{N}(\mathcal{T}_{n+1}(\mathbb{R})) \cap \mathcal{N}(\mathcal{S}_{n+1}^-(\mathbb{R})),$$

de sorte que l'on aboutit au problème suivant :

$$\begin{array}{cc} \min\limits_{X \in \mathcal{S}_{n+1}(\mathbb{R})} & \|X - R\|_F^2 \\ & X \in \mathcal{T}_{n+1}(\mathbb{R}) \cap \mathcal{N}(\mathcal{S}_{n+1}^-(\mathbb{R})). \end{array} \tag{II.19}$$

Pour calculer la projection d'un point sur $\mathcal{C}_{n+1}^\circ$, il faut donc calculer *la projection de ce point sur l'intersection d'un sous-espace* $(\mathcal{T}_{n+1}(\mathbb{R}))$ *et d'un cône convexe* $(\mathcal{N}(\mathcal{S}_{n+1}^-(\mathbb{R}))$, qui est convexe en tant qu'image par une application linéaire d'un cône convexe). Si l'on sait déterminer (facilement) les projections d'un point sur $\mathcal{T}_{n+1}(\mathbb{R})$ et $\mathcal{N}(\mathcal{S}_{n+1}^-(\mathbb{R}))$ séparément, il est possible d'utiliser des algorithmes comme celui de Boyle-Dysktra, pour calculer la projection sur l'intersection $\mathcal{T}_{n+1}(\mathbb{R}) \cap \mathcal{N}(\mathcal{S}_{n+1}^-(\mathbb{R}))$.

La projection sur le cône $\mathcal{S}_{n+1}^-(\mathbb{R})$ est explicite et bien connue (c.f.[35]) , elle utilise la décomposition de Moreau suivant les cônes mutuellement polaires $\mathcal{S}_{n+1}^+(\mathbb{R})$ et $\mathcal{S}_{n+1}^-(\mathbb{R})$; seulement, ici, l'application $\mathcal{N}$ perturbe la méthodologie. Pour disposer d'une procédure analogue, il faudrait connaître la décomposition

$$\mathcal{S}_{n+1}(\mathbb{R}) = \mathcal{N}(\mathcal{S}_{n+1}^-(\mathbb{R})) + \mathcal{N}^{-1}(\mathcal{S}_{n+1}^+(\mathbb{R})),$$

car on démontre facilement que $\left[\mathcal{N}(\mathcal{S}_{n+1}^-(\mathbb{R}))\right]^\circ = \mathcal{N}^{-1}(\mathcal{S}_{n+1}^+(\mathbb{R}))$ du fait que $\mathcal{N}$ est auto-adjoint et grâce

Lemme II.6: *Soit* $A \in L(\mathbb{R}^n)$ *inversible et* $K \subset \mathbb{R}^n$ *non vide ; alors*

$$\left[A^{-1}(K)\right]^\circ = A^\star(K^\circ).$$

Malheureusement, il n'y a, à notre connaissance, aucune formule analytique de décomposition basée sur les valeurs propres comme dans le cas de $\mathcal{S}_{n+1}^+(\mathbb{R})$ et $\mathcal{S}_{n+1}^-(\mathbb{R})$. Le seul cas qui se généralise directement est le cas où $\mathcal{N}$ est une involution ou un endomorphisme orthogonal.

II.4.4 Généralisations unidimensionelles

Comme nous le verrons au chapitre 4, il existe des généralisations possibles pour des signaux multidimensionels du cône $\mathcal{C}_{n+1}$. Cependant, avant de considérer de telles extensions, il est nécessaire de s'intéresser à des extensions simples unidimensionnelles et particulièrement utiles pour le traitement du signal. Ainsi dans le chapitre 1, nous avons parlé de problème de filtrage ou il apparaissait nécessaire de pouvoir traiter une contrainte du type

$$x_0 + 2\sum_{k=0}^{n} x_k \cos k\omega \geqslant 0 \text{ pour } \omega \in [\alpha, \beta]. \tag{II.20}$$

L'article [1] traite ainsi cette question en utilisant les polynômes de Chebyshev combinés à deux changements de variable affine : ainsi d'un polynôme trigonométrique pair positif, on passe à un polynôme positif sur $[-1,-1]$, puis positif sur $[\cos\alpha, \cos\beta]$. En désignant par $M(\alpha,\beta)$ la composition de ces deux transformations linéaires (changement de base dans l'espace des polynômes), la formulation semi-infinie (II.20) devient simplement

$$M(\alpha,\beta)x \in \mathcal{C}_{n+1}.$$

Cependant les auteurs de [1], ont souligné le mauvais conditionnement de $M(\alpha,\beta)$ qui apparaissait dans certains cas et qui rendait sa résolution numérique difficile. A l'aide de cette technique, en utilisant des produits cartésiens de cônes, on peut alors traiter des problèmes plus complexes du type

$$a_k \leqslant \langle x, \nu(\omega)\rangle \leqslant b_k \text{ pour } x \in [\alpha_k, \beta_k] \text{ avec } k = 1, \cdots, m.$$

On appelle ce type de contraintes, en Traitement du signal, des *contraintes de masque de spectre*. L'article très récent de Faybusovich [20] traite explicitement ce type de contrainte en introduisant un cône *ad hoc* qu'il contruit par récurrence.

L'autre extension possible est de ne pas se restreindre à des filtres à coefficients réels, mais plutôt complexes, en imposant toutefois que $x_{-k} = \overline{x_k}$ afin que l'on puisse encore parler de polynôme trigonométrique positif. L'ouvrage [33] traite évidemment ce cas, sans toutefois appliquer les résultats au traitement du signal.

Chapitre III

Résolution numérique de problèmes d'optimisation avec contraintes d'autocorrélation

Même si les problèmes impliquant $\mathcal{C}_{n+1}$ peuvent être de natures diverses (identification ou synthèse en Traitement du signal), leurs solutions se modélisent souvent de la même façon, et que l'on soit confronté à un coût linéaire ou un coût quadratique convexe, la question reste toujours la même : comment prend-on en compte de manière pratique une contrainte du type $x \in \mathcal{C}_{n+1}$? Les deux approches connues avant cette thèse étaient les méthodes de grille ou de programmation semi-infinie, et les méthodes de *points intérieurs* appelées aussi algorithmes de suivi de chemin. Les premières sont assez intuitives à mettre en œuvre - on remplace une infinité non dénombrable de contraintes par un nombre fini d'entre elles- mais conduisent souvent à un résultat sous-optimal. Les secondes sont beaucoup plus précises et convergent vers le minimum global, mais présentent des temps de calcul parfois prohibitif pour des grandes dimensions.

Dans ce chapitre nous allons présenter ces deux méthodes ainsi que deux autres approches possibles pour traiter plus spécifiquement les problèmes d'approximation faisant intervenir $\mathcal{C}_{n+1}$.

III.1 Optimisation semi-infinie

On qualifie de *problème d'optimisation semi-infinie* un problème ayant une infinité -souvent non-dénombrable - de contraintes. De nombreuses méthodes de résolution ad hoc ont vu le jour pour ces type de problèmes ; une bonne revue des différentes approches est l'article-revue de Hettich et Kortanek [29]. Une méthode très simpliste et intuitive pour traiter ce genre de problème est la suivante. Puisque $x \in \mathcal{C}_{n+1}$ s'écrit aussi

$$\langle x, \nu(\omega) \rangle = x_0 + 2\sum_{k=0}^{n} x_k \cos k\omega \geqslant 0 \quad \text{pour tout } \omega \in [0, \pi], \tag{III.1}$$

alors avec un certain degré d'approximation, en choisissant *une grille* de points $\{\omega_i\}_{i=1}^M \in [0,\pi]^M$ (le choix le plus souvent rencontré consiste à prendre une grille équirépartie $\omega_k = k\pi/M$), on discrétise alors la contrainte (III.1) en

$$\langle x, v(\omega_i)\rangle \geqslant 0 \text{ pour } i = 1, \cdots, M.$$

Si la fonction-objectif est par exemple linéaire, on peut résoudre ce genre de problème d'optimisation, avec un algorithme du simplexe ou un algorithme très efficace de points intérieurs. Évidemment la solution obtenue n'étant pas en général la solution optimale puisque l'on a agrandi le domaine en relâchant une infinité de contraintes et, de ce fait, la solution n'est pratiquement jamais *réalisable*. Si l'on s'arrête à ce niveau-là, on a une méthode un peu grossière mais qui dans certains cas donne des résultats suffisants dans les applications. C'est par exemple la méthode préconisée dans l'ouvrage [5] (p. 214) pour traiter des problèmes de filtre passe-bas. Les auteurs de [11, 48] utilisent aussi des méthodes de programmation semi-infinie pour résoudre leur problème. Dans l'article [38], les auteurs traitent justement par des méthodes d'optimisation semi-infinie un problème voisin :

$$\begin{array}{ll} \min\limits_{a\in\mathbb{R}^{2N+1}} & \langle r, a\rangle \\ \text{tel que} & a \in \mathcal{C}_{2N+1} \\ & a_{2k} = 0 \text{ pour } k = 1, \cdots, N \\ & a_1 = 0. \end{array}$$

Pour résoudre ce problème ils utilisent deux méthodes : une première, où ils discrétisent $\omega \in [0,\pi]$ comme précédemment, puis à partir de la solution non-réalisable $\widetilde{a}$, ils calculent un $\hat{a}$ réalisable. Evidemmment, cette méthode est sous-optimale, et des problèmes de conditionnement apparaissent dans les programmes linéaires lorsque la discrétisation est trop fine. Il faut noter qu'ils utilisent une discrétisation de l'ordre de $M = 20N$, ce qui leur permet de borner l'erreur commise par 1.3% mais autorise seulement des N pas trop grands ($N \approx 10$). La seconde méthode utilisée est une méthode de plans coupants pour l'optimisation semi-infinie. L'idée générale est de produire de nouvelles contraintes pour approcher au mieux l'ensemble contrainte près de l'optimum. Cet algorithme permet, tout en convergeant vers l'optimum global, d'oublier des contraintes générées précédemment afin de garder une taille raisonnable pour le programme linéaire courant.

Nous n'insisterons pas plus longuement sur les méthodes d'optimisation semi-infinie, qui fournissent néanmoins une méthode simple mais grossière pour résoudre ces problèmes pour des petites dimensions.

III.2 Algorithmes de suivi de chemin

Les méthodes de point intérieurs ont été utilisées dans [1, 27] pour prendre en compte des contraintes d'autocorrélation. L'idée générale commune aux deux articles est de travailler avec le cône polaire qui admet une formulation LMI simple en au plus $n+1$ variables ; à cela s'ajoute l'avantage numérique de pouvoir utiliser une FFT dans les calculs, pour gagner un ordre de

complexité. La présentation que nous ferons ici s'inspire plus de [1] : la différence fondamentale dans notre approche pour le problème de projection est que l'on utilise directement le théorème de décomposition de Moreau pour calculer la projection sans passer par une formulation duale conique du problème.

III.2.1 Introduction

On ne fera pas ici un exposé approfondi de ces méthodes, sachant qu'il existe déjà de très bonnes références sur le sujet : à ce propos, la référence classique est la monographie [42] de Nesterov et Nemirovski qui ont donné le bon formalisme de l'auto-concordance des fonctions barrières (cf. la définition III.1) pour traiter les méthodes de points intérieurs pour l'optimisation conique. A cette référence théorique, on pourra préférer les présentations moins formelles suivantes [5, 41, 39] ; enfin, dans un souci de mise en œuvre pratique, l'ouvrage [7] est particulièrement intéressant, car il développe de nombreux cas d'application, en détaillant l'art de l'ingénieur pour la programmation de ces algorithmes. On se bornera donc ici à présenter un seul algorithme simple et assez général, qui appartient à la catégorie des algorithmes dits *à petits pas* (Short Step Path Following Method), et qui par conséquent ne peut pas théoriquement rivaliser avec des algorithmes primaux-duaux du type Predicteur-Correcteur : pour mettre en place un tel algorithme il faut pouvoir borner l'erreur commise entre l'estimé courant et le minimum du problème de centrage ; comme souligné p. 113 dans [39], cela nécessite d'utiliser une information duale. Or dans notre problème, seul $\mathcal{C}_{n+1}^{\circ}$ a une barrière calculable de manière "économique". Si l'on utilise aussi $\mathcal{C}_{n+1}$, on va grever fortement les performances de notre algorithme. C'est pourquoi, notre choix s'est porté sur cet algorithme, qui donne de bons résultats si on l'utilise de manière pragmatique, c'est-à-dire en relâchant les bornes théoriques sur l'augmentation du paramètre de pénalité, sans pour autant observer de problèmes de convergence. Supposons que l'on souhaite résoudre un problème d'optimisation convexe sous la forme générique suivante :

$$(\mathcal{P}) \begin{cases} \min\limits_{x} & \langle c, x \rangle \\ & x \in G, \end{cases} \tag{III.2}$$

où l'ensemble-contrainte G est convexe, *borné*, fermé. Dans un contexte pratique, l'hypothèse de bornitude n'est pas très restrictive, puisqu'il suffit d'ajouter dans le problème une contrainte radiale ($\|x\| \leqslant r$) avec un rayon suffisamment grand, et si cette contrainte se sature à l'optimum, alors le problème peut être considéré comme non borné.

On peut arguer que le coût linéaire n'est pas très général, mais dans le cas d'un coût convexe $f(x)$, on rajoute une variable $y \in \mathbb{R}$ et on s'intéresse à

$$(\mathcal{P}') \begin{cases} \min\limits_{x,y} & y \\ & (x, y) \in (G \times \mathbb{R}) \cap \{(x, y) | y \geqslant f(x)\}, \end{cases}$$

qui est un problème convexe à coût linéaire.

Considérons de plus que G est muni d'*une barrière* F *auto-concordante de paramètre* θ*, logarithmiquement convexe (*θ*-BALC), supposée de plus non-dégénérée, et calculable en temps polynomial*. Avant de décrire l'algorithme, voyons à quoi correspondent ces définitions, et leurs utilités pour l'algorithme.

Définition III.1: *Soit* F *une fonction d'un ouvert convexe non vide* $Q \subset \mathbb{R}^n$ *dans* $\mathbb{R}$ *et* $\theta > 0$ *; alors* F *est une* θ*-BALC lorsque :*

(i) F *est* C^3*, convexe sur* Q (régularité)

(ii) $\lim_{x \to \partial Q} F(x) = \infty$ (propriété de barrière)

(iii) $\forall x \in Q$, $\forall h \in \mathbb{R}^n$

$$|D^3F(x)[h,h,h]| \leqslant 2(D^2F(x)[h,h])^{\frac{3}{2}} \quad \text{(auto-concordance)}$$

(iv) $\forall x \in Q$, $\forall h \in \mathbb{R}^n$

$$|DF(x)[h]| \leqslant (\theta D^2F(x)[h,h])^{\frac{1}{2}} \quad \text{(bornitude de la différentielle)}$$

(v) $\forall t > 0$, $\forall x \in \mathbf{int}\, K$

$$F(tx) = F(x) - \theta \log t \quad \text{(logarithmicité)}$$

Parmi toutes ces propriétés, celles qui assurent la convergence polynomiale des points générés par l'algorithme vers un "point proche de l'optimum" sont (iii) et (iv). La propriété (iii) assure en fait le caractère lipschitzien de l'opérateur hessien par rapport à la "norme locale" $\|h\|_x^2 = D^2F(x)[h,h]$, et permet ainsi de s'affranchir d'un choix d'une norme pour la preuve de convergence de l'algorithme de Newton. La propriété (ii) quant à elle garantit que l'on reste à l'intérieur du domaine réalisable sans trop s'approcher de la frontière.

Remarque III.1: *Comme on l'a précisé dans la définition précédente, les BALC sont toujours définies sur des ouverts, en raison de la propriété de barrière qui les contraint à prendre des valeurs infinies sur la frontière. Pourtant, dans la suite, on va associer à chaque cône convexe* **fermé** *une BALC. Il sera systématiquement sous-entendu que ces BALC sont évidemment définies sur* **l'intérieur** *de ces cônes.*

Sous ces hypothèses, le principe de l'algorithme est *grosso modo* le suivant : pour traiter la contrainte "$x \in G$", on la pénalise en utilisant la θ-BALC F et on résout la suite (C_t) de problèmes de centrage suivants

$$(C_t) \quad \min_x \quad F_t(x) = t\langle c, x\rangle + F(x),$$

qui présentent l'avantage d'être sans contraintes ! On trace ainsi la trajectoire centrale $\mathscr{C} = \{x^*(t)\}_{t \geqslant 0}$ (où $x^*(t) = \arg\min_x F_t(x)$) qui est l'ensemble continu des minimiseurs dans (C_t) (sous l'hypothèse de non dégénérescence, le minimum dans (C_t) est unique, et par conséquence $\arg\min$ est bien une application). Le but étant de trouver la valeur optimale $\langle c, x^*\rangle$ de $(\mathcal{P})$, on espère

$$\lim_{t \to +\infty} x^*(t) = x^*.$$

Numériquement, on ne construit pas exactement la trajectoire (dite) centrale mais plutôt une suite de points dans le voisinage de la trajectoire centrale, et l'on utilise le *décrément de Newton*

$$\lambda(F_t, x) = \max\{DF_t(x)[h] | D^2F(x)[h,h] \leqslant 1\} = \sqrt{\nabla_x F_t(x)^\top \left[\nabla_x^2 F(x)\right]^{-1} \nabla_x F_t(x)}$$

comme mesure de proximité à la trajectoire centrale.

III.2.2 Schéma algorithmique

On considère l'algorithme décrit de la manière suivante :

Algorithme 1 Algorithme de suivi de chemin élémentaire

Entrée : $x_0 \in \mathbf{int}G$, $t_0 > 0$, $\varepsilon > 0$, $\frac{1}{12} > \gamma > 0$ et $\kappa > 0$
  $x \leftarrow x_0$, $t \leftarrow t_0$
  Tant que $\theta > t\varepsilon$ **Faire**
    Tant que $\lambda(F_t, y) > \kappa$ **Faire**
      $y \leftarrow y - \frac{1}{1+\lambda(F_t,y)} \left[\nabla_x^2 F(y)\right]^{-1} \nabla_x F_t(y)$
    Fin Tant que
    $x \leftarrow y$, $t \leftarrow (1 + \frac{\gamma}{\sqrt{\theta}})t$
  Fin Tant que

Au vu du pseudo-code, on observe que l'algorithme comprend deux boucles imbriquées : les itérations internes en y correspondent à un algorithme de Newton relaxé (ce qui évite le recours à une recherche linéaire) appliqué au problème de centrage ; la boucle externe, quant à elle, correspond à la variation du paramètre t de pénalité ; on construit ainsi la suite des minimiseurs qui tend vers la solution optimale. Sous les hypothèses précédentes, on pourra trouver dans [39] une preuve de convergence, ainsi qu'une analyse de la complexité.

III.2.2.1 Complexité

L'algorithme présenté ici est un algorithme **itératif**, i.e. même s'il converge vers la solution du problème original, on obtient après un nombre fini d'itérations qu'une **approximation** , éventuellement trés précise, de la solution du problème original. Ainsi, parler de complexité pour un tel algorithme semble *a priori* inadapté, puisque le nombre d'itérations n'étant pas fini, tout au plus, pourrions parler de vitesse de convergence. Cependant, on peut généraliser la notion classique [1] de complexité grâce à la notion de solution ε*-sous-optimale* :

$$\overline{x} \ \varepsilon\text{-sous-optimale} \ \Leftrightarrow f(\overline{x}) \leqslant f^* + \varepsilon,$$

[1] Nombre d'opérations élémentaires réalisé par un algorithme, voir [25]

où f^* représente la valeur optimale du problème. On ne va donc pas chercher la solution optimale de $(\mathcal{P})$, mais une solution ε-sous-optimale. On peut définir alors la complexité $\mathcal{N}(\varepsilon, p)$ de l'algorithme comme *le nombre d'opérations de base* à effectuer sur une instance p du problème $(\mathcal{P})$ (c'est-à-dire que p est un problème concret où l'on a donné des valeurs précises à chacun des paramètres du problème $(\mathcal{P})$), pour obtenir une solution ε-sous-optimale. Concernant l'algorithme précédent, on peut démontrer comme dans [39] qu'il a une complexité de

$$\mathcal{N}(\varepsilon, p) \in \mathcal{O}(1)\sqrt{\theta(p)}\ln\left(\frac{\theta(p)}{t_0\varepsilon}\right),$$

où la constante $\mathcal{O}(1)$ ne dépend que de γ et κ et $\theta(p)$ est le paramètre de la BALC associée à la résolution de l'instance p de $(\mathcal{P})$.

III.2.3 Application au problème de projection sur $\mathcal{C}_{n+1}$

Considérons le problème de projection sur $\mathcal{C}_{n+1}$:

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^{n+1}} & \|x-r\|_2^2 \\ & x \in \mathcal{C}_{n+1}. \end{array} \tag{III.3}$$

Grâce à la décomposition de Moreau, ceci est équivalent au problème de projection sur $\mathcal{C}_{n+1}^\circ$:

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^{n+1}} & \|x-r\|_2^2 \\ & x \in \mathcal{C}_{n+1}^\circ. \end{array} \tag{III.4}$$

Réécrivons ce problème dans le formalisme précédent :

$$(\mathcal{P}) \begin{cases} \min & y \\ & \|x-r\|_2 \leqslant y \\ & x \in \mathcal{C}_{n+1}^\circ, \end{cases} \tag{III.5}$$

où l'on est effectivement en présence d'un coût linéaire. Intéressons-nous à l'ensemble des contraintes

$$G = \{(x,y) \in \mathbb{R}^{n+2} : x \in \mathcal{C}_{n+1}, \|x-r\|_2 \leqslant y\}.$$

Si on désigne par $\mathcal{L}_{n+2}(\mathbb{R}) = \{(x,y) \in \mathbb{R}^{n+2} : \|x\|_2 \leqslant y\} = \mathtt{epi}\|\cdot\|_2$, l'épigraphe de la fonction norme euclidienne, appelé aussi *cône de Lorentz*, alors le problème au-dessus revient à

$$(\mathcal{P}) \begin{cases} \min & y \\ & \begin{pmatrix} x-r \\ y \end{pmatrix} \in \mathcal{L}_{n+2}(\mathbb{R}) \\ & x \in \mathcal{C}_{n+1}^\circ, \end{cases} \tag{III.6}$$

et l'ensemble des contraintes G prend la forme suivante

$$G = \mathcal{L}_{n+2}(\mathbb{R}) \cap (\mathcal{C}^{\circ}_{n+1} \times \mathbb{R}).$$

Nous devons maintenant trouver une fonction barrière pour G. Supposons que nous connaissions déjà des barrières auto-concordantes pour $\mathcal{L}_{n+2}(\mathbb{R})$ ($\psi_1(x,y)$) et pour $\mathcal{C}^{\circ}_{n+1} \times \mathbb{R}$ ($\psi_2(x)$), alors on peut prendre comme barrière pour G

$$\psi(x,y) = \alpha\psi_1(x,y) + \beta\psi_2(x),$$

du moment que $\alpha, \beta \geqslant 1$, ce qui garantit que ψ reste bien une BALC (car ψ_1 et ψ_2 le sont). Dans la suite, on prendra simplement $\alpha = \beta = 1$.

En choisissant ψ comme barrière, l'Algorithme 1 nécessite de calculer une direction de Newton à chaque itération interne. Si on pose

$$G_t(x,y) = \begin{pmatrix} \nabla_x\psi_1 \begin{pmatrix} x-r \\ y \end{pmatrix} + \nabla\psi_2(x) \\ \tau + \frac{\partial\psi_1}{\partial y} \begin{pmatrix} x-r \\ y \end{pmatrix} \end{pmatrix},$$

alors on cherche à annuler $G_t(x,y)$ grâce à la direction de Newton que voici

$$\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = -DG_t(x,y)^{-1}[G_t(x,y)],$$

où

$$DG_t(x,y) = \begin{pmatrix} \nabla^2_{xx}\psi_1 \begin{pmatrix} x-r \\ y \end{pmatrix} + \nabla^2\psi_2(x) & \frac{\partial}{\partial y}(\nabla_x\psi_1) \begin{pmatrix} x-r \\ y \end{pmatrix} \\ [\nabla_x(\frac{\partial\psi_1}{\partial y}) \begin{pmatrix} x-r \\ y \end{pmatrix}]^{\top} & \frac{\partial^2\psi_1}{\partial y^2} \begin{pmatrix} x-r \\ y \end{pmatrix} \end{pmatrix}.$$

Nous verrons plus loin comment on choisit ψ_1 et ψ_2, et comment on calcule leurs dérivées, mais continuons pour l'instant avec le schéma général de résolution. Calculer la direction de Newton nécessite donc d'inverser $DG_t(x,y)$. Or $G_t(x,y)$ correspond au gradient d'une fonction convexe - conditions d'optimalité d'ordre 1 obligent - donc, conséquence du théorème de Schwarz, la matrice $DG_t(x,y)$ est symétrique et, conséquences de la convexité de la fonction ainsi que de la non-dégénérescence de la barrière, $DG_t(x,y)$ est une matrice définie positive. Un moyen efficace et numériquement stable d'inverser cette matrice est de recourir à la décomposition de Cholesky : ainsi, pour une matrice A définie positive, on sait qu'il existe une matrice L triangulaire inférieure telle que $A = LL^{\top}$; de plus le calcul de la matrice L est de $\frac{1}{3}n^3$ flops, soit moitié moins que la décomposition LU ($\frac{2}{3}n^3$ flops).

La décomposition de Cholesky possède un avantage supplémentaire : l'Algorithme 1 nécessite de relaxer le pas de Newton par un facteur

$$\frac{1}{1+\lambda(F_t,(x,y))} = \frac{1}{1+\sqrt{\langle DG_t(x,y)^{-1}G_t(x,y), G_t(x,y)\rangle}}.$$

Pour calculer $DG_t(x,y)^{-1}[G_t(x,y)]$ grâce à la décomposition de Cholesky, on ne calcule *jamais* la matrice $DG_t(x,y)^{-1}$, mais on procède (comme dans la résolution d'un système LU) à deux résolutions de systèmes triangulaires : si $DG(x,y) = LL^\top$, alors $LL^\top\alpha = G(x,y)$. On commence par résoudre $Lw = G(x,y)$ (substitution avant), puis on résout le second système, $L^\top\alpha = w$ (substitution arrière). Alors,

$$\langle w,w\rangle = (L^{-1}G(x,y))^\top(L^{-1}G(x,y)) = [G(x,y)]^\top L^{-\top}L^{-1}G(x,y)$$

et comme

$$DG(x,y) = LL^\top \Leftrightarrow DG(x,y)^{-1} = (LL^\top)^{-1} = L^{-\top}L^{-1},$$

on a tout simplement :

$$\lambda(F_t,(x,y)) = \sqrt{\langle w,w\rangle}.$$

On a pratiquement tous les éléments en main pour implémenter l'Algorithme 1 en machine. Reste à décider maintenant des choix pour ψ_1 et ψ_2 et comment calculer efficacement leurs dérivées premières et secondes.

III.2.3.1 Barrière de $\mathcal{L}_{n+2}(\mathbb{R})$

$\mathcal{L}_{n+2}(\mathbb{R})$ est un cône convexe solide et pointé, qui est assez bien connu dans la littérature de l'Optimisation convexe. Il a une frontière complètement lisse et on peut le munir de la 2-BALC suivante

$$\psi_1(x,y) = -\log(y^2 - \|x\|^2).$$

On peut alors calculer facilement les dérivées premières et secondes de ψ_1 en $(x-r,y)$:

- dérivées premières :

$$\frac{\partial\psi_1}{\partial y}\begin{pmatrix}x-r\\ y\end{pmatrix} = -\frac{2y}{y^2-\|x-r\|^2}$$

et

$$\nabla_x\psi_1\begin{pmatrix}x-r\\ y\end{pmatrix} = \frac{2(x-r)}{y^2-\|x-r\|^2}.$$

- dérivées secondes :

$$\nabla^2_{xx}\psi_1\begin{pmatrix}x-r\\ y\end{pmatrix} = \frac{2I_n}{y^2-\|x-r\|^2} + \frac{4(x-r)(x-r)^\top}{(y^2-\|x-r\|^2)^2},$$

ainsi que

$$\frac{\partial}{\partial y}\nabla_x\psi_1\begin{pmatrix}x-r\\ y\end{pmatrix} = -\frac{4y(x-r)}{(y^2-\|x-r\|^2)^2}$$

et

$$\frac{\partial^2\psi_1}{\partial y^2}\begin{pmatrix}x-r\\ y\end{pmatrix} = \frac{4y^2}{(y^2-\|x-r\|^2)^2} - \frac{2}{y^2-\|x-r\|^2}.$$

En introduisant la matrice

$$J_{n+2} = \begin{pmatrix} I_{n+1} & 0 \\ 0 & -1 \end{pmatrix},$$

on peut réécrire $\psi_1(x,y) = -\log((x,y)^\top J_{n+2}(x,y))$ et avoir des écritures plus élégantes des dérivées premières et secondes de ψ_1, mais nous préférons ici la première forme car elle correspond mieux à la réalité de l'implémentation de l'algorithme pour lequel les produits matriciels constitueraient un coût supplémentaire inutile. La vérification que ψ_1 est bien une 2-BALC ne présente pas de difficultés majeures, elle est détaillée dans [5, 39].

III.2.3.2 Barrière de $\mathcal{C}^\circ_{n+1}$

On rappelle que

$$\mathcal{A}^\star(\mathcal{C}^\circ_{n+1}) = \mathcal{T}_{n+1}(\mathbb{R}) \cap \mathcal{S}^-_{n+1}(\mathbb{R}).$$

En supposant que l'on connaisse une barrière de $\mathcal{S}^-_{n+1}(\mathbb{R})$, alors grâce à une telle barrière on peut facilement forcer les images $\mathcal{A}^\star(x)$ à être dans $\mathcal{S}^-_{n+1}(\mathbb{R})$. Comme $(n+1)$-BALC bien connue de $\mathcal{S}^-_{n+1}(\mathbb{R})$, on peut prendre

$$-\ln\det(-X) \quad \text{pour } X \prec 0.$$

Il est par conséquent logique de choisir

$$\psi_2(x) = -\ln\det\mathcal{A}^\star(-x)$$

comme $(n+1)$-BALC de $\mathcal{C}^\circ_{n+1}$.

Les propriétés de régularité, de convexité, de barrière, d'auto-concordance, etc. proviennent directement du fait que $-\ln\det$ est déjà une "bonne barrière" pour le cône des matrices semidéfinies positives. Là encore on pourra se reporter aux références précédentes.

Intéressons-nous d'abord à l'expression des dérivées premières et secondes de ψ_2 : un peu de calcul différentiel nous donne pour $h \in \mathbb{R}^{n+1}$

$$D\psi_2(x)[h] = -\frac{1}{\det\mathcal{A}^\star(-x)}\langle\langle(\mathtt{com}\mathcal{A}^\star(-x))^\top, \mathcal{A}^\star(-h)\rangle\rangle = \langle\langle[\mathcal{A}^\star(-x)]^{-1}, \mathcal{A}^\star(h)\rangle\rangle,$$

(com désignant la comatrice) et, en redifférentiant, on aboutit pour tout $(h,k) \in (\mathbb{R}^{n+1})^2$ à

$$D^2\psi_2(x)[h,k] = \langle\langle\mathcal{A}^\star(h), [\mathcal{A}^\star(-x)]^{-1}\mathcal{A}^\star(k)[\mathcal{A}^\star(-x)]^{-1}\rangle\rangle.$$

Afin de calculer le pas de Newton, on souhaiterait disposer des expressions de ces dérivées de manière matricielle, c'est-à-dire le vecteur gradient $\nabla_x\psi_2$ ainsi que la matrice hessienne $\nabla^2_{xx}\psi_2$. En utilisant le fait que $(\mathcal{A}^\star)^\star = \mathcal{A}$, le gradient devient

$$\nabla\psi_2(x) = \mathcal{A}\left([\mathcal{A}^\star(-x)]^{-1}\right).$$

Pour la matrice hessienne H, il suffit de substituer à h et k les vecteurs e_i et e_j de la base canonique. On obtient alors

$$H_{ij} = \langle\langle \mathcal{A}^\star(e_i), [\mathcal{A}^\star(-x)]^{-1}\mathcal{A}^\star(e_j)[\mathcal{A}^\star(-x)]^{-1}\rangle\rangle = \langle\langle A^{(i)}, [\mathcal{A}^\star(-x)]^{-1}A^{(j)}[\mathcal{A}^\star(-x)]^{-1}\rangle\rangle.$$

Intéressons-nous à H et considérons que l'on a une décomposition de Cholesky $RR^\top$ de $[\mathcal{A}^\star(-x)]^{-1}$. Un calcul rapide de R, sachant que $\mathcal{A}^\star(x)$ est sous forme Toeplitz, peut s'obtenir grâce à l'algorithme de Levinson-Durbin (de complexité $\mathcal{O}(n)$) comme décrit dans [1], mais pour des raisons de stabilité numérique il est parfois préférable d'utiliser la décomposition de Cholesky directement. En désignant par $\{r_k\}_{k=0}^n$ les vecteurs colonnes de R, on a

$$[\mathcal{A}^\star(-x)]^{-1} = RR^\top = \sum_{k=0}^{n} r_k r_k^\top,$$

et par conséquent

$$H_{ij} = \sum_{0\leqslant k,l\leqslant n} \langle\langle A^{(i)}, r_k r_k^\top A^{(j)} r_l r_l^\top\rangle\rangle.$$

Dans cette expression, on remarque que

$$r_k^\top A^{(j)} r_l = \langle\langle A^{(j)}, r_l r_k^\top\rangle\rangle = \mathcal{A}(r_l r_k^\top)_j.$$

Il vient alors :

$$H_{ij} = \sum_{0\leqslant k,l\leqslant n} \mathcal{A}(r_k r_l^\top)_i \mathcal{A}(r_k r_l^\top)_j.$$

On reconnaît là une somme de matrices dyadiques ; H s'écrit donc comme suit :

$$H = \sum_{0\leqslant k,l\leqslant n} \mathcal{A}(r_k r_l^\top)[\mathcal{A}(r_k r_l^\top)]^\top.$$

En utilisant l'identité (II.4), on obtient finalement après simplication

$$H = \frac{1}{2}\sum_{0\leqslant k,l\leqslant n} \mathtt{corr}_a(r_k, r_l)(\mathtt{corr}_a(r_k, r_l) + \mathtt{corr}_a(r_l, r_k))^\top.$$

La décomposition de Cholesky nous donne aussi une expression simple pour le gradient :

$$g = \nabla\psi_2(x) = \mathcal{A}\left(\sum_{k=0}^{n} r_k r_k^\top\right) = \sum_{k=0}^{n} \mathtt{corr}_a(r_k, r_k).$$

Nous avons maintenant des expressions matricielles plus agréables du gradient et de la matrice hessienne. Malheureusement, le coût de calcul de H en utilisant directement cette formule reste élevé, $\mathcal{O}(n^4)$, ce qui peut être déjà prohibitif pour des valeurs moyennes de n (de

l'ordre de 100). Nous allons voir dès à présent, par une méthode astucieuse décrite dans [1], comment ramener cette complexité à $\mathcal{O}(n^3)$.

On a ramené le calcul de la hessienne à un calcul impliquant des corrélations. Or, on a expliqué dans la preuve de la proposition I.3, comment calculer une corrélation à un coût quasi-linéaire ($\mathcal{O}(N \log N)$) : en suivant cette démarche, on choisit un entier $N \geqslant 2n+1$, on désigne alors par $\tilde{r}_k = Pr_k$ et $\tilde{r}_l = Pr_l$ les vecteurs complétés par des zéros ($P : \mathbb{R}^{n+1} \to \mathbb{R}^N$ est une injection linéaire) ; on notera $R_k = W\tilde{r}_k$ et $R_l = W\tilde{r}_l$ leurs TFD respectives. La surjection linéaire $S : \mathbb{R}^N \to \mathbb{R}^{n+1}$ permet d'extraire ensuite les $n+1$ premières composantes. Alors

$$\mathtt{corr}_a(r_k, r_l) = \mathcal{F}^{-1}(\overline{\mathcal{F}(\tilde{r}_k)} \circ \mathcal{F}(\tilde{r}_l)) = \frac{1}{N} S\overline{W}(\overline{WPr_k} \circ WPr_l).$$

On peut donc réécrire H sous la forme

$$\begin{aligned} H &= \frac{1}{2N^2} \sum_{0 \leqslant k,l \leqslant n} S\overline{W}(\overline{R_k} \circ R_l)(S\overline{W}(\overline{R_k} \circ R_l + \overline{R_l} \circ R_k))^\top \\ &= \frac{1}{2N^2} S\overline{W} \sum_{0 \leqslant k,l \leqslant n} (\overline{R_k} \circ R_l)(\overline{R_k} \circ R_l + \overline{R_l} \circ R_k)^\top W^\star S^\top. \end{aligned}$$

Dans cette formule, si l'on utilise les identités triviales

$$(a \circ b)(c \circ d)^\top = ac^\top \circ bd^\top = ad^\top \circ bc^\top,$$

alors

$$H = \frac{1}{2N^2} S\overline{W} \sum_{0 \leqslant k,l \leqslant n} (\overline{R_k R_k}^\top) \circ (R_l R_l^\top) + (\overline{R_k} R_k^\top) \circ (R_l R_l^\star) W^\star S^\top.$$

Dans cette expression la somme double est séparable comme produit de sommes simples sur k et l, et les variables indicielles étant muettes, on peut réécrire H en en les permutant, ce qui donne

$$\nabla_{xx}\psi_2 = \frac{1}{2N^2} S\overline{W} \left[\left(\sum_{k=0}^{n} R_k R_k^\top \right) \circ \left(\overline{\sum_{k=0}^{n} R_k R_k^\top} \right) + \left(\sum_{l=0}^{n} R_l R_l^* \right) \circ \left(\sum_{l=0}^{n} R_l R_l^* \right)^* \right] W^\star S^\top.$$

L'avantage de cette forme est que l'on possède dorénavant une méthode "rapide" de calcul de H en $\mathcal{O}(n^3)$: en effet :

- le calcul des r_k coûte $\frac{1}{3}n^3$ (Cholesky) ;
- celui de l'ensemble des R_k coûte $\mathcal{O}(nN \log N)$ d'après le chapitre 1, et comme $N \in \mathcal{O}(n)$, ceci ne coûte finalement que $\mathcal{O}(n^2 \log n)$;
- le calcul de $\sum_{k=0}^{n} R_k R_k^\top$ coûte quant à lui $\mathcal{O}(n^3)$, ce n'est rien d'autre que le produit matriciel $BB^\top$ où $B = [R_1, \cdots, R_n]$;

– les produits de Hadamard coûtent $\mathcal{O}(n^2)$;
– et pour les mêmes raisons que précédemment, la FFT inverse coûte quant à elle $\mathcal{O}(n^2 \log n)$.

Finalement, la complexité de l'ensemble (celle de la partie la plus coûteuse en ordre de grandeur) est de $\mathcal{O}(n^3)$.

De même, on déduit par un calcul simple l'expression du gradient

$$\nabla_x \psi_2 = \frac{1}{N} \overline{W} \left(\sum_{k=0}^{n} \overline{R_k} \circ R_k \right),$$

dont le calcul a pour complexité $\mathcal{O}(n^2 \log n)$.

III.2.3.3 Mise en œuvre pratique

L'étude des algorithmes numériques pour la résolution de problèmes impliquant $\mathcal{C}_{n+1}$ a donné lieu à une implémentation numérique effective. Nous nous sommes intéressés à la résolution du problème d'approximation cité plus haut, dont la résolution requiert des temps de calcul encore raisonnables en grande dimension (par exemple, $n = 1000$). Notre implémentation n'est pas fondamentalement très différente de celle de [1]. En réalité, pour résoudre le problème d'approximation, nous utilisons directement la décomposition de Moreau et nous résolvons le problème "primal"

$$\begin{aligned} \min_{x \in \mathbb{R}^{n+1}} \quad & \|x - r\|_2^2 \\ & x \in \mathcal{C}_{n+1}^\circ, \end{aligned}$$

puis nous récupérons le projeté sur $\mathcal{C}_{n+1}$ par simple différence : $p_{\mathcal{C}_{n+1}}(x) = x - p_{\mathcal{C}_{n+1}^\circ}(x)$. L'approche préconisée dans [1] consiste à formuler le problème sous forme conique à l'aide du cône de Lorentz, puis à résoudre le problème dual

$$\begin{aligned} \max_{\mu \in \mathbb{R}^{n+1}} \quad & -\langle r, u \rangle \\ & -\mu \in \mathcal{C}_{n+1}^\circ \\ & \|\mu\| \leqslant 1, \end{aligned} \tag{III.7}$$

qui est traité par une Minimisation Séquentielle sans Contrainte (SUMT) du type

$$\begin{aligned} \max_{\mu \in \mathbb{R}^{N+1}} \quad & -\langle r, u \rangle - \frac{1}{2}\|\mu\|^2 \\ & -\mu \in \mathcal{C}_{n+1}^\circ, \end{aligned} \tag{III.8}$$

et l'on obtient une solution primale x^* orthogonale à μ suivant la formule

$$x^* = r - \langle r, \mu \rangle \mu,$$

en ayant au préalable renormalisé μ à 1.

L'implémentation a d'abord été effectuée en Scilab pour développer rapidement un prototype qui fonctionnait en petite dimension ($n \in \{2, \cdots, 15\}$). Ensuite, pour des raisons évidentes de rapidité d'exécution, nous nous sommes tournés vers une solution en C/C++ utilisant des bibliothèques mathématiques du domaine public, notamment ATLAS (Implémentation de BLAS) et la bibliothèque FFTW[21] spécialisée dans le calcul des transformées de Fourier Rapide. Nous n'avons pas pu comparer les temps de résolution de notre algorithme avec ceux de [1], car leur code en libre accès sur Internet est écrit pour un environnement Windows, et lors du portage du code, sous Linux, la seule partie de code que nous avons réussi à faire fonctionner correctement est l'évaluation de la barrière et de ses dérivées. A ce niveau toutefois, nous avons comparé les temps de calcul sur la même machine (Pentium XEON 4 processeurs avec 1 Go de Mémoire) avec les mêmes instances de problèmes, et nous avons constaté qu'ils étaient sensiblement les mêmes (cela nous a même permis d'optimiser notre code afin d'atteindre les mêmes performances de calcul), ce qui est plutôt logique vu que nous avons emprunté l'idée de la méthode d'évaluation de la barrière à [1]. Un détail qui a son importance quand on arrive dans les limites de temps de calculs raisonnables réside dans le calcul des FFT. Comme on l'a présenté précédemment, on peut voir la TFD comme un élément de $L(\mathbb{C}^N)$. Or ici $x \in \mathbb{R}^N$, et donc

$$\overline{X}(l) = \sum_{k=0}^{N-1} e^{\frac{2ikl}{N}} x(k) = X(-l).$$

En considérant les signaux comme périodiques, on voit que l'on peut se contenter de calculer seulement la moitié des transformées et obtenir les autres par conjugaison complexe. Ainsi, la bibliothèque FFTW fait parfaitement bien cela, et il suffit d'utiliser les FFTs pour des données réelles. Ainsi par exemple, dans le cas $n = 600$, nous avions avec la première méthode des matrices en mémoire de taille $T = 2048$, alors qu'avec des FFTs réelles la taille n'était plus que de $T = 1024$. On peut comprendre aisément, que si l'évaluation de la matrice hessienne se fait souvent, la différence devient non-négligeable.

Comme on l'a vu précédemment, l'Algorithme 1 nécessite un point réalisable pour démarrer. Comme point intérieur à $\mathcal{C}^\circ_{n+1}$, nous avons choisi $x_0 = (-1, 0, \cdots, 0)$ (car $\mathcal{A}^\star(x_0) = -I_n$) et nous prenons $y = \|x_0 - r\| + 1$ pour que le couple (x, y) soit un point strictement réalisable. Il faut noter que pour un problème avec des contraintes égalités, par exemple dans un problème de conception de filtre, il faudrait résoudre d'abord un premier problème (souvent désigné comme phase I en progammation mathématique) pour trouver un point réalisable.

III.2.3.4 Tests sur la dimension

Pour illustrer notre implémentation, nous avons réalisé des problèmes de dimensions différentes. Ainsi, pour chaque n, nous avons créé 5 vecteurs aléatoires grâce à une loi uniforme dans $[0, 1]^{n+1}$ (on peut contester ce choix, mais il arrive très souvent dans les applications pratiques que l'on normalise les données, ce qui justifie la nature des problèmes-test), puis on a effectué une moyenne sur les temps de calcul respectifs. On a pris $\varepsilon = \kappa = 10^{-3}$ et $\mu = 1 + \frac{\gamma}{\sqrt{\theta}} = 8$ comme paramètres fixes de l'algorithme. On a ainsi représenté sur la

FIG. III.1: Temps CPU en fonction de n

n+1	Nb. Ité.	Temps CPU
10	6	7ms
20	6	24.43ms
100	7	1.90s
200	7	14.39s
600	8	14m54s
1000	8	37m21s

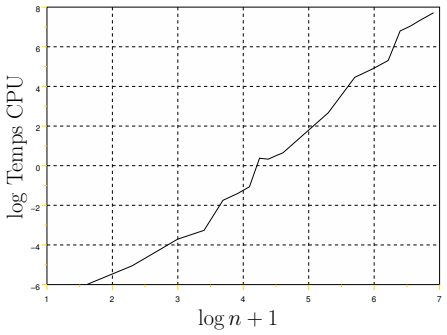


Figure III.1 le logarithme népérien du temps de calcul en fonction de celui de n. Théoriquement, on devrait observer une droite de pente environ 3.5, car la complexité du calcul est en $\mathcal{O}(n^3\sqrt{n}\log n)$. On constate plutôt les sauts dus aux changements de puissance de 2 dans la dimension des vecteurs de la FFT.

Pour assurer la convergence de l'algorithme vers un point x^* ε-sous-optimal, il faudrait vérifier que les trois conditions (II.17) sont vérifiées à ε près. Comme les conditions d'appartenance à $\mathcal{C}_{n+1}$ et $\mathcal{C}^\circ_{n+1}$ deviennent numériquement compliquées à vérifier et peu sûres en grande dimension, on s'est contenté seulement de vérifier à ε près, la condition d'orthogonalité suivante

$$\langle x^* - r, x^* \rangle \leqslant \varepsilon.$$

On constate que l'on peut atteindre des dimensions de problèmes assez grandes, par exemple ($n = 1000$) pour des temps de calcul encore raisonnables. Nous avons fait un test pour $n = 2000$, mais même si l'algorithme termine, le temps de calcul s'avère prohibitif pour des applications pratiques (plus d'un jour et 6 heures). A propos de la robustesse de l'algorithme aux données mal conditionnées, on peut noter deux points :

- quand les données ont des valeurs assez grandes : prendre par exemple un "signal rampe" ($x(i) = i$) avec $n > 50$; l'algorithme ne converge plus car les données manipulées sont trop élevées. Il faut noter que la principale application du problème d'approximation est de chercher la projection d'un signal qui est censé être proche d'un signal d'autocorrélation, ce qui n'est vraisemblablement pas le cas pour le signal rampe.
- un deuxième point intéressant du point de vue numérique réside dans l'utilisation des matrices Toeplitz pour décrire $\mathcal{C}^\circ_{n+1}$. Selon un article récent [4], les matrices Hankel définies positives présentent un très mauvais conditionnement (au sens classique où on l'entend : le conditionnement $\kappa(A) = \|A\| \cdot \|A^{-1}\|$ d'une matrice A inversible fournit une borne supérieure de l'erreur relative commise lors de la résolution d'un système du type $Ax = b$, si l'on perturbe b). On pourrait donc craindre *a priori* que ce soit aussi

le cas pour des matrices Toeplitz définies positives, car entre les matrices

$$T(x) = \begin{pmatrix} x_0 & x_1 & \cdots & x_n \\ x_1 & x_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & x_1 \\ x_n & \cdots & x_1 & x_0 \end{pmatrix} \text{ et } H(x) = \begin{pmatrix} x_0 & x_1 & \cdots & x_n \\ x_1 & \iddots & \iddots & \vdots \\ \vdots & x_n & \iddots & x_1 \\ x_n & \cdots & x_1 & x_0 \end{pmatrix}$$

on a la relation très simple $JT = TJ = H$ avec $J = H([0, \cdots, 0, 1])$, matrice de permutation, donc orthogonale. Par conséquent, $T(x)$ et $H(x)$ présentent le même conditionnement. Or on peut se convaincre facilement que $T(x) \succ 0$ n'implique pas $H(x) \succ 0$. En effet,

$$\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \succ 0 \text{ mais } \begin{pmatrix} -1 & 2 \\ 2 & -1 \end{pmatrix} \not\succ 0,$$

et par conséquent nos matrices $T(x)$ ne sont pas contraintes à subir la borne inférieure sur le conditionnement des matrices Hankel *définies positives* fournie dans [4].

III.2.3.5 Influence de μ

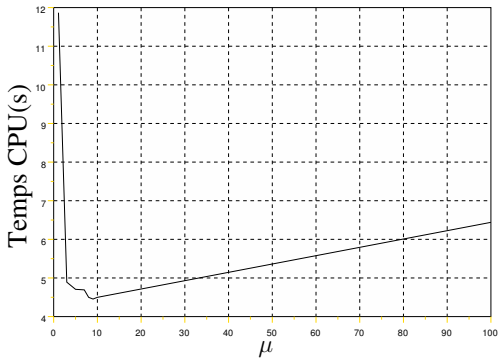
L'Algorithme 1 est souvent désigné dans la littérature comme un algorithme à petits pas : c'est-à-dire que pour garantir que l'on reste dans le voisinage de la trajectoire centrale, il faut augmenter lentement t à l'aide du paramètre

$$\mu = 1 + \frac{\gamma}{\sqrt{\theta}},$$

où $\gamma < 1/12$. Par exemple, dans notre cas, pour un problème de taille $n = 50$, on "devrait" prendre typiquement $\mu = 1 + 0.05/\sqrt{50} \approx 1.0007$, soit avec $t_0 = 1$ et $\varepsilon = 10^{-3}$, environ 1541 itérations externes et, dans la pratique, on peut se contenter de prendre μ beaucoup plus grand, par exemple entre 2 et 10, et l'algorithme converge alors seulement en moins de 10 itérations. Dans la plupart des cas, et c'est ce que nous avons toujours observé numériquement dans nos exemples, il y a convergence sans aucun problème. Pour avoir une convergence plus rapide, et de surcroît garantie par la théorie, il faudrait implémenter une stratégie à pas longs, où l'on doit prédire puis corriger la longueur du pas de Newton. Beaucoup d'implémentations utilisent souvent une information duale supplémentaire, qui n'est pas calculable efficacement ici, c'est aussi pourquoi, dans un souci de simplicité, nous nous sommes limités à cet algorithme qui marche en théorie mais que nous utilisons en pratique dans des "limites non prévues par le concepteur", et qui donne pourtant des bons résultats. La sélection du μ optimal peut se faire de manière empirique en traçant la courbe des temps de calcul en fonction de μ. Ainsi par exemple sur un problème de taille $N = 50$, on constate empiriquement que la valeur optimale est proche de $\mu = 9$ (voir Fig. III.2).

FIG. III.2: Temps CPU en fonction de μ

μ	It	Temps CPU
1.1	127	11.868s
3	12	4.892s
5	9	4.706s
7	8	4.692s
8	7	4.498s
9	7	4.454s
10	7	4.498s
100	4	6.44s



III.3 Un algorithme basé sur des projections alternées

Considérons encore une fois le problème d'approximation reformulé comme en (II.19) :

$$\min_{X \in \mathcal{S}_{n+1}(\mathbb{R})} \quad \|X - R\|_F^2$$
$$X \in \mathcal{T}_{n+1}(\mathbb{R}) \cap \mathcal{N}(\mathcal{S}_{n+1}^{-}(\mathbb{R})).$$

Ce problème comporte, il est vrai, une variable matricielle, cependant il présente une bonne structure. On cherche ainsi la projection orthogonale d'une matrice R sur l'intersection d'un sous-espace ($\mathcal{T}_{n+1}(\mathbb{R})$) et d'un cône convexe ($\mathcal{N}(\mathcal{S}_{n+1}^{-}(\mathbb{R}))$). Il existe dans la littérature un algorithme sur mesure pour ce genre de problèmes : l'algorithme de Boyle-Dykstra.

III.3.1 Algorithme de Boyle-Dykstra

On se contentera ici d'énoncer le théorème sur la convergence de l'algorithme BD et de présenter le schéma général de l'algorithme. Pour plus de détails sur les méthodes de projections alternées (algorithmes de Von Neumann, etc), ou les problèmes d'intersection d'ensembles convexes, avec notamment des applications au traitement d'image, on pourra se reporter à [14] ; si l'on préfère une introduction synthétique en langue française sur le sujet, on pourra consulter certains chapitres de la thèse [49]. On considère dans la suite un espace de Hilbert E, A et B deux convexes fermés de E, ce qui implique que les projections sur A et B sont bien des applications définies sur E. Le schéma de l'algorithme BD est inspiré d'un algorithme antérieur dû à Von Neumann, où l'on projette alternativement sur A et B, mais avec des termes correctifs par rapport à l'algorithme de Von Neumann, lequel est prévu seulement pour des sous-espaces (si on applique brutalement l'algorithme de Von Neumann à deux convexes fermés, et si leur intersection est non vide, on obtient à la limite seulement un point quelconque de l'intersection, mais pas forcément la projection sur l'intersection). Donnons tout de suite le schéma de l'algorithme BD :

Algorithme 2 Algorithme de Boyle-Dykstra

Entrée : p_A et p_B projections orthogonales sur A et B
Entrée : $x \in E$, $\varepsilon > 0$
  $a_0 \leftarrow 0$, $b_0 \leftarrow x$, $c_0 \leftarrow 0$, $d_0 \leftarrow 0$
  Tant que $\|b_n - a_n\| \geqslant \varepsilon$ **Faire**
    $a_{n+1} \leftarrow p_A(b_n + c_n)$
    $c_{n+1} \leftarrow b_n + d_n - a_{n+1}$
    $b_{n+1} \leftarrow p_B(a_{n+1} + d_n)$
    $d_{n+1} \leftarrow a_{n+1} + d_n - b_{n+1}$
  Fin Tant que

Les suites (a_n) et (b_n) représentent les projections alternées sur les parties A et B, et les c_n et d_n sont les termes correctifs (par rapport à l'algorithme originel de Von Neumann) qui correspondent géométriquement au déplacement arrière (c_{n+1}) entre l'itéré courant$(b_n + c_n)$ et son projeté (a_{n+1}).

Pour un résultat de convergence des suites (a_n) et (b_n) on a le

Théorème III.1: *Si* E *est un espace de Hilbert,* A, B *deux convexes fermés de* E, *alors les suites* $\{a_n\}_{n\geqslant 0}$ *et* $\{b_n\}_{n\geqslant 0}$ *définies suivant l'Algorithme 2 vérifient*

$$\lim_{n\to+\infty} b_n - a_n = \lim_{n\to+\infty} b_n - a_{n+1} = v,$$

où $v = p_{\overline{B-A}}(0)$ *et* $\|v\| = d(A,B)$. *De plus :*

(i) *si* $d(A,B)$ *n'est pas atteinte (i.e., la borne inférieure dans la définition de* $d(A,B)$ *n'est pas atteinte)*

$$\lim_{n\to+\infty} \|a_n\| = \lim_{n\to+\infty} \|b_n\| = +\infty;$$

(ii) *si* $d(A,B)$ *est atteinte,*

$$\lim_{n\to+\infty} a_n = p_F(x) \text{ et } \lim_{n\to+\infty} b_n = p_G(x),$$

où

$$F = \{a \in A \,:\, d(a,B) = d(A,B)\} \text{ et } G = \{b \in B \,:\, d(b,B) = d(A,B)\},$$

sont des convexes non vides tels que $F + v = G$.

Pour la preuve de ce théorème, voir [3]. Connaissant ce résultat, qui malheureusement n'est qu'un théorème de convergence sans précision sur la vitesse, on peut mettre en place la structure de l'algorithme appliqué à notre cas particulier en précisant les projections orthogonales sur $\mathcal{T}_{n+1}(\mathbb{R})$ et sur $\mathcal{N}(\mathcal{S}^-_{n+1}(\mathbb{R}))$.

III.3.2 Projection sur $\mathcal{T}_{n+1}(\mathbb{R})$

La projection sur $\mathcal{T}_{n+1}(\mathbb{R})$ est assez simple : les $A^{(i)}$ en constituant une base orthogonale, elle s'écrit donc

$$p_{\mathcal{T}_{n+1}(\mathbb{R})}(M) = \sum_{k=0}^{n} \langle\langle M, A^{(i)} \rangle\rangle \frac{A^{(i)}}{||A^{(i)}||^2},$$

dont le calcul a une complexité de $\mathcal{O}(n^3)$.

III.3.3 Projection sur $\mathcal{N}(\mathcal{S}_{n+1}^{-}(\mathbb{R}))$

L'idée de cet algorithme nous est venue du fait que l'on peut projeter orthogonalement (au sens de la norme associée à $\langle\langle \cdot, \cdot \rangle\rangle$) facilement sur $\mathcal{S}_{n+1}^{-}(\mathbb{R})$. En effet, pour une matrice A, si on dispose de ses valeurs propres et de vecteurs propres associés dans une base orthogonale (complexité du calcul des éléments propres : $\mathcal{O}(n^3)$ pour la boucle principal) sous la forme

$$A = P\, \mathtt{Diag}(\lambda_0, \cdots, \lambda_n)\, P^\top,$$

alors la projection sur $\mathcal{S}_{n+1}^{-}(\mathbb{R})$ est

$$A^- = P\, \mathtt{Diag}(\lambda_0^-, \cdots, \lambda_n^-)\, P^\top,$$

où $x^- = \min(0, x)$ (la démonstration élémentaire de ce résultat utilise la décomposition de Moreau, voir par exemple [49] p.20 ou [35]).

Malheureusement pour nous, on doit projeter sur le cône $\mathcal{N}(\mathcal{S}_{n+1}^{-}(\mathbb{R}))$, et même si l'application $\mathcal{N}$ a de bonnes propriétés (elle est auto-adjointe, inversible, de matrice diagonale, définie positive), elle ne rend pas pas évidente l'utilisation de ce résultat.

Pour l'instant, l'unique solution algorithmique que nous avons trouvée pour calculer la projection est un algorithme ... de points intérieurs ! Nous l'avons implémenté juste pour démontrer la convergence numérique de BD sur ce problème, mais il ne marche que pour des petites dimensions ($2 \leqslant n \leqslant 15$), alors que nous voulions l'utiliser comme alternative à un algorithme de suivi de chemin ; par conséquent il ne doit pas être basé lui-même sur un algorithme de points intérieurs, à moins que ce dernier soit de coût moindre, ce qui n'est précisément pas le cas ici.

La projection s'exprimant, elle aussi, comme un problème d'approximation

$$\begin{aligned} \min_{X \in \mathcal{S}_n(\mathbb{R})} \quad & ||X - R||_F^2 \\ & \mathcal{N}(X) \preceq 0, \end{aligned} \tag{III.9}$$

on peut utiliser comme barrière

$$-\log\det \mathcal{N}^{-1}(-X),$$

et reformuler le problème à l'aide du cône de Lorentz. Pour travailler avec moitié moins de variables, on optimise dans $\mathbb{R}^{n(n+1)/2}$ en utilisant l'isométrie $\mathtt{svec} : \mathcal{S}_n(\mathbb{R}) \to \mathbb{R}^{n(n+1)/2}$ définie par

$$\mathtt{svec}(A) = \sum_{i=1}^{n} A_{ii}\mathbf{e}_{ii} + \sqrt{2}\sum_{i=2}^{n}\sum_{j=1}^{i-1} A_{ij}\mathbf{e}_{j(j-1)/2+i},$$

qui consiste à déplier la moitié supérieure de la matrice A, tout en affectant d'un facteur $\sqrt{2}$ les coefficients extra-diagonaux. Avec cet algorithme, on obtient un coût de projection sur $\mathcal{N}(\mathcal{S}_n^+(\mathbb{R}))$ de complexité en $\mathcal{O}(n^4)\sqrt{n}\log(n/\epsilon)$.

Une série de tests avec des petites dimensions nous a donné les résultats suivants

n+1	Nb. Itérations	temps CPU
2	2	0.09s
3	30	4.8s
4	28	12.7s
6	21	38s
8	40	293.7s
10	46	804s

Un commentaire s'impose sur ces résultats : les temps de calcul sont beaucoup plus élevés que pour l'autre algorithme, et le nombre d'itérations ne semble pas dépendre trivialement de la taille du problème. On pourrait arguer que l'on a peut-être réalisé une implémentation lente (Scilab au lieu de C/C++) comparée à celle de l'autre algorithme, mais la plus grande complexité de cet algorithme explique parfaitement sa lenteur et, *tant qu'il n'y a pas de méthode "rapide"* (sous-entendu au plus en $\mathcal{O}(n^3)$) *pour le calcul de la projection sur* $\mathcal{N}(\mathcal{S}_{n+1}^-(\mathbb{R}))$*, cet algorithme a, de notre point de vue, peu d'intérêt pratique.*

III.3.4 Dans quels cas pourrait-on appliquer l'algorithme de Boyle-Dykstra ?

On l'a vu, pour l'instant, l'application de l'algorithme BD pour la résolution de problèmes de projection sur $\mathcal{C}_{n+1}$ ou sur son polaire s'avère inefficace numériquement. A l'inverse, on peut se demander dans quels problèmes de projection voisins du nôtre, cette approche s'avèrerait efficace. En fait, grâce à BD, le problème qu'on sait résoudre "efficacement" est le suivant :

$$(\mathcal{M}at)\begin{cases} \min\limits_{X\in\mathcal{S}_{n+1}(\mathbb{R})} & \|X-R\|_F^2 \\ & X \in \mathcal{S}_{n+1}^-(\mathbb{R}) \cap \mathcal{T}_{n+1}(\mathbb{R}). \end{cases} \tag{III.10}$$

En effet, la projection selon la norme de Frobenius sur $\mathcal{S}_{n+1}^-(\mathbb{R})$ devient calculable, dès que l'on possède un algorithme efficace de calcul de valeurs propres. Mais à quoi correspond

exactement ce problème dans l'espace $\mathbb{R}^n$? En fait on va montrer qu'il est équivalent au problème

$$(\mathcal{V}ec)\begin{cases}\min\limits_{x\in\mathbb{R}^{n+1}} & \|\sqrt{\mathcal{A}\mathcal{A}^\star}(x-r)\|^2\\ & X\in\mathcal{C}^\circ_{n+1},\end{cases} \tag{III.11}$$

avec r tel que $R=\mathcal{A}^\star(r)$, en admettant que R est Toeplitz. $\mathcal{A}\mathcal{A}^\star:\mathbb{R}^{n+1}\mapsto\mathbb{R}^{n+1}$ est un opérateur symétrique positif (il est même diagonal), il admet donc une racine carrée dont la matrice n'est autre que $\mathtt{diag}(\|A^{(0)}\|,\cdots,\|A^{(n)}\|)$.

Démontrons que si $\overline{X}$ est solution de $(\mathcal{M}at)$ alors $\bar{x}=[\mathcal{A}^\star]^{-1}(X)$ est solution de $(\mathcal{V}ec)$. Si $\overline{X}$ est solution de $(\mathcal{M}at)$ ($(\mathcal{M}at)$ admet une solution, la fonction-objectif est strictement convexe et coercive sur un ensemble convexe fermé), alors $X\in\mathcal{T}_{n+1}(\mathbb{R})$, et comme $\mathcal{A}^\star$ est surjective sur $\mathcal{T}_{n+1}(\mathbb{R})$, l'élément$\bar{x}$ existe. En posant $\|x\|_M=\|\sqrt{\mathcal{A}\mathcal{A}^\star}x\|$, vérifions les conditions du théorème de Moreau pour $(\mathcal{V}ec)$:

- $\bar{x}\in\mathcal{C}^\circ_{n+1}$: en effet $\overline{X}=\mathcal{A}^\star(X)\preceq 0$;
- $\langle f-\bar{x},\bar{x}\rangle_M=0$; pour cela calculons

$$\begin{aligned}\langle f-\bar{x},\bar{x}\rangle_M &= \langle\sqrt{\mathcal{A}\mathcal{A}^\star}(f-\bar{x}),\sqrt{\mathcal{A}\mathcal{A}^\star}(\bar{x})\rangle=\langle\mathcal{A}\mathcal{A}^\star(f-\bar{x}),\bar{x}\rangle\\ &= \langle\langle\mathcal{A}^\star(f-\bar{x}),\mathcal{A}^\star(\bar{x})\rangle\rangle=\langle\langle F-\overline{X},\overline{X}\rangle\rangle=0,\end{aligned}$$

  car $\overline{X}$ est la projection de F sur le convexe $\mathcal{S}^-_{n+1}(\mathbb{R})\cap\mathcal{T}_{n+1}(\mathbb{R})$;
- $f-\bar{x}\in(\mathcal{C}^\circ_{n+1})^\circ$ (le polaire au sens de $\langle\cdot,\cdot\rangle_M$) ; pour cela, on utilise la même condition dans le cas de $(\mathcal{M}at)$

$$\forall S\in\mathcal{S}^-_{n+1}(\mathbb{R})\cap\mathcal{T}_{n+1}(\mathbb{R})\ \ \langle\langle S,F-\overline{X}\rangle\rangle\leqslant 0;$$

  on peut écrire $S=\mathcal{A}^\star(s)$ avec $\mathcal{A}^\star(s)\preceq 0$ de sorte que

$$\langle\langle\mathcal{A}^\star(s),\mathcal{A}^\star(f-\bar{x})\rangle\rangle\leqslant 0,$$

  et par conséquent

$$\langle s,\mathcal{A}\mathcal{A}^\star(f-\bar{x})\rangle=\langle s,f-\bar{x}\rangle_M\leqslant 0 \text{ pour tout } s\in\mathcal{C}^\circ_{n+1}.$$

On déduit de cela que $f=\bar{x}+f-\bar{x}$ réalise la décomposition de Moreau selon $\langle\cdot,\cdot\rangle_M$ pour $\mathcal{C}^\circ_{n+1}$ et son cône polaire. Par conséquent $\bar{x}$ est bien la solution de $(\mathcal{V}ec)$. On peut voir alors que $\bar{y}=\sqrt{\mathcal{A}\mathcal{A}^\star}\bar{x}$ est solution de :

$$\begin{array}{ll}\min\limits_{y\in\mathbb{R}^{n+1}} & \|y-\sqrt{\mathcal{A}\mathcal{A}^\star}r\|^2\\ & y\in\sqrt{\mathcal{A}\mathcal{A}^\star}\mathcal{C}^\circ_{n+1};\end{array} \tag{III.12}$$

D'après le Lemme II.6, comme $\sqrt{\mathcal{A}\mathcal{A}^\star}$ est auto-adjoint et bijectif,

$$\left[\sqrt{\mathcal{A}\mathcal{A}^\star}\mathcal{C}^\circ_{n+1}\right]^\circ=\left[\sqrt{\mathcal{A}\mathcal{A}^\star}\right]^{-1}\mathcal{C}_{n+1},$$

et on est alors capable de résoudre le problème :

$$\begin{array}{c} \min\limits_{z\in\mathbb{R}^{n+1}} \quad \| \left[\sqrt{\mathcal{A}\mathcal{A}^\star}\right]^{-1} (z - \mathcal{A}\mathcal{A}^\star r)\|^2 \\ z \in \mathcal{C}_{n+1}. \end{array} \tag{III.13}$$

C'est le problème de projection sur $\mathcal{C}_{n+1}$ pour la norme euclidienne définie par

$$\|x\|_w = \sqrt{\sum_{i=0}^{n} \left(\frac{x_i}{n+1-i}\right)^2}.$$

Cette norme euclidienne donne d'autant plus de poids à une coordonnée que son indice est élévé. Nous ne savons pas actuellement si un tel problème est pertinent du point de vue du Traitement du Signal ; par contre, les poids attribuées à chacune des coordonnées par cette norme rendent compte d'une certaine manière d'une réalité géométrique de $\mathcal{C}_{n+1}$: en se plaçant sur le compact $\mathcal{U}_n$ défini par la proposition II.1, plus une coordonnée a un indice élevé, moins son amplitude maximale

$$\Delta_i = \max\{\mathcal{A}(xx^\top)_i \,|\, x \in \mathbb{R}^{n+1}\} - \min\{\mathcal{A}(zz^\top)_i \,|\, z \in \mathbb{R}^{n+1}\},$$

est élevée d'après le Théorème II.2.

III.4 Algorithme de relaxation non-convexe

D'après ce qui précède, l'algorithme de suivi de chemin est la méthode la plus précise et efficace pour calculer la projection sur le cône $\mathcal{C}_{n+1}$. Cependant, en grande dimension, les premières itérations sont tellement coûteuses que le temps de calcul devient prohibitif. Dans certaines applications, trouver exactement le minimum global n'est pas forcément l'objectif, et avoir une solution assez proche du minimum peut être déjà suffisamment intéressant.

Suivant la définition II.1, $x \in \mathcal{C}_{n+1}$ si et seulement si : il existe $y \in \mathbb{R}^{n+1}$ tel que $x = \mathcal{A}(yy^\top)$. En utilisant cette paramétrisation, le problème de projection devient

$$(\mathcal{NC}) \left\{ \begin{array}{ll} \min\limits_{y\in\mathbb{R}^{n+1}} & \frac{1}{2}\|\mathcal{A}(yy^\top) - c\|^2. \end{array} \right.$$

Considérons la fonction $f_c : y \mapsto \frac{1}{2}\|\mathcal{A}(yy^\top) - c\|^2$. **Alors le problème de projection revient à minimiser sans contraintes la fonction f_c sur $\mathbb{R}^{n+1}$.** Le problème majeur provient alors de la non-convexité de f_c qui peut faire apparaître des minima locaux.

III.4.0.1 Différentielles de f_c

f_c est une fonction polynomiale de plusieurs variables et, par conséquent, elle est $\mathcal{C}^\infty$ sur $\mathbb{R}^{n+1}$ et son développement de Taylor en tout point est fini. Pour calculer rapidement

les différentielles de f_c, il suffit de calculer $f_c(y+h)$ et d'identifier les différents termes correspondants aux puissances de h ; ainsi

$$\begin{aligned} f_c(y+h) &= \frac{1}{2}\|\mathcal{A}((y+h)(y+h)^\top) - c\|^2 \\ &= \frac{1}{2}\|\mathcal{A}(yy^\top) - c + 2\mathcal{A}(yh^\top) + \mathcal{A}(hh^\top)\|^2 \\ &= f_c(y) + \langle \mathcal{A}(yy^\top) - c, 2\mathcal{A}(yh^\top) + \mathcal{A}(hh^\top)\rangle + \frac{1}{2}\|2\mathcal{A}(yh^\top) + \frac{1}{2}\mathcal{A}(hh^\top)\|^2 \\ &= f_c(y) + 2\langle \mathcal{A}^\star(\mathcal{A}(yy^\top) - c)y, h\rangle + \langle \mathcal{A}^*(\mathcal{A}(yy^\top) - c)h, h\rangle + 2\|\mathcal{A}(yh^\top)\|^2 + R(h), \end{aligned}$$

où $R(h)$ contient uniquement des termes en h d'ordre supérieur ou égal à 3. Dans ce développement, on peut reconnaître facilement le gradient (en y)

$$\nabla f_c(y) = 2\mathcal{A}^\star(\mathcal{A}(yy^\top) - c)y,$$

et la matrice hessienne de f_c (toujours en y)

$$D^2 f_c(y)[h,h] = \langle \mathcal{A}^\star(\mathcal{A}(yy^\top) - c)h, h\rangle + 2\|\mathcal{A}(yh^\top)\|^2.$$

En utilisant la définition de $\mathcal{A}$, il vient

$$\|\mathcal{A}(yh^\top)\|^2 = \sum_{i=0}^{n} \langle A^{(i)}y, h\rangle^2,$$

et puisque

$$\langle A^{(i)}y, h\rangle^2 = h^\top A^{(i)}yy^\top A^{(i)}h = \langle (A^{(i)}y)(A^{(i)}y)^\top h, h\rangle,$$

on en vient à conclure que

$$\|\mathcal{A}(yh^\top)\|^2 = \langle \mathcal{V}(y)\mathcal{V}(y)^\top h, h\rangle,$$

où $\mathcal{V}$ est un opérateur défini par

$$\mathcal{V}(y) = \left[A^{(0)}y, A^{(1)}y, \cdots, A^{(n)}y\right]$$

lequel peut se décomposer en $\mathcal{V}(y) = \frac{1}{2}(T(y) + H(y))$, avec

$$T(y) = \begin{pmatrix} y_0 & 0 & \cdots & 0 \\ y_1 & y_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ y_n & \cdots & y_1 & y_0 \end{pmatrix} \text{ et } H(y) = \begin{pmatrix} y_0 & \cdots & y_{n-1} & y_n \\ \vdots & \iddots & \iddots & 0 \\ y_{n-1} & y_n & \iddots & \vdots \\ y_n & 0 & \cdots & 0 \end{pmatrix}.$$

La matrice hessienne s'écrit donc alors

$$\nabla^2 f_c(y) = \mathcal{A}^\star(\mathcal{A}(yy^\top) - c) + 2\mathcal{V}(y)\mathcal{V}(y)^\top.$$

III.4.0.2 Propriétés de f_c

Comme on l'a vu, f_c est une fonction polynomiale de plusieurs variables, mais du fait de sa définition particulière, elle présente des propriétés et des symétries additionnelles intéressantes :

Proposition III.1: **i)** f_c *est invariante par symétrie centrale, i.e.* $f_c(-x) = f_c(x)$;

ii) f_c *est invariante par retournement, i.e.* $f_c(x_0, \cdots, x_n) = f_c(x_n, x_{n-1}, \cdots, x_0)$;

iii) *si* $c \in \mathcal{C}^\circ_{n+1}$ *alors* f_c *est convexe et* 0 *est son unique mimimum global.*

Démonstration. **i)** On a $f_c = \frac{1}{2}\|\cdot - c\|^2 \circ g$ avec $g(y) = \mathcal{A}(yy^\top)$; or $g(-y) = \mathcal{A}((-y)(-y)^\top) = \mathcal{A}(yy^\top) = g(y)$.

ii) On a de même $g(Jy) = g(y)$, où J est la matrice d'échange ou de retournement

$$J = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ \vdots & \iddots & \iddots & 0 \\ 0 & \iddots & \iddots & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix}.$$

car les $A^{(i)}$ étant des Toeplitz symétriques, elles sont par conséquent *centro-symétriques* ; par définition une matrice M est centro-symétrique si et seulement si elle commute avec la matrice d'échange J, i.e.

$$JM = MJ;$$

Certains auteurs [9, 52] ont étudié les propriétés de ces matrices, et présentent les matrices Toeplitz symétriques comme exemples particuliers de cette classe de matrice. Ceci implique que $\langle A^{(i)}Jy, Jy\rangle = \langle A^{(i)}y, y\rangle$ car J est symétrique.

On peut démontrer au passage grâce à l'article [12] de Chu, que les seules transformations linéaires F telles que $g \circ F = g$ sont I, $-I$, J et $-J$, car il y démontre que le groupe de stabilité de $\mathcal{T}_{n+1}(\mathbb{R})$ a huit éléments exactement (dont les quatre précédents).

iii) Supposons que $c \in \mathcal{C}^\circ_{n+1}$; alors $\mathcal{A}^\star(-c) \succeq 0$. Ainsi, pour $h \in \mathbb{R}^{n+1}$, on peut calculer

$$\begin{aligned} \langle \nabla^2 f_c(y)h, h\rangle &= \langle \mathcal{A}^\star(\mathcal{A}(yy^\top) - c)h, h\rangle + 2\langle \mathcal{V}(y)\mathcal{V}(y)^\top h, h\rangle \\ &= \langle \mathcal{A}(yy^\top), \mathcal{A}(hh^\top)\rangle + \underbrace{2\langle \mathcal{V}(y)\mathcal{V}(y)^\top h, h\rangle}_{\geqslant 0} + \underbrace{\langle \mathcal{A}^\star(-c)h, h\rangle}_{\geqslant 0}, \end{aligned}$$

et voir que le premier terme est un produit scalaire entre deux éléments de $\mathcal{C}_{n+1}$, qui est donc positif d'après la proposition II.8 ; on conclut que pour tout $y \in \mathbb{R}^{n+1}$ la matrice hessienne $\nabla^2 f_c(y)$ est semidéfinie positive, et que donc f_c est convexe. Tout minimum local $\bar{y}$ de f_c est un minimum global. L'origine vérifie trivialement la condition d'optimalité $\nabla f_c(0) = 0$, par conséquent c'est un minimum local et global. Il est unique, car $\frac{1}{2}\|\cdot - c\|^2$ est strictement convexe et l'unique antécédent de 0 par g c'est 0. □

III.4.0.3 Conditions d'optimalité de $\mathcal{NC}$

Le problème d'optimisation ($\mathcal{NC}$) peut être vu comme une version affaiblie du problème de projection (III.3). En effet, on peut écrire les conditions d'optimalité du problème de projection, ce qui donne

$$0 \in \partial(\frac{1}{2}\|\cdot - c\|^2))(\bar{x}) + N(\mathcal{C}_{n+1}, \bar{x}),$$

conduisant aux conditions de la décomposition de Moreau

$$\begin{cases} \bar{x} \in \mathcal{C}_{n+1} \\ c - \bar{x} \in \mathcal{C}_{n+1}^{\circ} \\ \langle c - \bar{x}, \bar{x} \rangle = 0. \end{cases}$$

Ce qui est au-dessus équivaut à : $\exists y \in \mathbb{R}^{n+1}$ tel que

$$\begin{cases} \bar{x} = \mathcal{A}(yy^\top) \\ \mathcal{A}^\star(\mathcal{A}(yy^\top) - c) \succeq 0 \\ \langle c - \mathcal{A}(yy^\top), \mathcal{A}(yy^\top) \rangle = 0. \end{cases}$$

Ecrivons maintenant les conditions d'optimalité du premier et du second ordre vérifiées en un minimiseur local $\tilde{y}$ de ($\mathcal{NC}$) :

$$\begin{cases} \mathcal{A}^\star(\mathcal{A}(\tilde{y}\tilde{y}^\top) - c)\tilde{y} = 0 \\ \mathcal{A}^\star(\mathcal{A}(\tilde{y}\tilde{y}^\top) - c) + 2\mathcal{V}(\tilde{y})\mathcal{V}(\tilde{y})^\top \succeq 0. \end{cases}$$

Alors, si on pose $\tilde{x} = \mathcal{A}(\tilde{y}\tilde{y}^\top)$, on voit que $\tilde{x}$ vérifie la première et la troisième condition d'optimalité du problème de projection, et, par contre, la seconde est remplacée par

$$\mathcal{A}^\star(\tilde{x} - c) \succeq -2\mathcal{V}(\tilde{y})\mathcal{V}(\tilde{y})^\top.$$

C'est en ce sens que l'on qualifie ce problème ($\mathcal{NC}$) de *relaxation non-convexe* du problème original ($\mathcal{C}$).

III.4.1 Mise en œuvre pratique de la relaxation non-convexe

Une première approche possible consiste à utiliser un algorithme de descente du type Quasi-Newton pour résoudre l'équation $\nabla f_c(y) = 0$. Les avantages sont la simplicité de mise en œuvre et la rapidité de l'algorithme. Pour une bonne description des Méthodes Quasi-Newton, appelées aussi parfois méthodes à métrique variable comme dans [41], on pourra consulter en français la première partie de [6]. L'inconvénient majeur reste la convergence éventuelle vers des minima locaux. Pour le calcul du gradient, on peut profiter de la bonne structure de f_c et de son gradient

$$\nabla f_c = \mathcal{A}^\star(\mathcal{A}(yy^\top) - c)y.$$

Comme cela a été vu lors de la définition de l'opérateur $\mathcal{A}$, on sait que

$$\mathcal{A}(xy^\top) = \frac{1}{2}(\mathtt{corr}_a(x,y) + \mathtt{corr}_a(y,x)),$$

ce qui permet de calculer rapidement f_c en utilisant une corrélation par FFT. Mais en exploitant la structure du gradient, on peut aussi tirer parti du fait que $\mathcal{A}^\star(x)$ est une matrice Toeplitz. La multiplication matrice-vecteur rapide ($\mathcal{O}(n\log n)$) pour une matrice Toeplitz est une chose bien connue, mais dans notre cas, on peut démontrer directement le résultat que voici.

Lemme III.1: *Pour tout* $x, y \in \mathbb{R}^{n+1}$, $\mathcal{A}^\star(x)y = \frac{1}{2}\left(J\mathtt{corr}_a(x, Jy) + \mathtt{corr}_a(x,y)\right)$, *où* $J_{kl} = [k+l=n]$ *est la matrice d'échange.*

Démonstration. . On a :

$$\mathcal{A}^\star(x)y = \sum_{i=0}^{n} x_i A^{(i)} y = \frac{1}{2}\sum_{i=0}^{n} x_i E^i y \;+\; x_i (E^i)^\top y;$$

or

$$\left((E^i)^\top y\right)_j = \sum_{k=0}^{n} [k = j+i] y_k = y_{j+i}[j+i \leqslant n],$$

d'où

$$\left(\sum_{i=0}^{n} x_i (E^i)^\top y\right)_j = \sum_{i=0}^{n-j} x_i y_{j+i} = \mathtt{corr}_a(x,y)_j.$$

Si on considère la matrice d'échange $J = [J_{kl}]_{k=0,l=0}^{n}$ avec le terme général précédent, alors

$$J(E^i)^\top J = E^i \tag{III.14}$$

et on en déduit :

$$\begin{aligned}\mathcal{A}^\star(x)y &= \frac{1}{2}\sum_{i=0}^{n} x_i J(E^i)^\top Jy + \sum_{i=0}^{n} x_i (E^i)^\top y \\ &= \frac{1}{2}(J\mathtt{corr}_a(x, Jy) + \mathtt{corr}_a(x,y)).\end{aligned}$$

□

Il s'ensuit directement une méthode d'évaluation du gradient ∇f_c et de f_c en $\mathcal{O}(n\log n)$:

1. on calcule $z = \mathcal{A}(yy^\top)$ grâce à deux FFTs à un coût de $\mathcal{O}(n\log n)$;
2. on calcule $x = z - c$ qui a un coût linéaire ;
3. on calcule $f_c(y)$ comme la norme au carré de x ($\mathcal{O}(n)$) ;
4. on calcule $\mathtt{corr}_a(x,y)$ par une FFT($\mathcal{O}(n\log n)$) ;
5. on calcule $J\mathtt{corr}_a(x, Jy)$ aussi par FFT (même coût).

Le tout a donc une complexité de $\mathcal{O}(n\log n)$.

III.4.2 Comparaison entre les algorithmes de suivi de chemin et celles traitant la relaxation non-convexe

La relaxation semble *a priori* une alternative en grande dimension aux algorithmes de suivi de chemin, car il n'y a pas de matrice hessienne à calculer et l'évaluation du gradient est vraiment peu coûteuse. Aussi il convient de tester son efficacité en dimension moyenne sur des exemples pour lesquels on connaît exactement la projection que l'on cherche.

Pour cela, on peut construire une solution artificielle de la façon suivante. D'après le théorème de décomposition de Moreau, un vecteur $c \in \mathbb{R}^{n+1}$ se décompose de manière unique sous la forme :

$$c = \underbrace{c_P}_{\in \mathcal{C}_{n+1}} + \underbrace{c_D}_{\in \mathcal{C}_{n+1}^\circ},$$

avec $\langle c_D, c_P \rangle = 0$. Le cas intéressant est lorsque $c \notin \mathcal{C}_{n+1} \cup \mathcal{C}_{n+1}^\circ$, avec les projections c_P et c_D qui appartiennent respectivement aux frontières $\partial\mathcal{C}_{n+1}$ et $\partial\mathcal{C}_{n+1}^\circ$. Alors, nous allons construire artificiellement c_D, puis c_P et la somme c. Considérons

$$x = (-2\sum_{i=1}^{n} |x_i| - 1, x_1, \cdots, x_n),$$

avec $(x_1, \cdots, x_n)$ choisis aléatoirement ; $\mathcal{A}^\star(x)$ est une matrice à diagonale dominante et

$$\lambda_{\max}(\mathcal{A}^\star(x)) = \max \operatorname{spec}(\mathcal{A}^\star(x))$$

est négative. Par conséquent

$$c_D = x - \lambda_{\max}(\mathcal{A}^\star(x))e_1,$$

est tel que

$$c_D \in \partial\mathcal{C}_{n+1}^\circ.$$

On choisit $y \in \ker \mathcal{A}^\star(c_D)$ et $c_P = \mathcal{A}(yy^\top)$; alors $c = c_P + c_D$ a pour projections c_P et c_D !

On a ainsi testé notre implémentation basée sur le code M1QN3 (cf. [23]), sur une série de problèmes de tailles faibles à moyennes, en évaluant l'erreur commise par chacune des méthodes :

n	err_{ipm}	err_{qn}	$\lambda_{max}(\mathcal{A}^\star(c - x_{qn}))$	$T_{ipm}(s)$	$T_{qn}(s)$
10	2.14e-10	4.07e-09	-1.88e-07	0.23	0.01
50	7.6e-11	5.33e-09	-1.13e-06	1.88	0.05
100	5.99e-11	5.77e-06	-0.001	9.95	0.06
150	1.20e-10	0.0003	-0.0352	51.33	0.07
200	6.37e-11	2.31e-07	2.90e-05	65.55	0.07
250	4.24e-11	4.94e-08	4.69e-06	150.39	0.09
300	5.55e-11	0.0042	-2.21	440.32	0.12

L'erreur affichée ici est une erreur relative, i.e.

$$\mathrm{err}_M = \|x_M - x_{opt}\|/\|x_M\|.$$

On constate que la méthode de points intérieurs est toujours très précise, par contre son temps de calcul augmente rapidement avec n, tandis que la méthode de Quasi-Newton présente des temps de calcul faibles. Les erreurs relatives peuvent paraître grandes, pour certains cas, mais dans un contexte applicatif, une erreur de 4/1000 est souvent acceptable. Pour le calcul de la plus grande valeur propre de $\mathcal{A}^\star(c-x)$, on a utilisé la fonction "spec" de Scilab qui est basée sur les fonctions DGEEV et ZGEEV de Lapack. Cependant, en grande dimension, $n > 1000$, la confiance numérique dans ces fonctions diminue, et il faudrait plutôt se tourner vers des méthodes itératives comme dans ARPACK (surtout en raison de la structure Toeplitz de la matrice).

Nous avons ensuite fait des tests avec uniquement l'algorithme de Quasi-Newton. On n'a pas comparé à des solutions artificielles comme précédemment, car leur construction en grande dimension nécessiterait de développer du logiciel spécifique. Par contre, nous avons comparé la valeur optimale $\bar{f}_c$ avec la valeur d'initialisation de l'algorithme pour voir dans quelle mesure la solution fournie par l'algorithme diminuait significativement la fonction-objectif. Nous avons aussi reporté la norme du gradient, afin d'estimer avec quelle précision la condition d'optimalité du premier ordre était vérifiée.

N	$\bar{f}_c$	$\|\nabla f_c\|$	$\bar{f}_c/f_c(x_0)$	T(s)
500	22.616042	0.000001	1.891e-05	1.01
1000	49.460434	0.000001	5.102e-06	2.75
1500	68.837221	0.000065	1.833e-06	4.78
2000	91.484743	0.000523	9.756e-07	6.19
2500	113.866738	0.000223	6.569e-07	8.15
3000	138.995922	0.000417	4.748e-07	9.22
3500	161.132715	0.000449	3.522e-07	10.74
4000	188.095904	0.004213	2.778e-07	12.25
4500	212.195746	0.014465	2.301e-07	15.65
5000	237.345775	0.018066	1.808e-07	16.69

Comme premier commentaire de ces résultats, on peut tout d'abord voir que les temps de calculs restent acceptables. En fait, même en dimension moyenne ($N = 500$), ils sont plus grands que précédemment, parce que l'on pas réglé le nombre maximum d'itérations en fonction de la dimension mais fixé ce nombre à une valeur arbitraire ($n_{max} = 600$). Le gradient prend des valeurs non négligeables pour $n = 5000$, de l'ordre de 2%, mais il faut voir que la dimension des vecteurs devient importante pour relativiser ce résultat. Enfin, le ratio $\bar{f}_c/f_c(x_0)$ montre que l'algorithme minimise quand même bien la fonction par rapport à la solution aléatoire brute $\tilde{f}_c = f_c(x_0)$.

Une question qui se pose après l'utilisation de cette méthode est la suivante : l'algorithme fournit un point réalisable qui, on l'espère, n'est pas trop loin de l'optimum cherché. Or, dans la méthode de Points Intérieurs, ce sont les premières itérations qui sont les plus coûteuses en

temps Processeur : serait-il alors possible d'utiliser la relaxation non-convexe pour initialiser la méthode de Points Intérieurs ? Nous avons essayé ainsi d'initialiser la méthode de Points intérieurs, avec le résultat de la relaxation non-convexe. Malheureusement, comme on pouvait l'anticiper, cela n'accélère pas les premières itérations. Une itération externe dans la Méthode de Points Intérieurs est initialisée avec un point approximativement solution (i.e. $\lambda(F_t, y) \leqslant \kappa$) du problème de centrage

$$(C_t) \quad \min_x \quad F_t(x) = t\langle c, x\rangle + F(x).$$

Le problème est donc de savoir à quelle valeur de t correspondrait une solution $\bar{x} = \mathcal{A}(\bar{y}\bar{y}^\top)$ du problème $(\mathcal{NC})$. Pour cela il faudrait au moins connaître des estimations sur la distance entre la solution $\bar{x}$ et l'optimum, estimations qui nous échappent pour l'instant.

Chapitre IV

Extensions Bidimensionnelles

Dans les précédents chapitres, on a considéré le cône des fonctions d'autocorrélations d'un signal discret à support fini. Ce signal étant indexé par $\mathbb{N}$ ou $\mathbb{Z}$, on parle alors de signal unidimensionnel. Par extension, on peut s'intéresser à des signaux bi-dimensionnels (indexés par $\mathbb{N}^2$) voire en dimension supérieure (indexés par $\mathbb{N}^d$). Dans ce cas, la notion de temps n'est plus pertinente mais, en contrepartie, on peut s'intéresser à des signaux plus complexes tels que des images : par exemple, on peut considérer une image en niveau de gris de taille $m \times n$, comme un élément de $\mathbb{C}^{\mathbb{N}^2}$ ou plus "réellement" de $[0,1]^{\{1,\cdots,m\}\times\{1,\cdots,n\}}$. Si cette extension élargit notablement le champ des applications de ce modèle, c'est au prix d'une complexification de la modélisation et de la résolution numérique de certains problèmes. Ainsi la première difficulté à laquelle nous allons être confrontés provient de la définition même du modèle, dont le choix peut s'effectuer, comme nous allons le voir, entre plusieurs possibilités. Enfin, comme précisé dans le titre, nous nous limiterons dans la suite au cas $d = 2$, principalement pour deux raisons : les énoncés sont plus simples, et il est encore possible d'effectuer les calculs ; ainsi certains auteurs donnent des formulations en dimension d quelconque, mais se limitent au cas $d = 2$ dans les applications, en raison d'une complexité trop grande dans les calculs numériques.

Comme rappelé précédemment, l'étude des polynômes trigonométriques positifs à plusieurs variables s'est relativement bien développée dans la dernière décennie avec, notamment, les travaux de Megretski [37], Woerdeman [36], Dritschel [17] pour la théorie, et Dumitrescu [18], Hachez [28] et Roh [51] pour les applications. Cette théorie constitue un pendant à la théorie des polynômes positifs qui serait en quelque sorte ce que sont les matrices Toeplitz aux matrices Hankel. On a ainsi une sorte de dualité entre ces deux théories et, de manière générale, on constate que le contexte des polynômes trigonométriques semble plus stable numériquement du fait de la compacité de $\mathbb{T}$ en regard du caractère non-compact de $\mathbb{R}$, de la même manière que les matrices Toeplitz semblent présenter un meilleur conditionnement numérique que les matrices Hankel.

IV.1 Quelle généralisation choisir ?

Dans le chapitre 1, nous avions donné deux définitions ou formulations de $\mathcal{C}_{n+1}$ que l'on peut voir *grosso modo* comme : pour la première, le fait pour un polynôme d'être un "carré", et pour l'autre le fait d'être positif. Grâce au théorème de Riesz-Féjer, il nous était possible de confondre ces deux objets. Malheureusement, en dimension supérieure, comme le théorème de D'Alembert n'a pas d'équivalent, il n'y pas de généralisation connue du théorème de Riesz-Féjer : on peut par exemple trouver des polynômes trigonométriques à deux variables positifs qui ne s'écrivent pas comme le module d'un carré[1]. On a donc le choix *a priori* entre deux généralisations possibles : l'ensemble des vecteurs dont les coefficients sont des fonctions d'autocorrélations et l'ensemble des vecteurs qui sont les coefficients de polynômes trigonométriques positifs. En réalité, comme nous allons le voir, il y aurait trois choix possibles de généralisations : les fonctions d'autocorrélations, les polynômes sommes de carrés et les polynômes positifs. On va maintenant détailler ces différentes possibilités

IV.1.1 Corrélation de deux matrices

Définition IV.1: *Soit* (X, Y) *deux matrices de* $\mathcal{M}_{m,n}(\mathbb{R})$ *; on définit alors* la corrélation acyclique *de* X *et* Y *comme un élément de* $\mathcal{M}_{m,n}(\mathbb{R})$ *dont le terme général est :*

$$\mathtt{corr}_a(X, Y)_{ij} = \sum_{k=0}^{m-1-i} \sum_{l=0}^{n-1-j} X_{kl} Y_{(k+i)(l+j)}.$$

Notons que si l'on utilise le produit de Kronecker de deux matrices défini par

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \dots & \ddots & \vdots \\ a_{n1}B & \cdots & a_{nn}B \end{pmatrix},$$

alors

$$\mathtt{corr}_a(X, Y)_{ij} = \langle\langle \left(E^{n,j} \otimes E^{m,i}\right)^\top, \mathtt{vec}(X)\mathtt{vec}(Y)^\top \rangle\rangle,$$

où : $E^{m,j} \in \mathcal{M}_m(\mathbb{R})$ est la matrice du $j^{\text{ème}}$ décalage à droite définie comme au chapitre 2 ; vec est l'isométrie identifiant $\mathcal{M}_{m,n}(\mathbb{R})$ et $\mathbb{R}^{mn}$ par concaténation des colonnes à la suite des autres dans un grand vecteur de taille mn :

$$x = \mathtt{vec}(X) = [x_{11}, \cdots, x_{m1}, x_{12}, \cdots, x_{m2}, \cdots, x_{1n}, \cdots, x_{mn}]^\top.$$

Comme il s'agit d'une isométrie, elle est bijective et son inverse $\mathtt{vec}^{-1}$ est bien définie elle aussi mais son expression est plus compliquée, car il faut utiliser des restes et des divisions entières

[1] Dans son livre [45] p. 209, Rudin démontre que le polynôme trigonométrique $1 + \delta(\cos\omega_1 + \cos\omega_2)$ avec $\delta < 1/2$ ne peut pas se factoriser.

pour trouver la position $(i(k), j(k))$ dans la matrice antécédent d'un coefficient d'indice k donné.

Il existe une identité très utile, que nous désignerons dans la suite comme l'*identité d'aplatissement du produit de Kronecker* permettant de ramener un double produit matriciel (à droite et a gauche) à un produit matrice-vecteur dans $\mathbb{R}^{m,n}$ en utilisant $\otimes$,

$$\mathrm{vec}(AXB) = (B^\top \otimes A)\mathrm{vec}X.$$

Pour une démonstration, voir [32] p. 254-255. Dans la suite, nous adopterons la convention suivante : on réservera la notation minuscule aux éléments $x \in \mathbb{R}^{mn}$ et la notation majuscule X à la matrice correspondante de $\mathcal{M}_{m,n}(\mathbb{R})$. On se permettra quelques fois l'abus de notation consistant à échanger x et X, comme argument ou image d'une application, étant entendu qu'il faut composer par vec ou vec^{-1} pour que cela fonctionne bien.

IV.1.2 Cône généralisé d'autocorrélation

La généralisation du cône d'autocorrélation que nous allons choisir correspond à la Définition II.1 que nous avions donnée dans le cas unidimensionnel.

Définition IV.2: *On appelle cône des matrices à composantes autocorrélées, le sous ensemble de $\mathcal{M}_{m,n}(\mathbb{R})$ défini par*

$$\mathcal{C}_{m,n} = \{\mathrm{corr}_a(Y, Y) \mid Y \in \mathcal{M}_{m,n}(\mathbb{R})\}.$$

Notons au passage que l'appellation de cône est justifiée par la bilinéarité de l'application $\mathrm{corr}_a(\cdot,\cdot)$.

Si on définit aussi l'opérateur $\mathcal{A} : \mathcal{M}_{mn}(\mathbb{R}) \to \mathcal{M}_{m,n}(\mathbb{R})$ par

$$\mathcal{A}(M) = \sum_{i=0}^{m-1}\sum_{j=0}^{n-1} \langle\langle A^{ij}, M\rangle\rangle e_{ij},$$

où

$$A^{ij} = \frac{1}{2}\left(E^{n,j} \otimes E^{m,i} + \left(E^{n,j} \otimes E^{m,i}\right)^\top\right),$$

la composante (p, q) de $\mathcal{A}(xx^\top)$ a donc pour valeur $\mathrm{corr}_a(x, x)_{pq}$ et, finalement,

$$\mathcal{C}_{m,n} = \{\mathcal{A}(yy^\top) \mid y \in \mathbb{R}^{mn}\}.$$

IV.1.3 Polynômes Sommes de Carrés (SOS, pour Sums of Squares)

L'approche de la positivé par des décompositions en somme de carrés est très présente dans le cadre des polynômes positifs classiques(cf. [34, 43]). Elle a son équivalent pour les polynômes trigonométriques. Cependant, avant de parler de somme de carrés, précisons ce que l'on entend par "polynôme trigonométrique".

Définition IV.3: *On appelle polynôme trigonométrique de degré* $(\max_{s\in S}|s_1|, \max_{s\in S}|s_2|)$ *toute fraction rationnelle définie pour* $z \in \mathbb{T}^2$ *par*

$$P_x(z) := \sum_{k\in S} x_k z^k = \sum_{(k_1,k_2)\in S} x_{(k_1,k_2)} z_1^{k_1} z_2^{k_2},$$

où S *est un ensemble fini de* $\mathbb{Z}^2$. *Si, de plus, pour tout* $k \in S$, $x_k = x_{-k}$ *alors on dit que* P_x *est* **pair**.

La condition de symétrie provient du cas complexe où $x_k = \bar{x}_{-k}$, ce qui permet d'avoir $P_x(z) \in \mathbb{R}$ pour $z \in \mathbb{T}^2$ et de pouvoir ainsi parler de positivité pour P : en effet, si $z = e^{i\omega_1 + i\omega_2}$ alors

$$P_x(z) = \sum_{k=-n}^{n} x_k e^{\langle k,\omega\rangle} = x_{00} + 2 \sum_{\substack{k=-n \\ k\neq 0}}^{n} x_k \cos\langle k, \omega\rangle \in \mathbb{R}.$$

Dans la suite, on écrira arbitrairement $P_x(z)$ ou $P_x(\omega)$ pour désigner un polynôme trigonométrique ; alors pour $\omega \in [0, 2\pi]^2$, en posant

$$\nu(\omega) = (1, 2\cos\omega_1, \cdots, 2\cos(m-1)\omega_1, 2\cos\omega_2, \cdots, 2\cos((m-1)\omega_1 + (n-1)\omega_2)),$$

on réécrira un polynôme trigonométrique pair de coefficients $\{x_{ij}\}_{i=0,j=0}^{i=m-1,j=n-1}$ sous la forme

$$P_x(\omega) = \langle x, \nu(\omega)\rangle.$$

Pour un polynôme trigonométrique somme de carrés, on donne la

Définition IV.4: *Un polynôme trigonométrique* P_x *est somme de carrés s'il existe* $F_1, \cdots, F_r$, *polynômes trigonométriques de la forme*

$$F_k(z) = \sum_{l=0}^{d_k} f_l z^l \text{ pour } k = 1, \cdots, r,$$

tels que

$$P_x(z) = \sum_{k=1}^{r} F_k(z)\overline{F_k(z)} \text{ pour tout } z \in \mathbb{T}^2.$$

Pour un argument sous forme exponentielle $z = e^{i\omega}$, on a l'écriture suivante sous forme de modules au carré,

$$P_x(\omega) = \sum_{k=1}^{r} |F_k(\omega)|^2.$$

Notons que dans cette définition le degré des F_k n'est pas *a priori* fixé : comme on le verra après, un polynôme strictement positif sur $\mathbb{T}^2$ admet toujours une écriture en somme de carrés mais avec des degrés qui peuvent être très élevés. Remarquons que si l'on fixe arbitrairement

le degré maximum de chaque F_k à $(m-1, n-1)$, alors chaque carré $|F_k(z)|^2$ correspond à une contribution

$$x^k = \mathcal{A}(y^k (y^k)^\top) \tag{IV.1}$$

dans le vecteur des coefficients $x = \sum_{k=1}^{r} x^k$. En effet, pour démontrer cela, commençons par le lemme suivant

Lemme IV.1: *Si l'on pose*

$$F_y(\omega) = \sum_{k=0}^{m-1} \sum_{l=0}^{n-1} y_{kl} e^{ik\omega_1 + il\omega_2},$$

alors on a l'équivalence

$$x = \mathcal{A}(yy^\top) \Leftrightarrow \langle x, \nu(\omega) \rangle = F_y(\omega)\overline{F_y(\omega)}.$$

Démonstration. Posons

$$u_m(\omega) := \left(1, e^{i\omega}, \cdots, e^{i(m-1)\omega}\right),$$

alors avec $(\omega_1, \omega_2) \in [0, \pi]^2$, si l'on calcule

$$(u_n(\omega_2) \otimes u_m(\omega_1))^\top y y^\top \overline{u_n(\omega_2) \otimes u_m(\omega_1)},$$

on peut voir que ce produit matriciel n'est en réalité que le produit d'un complexe z_0 et de son conjugué (donc égal au réel $|z_0|^2$), où

$$z_0 = u_n(\omega_2)^\top \otimes u_m(\omega_1)^\top y = \mathtt{vec}(u_m(\omega_1)^\top Y u_n(\omega_2)),$$

moyennant l'identité d'aplatissement du produit de Kronecker. Or z_0 se réécrit en

$$\sum_{k=0}^{m-1} \sum_{l=0}^{n-1} Y_{kl} e^{ik\omega_1 + il\omega_2},$$

et donc $z_0\overline{z_0}$ vaut

$$\left(\sum_{p \in K_{m,n}} Y_p e^{i\langle p, \omega \rangle} \right) \left(\sum_{q \in K_{m,n}} Y_p e^{-i\langle q, \omega \rangle} \right) = \sum_{s=-(m,n)}^{(m,n)} \mathtt{corr}_a(Y, Y)_{|s|} e^{i\langle s, \omega \rangle},$$

où $K_{m,n} = \{0, \cdots, m-1\} \times \{0, \cdots, n-1\}$. On reconnaît alors $F_y(\omega) = \sum_{p \in K_{m,n}} Y_p e^{i\langle p, \omega \rangle}$, d'où l'égalité

$$F_y(\omega)\overline{F_y(\omega)} = \sum_{s=-(m,n)}^{(m,n)} \mathcal{A}(yy^\top)_{|s|} e^{i\langle s, \omega \rangle} = \langle x, \nu(\omega) \rangle.$$

□

En utilisant le lemme précédent, pour chaque carré F_k, on voit qu'il existe y^k tel que (IV.1) ait lieu, et en posant $Q = \sum_{k=1}^{r} y^k (y^k)^\top$, on à l'écriture

$$x = \mathcal{A}(Q),$$

où Q est une matrice de Gram, donc semidéfinie positive, ce qui correspond à la version symétrisée de la paramétrisation par trace donnée dans [18], à savoir

$$r_k = \mathtt{Tr}(T_k \cdot Q),$$

avec $T_k = E^{k_1} \otimes E^{k_2}$. Dans la suite, quand nous parlerons de polynômes trigonométriques sommes de carrés, relatifs à notre problème, nous nous restreindrons aux polynômes dont les coefficients sont dans le cône convexe

$$\mathcal{A}(\mathcal{S}^+_{mn}(\mathbb{R})) = \{\mathcal{A}(M) \mid M \succeq 0\},$$

ce qui, dans la décomposition en somme de carrés, borne automatiquement le degré des F_k à $(m-1, n-1)$.

IV.1.4 Polynômes trigonométriques positifs

Parmi les polynômes trigonométriques, ceux qui semblent pertinents dans un contexte d'optimisation sont les polynômes trigonométriques **positifs** : en effet, supposons que l'on ait défini cette notion ($P_x \geqslant 0$), alors un problème du type

$$\min_{\omega \in [0,2\pi]^2} P_x(\omega)$$

pourrait se réécrire en

$$\begin{aligned} \min_{\gamma \in \mathbb{R}, x} \quad & \gamma \\ & \gamma - P_x \geqslant 0. \end{aligned}$$

Si l'on sait algorithmiquement caractériser de tels polynômes, alors on est en mesure de résoudre des problèmes d'optimisation polynômiale posés sur le tore à deux dimensions $\mathbb{T}^2$. La méthodologie préconisée par Megretski dans [37] peut même s'étendre à des contraintes plus compliquées où le domaine (de z) lui-même n'est plus $\mathbb{T}^2$ (où $[0, 2\pi]^2$ pour ω) mais également défini grâce à un polynôme trigonométrique : par exemple,

$$2\cos\omega_1 - \cos\omega_2 \geqslant c$$

correspond à une hélice dans le domaine des ω. Cette méthodologie qui permet donc de prendre en compte des contraintes est bien décrite dans [18, 51]. La notion de polynôme trigonométrique est particulièrement simple à définir :

Définition IV.5: *Un polynôme trigonométrique (pair) sera dit positif lorsque :*

$$\forall \omega \in [0, 2\pi]^2 \, , \, P_x(\omega) \geqslant 0.$$

On notera $\mathcal{P}^+_{m,n}(\mathbb{T})$ *l'ensemble des coefficients des polynômes trigonométriques positifs de degré* (m, n)*,*

$$\mathcal{P}^+_{m,n}(\mathbb{T}) = \{x \in \mathbb{R}^{mn} \mid \langle x, v(\omega)\rangle \geqslant 0 \, , \, \forall \omega \in [0, 2\pi]^2\}.$$

La relation entre polynôme somme de carrés et polynôme positif reste assez forte : ainsi s'il évident qu'un polynôme somme de carrés est positif, l'implication réciproque est plus compliquée : un polynôme strictement positif peut toujours s'écrire comme une somme de carrés (cf. [17] Corollaire 5.2), dont les degrés peuvent être arbitrairement grands, car le corollaire en question s'appuie sur un résultat d'approximation et le degré peut donc parfois être très élevé, mais par contre il existe des polynômes seulement positifs qui ne s'écrivent pas comme somme de carrés. Il en résulte l'inclusion stricte suivante entre $\mathcal{A}(\mathcal{S}^+_{mn}(\mathbb{R}))$ et $\mathcal{P}^+_{m,n}(\mathbb{T})$,

$$\mathcal{A}(\mathcal{S}^+_{mn}(\mathbb{R})) \subsetneq \mathcal{P}^+_{m,n}(\mathbb{T}).$$

IV.2 Propriétés et polarité sur les cônes introduits

Au vu des définitions précédentes, nous voyons que trois objets mathématiques distincts, $\mathcal{C}_{m,n}$, $\mathcal{A}(\mathcal{S}^+_{mn}(\mathbb{R}))$ et $\mathcal{P}^+_{m,n}(\mathbb{T})$, ne sont en réalité qu'un seul dans le cas unidimensionnel ; il importe maintenant de décrire quelques propriétés du cône $\mathcal{C}_{m,n}$ et ses relations avec les deux autres cônes.

IV.2.1 Quelques Propriétés de $\mathcal{C}_{m,n}$

Hormis la convexité, plusieurs propriétés de $\mathcal{C}_{n+1}$ se transposent bien à $\mathcal{C}_{m,n}$: ainsi,nous avons la

Proposition IV.1: *le cône* $\mathcal{C}_{m,n}$ *est fermé, d'intérieur non vide et saillant.*

Démonstration. Soit $h : \mathcal{M}_{m,n}(\mathbb{R}) \times \mathbb{R}^{mn} \to \mathcal{M}_{m,n}(\mathbb{R})$ définie par

$$h(X, y) := X - \mathcal{A}(yy^\top).$$

Il va de soi que h est de classe C^∞ car polynomiale. Alors

$$\mathcal{C}_{m,n} = h^{-1}(\{0\}) \cap (\mathcal{M}_{m,n}(\mathbb{R}) \times \{0\}).$$

Comme h est continue, $\mathcal{C}_{m,n}$ est bien fermé comme intersections de fermés. Pour montrer que l'intérieur est non vide, on utilise la paramétrisation que voici $\mathcal{C}_{m,n}$ par

$$\mathcal{C}_{m,n} = \Phi(\mathbb{R}^{mn}) \;\text{ avec }\; \Phi(y) = \mathcal{A}(yy^\top).$$

On a

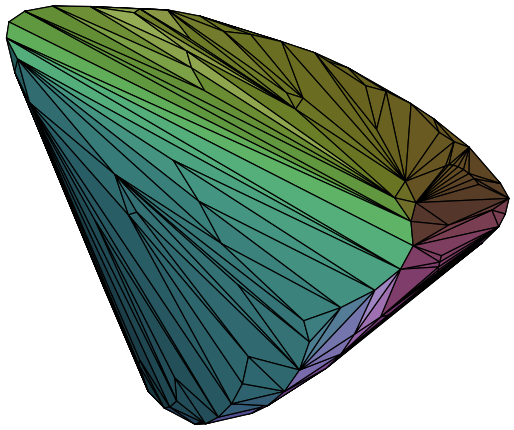
$$D_y\Phi[h] = 2\mathcal{A}(yh^\top) = \mathtt{corr}_a(y,h) + \mathtt{corr}_a(h,y),$$

et si on considère $y_0 = \mathrm{vec}(e_{00})$, alors

$$D_{y_0}\Phi[h] = 2\mathcal{A}(\mathtt{vec}(e_{00})h^\top) = \mathtt{corr}_a(y_0,h) + \mathtt{corr}_a(h,y_0) = h + h_{00}e_{00}.$$

L'application linéaire $D_{y_0}\Phi$ est clairement bijective car elle est diagonale dans la base des $\{e_{ij}\}_{i,j=0}^{i=m,j=n}$ et qu'aucun de ses termes diagonaux n'est nul. Par le théorème d'inversion locale, il existe $\mathcal{U}$ et $\mathcal{V}$ voisinages respectifs de y_0 et $\Phi(y_0)$ sur lesquels Φ réalise un difféomorphisme, ce qui assure que $\mathcal{C}_{m,n}$ contient au moins un point dont le voisinage est de "pleine dimension", et par conséquent $\mathcal{C}_{m,n}$ ne peut être d'intérieur vide. On montre ensuite que $\mathcal{C}_{m,n}$ est saillant comme à la proposition II.3. □

FIG. IV.1: Une approximation de la base $\mathcal{U}_{2,2}$



Comme $\mathcal{C}_{m,n}$ est saillant, on peut le décrire par le biais d'une base compacte $\mathcal{U}_{m,n}$ définie comme suit

$$\mathcal{U}_{m,n} = \{\mathcal{A}(yy^\top) \mid y \in \mathbb{S}_{mn-1}\}.$$

Là encore, le fait de fixer la composante $x_{00} = 1$ dans $x = \mathcal{A}(yy^\top)$ force y à être de norme un. Pour les mêmes raisons que précédemment, $\mathcal{U}_{m,n}$ est un compact de $\mathbb{R}^{mn}$, contenu dans un espace affine de dimension $mn-1$; alors, on peut en donner la plus petite approximation extérieure par des parallélépipèdes rectangles, en recherchant les plus grandes valeurs propres des A^{ij} grâce aux formulations variationnelles de Courant-Fisher pour les valeurs propres. Des tests numériques, font apparaître que les valeurs propres des A^{ij} seraient du même type que celles du théoreme II.2, et pourraient se décrire sous la forme $\cos(\frac{p}{q}\pi)$; les valeurs des

rationels $\frac{p}{q}$ qui conviennent, font l'objet de recherches actuellement de notre part. On peut tout de même visualiser une approximation de l'enveloppe convexe de $\mathcal{U}_{2,2}$ sur la figure IV.1

A la différence du cas unidimensionnel, il semble malheureusement que

$$\mathcal{A}(\mathcal{S}^+_{mn}(\mathbb{R})) \not\subset \mathcal{A}(\{xx^\top \mid x \in \mathbb{R}^{mn}\}),$$

car on ne peut plus utiliser le théorème de Riesz-Féjer pour montrer l'égalité. si on trouvait un contre-exemple illustrant cette non-inclusion, on démontrerait la non-convexité de $\mathcal{C}_{m,n}$, de par la proposition suivante qui décrit l'enveloppe convexe de $\mathcal{C}_{m,n}$.

Proposition IV.2: *L'enveloppe convexe conique de $\mathcal{C}_{m,n}$ est l'ensemble $\mathcal{A}(\mathcal{S}^+_{mn}(\mathbb{R}))$ des polynômes sommes de carrés, i.e.*

$$\mathtt{cone}(\mathcal{C}_{m,n}) = \mathcal{A}(\mathcal{S}^+_{mn}(\mathbb{R})).$$

Démonstration. Si $X \in \mathtt{cone}(\mathcal{C}_{m,n})$, alors il existe $(y_1, \cdots, y_p) \in \mathbb{R}^{mn}$ et p réels positifs $\alpha_1, \cdots, \alpha_p$ tels que

$$X = \sum_{i=1}^{p} \alpha_i \mathcal{A}(y_i y_i^\top) = \mathcal{A}\underbrace{\left(\sum_{i=1}^{p} \alpha_i y_i y_i^\top\right)}_{\in \mathcal{S}^+_{mn}(\mathbb{R})},$$

d'où la première inclusion. L'inclusion inverse est directe, en écrivant la décomposition spectrale d'un élément de $\mathcal{S}^+_{mn}(\mathbb{R})$ et en lui appliquant $\mathcal{A}$. □

On obtient alors la description *a priori* non-convexe de $\mathcal{C}_{m,n}$

$$\mathcal{C}_{m,n} = \{\mathcal{A}(M) \mid M \in \mathcal{S}^+_{mn}(\mathbb{R}) \, , \mathtt{rg}\,(M) = 1\}.$$

Notons que la relaxation $M = xx^\top \Rightarrow M \succeq 0$ est très courante en optimisation combinatoire : on appelle cela une relaxation SDP (cf. [5]).

IV.2.2 Propriétés de $\mathcal{A}(\mathcal{S}^+_{mn}(\mathbb{R}))$ et de $\mathcal{P}^+_{m,n}(\mathbb{T})$

Contrairement à $\mathcal{C}_{m,n}$, les deux cônes en question conservent le caractère convexe de la dimension un. De plus, il ont aussi les bonnes propriétés pour induire un ordre partiel (cf [5]), comme souligné dans la proposition suivante

Proposition IV.3: *Les cônes $\mathcal{A}(\mathcal{S}^+_{mn}(\mathbb{R}))$ et $\mathcal{P}^+_{m,n}(\mathbb{T})$ sont convexes, fermés, solides et saillants.*

Démonstration. Commençons par le caractère solide : comme ces deux cônes contiennent $\mathcal{C}_{m,n}$, ils sont comme lui d'intérieur non vide. Ensuite intéressons nous à $\mathcal{A}(\mathcal{S}^+_{mn}(\mathbb{R}))$. Il est évidemment convexe (l'image par une application linéaire d'un convexe est convexe). Il est fermé (on peut le démontrer facilement comme à la proposition IV.1). Il est saillant car si $x = \mathcal{A}(M)$, en écrivant

$$M = \sum_i \lambda_i v^i (v^i)^\top \text{ avec les } \lambda_i \geqslant 0$$

sous la forme d'une décomposition spectrale, il vient que $x_{00} = \sum_i \lambda_i \sum_j (v_j^i)^2 \geqslant 0$; mais si $x = \mathcal{A}(-S)$ avec $S \succeq 0$, alors, par le même raisonnement $x_{00} \leqslant 0$, et donc finalement

$$x_{00} = 0 = \sum_i \lambda_i \sum_j (v_j^i)^2,$$

ce qui implique que les v^i sont tous nuls et $x = \mathcal{A}(0) = 0$. Pour $\mathcal{P}^+_{m,n}(\mathbb{T})$ c'est plus simple, on démontre la convexité et le caractère fermé en même temps grâce à l'écriture

$$\mathcal{P}^+_{m,n}(\mathbb{T}) = \bigcap_{\omega \in [0,2\pi]^2} \{x \in \mathbb{R}^{mn} \mid \langle v(\omega), x\rangle \geqslant 0\};$$

en effet, une intersection quelconque de convexes fermés (les demi-espaces $\{x \in \mathbb{R}^{mn} | \langle v(\omega), x\rangle \geqslant 0\}$ ici) est convexe fermé. Pour le caractère saillant, on remarque que si $\langle x, v(\omega)\rangle \geqslant 0$ et $\langle x, v(\omega)\rangle \leqslant 0$, pour tout $\omega \in [0, 2\pi]^2$, alors $\langle x, v(\omega)\rangle = 0$, et l'on conclut que $x = 0$, car les fonctions $\omega \mapsto \cos(\langle k, \omega\rangle)$ sont linéairement indépendantes. □

Récapitulons les différentes relations entre les trois cônes en question ; on a

$$\mathcal{C}_{m,n} \subset \mathtt{cone}(\mathcal{C}_{m,n}) = \mathcal{A}(\mathcal{S}^+_{mn}(\mathbb{R})) \subsetneq \mathcal{P}^+_{m,n}(\mathbb{T}).$$

IV.2.3 Cône polaire de $\mathcal{C}_{m,n}$

L'enveloppe convexe de $\mathcal{C}_{m,n}$ étant l'image par l'application linéaire $\mathcal{A}$ du cône des matrices SDP, et le cône polaire de $\mathcal{C}_{m,n}$ étant le même que celui de son enveloppe convexe (les fonctions d'appui ne "voient" que l'enveloppe convexe (fermée) des ensembles), on obtient directement le cône polaire de $\mathcal{C}_{m,n}$ en calculant l'image réciproque de $\mathcal{S}^-_{mn}(\mathbb{R})$ par l'adjoint de $\mathcal{A}$ (cf. [31]).

IV.2.3.1 Opérateur adjoint de $\mathcal{A}$

Comme dans le cas monodimensionnel, la description du cône polaire de $\mathcal{C}_{m,n}$ se fait de manière très simple à l'aide de l'opérateur adjoint de $\mathcal{A}$. L'opérateur $\mathcal{A}^\star$ prend ses arguments dans $\mathbb{R}^{mn}$ (en toute rigueur ce serait $\mathcal{M}_{m,n}(\mathbb{R})$, mais grâce à vec nous identifions ces deux espaces) et prend ses valeurs dans l'espace $\mathcal{S}_{mn}(\mathbb{R})$ des matrices symétriques de tailles mn. Plus précisement, l'image par $\mathcal{A}^\star$ d'un élément de $\mathbb{R}^{mn}$ est une matrice doublement Toeplitz : c'est-à-dire qu'elle contient des blocs Toeplitz eux-mêmes agencés de manière Toeplitz. Ainsi

$$\mathcal{A}^\star(x) = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} x_{ij} A^{ij}$$

s'écrit par blocs

$$\mathcal{A}^\star([x^0, \cdots, x^{n-1}]) = \frac{1}{2} \begin{pmatrix} \mathcal{B}(x^0) & \mathcal{T}(x^1) & \cdots & \mathcal{T}(x^{n-1}) \\ \mathcal{T}(x^1)^\top & \mathcal{B}(x^0) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathcal{T}(x^1) \\ \mathcal{T}(x^{n-1})^\top & \cdots & \mathcal{T}(x^1)^\top & \mathcal{B}(x^0) \end{pmatrix},$$

où $\mathcal{T} : \mathbb{R}^m \to \mathcal{M}_m(\mathbb{R})$ a pour expression

$$\mathcal{T}(x) = \begin{pmatrix} x_0 & x_1 & \cdots & x_{m-1} \\ 0 & x_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & x_1 \\ 0 & \cdots & 0 & x_0 \end{pmatrix},$$

et $\mathcal{B}(x) = \mathcal{T}(x) + \mathcal{T}(x)^\top$. Comme précédemment on peut définir le sous-espace des matrices doublement Toeplitz

$$\mathcal{T}_{m,n}(\mathbb{R}) = \mathcal{A}^\star(\mathbb{R}^{mn}) = \{M \in \mathcal{S}_{mn}(\mathbb{R}) \mid \exists X \in \mathbb{R}^{mn} \text{ tel que } M_{ij} = X_{|\varphi(i)-\varphi(j)|}\},$$

où $\varphi : \{0, \cdots, mn-1\} \to K_{m,n}$ est la bijection définissant les indices de vec^{-1}. Pour s'en assurer, on peut calculer le terme général de $\mathcal{A}^\star(x)$:

Lemme IV.2: *Soit $x \in \mathbb{R}^{mn}$; alors le terme général $[\mathcal{A}^\star(x)]_{ij}$ de la matrice $\mathcal{A}^\star(x)$ ne dépend que de la différence $|\varphi(i) - \varphi(j)|$.*

Démonstration. On a

$$\mathcal{A}^\star(x) = \frac{1}{2} \sum_{ij} x_{ij} (E^{n,j} \otimes E^{m,i} + (E^{n,j} \otimes E^{m,i})^\top).$$

Considérons le terme général de $E^{n,j} \otimes E^{m,i}$: pour cela on compose à gauche et à droite par

$$e_s^\top E^{n,j} \otimes E^{m,i} e_t,$$

avec $k = (k_1, k_2) = \varphi(s)$ et $l = (l_1, l_2) = \varphi(t)$; donc $e_s = \mathrm{vec}(e_{k_1} e_{k_2}^\top)$ et

$$e_s E^{n,j} \otimes E^{m,i} e_t = \mathrm{vec}(E^{m,i} e_{l_1} (E^{n,j} e_{l_2})^\top)_s = [i + l_1 \leqslant m][j + l_2 \leqslant n] \mathrm{vec}(e_{i+l_1} e_{j+l_2}^\top)_s.$$

Compte tenu du fait que $\mathrm{vec}(uv^\top)_s = u_{k_1} v_{k_2}$, le terme général vaut

$$[i + l_1 \leqslant m][j + l_2 \leqslant n][i + l_1 = k_1][j + l_2 = k_2];$$

De même, comme

$$e_s^\top (E^{n,j})^\top \otimes (E^{m,i})^\top e_t = [l_1 \geqslant i][l_1 - i = k_1][l_2 \geqslant j][l_2 - j = k_2],$$

on en déduit

$$\mathcal{A}^\star(x)_{st} = \frac{1}{2} x_{k-l}[k \geqslant l] + x_{l-k}[l \geqslant k],$$

où $x \geqslant y$ est l'ordre sur $\mathbb{R}^2$ défini par $x - y \in (\mathbb{R}_+)^2$. En posant

$$y = \frac{1}{2}(2x_{00}, x_{10}, \cdots, x_{(m-1)(n-1)}),$$

pour tenir compte du cas d'égalité $l = k$, on voit que

$$\mathcal{A}^\star(x)_{st} = y_{|k-l|} = y_{|\varphi(s)-\varphi(t)|}.$$

□

IV.2.3.2 Cône polaire et cône normal de $\mathcal{C}_{m,n}$

Une fois bien décrit l'opérateur adjoint de $\mathcal{A}$, on obtient facilement une formulation par inégalités de $\mathcal{C}^\circ_{m,n}$

Proposition IV.4: *i)Le cône polaire de $\mathcal{C}_{m,n}$ admet la description sous forme d'inégalité linéaire matricielle (LMI) suivante :*

$$\mathcal{C}^\circ_{m,n} = \{X \in \mathcal{M}_{m,n}(\mathbb{R}) \mid \mathcal{A}^\star(X) \preceq 0\}.$$

ii) Le cône normal à $\mathcal{C}_{m,n}$ en Y est

$$\mathcal{N}(\mathcal{C}_{m,n}, Y) = \{X \in \mathcal{M}_{m,n}(\mathbb{R}) \mid \mathcal{A}^\star(X) \preceq 0, \langle\langle X, Y\rangle\rangle = 0\}.$$

Démonstration. Compte tenu de la proposition IV.2, la détermination du cône polaire se fait directement grâce à l'opérateur adjoint. L'obtention du cône normal en un point Y de ce cône consiste classiquement à couper le cône polaire par l'hyperplan d'équation $\{\langle\langle Y, \cdot\rangle\rangle = 0\}$. □

Pour l'instant, nous ne connaissons pas de formulation par générateurs de $\mathcal{C}^\circ_{m,n}$. Dans le cas unidimensionnel, on avait $\mathcal{C}^\circ_{n+1} = \text{cone}(\{-v(\omega) \mid \omega \in [0,\pi]\}$; dans le cas multivariable, c'est le cône $\mathcal{P}^+_{m,n}(\mathbb{T})^\circ \subset \mathcal{C}^\circ_{m,n}$ qui va hériter de la généralisation de cette expression.

IV.2.4 Cone polaire de $\mathcal{P}^+_{m,n}(\mathbb{T})$

On peut donner facilement une formulation par générateurs du cône $\mathcal{P}^+_{m,n}(\mathbb{T})^\circ$, vu que son expression à l'aide d'inégalités est très simple.

Proposition IV.5: $\mathcal{P}^+_{m,n}(\mathbb{T})^\circ$ *est l'***enveloppe conique convexe** *de*

$$S_{m,n} = \{-v(\omega) \mid \omega \in [0, 2\pi]^2\}.$$

Pour la démonstration de ce résultat, nous renvoyons à la proposition II.14 : vu les notations utilisées, c'est *mutatis mutandis* la même démarche.

A partir de la formulation précédente, il est aussi possible de construire une formulation de $\mathcal{P}^+_{m,n}(\mathbb{T})^\circ$ par des inégalités, utilisant pour cela des mesures positives. Cette formulation revient à paramétrer une mesure (un élément de l'espace dual de l'espace des fonctions continues sur un compact) par ses moments, et l'on obtient ainsi une suite infinie d'Inégalités Linéaires Matricielles (LMI). On utilise pour cela une généralisation de la matrice $\mathcal{A}^\star$. Etant donné $x \in \mathbb{R}^{mn}$, on considère la matrice doublement Toeplitz $T_r(x)$ d'ordre l dont le terme général est

$$[T_r(x)]_{ij} = y_{\varphi(i)-\varphi(j)},$$

où $\varphi : \mathbb{N} \mapsto \mathbb{Z}^2$ est une énumération de $\mathbb{Z}^2$ et y est une fonction de $\mathbb{Z}^2$ dans $\mathbb{C}$ construite à partir de x comme suit :

$$y(k_1,k_2) = \begin{cases} -x_{00} & \text{si} \quad k_1 = k_2 = 0 \\ -\frac{1}{2}x_{k_1 k_2} & \text{si} \quad 0 \leqslant k_1 \leqslant m\,,\ 0 \leqslant k_2 \leqslant n \\ -\frac{1}{2}x_{(-k_1)(-k_2)} & \text{si} \quad m \leqslant k_1 \leqslant 0\,,\ n \leqslant k_2 \leqslant 0 \\ 0 & \text{sinon.} \end{cases} \tag{IV.2}$$

Proposition IV.6: *Avec les notations précédentes, on peut caractériser le cône polaire de* $\mathcal{P}^+_{m,n}(\mathbb{T})$ *par*

$$\mathcal{P}^+_{m,n}(\mathbb{T})^\circ = \{x \in \mathbb{R}^{m,n} \mid T_r(x) \preceq 0, \forall r \in \mathbb{N}\}.$$

La démonstration de ce résultat est basée sur le théorème de Bochner. Cela nécessite d'utiliser des notions de la théorie de la *dualité de Pontryagin* qui généralise la Transformation de Fourier sur les groupes localement compact abéliens (GLCA). Pour une présentation de cette théorie, une référence classique est [45] ; une présentation en français plus tournée vers les applications est [46]. Avant de présenter le théorème de Bochner, nous avons besoin d'introduire la notion de fonction définie positive sur un groupe.

Définition IV.6: *Soit* G *un groupe abélien localement compact (GLCA), et* $\phi : G \to \mathbb{C}$ *; on dit que* ϕ *est définie positive si, et seulement si, pour tout* $r \in \mathbb{N}$,

$$\sum_{i=1}^{r}\sum_{j=1}^{r} \phi(x_i - x_j)c_i\bar{c_j} \geqslant 0,$$

quels que soient $x_1, \cdots, x_r \in G$ *et* $c_1, \cdots, c_r \in \mathbb{C}$.

Dans la dualité de Pontryagin, on associe à chaque GLCA G un groupe dual Γ constitué de l'ensemble des morphismes de G dans $(\mathbb{T}, \times)$ (on les appelle des caractères) ; ainsi on établit une dualité $(\cdot, \cdot)$ entre G et Γ par

$$(x, \gamma) = \gamma(x),$$

définie ainsi pour $x \in G$ et $\gamma \in \Gamma$. Si on munit G de sa mesure de Haar, alors on définit la transformée de Fourier d'une fonction f intégrable sur G par

$$\hat{f}(\gamma) = \int_G f(x)(-x, \gamma)dx.$$

Parmi les propriétés découlant de cette définition, une des plus importantes reste sans aucun doute que $f \mapsto \hat{f}$ transforme le produit de convolution sur l'algèbre $\mathbb{C}[G]$ entre un produit classique sur $\mathbb{C}[\gamma]$. On peut dès à présent énoncer le théorème de Bochner, tel qu'il est présenté dans [45].

Théorème IV.1: *Soit* G *un GLCA,* Γ *son dual, et* $(\cdot, \cdot)$ *la dualité correspondante. Une fonction continue* $\phi : G \to \mathbb{C}$ *est dite définie positive si, et seulement si, il existe une mesure positive* $\mu \geqslant 0$ *sur* Γ *telle que*

$$\phi(x) = \int_\Gamma (x, \gamma)d\mu(x).$$

Pour la démonstration, on pourra se reporter à [45] : la partie "difficile" est la condition nécessaire, celle où il faut construire la mesure positive μ. On va maintenant démontrer la Proposition IV.6.

Démonstration. Si l'on considère $x \in \mathbb{R}^{mn}$ tel que $T_r(x) \preceq 0$ pour tout $r \in \mathbb{N}$, alors pour $G = \mathbb{Z}^2$, $y : G \to \mathbb{C}$ telle que proposée dans (IV.2) est bien une fonction définie positive sur G, car la matrice étant définie positive pour tout r, on a : quel que soit le choix de $c_1, \cdots, c_p \in \mathbb{C}$, et de $k_1, \cdots, k_p \in G$, il existe un r suffisament grand tel que $\varphi(1, \cdots, r) \supset \{k_1, \cdots, k_p\}$, et donc le caractère semi-défini positif de $T_r(x)$ assure que la somme suivante

$$\sum_{s=1}^{p} \sum_{t=1}^{p} y(k_s - k_t) c_t \bar{c_s},$$

est positive. On peut donc légitimement appliquer le théorème de Bochner ; par conséquent il existe une mesure positive μ sur $\Gamma = \mathbb{T}^2$ telle que

$$y(k) = \int_{\mathbb{T}^2} (k, \gamma) d\mu(\gamma) \text{ pour tout } k \in \mathbb{Z}^2,$$

propriété qui est vraie en particulier pour $k \in K_{m,n}$. On obtient ainsi pour tout $k \in K_{m,n}$:

$$x_k = \begin{cases} -\displaystyle\int_{[0,2\pi]^2} d\mu(\omega) & \text{si } k = (0,0); \\ -2\displaystyle\int_{[0,2\pi]^2} e^{i\langle k,\omega\rangle} d\mu(\omega) & \text{sinon.} \end{cases}$$

Or x_k étant réel, si l'on prend la partie réelle du membre de droite au-dessus, on constate que x appartient à la "surface des moments" engendrée par

$$(-1, -2\cos\omega_1, \cdots, -2\cos(m-1)\omega_1, -2\cos\omega_2, \cdots, -2\cos((m-1)\omega_1 + (n-1)\omega_2)),$$

qui n'est autre que $\mathrm{cone}(S_{m,n})$. La réciproque est directe, tous les raisonnements faits peuvent servir dans l'autre sens. □

IV.3 Heuristique de projection sur $\mathcal{C}_{m,n}$

Dans l'article [28], Hachez et Woerdeman proposent une méthode qui permet, étant donné un polynôme somme de carrés donné, de chercher l'écriture $p(z) = |F_1(z)|^2$ la plus proche. Ils imposent de plus que F_1 soit *extérieur*, c'est-à-dire qu'il n'ait aucune racine dans $\overline{\mathbb{D}}^2$, où $\mathbb{D} = \{z \in \mathbb{C} \mid |z| \leqslant 1\}$. Pour calculer cette approximation, ils proposent une méthode du type SDP en deux étapes (primale et duale).

Cette question, très pertinente dans le contexte du Traitement du Signal, peut amener à considérer le cas où le polynôme n'est pas forcément une somme de carrés ni même positif ; cependant, on peut quand même chercher quelle serait la factorisation $p(z) = |F_1(\omega)|^2$ la plus proche, ce qui revient en fait, pour une matrice $X \in \mathcal{M}_{m,n}(\mathbb{R})$, à chercher la matrice autocorrélée la plus proche.

On souhaiterait donc résoudre le problème suivant

$$(\mathrm{App}) \begin{cases} \min\limits_{X \in \mathcal{M}_{m,n}(\mathbb{R})} & \|X - C\|^2 \\ & X \in \mathcal{C}_{m,n}, \end{cases} \tag{IV.3}$$

où la norme considérée $\|\cdot\|$ est celle de Frobenius. Comme on l'a vu précédemment, le cône $\mathcal{C}_{m,n}$ est fermé, mais non-convexe (sans doute), il existe donc des points minimiseurs dans le calcul de la distance d'une matrice X donnée à $\mathcal{C}_{m,n}$, mais *a priori* leur unicité n'est pas assurée. De plus, le problème d'optimisation, du fait de sa non-convexité peut comporter des minimiseurs locaux, et la convergence vers l'optimum global d'un algorithme d'optimisation est donc difficile du point de vue de la théorie de la complexité. Pour le moment, on peut tout au plus espérer mettre en œuvre de bonnes heuristiques qui approcheraient bien l'optimum global.

IV.3.0.1 Principe général de l'heuristique

L'idée mise en oeuvre ici va consister à combiner les deux approches vues dans le cas unidimensionnel. En effet, la relaxation non-convexe est une approche locale, qui fournit donc une borne supérieure de la distance recherchée. L'algorithme de points intérieurs quant à lui, résout un problème dual, donc (selon la théorie de la dualité lagrangienne), il va nous fournir une borne inférieure sur la valeur optimale du problème. À l'aide des deux valeurs obtenues, on aura donc un encadrement - éventuellement grossier - de la distance minimale à déterminer.

En s'inspirant de la relaxation non-convexe proposée au chapitre précédent on voit aisément que le premier problème à résoudre est le suivant

$$(\mathcal{NC}) \begin{cases} \min\limits_{y \in \mathbb{R}^{mn}} & \|\mathcal{A}(yy^\top) - C\|^2. \end{cases}$$

Pour cela, on peut encore utiliser un algorithme de type quasi-Newton, avec une implémentation du calcul du gradient efficace que l'on détaillera ultérieurement. Du fait de la non convexité de la fonction coût, on convergera *a priori* vers un minimum local. Ainsi, en désignant par $\mathtt{val}(\mathcal{NC}, x_0)$ la valeur calculée par notre algorithme, avec un point initial $x_0 \in \mathbb{R}^{mn}$, et par $\mathtt{val}(\mathrm{App})$ la valeur du minimum global de (App), alors nécessairement

$$\mathtt{val}(\mathrm{App}) \leqslant \mathtt{val}(\mathcal{NC}, x_0) \qquad \forall\, x_0 \in \mathbb{R}^{mn}.$$

Pour obtenir une minoration de la valeur de notre problème, on peut relâcher l'ensemble-contrainte de manière obtenir une valeur minimale plus basse. Pour cela, on considère l'enveloppe conique convexe de $\mathcal{C}_{m,n}$ et le problème relaxé

$$(\mathcal{R}) \begin{cases} \min\limits_{X \in \mathcal{M}_{m,n}(\mathbb{R})} & \|X - C\|^2 \\ & X \in \mathtt{cone}(\mathcal{C}_{m,n}). \end{cases}$$

On a nécessairement

$$\texttt{val}(\mathcal{R}) \leqslant \texttt{val}(\mathcal{A}pp).$$

Pour résoudre ce problème d'optimisation relaxé, on utilisera le problème parent de projection sur le cône polaire,

$$(\mathcal{D}) \left\{ \begin{array}{ll} \min\limits_{X \in \mathcal{M}_{m,n}(\mathbb{R})} & \|X - C\|^2 \\ & X \in \mathcal{C}^\circ_{m,n}, \end{array} \right.$$

et grâce au minimiseur $\overline{X}$ de $(\mathcal{D})$, on récupèrera $\texttt{val}(\mathcal{R})$ simplement par

$$\texttt{val}(\mathcal{R}) = \|\overline{X}\|^2.$$

On a donc finalement obtenu un encadrement de la valeur optimale recherchée

$$\texttt{val}(\mathcal{R}) \leqslant \texttt{val}(\mathcal{A}pp) \leqslant \texttt{val}(\mathcal{NC}, x_0) \qquad \forall\, x_0 \in \mathbb{R}^{mn}.$$

IV.3.1 Résolution numérique du problème $(\mathcal{D})$

L'idée est de calculer la projection sur le polaire $\mathcal{C}^\circ_{m,n}$ en utilisant la mn-BALC

$$\phi(y) = -\ln\det \mathcal{A}^\star(-y).$$

On notera au passage, vu la linéarité de $\mathcal{A}^\star$, le fait que ϕ soit effectivement une BALC ne pose aucun problème, il faut juste remarquer que son paramètre devient $m * n$.

Comme nous utilisons les mêmes notations qu'au chapitre précédent, on trouve encore pour le gradient

$$\nabla_y \phi = \mathcal{A}([\mathcal{A}^\star(-y)]^{-1}),$$

et pour la matrice hessienne

$$D^2_y\phi[u,v] = \langle\langle [\mathcal{A}^\star(-y)]^{-1}\, \mathcal{A}^\star(u)\, [\mathcal{A}^\star(-y)]^{-1}, \mathcal{A}^\star(v)\rangle\rangle.$$

Les calculs suivants se déroulent de manière similaire, excepté le nombre de vecteurs dans la base. Considérons ainsi que l'on connaisse une décomposition de Cholesky de $[\mathcal{A}^\star(-y)]^{-1}$ sous la forme

$$[\mathcal{A}^\star(-y)]^{-1} = RR^\top = \sum_{k=0}^{mn} r_k r_k^\top,$$

alors l'opérateur hessien appliqué à u, v s'écrit

$$D^2_y\phi[u,v] = \sum_{k=0}^{m}\sum_{l=0}^{mn} \langle r_k, \mathcal{A}^\star(u) r_l\rangle \langle r_k, \mathcal{A}^\star(v) r_l\rangle.$$

Si on l'applique aux vecteurs de base $\{e_{ij}\}$, on en déduit pour le terme général de la matrice hessienne $H_{ij,pq}$:

$$D^2_y\phi[e_{ij}, e_{pq}] = \frac{1}{4}\sum_{k,l=0}^{mn} (\texttt{corr}_a(r_k, r_l) + \texttt{corr}_a(r_l, r_k))_{ij} (\texttt{corr}_a(r_k, r_l) + \texttt{corr}_a(r_l, r_k))_{pq}.$$

Comme dans le cas unidimensionel, on reconnaît une somme de matrice dyadiques et, par symétrie, la matrice hessienne $H \in \mathcal{S}_{mn}(\mathbb{R})$ se simplifie en

$$H = \frac{1}{2} \sum_{0 \leqslant k,l \leqslant mn} \mathtt{corr}_a(r_k, r_l)(\mathtt{corr}_a(r_k, r_l) + \mathtt{corr}_a(r_l, r_k))^\top. \tag{IV.4}$$

Là encore, une méthode rapide de calcul se base sur la transformée de Fourier discrète mais cette fois bidimensionnelle. Ainsi, en désignant par W_2^N la transformée de Fourier unidimensionnelle de $\mathbb{C}^N$ dans lui-même, on aura une expression de la TFD bidimensionnelle comme suit :

$$\begin{array}{rcl} \mathcal{F}_2 : M_{M,N}(\mathbb{C}) & \to & M_{M,N}(\mathbb{C}) \\ X & \mapsto & W_1^M X (W_1^N)^\top \end{array}$$

De ce fait, en utilisant l'identité d'aplatissement du produit de Kronecker

$$W_2^{M,N} x = \mathtt{vec}(\mathcal{F}_2(X)) = (W_1^N \otimes W_1^M) x.$$

Bien sur, cette transformée de Fourier jouit de beaucoup de propriétés communes avec son pendant unidimensionnel. Ainsi, il existe aussi un théorème de corrélation discrète bidimensionnelle,

$$\mathtt{corr}_c(x, y) = W_2^{-1} \left(\overline{W_2 x} \circ W_2 y \right),$$

où l'on a omis volontairement les indices M et N en supposant que l'on sait à l'avance dans quel espace on travaille. Cette corrélation cyclique est définie par

$$\mathtt{corr}_c(x, y)_{ij} = \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} \bar{x}_{kl} y_{(k+i)(l+j)},$$

où $(k + i)$ (resp. $(l + j)$) est considéré comme un élément de $\mathbb{Z}/M\mathbb{Z}$ (resp. $\mathbb{Z}/N\mathbb{Z}$). En utilisant l'injection linéaire $P_{n,N} : \mathbb{R}^n \to \mathbb{R}^N$ (qui consiste à compléter par des zéros) et la surjection linéaire (projection canonique) $S_{N,n} : \mathbb{R}^N \to \mathbb{R}^n$, on définit $P_{n,N}^{m,M} = P_{n,N} \otimes P_{m,M}$ et $S_{N,n}^{M,m} = S_{N,n} \otimes S_{M,m}$; alors

$$\mathtt{corr}_a(x, y) = S_{N,n}^{M,m} W_2^{-1} \left[\left(\overline{W_2 P_{n,N}^{m,M} x} \right) \circ \left(W_2 P_{n,N}^{m,M} y \right) \right].$$

Dans la suite, on allègera les notations en utilisant les conventions suivantes : $P_2 = P_{n,N}^{m,M}$ et $S_2 = S_{N,n}^{M,m}$.

Avec l'identité , où l'on note $R^k = P_2 r_k$, on déduit

$$H = \frac{1}{2} \sum_{k=0}^{mn} \sum_{l=0}^{mn} S_2 W_2^{-1} \left(\overline{R^k} \circ R^l \right) \left(\overline{R^k} \circ R^l + \overline{R^l} \circ R^k \right)^\top W_2^\top S_2^\top.$$

En développant cette expression, et en utilisant le fait que

$$(x \circ y)(b \circ c)^\top = (x b^\top) \circ (y c^\top) = (x c^\top) \circ (y b^\top),$$

puis en séparant les sommes sur k et l comme produits de facteurs, on obtient pour l'expression de H :

$$H = \frac{1}{2} S_2 W_2^{-1} \left[\left(\overline{\sum_{k=0}^{mn} R_k R_k^\top} \right) \circ \left(\sum_{l=0}^{mn} R_l R_l^\top \right) + \left(\overline{\sum_{k=0}^{mn} R_k \overline{R_k}^\top} \right) \circ \left(\sum_{l=0}^{mn} R_l \overline{R_l}^\top \right) \right] W_2^{-\top} S_2^\top.$$

Le gradient s'écrit quant à lui :

$$g = \sum_{k=0}^{mn} \mathtt{corr}_a(r_k, r_k) = S_2 W_2^{-1} (R_k \circ R_k).$$

Le calcul de la complexité de l'évaluation de H et de g se déduit facilement de celui du chapitre précédent. Sachant que la complexité d'une transformée de Fourier Bidimensionnelle de taille (m, n) coûte $\mathcal{O}(mn \log(mn))$, on en déduit que l'évaluation par cette méthode coûte au plus $\mathcal{O}((mn)^3)$; ceci permet de gagner un ordre de grandeur par rapport à l'utilisation de l'évaluation classique d'une BALC du cône des matrices SDP qui serait en $\mathcal{O}((mn)^4)$.

IV.3.2 Une approche numérique pour $(\mathcal{NC})$

Ici encore, nous avons volontairement choisi d'utiliser les mêmes notations qu'au chapitre précédent car elles permettent souvent de calquer les raisonnements et les algorithmes concernant $\mathcal{C}_{m,n}$ sur ceux de $\mathcal{C}_{n+1}$ sans nécessiter de grands changements. Ainsi, avec $f_c : \mathbb{R}^{mn} \to \mathbb{R}$ définie comme précedemment par

$$f_c(y) = \frac{1}{2} \|\mathcal{A}(yy^\top) - c\|^2,$$

où $c = \mathtt{vec} C$, on va nécessairement obtenir la même expression synthétique pour le gradient et l'opérateur hessien. On va observer des différences seulement dans leurs estimations. Par exemple, pour la méthode d'évaluation rapide du gradient ∇f_c de f_c, il suffit de remplacer la matrice d'échange J_n par $J_{mn} = J_n \otimes J_m$, comme le précise le

Lemme IV.3: *Soient x, y dans $\mathbb{R}^{mn}$; alors*

$$\mathcal{A}^\star(x) y = \frac{1}{2} \left(J_{mn} \mathtt{corr}_a(x, J_{mn} y) + \mathtt{corr}_a(x, y) \right).$$

Démonstration. Par définition,

$$\mathcal{A}^\star(x) y = \frac{1}{2} \left[\sum_{kl} x_{kl} E_n^l \otimes E_m^k y + \sum_{kl} x_{kl} (E_n^l \otimes E_m^k)^\top y \right].$$

Considérons la seconde somme ; on a

$$(E_n^l)^\top \otimes (E_m^k)^\top y = \mathtt{vec} \left((E_m^k)^\top Y E_n^l \right).$$

Or, comme on l'a vu au chapitre 2, multiplier à gauche par $(E_m^k)^\top$ va décaler les lignes de Y de k indices vers le bas, et multiplier à droite Y par E_n^l décale les indices des colonnes de l indices vers la droite ; donc

$$\left[(E_m^k)^\top Y E_n^l\right]_{ij} = Y_{(i+k)(j+l)}[i \leqslant m-k][j \leqslant n-l].$$

Par conséquent, la composante suivant (i,j) dans $\mathbb{R}^{mn}$ pour la seconde somme vaut

$$\sum_{k=0}^{m}\sum_{l=0}^{n} X_{ij}Y_{(i+k)(j+l)}[i \leqslant m-k][j \leqslant n-l] = \sum_{k=0}^{m-i}\sum_{l=0}^{n-j} X_{kl}Y_{(i+k)(j+l)} = \mathtt{corr}_a(x,y).$$

Pour la première somme, si l'on tient compte du fait que

$$J_{mn}(E_n^l)^\top \otimes (E_m^k)^\top J_{mn} = (J_n \otimes J_m)((E_n^l)^\top \otimes (E_m^k)^\top)(J_n \otimes J_m),$$

et que (cf. [32] par exemple),

$$(A \otimes B)(C \otimes D) = (A \otimes C)(B \otimes D),$$

(dans le cas où les dimensions sont compatibles bien sûr), il vient alors

$$(J_n(E_n^l)^\top J_n) \otimes (J_m(E_m^k)^\top J_m) = E_n^l \otimes E_m^k,$$

d'après (III.14). On en conclut que la première somme n'est rien d'autre que

$$J_{mn}\mathtt{corr}_a(x, J_{mn}y).$$

□

A l'aide de ce lemme, on déduit que la complexité d'évaluation du gradient et de la fonction objectif se fait en un coût maximum de $\mathcal{O}(mn\log(mn))$. Nous avons donc mis en œuvre cette méthode à l'aide d'un code (Scilab/C++/BLAS/FFTW) pour la partie duale $(\mathcal{D})$ et un code Scilab/M1QN3 pour la partie primale $(\mathcal{NC})$. Il faut noter que l'évaluation de f_c et de son gradient n'ont pas été codés en C mais en Scilab, cependant leur temps d'évaluation restent faibles en regard de ceux de l'algorithme de points intérieurs que nous avons pourtant chercher à optimiser. Nous avons réporté dans le tableau suivant, les résultats de quelques experiences numériques,

m	n	$val(\mathcal{NC}, x_0)$	$val(\mathcal{R})$	$T_{qn}(s)$	$T_{IPM}(s)$
5	2	0.839551	0.839644	0.01	0
5	5	0.910695	0.910783	0.02	2
10	5	3.281424	3.281727	0.03	14
10	10	5.350762	5.349270	0.05	118
15	7	5.442012	5.433882	0.05	39
15	15	14.236943	14.180545	0.13	352
20	10	11.984636	11.981227	0.11	1163
20	20	23.515661	23.311975	0.39	12297

où la précision de l'algorithme de points intérieurs était fixée à 10^{-4} et l'augmentation de la pénalité à 2 ; pour M1QN3, nous avions fixé le nombre maximum d'itérations à 500 (bien que ce nombre n'ait jamais été atteint, vu que la convergence se faisait en au plus 150 itérations), et la précision sur le gradient à 10^{-10}. Ces résultats méritent quelques commentaires : la première remarque que l'on peut faire, c'est que l'on ne peut malheureusement traiter que des problèmes de taille modeste en comparaison de ceux que l'on pouvait traiter en dimension un. Ainsi, pour l'instant, le plus gros problème que l'on ait reussi à traiter est de dimension 20x20, pour un temps CPU d'environ 3h ; il faut se rendre compte que, dans ce cas, on manipule déjà des matrices d'ordre 400, et donc avec une FFT on passe à des matrices d'ordre 4096, ce qui devient prohibitif en temps de calcul pour la partie concernant le problème dual. On retrouve ainsi une caractéristique générale des méthodes numériques en Traitement d'Images : avec le *fléau de la dimension*, les méthodes d'optimisation d'ordre deux deviennent rapidement inefficaces. En comparaison, la méthode Quasi-Newton parfois surnommée d'"ordre 1.5" (car on approxime de manière moins coûteuse la matrice hessienne) présente des temps de calcul nettement inférieurs et qui n'augmentent que très lentement si l'on compare à la méthode de points intérieurs. Pour des problèmes de dimension plus grande, il est clair que c'est la seule viable. A propos de l'encadrement obtenu, on voit qu'il n'est pas mauvais (de l'ordre de 10^{-2}), et si l'on regarde attentivement les deux premiers résultats, on voit que $\mathrm{val}(\mathcal{R}) > \mathrm{val}(\mathcal{NC}, \mathrm{x}_0)$, ce qui contredit apparemment la théorie ! En réalité, on constate que la différence $\mathrm{val}(\mathcal{R}) - \mathrm{val}(\mathcal{NC}, \mathrm{x}_0)$ - quand elle est strictement positive - reste de l'ordre de 10^{-4}, ce qui est exactement la précision que nous avons choisie pour l'algorithme de points intérieurs ; vu que ce dernier fournit une solution ε-sous-optimale, ceci finalement ne contredit en rien la théorie.

Conclusion

Comme cela a été dit dans l'introduction de ce mémoire de thèse, un de nos objectifs était de dégager des propriétés nouvelles du cône $\mathcal{C}_{n+1}$ des vecteurs à composantes auto-corrélées, afin d'en faire un objet plus familier de la théorie des cônes convexes. Au regard des propriétés additionnelles établies par rapport aux travaux précédents [1, 33], nous pensons avoir contribué à ce problème, en fournissant quelques propriétés intéressantes du cône. L'appréhension ou l'intuition que l'on peut avoir d'un objet mathématique peut s'apprécier aussi à l'aune des moyens algorithmiques mis en œuvre afin de résoudre des problèmes où cet objet apparaît. Ainsi, le problème de projection qui à servi de test tout au long de cette thèse, de par sa nature même, nous informe sur $\mathcal{C}_{n+1}$, puisque le résultat des différents algorithmes de projection consiste justement en des approximations d'éléments de $\mathcal{C}_{n+1}$. Dans le même esprit, il serait dommage de définir des objets mathématiques comme par exemple, π ou $\sqrt{2}$ sans pouvoir en donner des algorithmes d'approximation. De surcroit, concernant le problème de projection traité, les algorithmes proposés répondent ainsi au deuxième exemple issu du Traitement du Signal que nous présentons dans le premier chapitre. Le troisième chapitre à été l'occasion pour nous, de présenter différents algorithmes pour résoudre le problème d'approximation ou de projection. La distinction sur la pertinence d'utiliser tel ou tel algorithme dans différents cas pourrait se faire en se basant sur la nature du problème de Traitement du Signal considéré, c'est-à-dire synthèse ou identification, mais on remarquera que pour des problèmes de synthèse, pour des raisons physiques ou économiques, il est généralement préférable de se limiter à peu de variables et, dans ce cas, l'algorithme de suivi de chemin qui est assez similaire à l'approche dans [1] suffit amplement, et c'est pourquoi nous n'avons pas développé ce cas dans notre étude. La réponse au problème d'identification induit, quant à elle, un distinguo selon la taille des problèmes. Dans le cas de processus stationnaires, où le nombre de mesures reste faible, là encore la solution par points intérieurs s'avère incontournable puisqu'elle est précise et relativement rapide (pour des dimensions raisonnables) ; en revanche, pour des problèmes de grande taille, pour lequels on a besoin d'une approximation assez bonne dans un temps limité, l'approche par une méthode de Quasi-Newton prend tout son sens. Enfin, concernant les problèmes bi-dimensionnels introduits dans le dernier chapitre, il s'agit d'un domaine actuel de recherche active, avec les développements parallèles concernant les polynômes positifs à plusieurs variables ; ceci a donc été pour nous l'occasion de voir dans quelle mesure les techniques développées pour $\mathcal{C}_{n+1}$ pouvaient s'y appliquer.

En ce qui concerne les suites possibles à donner à ce travail, on insistera sur la démonstration ou l'infirmation de la Conjecture II.1 à propos des faces du cône $\mathcal{C}_{n+1}$, qui serait un

prolongement intéressant, en étudiant par exemple au préalable le cas $n+1=4$; la preuve de la non-convexité de $\mathcal{C}_{m,n}$ serait aussi intéressante, à traiter même s'il semble qu'il faille aller au-delà de la dimension 4 d'après Dritschel [17] pour trouver un contre-exemple valable. La recherche des valeurs propres des A^{ij} pour le cas bidimensionnel, nous paraît être également une piste de travail possible dont l'issue ne semble pas trop éloignée. Enfin, concernant l'étude algorithmique, comme alternative à la relaxation non-convexe il serait sans doute fructueux d'envisager une approche par Newton tronqué et la résolution du problème dual par d'autres méthodes que l'algorithme de suivi de chemin se doit d'être exploré.

Références

[1] Alkire, B., and Vanderberghe, L. Convex Optimization problems involving finite autocorrelation sequences. *Mathematical Programming 93* (2002), 331–359.

[2] Barvinok, A. *A course in Convexity*, vol. 54 of *Graduate Studies in Mathematics*. American Mathematical Society, 2002.

[3] Bauschke, H. H., and Borwein, J. M. Dysktra's Alternating Projection Algorithm for two sets. *Journal of Approximation Theory 79* (1994), 418–443.

[4] Beckermann, B. The Condition Number of real Vandermonde, Krylov and positive definite Hankel matrices. *Numer. Mathematik 85* (2000), 553–577.

[5] Ben-Tal, A., and Nemirovski, A. *Lectures on Modern Convex Optimization*. MPS-SIAM Series on Optimization, 2002.

[6] Bonnans, F., Gilbert, J., Lemaréchal, C., and Sagastizábal, C. *Optimisation Numérique : aspects théoriques et pratiques*. Springer Verlag, 1997.

[7] Boyd, S., and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004.

[8] Brigham, E. O. *The Fast Fourier TransForm and Its Applications*. Prentice Hall, 1988.

[9] Butler, P., and Cantoni, A. Eigenvalues and eigenvectors of symmetric centrosymmetric matrices. *Linear Algebra and Its Applications 13* (1976), 275–288.

[10] Byrnes, C. I., and Lindquist, A. A Convex Optimization Approach to Generalized Moment Problems. In *Control and modeling of complex systems (Tokyo, 2001)*, Trends Math. Birkhäuser Boston, Boston, MA, 2003, pp. 3–21.

[11] Cadzow, A., and Sun, Y. Sequences with Positive Semidefinite Fourier Transforms. *IEEE Transactions On Acoustics, Speech, and Signal Processing 34*, 6 (1986), 1502–1510.

[12] Chu, M. The stability group of symmetric Toeplitz matrices. *Linear Algebra and Its Applications 185* (1993), 119–123.

[13] Chu, M., and Golub, H. *Inverse Eigenvalue Problems*. Numerical Mathematics and Scientific Computation. Oxford Science Publications, 2005.

[14] Combettes, P. The convex feasibility problem in image recovery. In *Advances in Imaging and Electron Physics*, P. Hawkes, Ed. New York Academic Press, 1996, pp. 155–270.

[15] Davidson, T., Luo, Z., and Sturm, J. Linear matrix inequality formulation of spectral mask constraints with applications to FIR filter design. *IEEE Transactions on Signal Processing 50*, 11 (2002), 2702–2715.

[16] Delsarte, P., and Genin, Y. Spectral properties of finite Toeplitz matrices. In *Mathematical theory of networks and systems (Beer Sheva, 1983)*, vol. 58 of *Lecture Notes in Control and Inform. Sci.* Springer, London, 1984, pp. 194–213.

[17] Dritschel, M. On Factorization of Trigonometric Polynomials. *Integral Equations and Operator Theory 49* (2004), 11–42.

[18] Dumitrescu, B. Trigonometric Polynomials Positive on Frequency Domains and Applications to 2-D FIR Filter Design. *IEEE Trans. Signal Processing* (2005).

[19] Dumitrescu, B., Tabus, I., and Stoica, P. On the Parametrization of Positive Real Sequences and *MA* Parameter Estimation. *IEEE Transactions on Signal Processing 49* (2001), 2630–2639.

[20] Faybusovich, L. Semidefinite descriptions of cones defining spectral mask constraints. *European journal of operational research* (2006).

[21] Frigo, M., and Johnson, S. G. The Design and Implementation of FFTW3. *Proceedings of the IEEE 93*, 2 (2005), 216–231. special issue on "Program Generation, Optimization, and Platform Adaptation".

[22] Fuentes, M. Diagonalization of the symmetrized discrete i^{th} right shift operator : an elementary proof. *Numerical Algorithms 44* (2007), 29–43.

[23] Gilbert, J.-C., and Lemaréchal, C. Some numerical experiments with variable-storage quasi-Newton algorithms. *Mathematical Programming 45* (1989), 407–435.

[24] Gorsich, D., Genton, M., and Strang, G. Eigenstructures of Spatial Design Matrices. *Journal of Multivariate Analysis 80* (2002), 138–165.

[25] Graham, R., Knuth, D., and Patashnik, O. *Concrete Mathematics*. Addison-Wesley, 1994.

[26] Grenander, U., and Szegő, G. *Toeplitz forms and their applications*, second ed. Chelsea Publishing Co., New York, 1984.

[27] Hachez, Y. *Convex Optimization over Non-Negative Polynomials : Structured Algorithms and Applications*. PhD thesis, Université Catholique de Louvain, 2003.

[28] Hachez, Y., and Woerdeman, H. Approximating sums of squares with a single square. *Linear Algebra and its Applications 399* (2005), 187–201.

[29] Hettich, R., and Kortanek, K. Semi-Infinite Programming : Theory, Methods, and Applications. *SIAM Review 35* (1993), 380–429.

[30] Hiriart Urruty, J.-B. *Optimisation et analyse convexe*. Presses Universitaires de France, PARIS, 1998.

[31] Hiriart-Urruty, J.-B., and Lemaréchal, C. *Fundamentals of Convex Analysis*. Springer, 2001, pp. 46–51.

[32] Horn, R., and Johnson, C. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

[33] Krein, M., and Nudelman, A. *The Markov Moment Problem and extremal problems*, vol. 50 of *Translations of Mathematical Monographs*. American Mathematical Society, 1977.

[34] Lasserre, J. Global Optimization with Polynomials and The Problem of Moments. *Siam Journal on Optimization 11* (2001), 796–817.

[35] Malick, J. A dual approach to semidefinite least-squares problems. *SIAM Journal on Matrix Analysis and Applications 26*, 1 (2004), 272–284.

[36] McLean, J. W., and Woerdeman, H. J. Spectral factorizations and sums of squares representations via semidefinite programming. *SIAM Journal on Matrix Analysis and Applications 23*, 3 (2002), 646–655.

[37] Megretski, A. Positivity of Trigonometric Polynomials. In *Proceedings of the* 42^{th} *IEEE Conference on Decision and Control* (2003).

[38] Moulin, P., Anitescu, M., Kortanek, K., and Potra, F. The Role of Linear Semi-Infinite Programming in Signal-Adapted QMF Bank Design. *IEEE Transactions on Signal Processing 45* (September 1997).

[39] Nemirovski, A. Interior point polynomial time methods in convex programming. `http://iew3.technion.ac.il/Labs/Opt/opt/LN/ipms.pdf`, 1996.

[40] Nesterov, Y. Squared functional systems and optimization problems. In *High Performance Optimization*. Kluwer Academic Publishers, 2000, pp. 405–440.

[41] Nesterov, Y. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, 2004.

[42] Nesterov, Y., and Nemirovski, A. *Interior-Point Polynomial Methods in Convex Programming*, vol. 13 of *SIAM Studies in Applied Mathematics*. Society for Industrial and Applied Mathematics(SIAM), Philadephia, PA, 1994.

[43] Parillo, P. A. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. PhD thesis, California Institute of Technology, Pasadena, 2000.

[44] Peyré, G. *L'Algèbre Discrête de la Transformée de Fourier*. Ellipses, 2004, pp. 65–79.

[45] Rudin, W. *Fourier analysis on groups*. Interscience Tracts in Pure and Applied Mathematics, No. 12. Interscience Publishers (a division of John Wiley and Sons), New York-London, 1962.

[46] Samuelides, M., and Touzillier, L. *Analyse Harmonique*. Cépadues, 1990.

[47] Shor, N. *Nondifferential Optimization and Polynomial Problems*, vol. 24 of *Nonconvex Optimization and Its Applications*. Kluwer Academic Publishers, 1998.

[48] Stoica, A., Moses, R., and Stoica, P. Enforcing positiveness on estimated spectral densities. *Electronics Letters 29* (1993).

[49] Takouda, P. L. *Problèmes d'approximation matricielle linéaires coniques : Approches par Projections et via Optimisation sous contraintes de semidéfinie positivité*. PhD thesis, Université Paul Sabatier, Toulouse III, 2003.

[50] Tarazaga, P. Faces of the cone of Euclidean distance matrices : characterizations, structure and induced geometry. *Linear Algebra and its Applications 408* (2005), 1–13.

[51] Toh, R., Dumitrescu, B., and Vandenberghe, L. Interior-Point Algorithms for Sum-Of-Squares Optimization of Multidimensional Trigonometric Polynomials. *Proceedings of ICASSP* (2007).

[52] Yasuda, M. Spectral characterizations for hermitian centrosymmetric k-matrices and hermitian skew centrosymmetric k-matrices. *SIAM Journal on Matrix Analysis and Applications 25* (2003), 601–605.

Résumé: Dans ce travail de thèse, nous étudions, dans un contexte d'analyse convexe et d'optimisation, la prise en compte des contraintes dites d'autocorrélation, c'est-à-dire : nous considérons les situations où les vecteurs représentant les variables à optimiser sont contraintes à être les coefficients d'autocorrélation d'un signal discret à support fini. Cet ensemble des vecteurs à composantes autocorrélées se trouve être un cône convexe ; nous essayons d'en établir le plus de propriétés possibles : concernant sa frontière (lisse ou polyédrale), ses faces, l'acuité, l'expression du cône polaire, l'évaluation du cône normal en un point,etc. Ensuite, nous étudions divers algorithmes pour résoudre des problèmes d'optimisation où le cône des vecteurs à composantes autocorrélées entre en jeu. Notre principal objet d'étude est le problème de la projection sur ce cône, dont nous proposons la résolution par trois algorithmes différents : algorithmes dits de suivi de chemin, celui des projections alternées, et via une relaxation non-convexe. Enfin, nous abordons la généralisation de la situation d'autocorrélation au cas de signaux bi-dimensionnels, avec toute la complexité que cela engendre : multiples définitions possibles, non-convexité des problèmes résultants, et complexité calculatoire accrue pour les algorithmes.

Mot-clefs: Analyse Convexe, Optimisation, Autocorrélation, Polynômes Trigonométriques Positifs, Contrainte de Semi-Définie Positivité.

Abstract: In this work, we study how to take into account, from the convex analysis and optimization viewpoint, constraint sets of the following type : sets of vectors whose components are autocorrelations lags of finite discrete signals. A set of vectors with autocorrelated components turns out to be a convex cone, for which we etablish many basic properties such as : smoothness or not of the boundary, structure of faces, acuteness, expression of the polar cone, evaluation of the normal cone at a point, etc. Next, we propose some algorithms to solve optimization problems where this type of constraint set appears ; in particular we consider the problem of projecting a point on the convex cone of vectors with autocorrelated constraints. For these purposes, we study three different algorithms: an interior point method, one using alternating projections, and one via a non-convex relaxation of the original problem. Finally, we suggest extensions to the bi-dimensional signals case ; we outline the main difficulties which therefore appear : various possible new definitions, non-convexity of occuring problems, and increase in the computational complexity of the algorithmic procedures.

Keywords : Convex Analysis, Optimization, Autocorrelation, Nonnegative Trigonometric Polynomials, SDP constraints.